
GAZİ ÜNİVERSİTESİ
Gazi
1926

FÖRDERUNG DER SCHREIBKOMPETENZ
DURCH KÜNSTLICHE INTELLIGENZ IM DAF-UNTERRICHT

Muhammed Mustafa Uçar

DISSERTATION

ABTEILUNG FÜR DIDAKTIK DER DEUTSCHEN SPRACHE

GAZI UNIVERSITÄT

INSTITUT FÜR ERZIEHUNGSWISSENSCHAFTEN

DEZEMBER, 2025

TELİF HAKKI VE TEZ FOTOKOPİ İZİN FORMU

YAZARIN

Adı :Muhammed Mustafa

Soyadı :Uçar

Bölümü :Alman Dili Eğitimi

İmza :

Teslim Tarihi :02/12/2025

TEZİN

Almanca Adı : Förderung der Schreibkompetenz durch künstliche Intelligenz im DaF-Unterricht

Türkçe Adı : Yabancı Dil Olarak Almanca Öğretiminde Yazma Becerisinin Yapay Zeka Aracılığıyla Geliştirilmesi

İngilizce Adı : Promoting writing skills through artificial intelligence in German as a foreign language lessons

NOT: Tez enstitüye teslim edilirken bu bilgiler doldurulmalıdır.

ETİK İLKELERE UYGUNLUK BEYANI

Tez yazma sürecinde bilimsel ve etik ilkelere uyduğumu, yararlandığım tüm kaynakları kaynak gösterme ilkelerine uygun olarak kaynakçada belirttiğimi ve bu bölümler dışındaki tüm ifadelerin şahsıma ait olduğunu beyan ederim.

Yazar Adı Soyadı: Muhammed Mustafa Uçar

İmza:

JÜRİ ONAY SAYFASI

Muhammed Mustafa Uçar tarafından hazırlanan ”Förderung der Schreibkompetenz durch künstliche Intelligenz im Daf-Unterricht“ adlı tez çalışması aşağıdaki jüri tarafından oy birliği / oy çokluğu ile Gazi Üniversitesi Alman Dili Eğitimi Anabilim Dalı‘nda Doktora tezi olarak kabul edilmiştir.

Danışman: Prof. Dr. Muhammet KOÇAK
(Alman Dili Eğitimi / Gazi Üniversitesi) ……

Başkan: Prof. Dr. Erkan ZENGİN
(Alman Dili ve Edebiyatı / Hacettepe Üniversitesi) ……

Üye: Prof. Dr. Aylin SEYMEN
(Alman Dili Eğitimi / Gazi Üniversitesi) ……

Üye: Doç. Dr. Hasan Kazım KALKAN
(Alman Dili Eğitimi / Gazi Üniversitesi) ……

Üye: Doç. Dr. Meltem EKTİ
(Alman Dili ve Edebiyatı / Hacettepe Üniversitesi) ……

Tez Savunma Tarihi: 02/12/2025

Bu tezin Alman Dili Eğitimi Anabilim Dalı’nda Doktora tezi olması için şartları yerine getirdiğini onaylıyorum.

Prof. Dr. Osman ÇİMEN
Eğitim Bilimleri Enstitüsü Müdürü ……

Meiner wertvollen Familie, die mein Ein und Alles ist…

DANKSAGUNG

Dieser lange und beschwerliche Weg zur Promotion ist nicht nur das Ergebnis individueller Anstrengung, sondern auch das Produkt unschätzbarer Unterstützung. Allen, die mir von Anfang bis Ende dieser Arbeit mit ihrer Anwesenheit Kraft gaben, mich stets unterstützt haben und an der Fertigstellung dieser Arbeit beteiligt waren, möchte ich meinen aufrichtigsten Dank aussprechen.

Meinem geschätzten Betreuer, Prof. Dr. Muhammet Koçak, der mich in jeder Phase meiner Dissertation mit seiner wertvollen Anleitung, seinem profunden Wissen und seiner Unterstützung begleitet hat, gilt mein unendlicher Dank. Seine Geduld, seine wegweisende Betreuung, seine konstruktive Kritik und sein Vertrauen in mich waren die wichtigste Stütze für die Fertigstellung dieser Arbeit.

Meinen herzlichen Dank richte ich an Prof. Dr. Aylin Seymen, Prof. Dr. Erkan Zengin und Assoc. Prof. Dr. Hasan Kazım Kalkan, von denen ich während meines Master- und Promotionsstudiums lernen durfte und von deren wertvollem Wissen und Erfahrungen ich profitieren konnte. Jeder von ihnen hat einen überaus wertvollen Beitrag zu meiner akademischen Entwicklung geleistet.

Ein besonderer Dank gilt meinem hochgeschätzten Dozenten Dr. Fatih Çolak, der mein Leben geprägt und die Grundlagen für meine heutige akademische Identität gelegt hat. Er hat mir nicht nur Wissen vermittelt, sondern auch eine Vision mit auf den Weg gegeben. Ohne sein Vertrauen in mich und seine Anleitung wäre ich heute nicht an diesem Punkt.

Meiner lieben Familie, die alle Hürden dieses beschwerlichen Weges mit mir gemeinsam getragen hat und mir durch ihre Anwesenheit stets Kraft gab... Für ihre unendliche Geduld,

ihr Verständnis und ihre bedingungslose Liebe kann ich ihnen nicht genug danken. Ihr Dasein war meine größte Motivationsquelle auf dieser Reise.

Und meiner lieben Tochter Gökçe... Ich danke dir für jeden Augenblick, den ich dir aufgrund dieser Arbeit nicht widmen konnte, und für die große Geduld und das Verständnis, das dein kleines Herz aufgebracht hat. Die Wehmut, dich während dieses intensiven Arbeitspensums zeitweise vernachlässigt zu haben, war mein ständiger Begleiter. Dieser Erfolg ist ebenso deiner wie meiner.

Zuletzt gedenke ich voller Dankbarkeit und mit großer Wehmut meines lieben Freundes, Kollegen und Weggefährten Dursun Karadavut, der die Fertigstellung dieser Arbeit leider nicht mehr miterleben durfte. Sein Andenken und seine Freundschaft haben mich in den schwersten Momenten dieser Reise begleitet.

YABANCI DİL OLARAK ALMANCA ÖĞRETİMİNDE YAZMA BECERİSİNİN YAPAY ZEKA ARACILIĞIYLA GELİŞTİRİLMESİ

(Doktora Tezi)

Muhammed Mustafa Uçar
GAZİ ÜNİVERSİTESİ
EĞİTİM BİLİMLERİ ENSTİTÜSÜ
Aralık, 2025

ÖZ

Eğitimde yapay zeka teknolojilerinin artan rolü, yabancı dil olarak Almanca öğretiminde yazma becerilerinin geliştirilmesini merkezi bir uygulama alanı haline getirmektedir. Yazılı metinlerin düzeltme sürecinin öğretmenler açısından zaman alıcı olması, bireyselleştirilmiş öğrenme fırsatlarını kısıtlayan bir faktör olarak öne çıkmaktadır. İngilizce öğretimi bağlamında yapay zeka kullanımına ilişkin literatür zengin olmasına karşın, yabancı dil olarak Almanca alanı özelindeki eksiklik dikkate değer bir araştırma boşluğuna işaret etmektedir. Bu bağlamda, yürütülen çalışmanın temel amacı, önde gelen üç yapay zeka sohbet robotunun (ChatGPT, DeepSeek ve Gemini) Almanca metinlerde yer alan hataları tespit etme, düzeltme ve bu düzeltmelere ilişkin geri bildirim sunma performanslarını nicel bir perspektiften değerlendirmek ve mukayese etmektir. Araştırmada yarı-deneysel desen benimsenmiştir. Çalışma materyalini, A1, A2 ve B1 seviyelerinde yapay zeka ile üretildikten sonra Corder'in hata analizi çerçevesi temel alınarak 8 spesifik gramer ve leksik-semantik hata türünü barındıracak biçimde sistematik olarak manipüle edilen 60 Almanca metin oluşturmaktadır. Söz konusu metinler, üç sohbet robotuna sunularak, birincil sonuç ölçütleri olarak her bir modelin üç ardışık aşamadaki (Hata Tespiti, Hata Düzeltmesi, Geri Bildirim) başarı oranları tespit edilmiştir. Elde edilen bulgular, hiçbir modelin tüm görevlerde mutlak bir üstünlük sergilemediğini göstermiştir. Gemini, tespit ettiği hataları düzeltme bağlamında (%0.4'lük bir doğrudan düzeltme hata oranı ile) mükemmele yakın bir performans

sergilemekle birlikte, bu performans modelin -bilhassa leksik-semantik kategorilerdeki- başlangıç tespit kapasitesiyle sınırlı kalmıştır. DeepSeek, en dengeli genel performansı sergilemiş; ancak B1 seviyesindeki leksik-semantik ekleme hatalarını düzeltmede %30'luk yüksek bir doğrudan hata oranı ile spesifik bir zafiyet göstermiştir. ChatGPT, gramer hatalarını saptamada (%100'e yakın başarı) uzmanlaşmış bir profil çizerken, leksik-semantik hataları (özellikle yanlış kelime seçimi kategorisinde; %70 genel, B1'de %55) tespit etme ve pedagojik geri bildirim sağlama konularında en düşük performansı kaydetmiştir. Genel bir eğilim olarak, görev ve metin karmaşıklığı arttıkça tüm modellerin performansında bir düşüş gözlemlenmiş; gramer temelli hatalar, leksik-semantik hatalara kıyasla daha yüksek bir başarı oranıyla ele alınmıştır. Sonuç olarak, değerlendirmeye alınan yapay zeka sohbet robotlarının, mevcut yetenekleri doğrultusunda, yabancı dil olarak Almanca öğrenenler için etkili dilbilgisi denetleyicileri veya düzeltmenler olarak işlev görebileceği; ancak henüz güvenilir pedagojik aktörler olarak nitelendirilemeyeceği anlaşılmıştır. Modellerin yapısal ve kural tabanlı hatalara yönelik sağladığı destek güçlü olmakla birlikte, bağlamsal nüansları kavramayı gerektiren karmaşık anlamsal görevleri yönetme ve isabetli pedagojik açıklamalar sunma konusundaki yeterlilikleri sınırlıdır. Bu durum, söz konusu teknolojilerin eğitim ortamlarında dikkatli ve denetimli bir kullanımını zorunlu kılmaktadır.

Anahtar Kelimeler : Hata Analizi, Yapay Zeka, Geri Bildirim, Yazma Becerisi, Yabancı Dil olarak Almanca
Sayfa Sayısı : XVIII+225
Danışman : Prof. Dr. Muhammet KOÇAK

FÖRDERUNG DER SCHREIBKOMPETENZ DURCH KÜNSTLICHE INTELLIGENZ IM DAF-UNTERRICHT

(Dissertation)

Muhammed Mustafa Uçar

GAZI UNIVERSITÄT

INSTITUT FUR ERZIEHUNGSWISSENSCHAFTEN

Dezember, 2025

ZUSAMMENFASSUNG

Die wachsende Rolle von Technologien der künstlichen Intelligenz (KI) im Bildungswesen rückt die Entwicklung von Schreibkompetenzen im Deutsch als Fremdsprache (DaF)-Unterricht in den Mittelpunkt der Anwendung. Der für Lehrende zeitaufwändige Korrekturprozess schriftlicher Texte erweist sich als ein Faktor, der individualisierte Lernmöglichkeiten einschränkt. Während die Literatur zur Nutzung KI im Kontext des Englischunterrichts reichhaltig ist, verweist der Mangel im DaF-Bereich auf eine bemerkenswerte Forschungslücke. In diesem Zusammenhang besteht das Hauptziel der durchgeführten Studie darin, die Leistung von drei führenden KI-Chatbots (ChatGPT, DeepSeek und Gemini) bei der Erkennung und Korrektur von Fehlern in deutschen Texten sowie bei der Bereitstellung von Feedback zu diesen Korrekturen aus einer quantitativen Perspektive zu bewerten und zu vergleichen. In der Untersuchung wurde ein quasi-experimentelles Design angewendet. Das Studienmaterial besteht aus 60 deutschen Texten auf den Niveaustufen A1, A2 und B1, die zunächst von der KI generiert und anschließend auf der Grundlage von Corders Fehleranalyse-Framework systematisch manipuliert wurden, um 8 spezifische grammatikalische und lexikalisch-semantische Fehlerarten zu enthalten. Diese Texte wurden den drei Chatbots vorgelegt, wobei als primäre Ergebnismaße die Erfolgsquoten jedes Modells in drei aufeinanderfolgenden Phasen (Fehlererkennung, Fehlerkorrektur, Feedback) ermittelt wurden. Die erzielten Befunde zeigten, dass keines der

Modelle in allen Aufgaben eine absolute Überlegenheit aufwies. Während Gemini bei der Korrektur der erkannten Fehler eine nahezu perfekte Leistung erbrachte (mit einer direkten Fehlerquote bei der Korrektur von 0.4%), blieb diese Leistung durch die anfängliche Erkennungskapazität des Modells – insbesondere in den lexikalisch-semantischen Kategorien – begrenzt. DeepSeek zeigte die ausgewogenste Gesamtleistung, wies jedoch mit einer hohen direkten Fehlerquote von 30% eine spezifische Schwäche bei der Korrektur von lexikalisch-semantischen Hinzufügungsfehlern auf dem B1-Niveau auf. ChatGPT zeigte ein spezialisiertes Profil bei der Erkennung grammatikalischer Fehler (nahezu 100% Erfolg), verzeichnete jedoch die niedrigste Leistung bei der Erkennung lexikalisch-semantischer Fehler (insbesondere in der Kategorie falsche Wortwahl; 70% gesamt, 55% auf B1) und bei der Bereitstellung pädagogischen Feedbacks. Als allgemeiner Trend wurde bei allen Modellen ein Leistungsabfall mit zunehmender Aufgaben- und Textkomplexität beobachtet; grammatikbasierte Fehler wurden mit einer höheren Erfolgsquote als lexikalisch-semantische Fehler behandelt. Zusammenfassend lässt sich feststellen, dass die evaluierten KI Chatbots gemäß ihren derzeitigen Fähigkeiten als effektive Grammatikprüfer oder Korrektoren für Lernende des Deutschen als Fremdsprache fungieren können, jedoch noch nicht als zuverlässige pädagogische Akteure eingestuft werden können. Während ihre Unterstützung bei strukturellen und regelbasierten Fehlern stark ist, sind ihre Kompetenzen bei der Bewältigung komplexer semantischer Aufgaben, die das Erfassen kontextueller Nuancen erfordern, sowie bei der Bereitstellung treffender pädagogischer Erklärungen begrenzt. Dieser Umstand erfordert eine sorgfältige und beaufsichtigte Nutzung dieser Technologien in Bildungsumgebungen.

Schlüsselwörter : Fehleranalyse, Künstliche Intelligenz, Feedback, Schreibkompetenz, Deutsch als Fremdsprache
Seitenanzahl : XVIII+225
Betreuer : Prof. Dr. Muhammet KOÇAK

INHALTSVERZEICHNIS

TABELLENVERZEICHNIS

ABKÜRZUNGSVERZEICHNIS

KI	Künstliche Intelligenz
ChatGPT	Chat Generative Pre-Trained Transformer
DaF	Deutsch als Fremdsprache
FSU	Fremdsprachenunterricht
EaF	Englisch als Fremdsprache
GenKI	Generative Künstliche Intelligenz
GPT	Generative Pre-Trained-Transformer
P	Phase
P1	Phase 1
P2	Phase 2
P3	Phase 3
P4	Phase 4
GE	Gesamt Erfolgreich
GF	Gesamt Fehlgeschlagen
GBF	Gesamt Bedingt Fehlgeschlagen
GES	Gesamt

GR	Grammatikalische Ebene
LS	Lexikalische/Semantische Ebene
AL	Auslassung
HF	Hinzufügung
AW	Auswahl
UM	Umstellung
C	ChatGPT
D	DeepSeek
G	Gemini

TEIL I

EINLEITUNG

Mit der zunehmenden Verbreitung von künstlicher Intelligenz (KI) wird in vielen Bereichen Software entwickelt, die das Leben erleichtert. Auch die Bildungs- und Ausbildungsaktivitäten wurden von der KI beeinflusst, und es werden verschiedene Software, Programme usw. in diesem Bereich eingesetzt. KI spielt inzwischen eine wichtige Rolle bei der Vorbereitung der Materialien, die Lehrpersonen in ihrem Unterricht verwenden, und dient auch als Hilfsmittel, um das Lernen zu unterstützen. Es ist vorstellbar, dass in naher Zukunft immer mehr Lehrpersonen und Lernenden von der KI profitieren werden. Intelligenzprogramme mit verschiedenen Zwecken sind heute leicht zu finden. Insbesondere die Chatbots ChatGPT (von OpenAI), DeepSeek (von DeepSeek) und Gemini (von Google) haben die Wertschätzung vieler Menschen gewonnen. Was also ist ChatGPT? ChatGPT gibt die Antwort auf diese Frage wie folgt:

> „[…] Ich bin ChatGPT, ein Sprachmodell, das von OpenAI entwickelt wurde. Ich basiere auf der GPT-3.5-Architektur, was für ‚Generative Pre-trained Transformer 3.5' steht. Ich wurde darauf trainiert, menschenähnliche Texte basierend auf den Eingaben zu generieren, die ich erhalte. Mein Ziel ist es, auf eine Vielzahl von Fragen zu antworten, Unterstützung zu bieten, Informationen bereitzustellen und in allgemeinen Konversationen zu interagieren. […] Ich wurde von OpenAI entwickelt, einem Unternehmen für KI mit dem Ziel, fortschrittliche KI-Modelle zu erforschen und zu entwickeln. OpenAI wurde von einer Gruppe von Technologieunternehmern und Forschern gegründet, darunter Elon Musk, Sam Altman und

andere. Das Unternehmen hat sich darauf konzentriert, die Fortschritte im Bereich der KI sicher und nachhaltig zu gestalten. GPT-3, die Architektur, auf der ich basiere, ist eines ihrer bemerkenswerten Projekte“ (ChatGPT, persönliche Kommunikation, 12 November 2023).

Diese Antwort gibt DeepSeek für dieselbe Frage:

„Der KI-Chatbot DeepSeek ist ein fortschrittlicher KI-Assistent, der natürliche Sprachverarbeitungstechnologien nutzt, um mit Benutzern zu interagieren. DeepSeek wurde entwickelt, um die Fragen der Benutzer zu verstehen, Informationen bereitzustellen, verschiedene Aufgaben zu erledigen und allgemein das Benutzererlebnis zu verbessern. Zu den Hauptfunktionen von DeepSeek können gehören:

1. Natürliches Sprachverständnis: Versteht die Fragen und Befehle der Benutzer und ermöglicht so eine menschenähnliche Interaktion.

2. Informationsbereitstellung: Greift auf eine umfangreiche Wissensdatenbank zu und liefert den Benutzern genaue und aktuelle Informationen.

3. Aufgabenmanagement: Kann verschiedene Aufgaben wie Terminvereinbarungen, Erinnerungen erstellen und E-Mails versenden erledigen.

4. Personalisierung: Bietet personalisierte Antworten und Empfehlungen basierend auf den Vorlieben und früheren Interaktionen der Benutzer.

5. Mehrsprachige Unterstützung: Kann in verschiedenen Sprachen interagieren und spricht somit eine breite Benutzergruppe an.

DeepSeek kann von Unternehmen genutzt werden, um ihren Kundenservice zu verbessern, von Einzelpersonen, um alltägliche Aufgaben zu erleichtern, und allgemein, um die Nutzung von KI-Technologien zu verbreiten“ (DeepSeek, persönliche Kommunikation, 24 März 2025).

Und Gemini gab die folgende Antwort auf dieselbe Frage.

„Google Gemini ist eine von Google AI entwickelte, multimodale (multimodal) Familie großer Modelle der künstlichen Intelligenz (KI). Seine grundlegendste Eigenschaft ist die Fähigkeit, nicht nur Text, sondern gleichzeitig auch verschiedene Arten von Informationen wie Bilder, Audio, Video und Code zu verstehen und zu verarbeiten. Gemini ist in verschiedenen Größen (z. B. Ultra, Pro und Nano) verfügbar, die für unterschiedliche Aufgaben und Plattformen (von Rechenzentren bis hin zu mobilen Geräten) optimiert sind. Der Chatbot, mit dem Sie gerade interagieren (also ich), wird ebenfalls von einem der Google Gemini-Modelle angetrieben.“ (Gemini, persönliche Kommunikation, 7 August 2025).

Das wichtigste Element der Zitate, die diese Studie betreffen, ist, dass die Chatbots darauf trainiert sind, menschenähnliche Texte zu produzieren. Sie sollen auch verschiedene Fragen

beantworten, Unterstützung anbieten und sogar Informationen liefern. Ist das Ziel einer Lehrperson bei der Entwicklung der Schreibfähigkeiten der Lernenden nicht sehr ähnlich, wenn nicht sogar dasselbe? Eine Lehrperson für Deutsch als Fremdsprache (DaF) beantwortet auch verschiedene Fragen, unterstützt die Lernenden und versorgt sie mit Informationen. Daher unterstützt diese Situation die Problemstellung dieser Studie auch.

Problemstellung

Es wäre sinnvoll, ein paar Sätze zu schreiben, um das Problem der Studie besser zu umreißen und die Studie einzugrenzen. Zunächst einmal wäre es gut, die Studie auf eine der vier grundlegenden Sprachfertigkeiten zu beschränken, die im Fremdsprachenunterricht (FSU) häufig genannt werden: Hören, Lesen, Schreiben und Sprechen. Man kann argumentieren, dass das Schreiben die Fähigkeit ist, bei der Lehrpersonen und Lernenden die meiste Hilfe brauchen. Das liegt daran, dass Lernende rezeptive Fähigkeiten (Lesen und Hören) selbst entwickeln können, aber für produktive Fähigkeiten (Sprechen und Schreiben) brauchen sie jemanden, der sie überprüfen und korrigieren kann. Da Chatbots -zumindest im Moment- noch nicht über Sprach- und Stimmwahrnehmungsfähigkeiten verfügen[1], wäre es eine realistischere Einschränkung, sich auf die Schreibfähigkeiten zu konzentrieren. Man kann davon ausgehen, dass die Korrektur schriftlicher Produkte durch Chatbots und die Rückmeldung über Fehler durch Chatbots die Schreibfähigkeiten der Lernenden verbessern kann. Das ist ein Thema, bei dem auch die Lehrkräfte Hilfe brauchen, denn je voller eine Fremdsprachenklasse ist, desto schwieriger wird es, gute Fortschritte in den Schreibfähigkeiten zu erzielen (genauso wie in anderen Fertigkeiten). Außerdem ist es den Lehrpersonen nicht immer möglich, schriftliche Arbeiten zu korrigieren und den Lernenden Feedback zu geben, und selbst wenn dies möglich ist, dauert es je nach Klassengröße sehr

[1]Diese Sätze wurden Ende 2023 geschrieben und damals war diese Situation gültig. Die schnelle Entwicklung der Technologie hat diesen Satz sehr schnell veralten lassen.

lange. Diese Situation untergräbt auch den Aspekt des unabhängigen/individuellen Lernens und wirkt sich negativ auf das individuelle Lerntempo der Lernenden aus. An diesem Punkt werden Chatbots sowohl den Lernenden als auch den Lehrpersonen helfen, wenn sie die schriftlichen Produkte wie eine Lehrperson überprüfen und nach der Entdeckung von Fehlern korrigierendes Feedback geben können. Chatbots sind derzeit in der Lage, dies zu tun, aber wie valid und zuverlässig ist dieses Verfahren? Genau das ist die Frage der Studie. Können Chatbots Korrektur-Feedback so genau und zuverlässig wie ein/e Deutschlehrer/in geben? Antworten auf diese Fragen werden in dieser Arbeit gesucht. Es wird davon ausgegangen, dass die Ergebnisse dieser Studie Lehrpersonen und Lernenden dabei helfen werden, die Schreibfähigkeiten im DaF-Unterricht zu verbessern. Obwohl viele Studien über KI im Englisch als Fremdsprache (EaF) Unterricht durchgeführt wurden, ist es besorgniserregend, dass es nicht genügend Forschung zur KI im DaF-Unterricht gibt. Daher soll diese Studie eine Lücke in diesem Bereich schließen.

Es ist also möglich, KI im FSU einzusetzen. In der Literatur sind verschiedene Studien zu diesem Thema zu finden. Bei der Betrachtung der veröffentlichten Studien wurde festgestellt, dass viele Untersuchungen zu KI und FSU existieren. Während insbesondere viele Studien zum EaF existieren, sind die Studien zum DaF unzureichend.

Die Tatsache, dass es nur wenige Studien über ChatGPT, DeepSeek und Gemini im Bereich DaF und viele Studien über EaF gibt, bildet die Grundlage für diese Untersuchung. Es wurden verschiedene Programme der KI im EaF-Unterricht für unterschiedliche Zwecke untersucht und geprüft. Es stellt sich die Frage, wie diese Programme, die im Fach EaF gewisse Fortschritte gemacht haben, im Fach DaF zum Erfolg führen können.

Das zentrale Problem der Studie besteht darin, herauszufinden, welcher Chatbot in bestimmten Bereichen erfolgreicher ist als ein anderer. Die Forschungsfragen der Studie sind folgende:

1. Wie viele Fehler in schriftlichen Produkten des DaF-Unterrichts können die KI-gestützten Chatbots ChatGPT, DeepSeek und Gemini identifizieren?
2. Wie viele der identifizierten Fehler in schriftlichen Produkten des DaF-Unterrichts können die KI-gestützten Chatbots ChatGPT, DeepSeek und Gemini korrekt korrigieren?
3. Wie viele der korrekt korrigierten Fehler in schriftlichen Produkten des DaF-Unterrichts können die KI-gestützten Chatbots ChatGPT, DeepSeek und Gemini mit angemessenem Feedback versehen?
4. Inwieweit können die KI-gestützten Chatbots ChatGPT, DeepSeek und Gemini erfolgreich eingesetzt werden, um die Schreibfertigkeit im DaF-Unterricht zu fördern?

Ziel der Studie

Ziel dieser Studie ist es, den potenziellen Einfluss von KI-gestützten Chatbots auf die Verbesserung der Schreibkompetenzen von Lernenden im Bereich DaF aufzuzeigen. Im Rahmen der Untersuchung werden experimentelle Methoden eingesetzt, um das von den Chatbots bereitgestellte sofortige Feedback, ihren Beitrag zur Fehlerkorrektur sowie ihre Fähigkeit zur Integration in den allgemeinen Lernprozess detailliert zu analysieren. Dadurch sollen wichtige Erkenntnisse darüber gewonnen werden, wie technologische Hilfsmittel effizient im Sprachunterricht eingesetzt werden können, welche Vorteile sie bei der Steigerung der Lernleistungen bieten und wie sie den Lehrmethoden eine neue Dimension verleihen können.

Relevanz der Studie

Diese Studie legt anhand konkreter Daten aus experimentellen Untersuchungen die kritische Rolle dar, die KI-gestützte Chatbots bei der Verbesserung der Schreibkompetenz im DaF-Unterricht spielen. Die durchgeführten experimentellen Anwendungen verdeutlichen den Einfluss des von den Chatbots bereitgestellten Feedbacks bei der Identifikation und Korrektur von Fehlern in schriftlichen Texten. Somit kann die Entwicklung der Schreibkompetenz dank der technologischen Möglichkeiten der KI effizienter und nachhaltiger erreicht werden.

Andererseits existieren im EaF-Unterricht zahlreiche Studien, die sich mit Fehleranalysen, Feedback und der Förderung von Schreibkompetenzen befassen, während die Forschung im Bereich des DaF-Unterrichts in diesem Kontext weitgehend begrenzt ist. Insbesondere in Untersuchungen, die sich auf die Verbesserung der Schreibkompetenz fokussieren und dabei Fehleranalysen sowie Feedbackprozesse evaluieren, wird eine deutliche Lücke in der Literatur festgestellt. Dies zeigt, dass trotz der vermehrt auftretenden Studien zu populären KI-Tools wie ChatGPT nahezu keine Untersuchungen vorliegen, die sich auf vergleichsweise neue Technologien wie DeepSeek und Gemini konzentrieren.

Darüber hinaus befasst sich diese Studie nicht nur mit Anwendungen von KI im DaF-Unterricht, sondern leistet auch einen wichtigen Beitrag zu innovativen Ansätzen an der Schnittstelle von FSU und KI-Technologien. Die Integration KI-gestützter Chatbots in den Bildungsprozess steigert sowohl die individuelle Leistung der Lernenden als auch die didaktischen Methoden, indem sie neue und effektive Strategien einführt. Es wird erwartet, dass die gewonnenen Ergebnisse zukünftigen Forschungen als Wegweiser dienen und eine umfassendere Bewertung der Rolle von Sprachunterricht sowie der Anwendung von KI im Bildungssektor ermöglichen.

Beschränkungen der Studie

Diese Studie ist auf persönliche Gespräche mit den KI-Chatbots ChatGPT, DeepSeek und Gemini beschränkt. Die in der Studie verwendeten Daten bestanden aus Texten, die von KI erstellt, von KI so bearbeitet wurden, dass sie Fehler enthielten, und wiederum von KI korrigiert wurden. Insbesondere wurde das Fehlerklassifikationsmodell von Corder verwendet, um die absichtlich in den Text eingefügten Fehler zu identifizieren. Dieses Modell ermöglicht eine systematische Klassifikation der Fehler nach Typen und fungiert als ein wichtiges Instrument, das die methodologischen Grenzen der Untersuchung festlegt.

Die Daten aus dem experimentellen Teil dieser Studie wurden in der zweiten Hälfte des Jahres 2025 erhoben. Die fehlerhaften Texte wurden am 01.10.2025 zur Korrektur in die KI-basierten Chatbots eingegeben. Angesichts der rasanten und ständigen technologischen Fortschritte im Bereich der KI wurde die Arbeit auf einen bestimmten Zeitraum von 2021 bis 2025 begrenzt. Diese Einschränkung dient dazu, das Potenzial der Ergebnisse, aufgrund neuer Paradigmen und Algorithmen in der Zukunft an Aktualität zu verlieren, zu minimieren.

Die Ergebnisse dieser Arbeit spiegeln die Entwicklung der KI innerhalb des festgelegten Zeitraums wider und können als Referenzpunkt für zukünftige ähnliche oder vergleichende Studien dienen. Die gewonnenen Daten sind eine entscheidende Ressource, um die Grundlage für zukünftige Forschungen zu schaffen und die Evolution der KI zu verstehen. Dementsprechend zielt diese Studie darauf ab, einen wichtigen Beitrag zur Literatur zu leisten, indem sie zuverlässige und vergleichbare Daten über den Stand der KI in einem bestimmten Zeitfenster bereitstellt.

TEIL II

THEORETISCHER TEIL

Dieser Abschnitt behandelt die wesentlichen Konzepte, die das Fundament der vorliegenden Studie bilden. Diese Konzepte werden unter den drei Hauptthemen Fehleranalyse, Feedback und KI in eigenen Unterabschnitten detailliert erläutert.

Zunächst werden die konzeptionellen und methodologischen Erklärungen zur Fehleranalyse dargelegt und deren Rolle sowie Bedeutung im Gesamtzusammenhang der Untersuchung hervorgehoben. Diese Analyse zielt darauf ab, potenzielle Abweichungen, Fehler oder Mängel systematisch zu identifizieren und auf dieser Grundlage Schlussfolgerungen zu ziehen.

Zweitens wird die Definition und Funktionalität von Feedback-Mechanismen beleuchtet. Insbesondere wird die entscheidende Rolle der Rückmeldung in den Trainings- und Optimierungsprozessen von KI-Systemen behandelt. In diesem Unterabschnitt werden verschiedene Arten von Feedbacks und ihre Auswirkungen auf die Datenqualität, die Modellleistung und die endgültigen Ergebnisse diskutiert.

Abschließend werden unter dem Thema KI grundlegende Definitionen und oberflächliche Erläuterungen gegeben, die sich auf den Umfang und den Kontext der Studie beschränken. Angesichts der Komplexität der KI wird in diesem Abschnitt angestrebt, das Mindestmaß an

Informationen zu vermitteln, das für die Erreichung der Forschungsziele erforderlich ist. Dieser Ansatz gewährleistet, dass die Grundprinzipien der KI verständlich sind, ohne vom Hauptfokus der Arbeit abzuweichen.

Fehleranalyse

Das Erlernen einer Fremdsprache ist von Natur aus ein nicht-linearer, komplexer und dynamischer Prozess. Es ist unvermeidlich, dass Lernende in diesem Prozess Äußerungen produzieren, die von den Normen der Zielsprache abweichen, und diese Abweichungen spielen eine entscheidende Rolle beim Verständnis der Spracherwerbsmechanismen. In der Fachliteratur werden diese Abweichungen im Allgemeinen in zwei Hauptkategorien unterteilt: den Fehler[2] und den Irrtum[3]. Die Klärung der Unterscheidung zwischen diesen beiden Begriffen ist von grundlegender Bedeutung sowohl für die Entwicklung von Lehrmethoden als auch für das Verständnis der Natur des Lernprozesses.

Der grundlegende Unterschied zwischen Fehler und Irrtum liegt darin, von welcher Ebene der kognitiven Prozesse des Lernenden er ausgeht. Diese Unterscheidung kann auf der von Chomsky (1965, zitiert nach Botley, 2015, s. 86) eingeführten Dichotomie von Kompetenz und Performanz begründet werden. Kompetenz bezeichnet das abstrakte und internalisierte Wissen eines Individuums über eine Sprache, während Performanz die Anwendung dieses Wissens in der Echtzeit-Handlung des Sprechens oder Schreibens ist.

In diesem Kontext resultiert der Fehler aus einem Mangel oder Defekt in der Kompetenz des Lernenden, also in dem mental konstruierten Sprachsystem. Brown (2007, s. 257f) definiert den Fehler als Ergebnis der systematischen Kompetenz eines Individuums; mit anderen Worten, das System im Kopf des Lernenden ist selbst fehlerhaft. In ähnlicher Weise stellt

[2] Eng. *error*

[3] Eng. *mistake*

Botley (2015, s. 83) fest, dass Fehler die systematische Abweichung des Lernenden von den grammatikalischen Regeln der Zielsprache widerspiegeln und ein Indikator für seine noch unvollständige Kompetenz sind. Erdoğan (2005, s. 263) unterstützt diese Ansicht und betont, dass ein Fehler auftritt, wenn ein Sprachelement unvollständig oder fehlerhaft gelernt wurde und vom Lernenden nicht selbstständig korrigiert werden kann, da er die korrekte Form nicht kennt. Ein Fehler ist daher, wie auch Gass und Selinker (2008, s. 102) feststellen, systematisch, neigt zur Wiederholung und wird vom Lernenden nicht als Abweichung wahrgenommen, da er Teil seines eigenen internalisierten Sprachsystems geworden ist. Lennon (1991, s. 182) betrachtet den Fehler aus einer breiteren Perspektive und definiert ihn als sprachliche Formen, die Muttersprachler unter denselben Kontexten und Bedingungen nicht produzieren würden.

Im Gegensatz dazu ist der Irrtum eine momentane und nicht-systematische Abweichung, die sich in der Performanz widerspiegelt, obwohl kein Problem in der Kompetenz des Lernenden vorliegt. Der Lernende kennt eigentlich die richtige Regel oder Struktur, kann sein Sprachwissen aber aufgrund physischer oder psychologischer Faktoren wie Müdigkeit, Unaufmerksamkeit, Gedächtnislücken oder Aufregung nicht korrekt anwenden (Brown, 2007; Corder, 1982). Botley (2015) beschreibt Irrtümer im Allgemeinen als unbeabsichtigte, momentane Lapsus und schlägt vor, dass sie auf Faulheit, Vergesslichkeit oder nicht ausreichend internalisierte Regeln zurückzuführen sein können. Das hervorstechendste Merkmal von Irrtümern ist, dass sie nicht-systematisch sind und in der Regel einmalige Ereignisse darstellen (Gass & Selinker, 2008, s. 102).

Eines der praktischsten Kriterien zur Unterscheidung dieser beiden Konzepte ist das Potenzial des Lernenden, seine eigene Produktion zu korrigieren. James (1998, s. 78) argumentiert, dass es sich um einen Irrtum handelt, wenn ein Lernender eine von ihm produzierte fehlerhafte Äußerung korrigieren kann, sei es nach einem Hinweis oder von selbst. Denn dies zeigt, dass der Lernende die korrekte Form eigentlich kennt, aber eine nicht beabsichtigte Form gewählt hat. Andererseits zeigt die Unfähigkeit oder der Unwille des

Lernenden, eine Korrektur vorzunehmen, dass er glaubt, die von ihm verwendete Form sei die richtige Form in seinem eigenen Kopf, was auf einen Fehler hindeutet.

Im FSU, insbesondere in den bis Ende der sechziger Jahre vorherrschenden behavioristischen Ansätzen, wurden Fehler als negativer Transfer von Muttersprachgewohnheiten und als schlechte Angewohnheiten betrachtet, die beseitigt werden müssen (Erdoğan, 2005, s. 262). Mit den von Corder (1982, s. 10f) angestoßenen kognitiven Ansätzen hat sich die Perspektive auf Fehler jedoch radikal verändert. Nach diesem modernen Ansatz sind Fehler keine Anzeichen von Versagen, sondern vielmehr ein natürlicher, unvermeidlicher und sogar nützlicher Teil des Lernprozesses.

Corder (1982) betont die Bedeutung von Fehlern in dreierlei Hinsicht:

Für den Lehrenden: Fehler sind ein diagnostisches Werkzeug. Sie liefern wertvolle Daten darüber, in welchen Bereichen der Lernende Schwierigkeiten hat, welche Aspekte der Sprache er noch nicht verstanden hat und an welchem Punkt im Lernprozess er sich befindet. Wie von Hendrickson (1978, s. 389) zitiert, stellt Corder (1973) fest, dass Fehler dem Lehrer Feedback über die Wirksamkeit seiner eigenen Lehrmaterialien und -techniken geben und aufzeigen, welche Teile des Lehrplans überarbeitet werden müssen. Erdoğan (2005, s. 266) fügt hinzu, dass die korrekte Analyse der Fehlerquellen es dem Lehrer ermöglicht, seine Lehransätze zu verbessern.

Für den Forschenden: Fehler öffnen ein Fenster zum Verständnis, wie Sprache gelernt wird. Sie liefern Forschern Belege für die kognitiven Strategien und Hypothesen, die der Lernende verwendet, um die Zielsprache zu verstehen. Dies ist eine grundlegende Datenquelle für die Entwicklung von Spracherwerbstheorien.

Für den Lernenden: Vielleicht am wichtigsten ist, dass Fehler der Lernprozess selbst sind. Der Lernende bildet Hypothesen über die Zielsprache, produziert Sprache unter Verwendung dieser Hypothesen und testet diese durch die Fehler, die er macht. Mit dem

Feedback aus seiner Umgebung formt er diese Hypothesen neu und nähert sich schrittweise dem Ziel an (Brown, 2007; Corder, 1982).

Aus dieser Perspektive haben die Fehler der Lernenden auch eine wegweisende Funktion für Pädagogen, Lehrplan- und Materialentwickler sowie Testersteller. Die systematische Analyse von Fehlern bereitet den Boden für die Entwicklung effektiverer Materialien, Lehrtechniken und Mess- und Bewertungsinstrumente (Erdoğan, 2005, s. 267).

Die Anerkennung von Fehlern als so wichtige Datenquelle führte zur Entstehung eines Forschungsbereichs namens Fehleranalyse. Dieser Analyseprozess zielt darauf ab, die Fehler in der Lernersprache systematisch zu untersuchen, um sowohl die Fehlerquellen zu erklären als auch Rückschlüsse auf die Lernprozesse zu ziehen. Die von Corder (zitiert nach Ellis, 1994, s. 48) für die Fehleranalyse vorgeschlagenen Schritte gelten in diesem Bereich als Standardmethodik:

1. Sammlung einer Sprachprobe
2. Identifizierung der Fehler
3. Beschreibung der Fehler
4. Erklärung der Fehler
5. Bewertung der Fehler.

Dieser Abschnitt kann kurz wie folgt zusammengefasst werden: Die Unterscheidung zwischen Fehler und Irrtum im FSU ist nicht nur eine terminologische Präferenz, sondern spiegelt auch eine grundlegende philosophische Haltung zum Lern- und Lehrprozess wider. Irrtümer sind momentane Performanzprobleme, während Fehler die natürlichsten und aufschlussreichsten Belege für das sich entwickelnde Sprachsystem (die Übergangskompetenz) des Lernenden sind. Daher bildet die Anerkennung von Fehlern nicht als Hindernisse, sondern als Stufen auf dem Lernweg und als wertvollen Feedback-

Mechanismus für Lernende und Lehrende gleichermaßen einen der Grundpfeiler des modernen Sprachlehrverständnisses.

Arten der Fehleranalyse

Die Fehleranalyse, die in den 1960er Jahren im Bereich des Zweitspracherwerbs als Reaktion auf die in der Sprachlehre vorherrschende Theorie der Kontrastivanalyse entstand, hat die Perspektive auf Lernfehler grundlegend verändert. Im Gegensatz zum Ansatz der Kontrastivanalyse, der die Hauptursache von Fehlern auf strukturelle Unterschiede zwischen Mutter- und Zielsprache zurückführte und sich auf den muttersprachlichen Transfer konzentrierte, bot die Fehleranalyse eine weitaus umfassendere Perspektive. Nach diesem neuen Ansatz sind die Fehler der Lernenden nicht nur eine Reflexion der Muttersprache, sondern auch ein Indikator für universelle Sprachlernstrategien und die im Gehirn des Lernenden ablaufenden kognitiven Prozesse. Erdoğan (2005, s. 262f) stellt fest, dass sich die Fehleranalyse in diesem Sinne auf die Untersuchung der kognitiven Prozesse des Lernenden konzentriert und Fehler als wertvolle Belege für den Zweitspracherwerbsprozess betrachtet. In diesem Kontext wurden Fehler nicht mehr als Anzeichen von Versagen gesehen, sondern als Teil eines dynamischen Prozesses, in dem der Lernende die Regeln der Zielsprache aktiv testet und zu internalisieren versucht.

Diese neue Perspektive, die die Fehleranalyse bot, führte zu wesentlichen Änderungen in der Sprachlehr- und -lernpraxis. Dulay, Burt und Krashen (1982, s. 138) betonen, dass eine der bahnbrechendsten Erkenntnisse der Fehleranalyse darin bestand, aufzuzeigen, dass die große Mehrheit der Grammatikfehler von Zweitsprachenlernern nicht auf den Einfluss der Muttersprache zurückzuführen ist. Im Gegenteil, sie stellten fest, dass diese Fehler große Ähnlichkeiten mit den Entwicklungsfehlern aufweisen, die kleine Kinder beim Erlernen ihrer Muttersprache machen. Dies wurde als starker Beweis dafür interpretiert, dass auch Zweitsprachenlerner, genau wie Kinder, die ihre Muttersprache erwerben, das Regelsystem

der Zielsprache schrittweise und nach ihren eigenen mentalen Schemata aufbauen. Folglich wurde die Fehleranalyse zu einem unverzichtbaren Werkzeug, um das Sprachsystem oder die Interlanguage des Lernenden zu verstehen. Brown (2007, s. 227) führt ebenfalls aus, dass die Beobachtung, Analyse und Klassifizierung der Fehler von Studierenden aus dem Bestreben, ihre inneren Sprachsysteme zu verstehen, das Forschungsfeld der Fehleranalyse hervorgebracht hat.

Das Hauptziel dieses Ansatzes ist es, laut Jabeen, Kazemian und Shahbaz (2015, s. 53), den Sprachlernprozess aus einer tieferen und detaillierteren Perspektive zu verstehen. Dieser Verstehensversuch beschränkt sich nicht nur auf das Auflisten von Fehlern. Altıparmak Yılmaz und Demir (2020, s. 106) geben an, dass die Fehleranalyse darauf abzielt, die von Sprachlernern gemachten Fehler zu identifizieren und die Häufigkeit dieser Fehler systematisch zu analysieren. Das Endziel dieses Prozesses ist jedoch, wie Abbott (1980, s. 124) betont, eine verlässliche psychologische Erklärung für die zugrunde liegenden Ursachen der Fehler zu liefern. Denn ohne das Verständnis der Ursache eines Fehlers kann keine dauerhafte und wirksame Lösung entwickelt werden. Wie Corder (1982, s. 52) feststellt, ist es erst möglich, einen Fehler systematisch zu korrigieren, wenn seine Ursache bekannt ist. Khansir (2012, s. 1029) definiert die Fehleranalyse als eine Art linguistischer Analyse, die sich auf Lernfehler konzentriert, wobei die Fehler mit der Zielsprache selbst verglichen werden und betont wird, dass der Einfluss der Muttersprache nicht die einzige Ursache für Fehler ist.

Die Fehleranalyse ist ein Ansatz, der vielseitige Vorteile für Lehrer, Forscher und sogar Schüler bietet (Subekti, 2018, s. 189) und eine entscheidende Rolle bei der Identifizierung der Schwierigkeiten und Bedürfnisse von Lernenden in einer bestimmten sprachlichen Entwicklungsphase spielt (Erdoğan, 2005, s. 267). Dieser Ansatz wird, wie Keshavars (zitiert nach Erdoğan, 2005, s. 262f) und Corder (1982, s. 45) feststellen, in zwei Hauptzweige unterteilt: die theoretische Fehleranalyse und die praktische (angewandte) Fehleranalyse.

Der theoretische Aspekt der Fehleranalyse ist ein forschungsbasiertes Untersuchungsfeld, das sich auf das Verständnis der Natur und Funktionsweise des Sprachlernprozesses konzentriert. Corder (1982, s. 45) erklärt die Hauptfunktion der theoretischen Fehleranalyse darin, als Methode zur Untersuchung des Sprachlernprozesses zu dienen und in diesem Prozess das vorhandene Wissen des Lernenden über die Zielsprache zu einem bestimmten Zeitpunkt zu beschreiben und mit der erhaltenen Ausbildung in Beziehung zu setzen. In diesem Sinne sucht die theoretische Fehleranalyse nach einer Antwort auf die Frage: *Was geht im Kopf des Lernenden vor?*

Dieser Forschungsprozess untersucht insbesondere die von Sprachlernern verwendeten mentalen Strategien und deren Ähnlichkeiten mit dem Erstspracherwerb. Laut Erdoğan (2005, s. 263) analysiert die theoretische Fehleranalyse Strategien der Lernenden wie Übergeneralisierung (z. B. die Anwendung einer Grammatikregel auf alle Situationen ohne Berücksichtigung von Ausnahmen) oder Vereinfachung (der Ausdruck einer komplexen Struktur in einer grundlegenderen Form). Die theoretische Analyse befasst sich hauptsächlich mit den Prozessen und Strategien des Spracherwerbs und deren Ähnlichkeiten mit dem Erstspracherwerb. Diese Untersuchungen zielen darauf ab, zu erforschen, ob der Sprachlernprozess eine universelle Struktur aufweist und ob es einen inneren Lehrplan gibt, dem die Lernenden bewusst oder unbewusst folgen (Erdoğan, 2005). Jabeen et al. (2015, s. 54) fassen diesen Prozess als ein Bestreben zusammen, die Ursachen von Fehlern zu verstehen, indem die Probleme beim Sprachenlernen und die dabei wirkenden grundlegenden Strukturen untersucht werden. Kurz gesagt, die theoretische Fehleranalyse betrachtet Fehler weniger als pädagogisches Problem, sondern vielmehr als eine Datenquelle, die die psycholinguistischen Mechanismen des Spracherwerbs beleuchtet.

Die aus der theoretischen Fehleranalyse gewonnenen Erkenntnisse dienen als Fahrplan für Praktiker und speisen den praktischen Aspekt der Fehleranalyse. Laut Corder (1982, s. 45) ist das Hauptziel des praktischen Aspekts der Fehleranalyse, die notwendigen Korrekturmaßnahmen und pädagogischen Interventionen zu bestimmen, um Lernsituationen

zu verbessern, mit denen der Lernende oder der Lehrende unzufrieden ist. Dies ist die Phase, in der theoretische Erkenntnisse direkt in die Unterrichtspraxis umgesetzt werden.

Die angewandte Fehleranalyse befasst sich auf der Grundlage der Erkenntnisse der theoretischen Analyse mit Themen wie der Organisation von Förderunterricht, der Gestaltung geeigneter Materialien und der Entwicklung effektiver Lehrstrategien. Erdoğan (2005, s. 263) stellt ebenfalls fest, dass die praktische Fehleranalyse darauf abzielt, ausgehend von theoretischen Erkenntnissen Förderkurse zu entwerfen, geeignete Materialien zu entwickeln und effektive Lehrstrategien zu formulieren. Wenn also theoretische Analysen ergeben, dass eine Gruppe von Lernenden eine bestimmte Grammatikstruktur systematisch falsch verwendet, übernimmt die praktische Fehleranalyse die Aufgabe, zusätzliche Übungen, differenzierte Lehrmethoden oder visuelle Materialien zu entwickeln, um dieses Problem zu lösen (Jabeen et al., 2015, s. 54). Dies gewährleistet, dass die Forschung einen konkreten Nutzen im Klassenzimmer hat.

Corder (1982, s. 47f) detailliert den Entscheidungsprozess dieses praktischen Interventionsprozesses weiter. Seiner Meinung nach sollten bei der Entscheidung für eine Korrekturmaßnahme zwei grundlegende Fragen gestellt werden: *Ist eine Intervention notwendig?* und *wenn ja, welche Art von Intervention sollte es sein?* Um diese Fragen zu beantworten, schlägt er vor, den Grad der Inkongruenz zwischen dem vorhandenen sprachlichen Wissen des Lernenden und den Anforderungen der Situation zu analysieren, und klassifiziert dies in drei Stufen:

Inkongruenz ersten Grades: Auf dieser Ebene gibt es kleine und unbedeutende Unterschiede zwischen dem sprachlichen Wissen und den Anforderungen der Situation. Diese Fehler behindern die Kommunikation in der Regel nicht und sind Situationen wie Lapsus, mit denen der Lernende selbst fertig werden kann. Daher erfordern sie meist keine pädagogische Intervention.

Inkongruenz zweiten Grades: Auf dieser Ebene geht es um Situationen, in denen der Lernende nicht über das erforderliche sprachliche Wissen verfügt, aber das Potenzial und die Motivation hat, dieses Wissen zu erlernen. Genau diese Ebene ist der fruchtbarste Bereich für korrigierende Interventionen. Das Feedback, die Erklärungen und die zusätzlichen Übungen des Lehrers sind an diesem Punkt sinnvoll und effektiv.

Inkongruenz dritten Grades: Auf dieser Ebene ist der Unterschied zwischen dem Wissen des Lernenden und den Anforderungen der Situation so groß, dass eine korrigierende Maßnahme höchstwahrscheinlich wirkungslos bleiben würde. In diesem Fall könnte es eine bessere Strategie sein, den Lernenden aus der Situation oder Aufgabe zu entfernen, die ihn zu diesem Fehler zwingt, d.h. das Niveau der Aufgabe an die aktuelle Kompetenz des Lernenden anzupassen. Dies spiegelt einen strategischeren und lernerzentrierten Ansatz wider, der der weit verbreiteten Annahme widerspricht, dass jeder Fehler zu jeder Zeit korrigiert werden müsse.

Klassifikation von Fehlern

Eine effektive Analyse der gemachten Fehler erfordert zunächst deren systematische Klassifikation. Diese Klassifikation kann nach verschiedenen Ansätzen erfolgen, die sowohl die äußere Erscheinung des Fehlers als auch seine Auswirkung auf die Kommunikation berücksichtigen. Auf diese Weise können Lehrende und Forschende die Problembereiche in der Sprachentwicklung des Lernenden klarer diagnostizieren.

Einer der gebräuchlichsten Ansätze zur Klassifikation von Fehlern ist die Kategorisierung nach linguistischen Ebenen. Laut Brown (2007, s. 231f) können Fehler auf verschiedenen linguistischen Ebenen untersucht werden, wie Phonologie (Lautlehre), Orthographie (Rechtschreibung), Lexikon (Wortschatz), Grammatik und Diskurs. Daneben existiert auch eine allgemeinere Klassifikation, die sich auf die Veränderungen in der Oberflächenstruktur

der Fehler konzentriert. Diese grundlegenden Kategorien sind in der Regel Hinzufügung, Auslassung, Auswahl und Umstellung.

Zwei weitere wichtige Konzepte, die verwendet werden, um die Art des Fehlers genauer zu verstehen, sind der Bereich[4] und der Umfang[5]. Brown (2007, s. 231f) und Ellis (1994, s. 70) nach Lennon verwenden diese Begriffe, um den analytischen Rahmen des Fehlers zu skizzieren. Der Bereich bezieht sich auf die Weite des Kontextes, der untersucht werden muss, um festzustellen, ob eine Äußerung fehlerhaft ist; dieser Kontext kann ein einzelnes Wort, ein Satz oder sogar der gesamte Text sein. Der Umfang definiert die Größe der sprachlichen Einheit, die geändert werden muss, um den Fehler zu korrigieren. Diese beiden Konzepte ermöglichen es uns zu verstehen, nicht nur was der Fehler ist, sondern auch, wo und in welchem Ausmaß er auftritt.

Eine weitere entscheidende Unterscheidung bei der Klassifikation von Fehlern wird nach ihrer Auswirkung auf die Kommunikation getroffen. In diesem Kontext werden Fehler in *globale* und *lokale* Fehler unterteilt. Burt (1975, s. 56f) definiert Fehler, die die Kommunikation erheblich behindern und die Bedeutung durch die Störung der Gesamtstruktur des Satzes unklar machen, als globale Fehler. Fehler wie eine falsche Wortstellung oder die fehlende bzw. falsche Verwendung von Konjunktionen, die den logischen Fluss des Satzes sicherstellen, fallen in diese Kategorie. Solche Fehler können es dem Hörer oder Leser nahezu unmöglich machen, die Botschaft wie beabsichtigt zu verstehen. Im Gegensatz dazu sind lokale Fehler solche, die nur ein bestimmtes Element des Satzes betreffen und in der Regel die allgemeine Verständlichkeit der Nachricht nicht beeinträchtigen (Brown, 2007, s. 231f). Kleinere grammatikalische Fehler wie Ungenauigkeiten bei der Konjugation von Verben und Deklination von Substantiven oder das Fehlen von Artikeln oder Hilfsverben werden als lokale Fehler betrachtet (Burt, 1975, s.

[4] Eng. *Domain*

[5] Eng. *Extent*

56f). Aus pädagogischer Sicht ist diese Unterscheidung äußerst wichtig, da der Korrektur von globalen Fehlern, die die Kommunikation vollständig unterbrechen, eine höhere Priorität eingeräumt wird als der von lokalen Fehlern.

Die Klassifikation von Fehlern nach ihrer äußeren Erscheinung ist eine der am häufigsten verwendeten Taxonomien. Diese Klassifikation konzentriert sich auf die beobachtbaren Strategien, die der Lernende bei der Fehlerproduktion anwendet. Forscher wie Subekti (2018, s. 190) und Corder (1982, s. 36) unterteilen diese Klassifikation in grundlegende Kategorien:

Auslassung[6]: Der Fall, in dem ein Element, das in einem grammatikalisch korrekten Satz vorhanden sein sollte (z. B. ein Hilfsverb, ein Artikel oder eine Pluralendung), im Satz fehlt.

Hinzufügung[7]: Die Verwendung eines unnötigen Elements, das nicht im Satz vorhanden sein sollte. Beispielsweise fällt die Verwendung einer zusätzlichen Präposition aufgrund von muttersprachlichem Einfluss oder ein unnötiges Pronomen in diese Kategorie.

Falschbildung / Auswahl[8]: Diese Kategorie umfasst Fälle, in denen ein falsches Element anstelle eines korrekten ausgewählt wird oder ein Element formal fehlerhaft verwendet wird. Zum Beispiel sind die Wahl einer falschen Zeitform (Auswahl) oder die falsche Verwendung der Endung für die dritte Person Singular bei einem Verb (Falschbildung) solche Fehler.

Umstellung[9]: Die Anordnung der Elemente eines Satzes in einer Weise, die nicht den Regeln der Satzstruktur entspricht, obwohl die Elemente selbst korrekt sind. Dies ist eine häufige Fehlerart, insbesondere in Fragesätzen oder bei der Anordnung von Adjektiven.

---

[6] Eng. *Omission*

[7] Eng. *Addition*

[8] Eng. *Misformation / Selection*

[9] Eng. *Ordering*

Tabelle 1.

Klassifikation von Fehlern nach Corder

Fehlertyp	Fehlerebene		
	Graphologisch / Phonologisch	Grammatisch	Lexikalisch-Semantisch
Auslassung	✓	✓	✓
Hinzufügung	✓	✓	✓
Auswahl	✓	✓	✓
Umstellung	✓	✓	✓

Quelle: Corder, S. P. (1982). Error analysis and interlanguage. Oxford: Oxford University

Corder (1982, s. 36) betont, dass diese Klassifikation der Oberflächenstrategie allein nicht ausreicht, da diese Fehlertypen auf verschiedenen linguistischen Ebenen auftreten können. Wie in Tabelle 1 zusammengefasst, kann ein Auslassungsfehler beispielsweise auf graphologischer Ebene als fehlender Buchstabe, auf grammatischer Ebene als fehlende Konjugationsendung oder auf lexikalisch-semantischer Ebene als fehlendes Wort zur Vervollständigung der Bedeutung auftreten. Dies zeigt, dass bei der Fehleranalyse sowohl die Art des Fehlers als auch die linguistische Ebene, auf der er auftritt, gemeinsam bewertet werden müssen.

Ein Ansatz zur Klassifikation von Fehlern, der tiefer geht als die Oberflächenform, wurde ebenfalls von Corder entwickelt. Wie Ellis (1994, s. 56) berichtet, untersucht Corder Fehler auf drei Ebenen, die sich nach dem Entwicklungsstadium des vom Lernenden internalisierten Regelsystems (seiner Interlanguage) richten. Diese Klassifikation konzentriert sich auf *das Verständnis des kognitiven Zustands hinter dem Fehler*:

Präsystematische Fehler: Diese Fehler treten in der Entdeckungsphase auf, in der der Lernende sich der Existenz einer bestimmten Regel in der Zielsprache noch nicht bewusst ist. Da der Lernende die korrekte Struktur nicht kennt, befindet er sich quasi auf dem Weg von Versuch und Irrtum, und seine Fehler sind inkonsistent und zufällig. In der Regel kann er nicht erklären, warum er diesen Fehler gemacht hat.

Systematische Fehler: Diese Phase zeigt, dass der Lernende nun eine Regel entdeckt hat, diese Regel aber nicht den Normen der Zielsprache entspricht. Der Lernende hat, wenn auch fehlerhaft, eine Hypothese gebildet und wendet diese konsequent an (z. B. durch

Übergeneralisierung). In dieser Phase kann der Lernende seine eigene falsche Regel erklären, ist aber in der Regel nicht in der Lage, seinen Fehler selbst zu korrigieren. Diese Fehler sind die wichtigsten Indikatoren dafür, dass der Lernprozess aktiv abläuft.

Postsystematische Fehler: Fehler auf dieser Ebene werden als Lapsus oder Mängel betrachtet, die der Lernende macht, obwohl er die richtige Regel eigentlich kennt, was auf Performanzfaktoren wie Unaufmerksamkeit, Müdigkeit oder Druck im Moment des Sprechens zurückzuführen ist. Der Lernende bemerkt diese Fehler in der Regel und kann sie selbst korrigieren, wenn er von jemand anderem darauf hingewiesen wird.

Ellis (1994, s. 56) weist darauf hin, dass Corders Klassifikation zwar theoretisch sehr nützlich ist, um die Sprachentwicklung des Lernenden zu verstehen, ihre praktische Anwendung jedoch schwierig ist. Um eindeutig zu bestimmen, in welche Kategorie ein Fehler fällt, erfordert es in der Regel ein persönliches Gespräch mit dem Lernenden, um dessen Denkprozess zu verstehen, was nicht immer möglich ist.

Identifikation und Erklärung von Fehlern

Das Endziel der Fehleranalyse ist, wie Corder (1982, s. 36) feststellt, eine zufriedenstellende Erklärung für die gemachten Fehler zu liefern. Die Voraussetzung für das Erreichen dieses Ziels ist jedoch die korrekte und umfassende Beschreibung und Interpretation der Fehler. Dieser Prozess erfordert eine weitaus komplexere Anstrengung des Verstehens und der Rekonstruktion als nur die Feststellung grammatikalischer Fehler. Um die Natur eines Fehlers zu verstehen, ist es unerlässlich, über die Oberflächenstruktur hinauszugehen und sich auf die Intention des Lernenden sowie den Kontext der Äußerung zu konzentrieren.

Nach Corders Modell müssen bei der Beschreibung von Fehlern zwei grundlegende Kategorien berücksichtigt werden: offenkundige[10] und verdeckte[11] Fehler. Brown (2007, s.

[10] Eng. *overt*

[11] Eng. *covert*

229) stellt diese Unterscheidung klar dar. Offenkundige Fehler sind Fehler, die bei der Betrachtung auf Satzebene offensichtlich gegen grammatische Regeln verstoßen und leicht zu erkennen sind. Ein Satz wie *Er gestern zur Schule gehen* enthält beispielsweise einen strukturell offensichtlichen Fehler. Im Gegensatz dazu erfordern verdeckte Fehler eine weitaus subtilere Analyse. Solche Äußerungen können auf Satzebene grammatikalisch vollkommen korrekt sein, sind aber im Kontext oder Diskurs, in dem sie verwendet werden, semantisch unpassend oder bedeutungslos.

Corder (1982, s. 40) betont diesen Sachverhalt und stellt fest, dass selbst ein Lernender, der einen grammatikalisch einwandfreien Satz bildet, einen Fehler machen kann. Wenn der gebildete Satz nicht zur Situation oder zum Kontext passt, in dem er geäußert wird, muss er trotz seiner strukturellen Korrektheit als Fehler gewertet werden. Zum Beispiel wäre die Äußerung *Was für ein fröhlicher Tag!* bei einer Beerdigung, obwohl grammatikalisch korrekt, ein schwerwiegender kontextueller Fehler. Diese Unterscheidung zeigt, dass die Fehleranalyse nicht nur auf der Satzebene, sondern auch auf der Diskursebene durchgeführt werden muss.

Die Grundlage für die Bestimmung der Existenz und Art eines Fehlers liegt in der korrekten Interpretation der Intention des Lernenden. Laut Corder (1982, s. 37) ist die Fehlererkennung ein Prozess des Vergleichs zwischen der tatsächlich geäußerten Aussage des Lernenden und der Bedeutung, die er eigentlich ausdrücken wollte. Der Analytiker führt eine Rekonstruktion durch, indem er die Frage stellt: *Wie würde ein Muttersprachler dies sagen, wenn er diese Bedeutung korrekt ausdrücken wollte?* Folglich wird die fehlerhafte Äußerung des Lernenden mit einer hypothetischen korrekten Äußerung verglichen, die dieselbe semantische Intention trägt. Der Erfolg dieses Prozesses hängt direkt davon ab, wie treffend der Analytiker die Absicht des Lernenden interpretieren kann.

An diesem Punkt wird deutlich, dass die Fehleranalyse kein rein mechanischer Korrekturvorgang ist, sondern eine tiefgreifende hermeneutische Anstrengung erfordert. Der

Analytiker versucht, in die mentale Welt des Lernenden einzutreten und die Bedeutung aus dessen Perspektive zu rekonstruieren.

Eine weitere komplexe Dimension bei der Bewertung von Fehlern ist die soziale Angemessenheit. Corder (1982, s. 41) betont, dass der Sprachgebrauch nicht nur von grammatischen Regeln, sondern auch von der sozialen Situation sowie der Stil- und Registerwahl (formell, informell etc.) abhängt. Der sozial normgerechte Gebrauch von Sprache ist ein Kompetenzbereich, der über die grammatische Korrektheit hinausgeht. In Corders Beispiel *Na, wie geht's uns denn heute, alter Mann?*[12] wäre es, obwohl grammatikalisch akzeptabel, sozial äußerst unpassend, wenn ein Lernender diesen Satz zu seinem Lehrer sagen würde.

Um diese Situation zu systematisieren, präsentiert Corder (1982, s. 41) ein Modell, in dem Äußerungen auf zwei Achsen bewertet werden können: grammatische Akzeptabilität und soziale Angemessenheit. Diese beiden Achsen ergeben vier mögliche Fälle:

1. Äußerungen, die sowohl grammatisch als auch sozial korrekt sind (akzeptabel und angemessen).
2. Äußerungen, die grammatisch fehlerhaft, aber von der sozialen Intention her korrekt sind (inakzeptabel, aber angemessen).
3. Äußerungen, die grammatisch korrekt, aber sozial fehlerhaft sind (akzeptabel, aber unangemessen).
4. Äußerungen, die sowohl grammatisch als auch sozial fehlerhaft sind (sowohl inakzeptabel als auch unangemessen).

Diese Klassifikation zeigt deutlich die mehrdimensionale Natur von Fehlern und die Bedeutung der pragmatischen Kompetenz im Spracherwerb.

[12] Eng. *Well, how are we today, old man?*

Im Prozess des Verstehens, was der Lernende meinte, stehen dem Analytiker zwei grundlegende Methoden zur Verfügung. Corder (1982, s. 37f) bezeichnet diese Methoden als *autoritative Interpretation* und *plausible Interpretation*:

Autoritative Interpretation: Wenn die Möglichkeit besteht, den Lernenden direkt zu kontaktieren, kann der Analytiker den Lernenden fragen, was er meinte, und ihn bei Bedarf bitten, es in seiner Muttersprache auszudrücken. Dies ist die zuverlässigste Methode, um die Absicht des Lernenden zu verstehen. Die einzige Schwierigkeit bei dieser Methode besteht darin festzustellen, ob der (vielleicht zufällig korrekte) Satz des Lernenden ein bewusstes Konstrukt oder ein Zufallstreffer war.

Plausible Interpretation: In Fällen, in denen der Lernende nicht erreichbar ist (z. B. bei der Analyse eines schriftlichen Textes), muss der Analytiker auf der Grundlage des sprachlichen und situativen Kontexts des Satzes eine logische Interpretation vornehmen. Diese Situation ist weitaus komplexer und birgt verschiedene Unsicherheiten.

Corder (1982, s. 42ff) detailliert die komplexen Szenarien, die im plausiblen Interpretationsprozess auftreten können. Selbst wenn der vom Lernenden gebildete Satz strukturell korrekt ist, kann der Analytiker diesen Satz richtig interpretieren, falsch interpretieren, ihn für mehrdeutig halten oder ihm im Kontext keine Bedeutung beimessen. Wenn der Satz des Lernenden offensichtlich fehlerhaft (strukturell defekt) ist, werden die Szenarien noch komplizierter: Der Analytiker kann aus diesem fehlerhaften Satz die ursprüngliche Absicht richtig erraten, eine falsche Bedeutung ableiten, zwischen mehreren möglichen Bedeutungen unentschlossen bleiben oder den Satz als völlig bedeutungslos empfinden.

Wie Corder (1982, s. 44) unterstreicht, ist das alleinige Kriterium in der Fehleranalyse nicht die strukturelle Korrektheit des Satzes. Das Wichtige ist, ob die Äußerung des Lernenden den kontextuellen und situativen Erwartungen der Zielsprache entspricht. Aus diesem Grund birgt selbst jeder gut strukturierte Satz potenziell einen Fehler, bis die ursprüngliche Absicht

des Lernenden sorgfältig geprüft wurde. Der Erfolg der Fehleranalyse hängt vollständig von der Treffsicherheit der Interpretationen ab und davon, wie genau der Analytiker die Absicht des Lernenden rekonstruieren kann.

Arten der Fehler

Die Analyse der im Sprachlernprozess auftretenden Fehler erfordert deren korrekte Klassifikation. In der Literatur werden Fehler nach verschiedenen Kriterien unterschieden, die sowohl ihre Quelle als auch ihre Auswirkung auf die Kommunikation berücksichtigen. Zu den grundlegendsten dieser Unterscheidungen gehören die zwischen Kompetenz- und Performanzfehlern sowie zwischen globalen und lokalen Fehlern. Das Verständnis dieser Konzepte ist von entscheidender Bedeutung für eine tiefere und präzisere Fehleranalyse.

Nach der von Chomsky (1965, s. 4) in die Linguistik eingeführten grundlegenden Unterscheidung bezeichnet die Kompetenz das abstrakte und internalisierte Wissenssystem eines Sprechers über seine Sprache, während die Performanz die tatsächliche Anwendung dieses Wissens in konkreten Situationen ist. Auf der Grundlage dieses Unterschieds können auch Fehler klassifiziert werden.

Performanzfehler resultieren nicht aus einem Mangel im grammatischen System des Lernenden, sondern aus Problemen, die bei der Anwendung dieses Wissens auftreten. Corder (1967, s. 166) stellt fest, dass solche Fehler charakteristischerweise nicht-systematisch sind, d.h. sie treten zufällig und inkonsistent auf. Forscher wie Richards (1971, s. 12) und Touchie (1986, s. 76) betonen, dass Performanzfehler Teil des normalen Sprachgebrauchs sind, den selbst Muttersprachler in vorübergehenden Zuständen wie Müdigkeit, Eile oder Unaufmerksamkeit machen können. Daher deuten solche Fehler in der Regel nicht auf einen schwerwiegenden Lernmangel hin und können vom Lernenden leicht korrigiert werden, wenn sie bemerkt werden.

Ellis (1994, s. 58) unterteilt Performanzfehler in zwei Unterkategorien: Erstens, Fehler wie Lapsus, die aus verschiedenen mentalen Verarbeitungsproblemen im Moment des Sprechens resultieren. Zweitens, Fehler, die aus den Kommunikationsstrategien des Lernenden entstehen, um Wissenslücken zu kompensieren (z. B. der Versuch, ein unbekanntes Wort zu umschreiben). Lott (1983, s. 257) legt nahe, dass Performanzfehler Ähnlichkeiten mit Corders postsystematischen Fehlern aufweisen; beide beschreiben Situationen, in denen die Regel bekannt ist, aber in der Anwendung Mängel auftreten.

Im Gegensatz zu Performanzfehlern spiegeln Kompetenzfehler einen Mangel oder eine Ungenauigkeit im mentalen System des Lernenden über die Zielsprache wider. Laut Touchie (1986, s. 76) sind diese Fehler weitaus ernster als Performanzfehler, da sie ein Indikator für unzureichendes Lernen sind. Corder (1967, s. 166) gibt an, dass das Hauptmerkmal von Kompetenzfehlern ihre Systematik ist. Das heißt, der Lernende wiederholt denselben Fehler konsequent, weil er eine bestimmte Regel falsch gelernt hat. Lott (1983, s. 257) stellt ebenfalls fest, dass solche Fehler mit Corders Konzept der systematischen Fehler übereinstimmen.

Richards hat drei Hauptquellen identifiziert, die Kompetenzfehlern zugrunde liegen (zitiert nach Ellis, 1994, s. 58):

1. Interferenzfehler: Das Auftreten von Strukturen oder Regeln aus der Muttersprache des Lernenden bei der Verwendung der Zielsprache, was zu Fehlern führt. Dies ist auch als klassischer Einfluss der Muttersprache bekannt.
2. Intralinguale Fehler: Diese Fehler stammen nicht aus der Muttersprache, sondern aus der inneren Struktur der zu lernenden Sprache selbst. Fälle wie die falsche Verallgemeinerung einer Regel, die der Lernende in der Zielsprache gelernt hat (z. B. das Anhängen der gleichen Vergangenheitsendung an alle Verben), die unvollständige Anwendung von Regeln oder das Unvermögen, die genauen Bedingungen für die Gültigkeit einer Regel zu lernen, fallen in diese Kategorie.

3. Entwicklungsfehler: Fehler, die als natürlicher Teil des Spracherwerbsprozesses entstehen, wenn der Lernende auf der Grundlage seines begrenzten Sprachwissens falsche Hypothesen über die Zielsprache bildet. Diese Fehler zeigen Ähnlichkeiten mit den Fehlern, die Kinder beim Erlernen ihrer Muttersprache machen.

Ein weiterer wichtiger Ansatz zur Klassifizierung von Fehlern konzentriert sich nicht auf ihre Quelle, sondern auf ihre Auswirkung auf die Kommunikation. Burt und Kiparsky (1975) haben auf dieser Grundlage eine hierarchische Unterscheidung von Fehlern in globale und lokale Fehler vorgenommen (zitiert nach Corder, 1975, s. 207).

Globale Fehler sind, wie der Name schon sagt, Fehler, die die Bedeutung ernsthaft beeinträchtigen, indem sie die Gesamtstruktur des Satzes oder die Beziehungen zwischen den Hauptbestandteilen stören (Burt & Kiparsky, 1974, zitiert nach Lennon, 1991, s. 183). Richards und Schmidt (2011, s. 247) stellen fest, dass solche Fehler einen Satz oder eine Äußerung schwer oder unmöglich zu verstehen machen. Laut Burt (1975, s. 6) behindern diese Fehler die Kommunikation erheblich, weil sie den Hörer oder Leser veranlassen, die Botschaft falsch zu interpretieren (Burt, 1975, s. 7f). Fehler, die die allgemeine Organisation des Satzes betreffen, wie z. B. eine falsche Wortstellung, fallen in diese Kategorie und haben einen deutlichen Einfluss auf die Verständlichkeit (Ellis, 1994, s. 704). Kurz gesagt, globale Fehler können die Integrität der Botschaft stören und zu einem Abbruch der Kommunikation führen.

Lokale Fehler sind Fehler, die nicht die Gesamtstruktur des Satzes stören, sondern nur ein einzelnes Element oder einen kleinen Teil des Satzes betreffen (Burt, 1975, s. 7). Laut Richards und Schmidt (2011, s. 247) führen diese Fehler in der Regel nicht zu Verständlichkeitsproblemen. Fehler bei der Deklination von Substantiven und Konjugation von Verben oder bei der Verwendung von Artikeln oder Hilfsverben können als Beispiele für diese Kategorie angeführt werden (Burt, 1975, s. 7; Ellis, 1994, s. 713). Wie Brown (2007, s. 263) unter Berufung auf Burt und Kiparsky berichtet, behindern lokale Fehler das

Verständnis der Botschaft nicht, weil sie in der Regel nur einen kleinen Teil des Satzes verletzen, was es dem Hörer/Leser ermöglicht, eine korrekte Vermutung über die beabsichtigte Bedeutung anzustellen.

Burt und Kiparsky fügen hinzu, dass diese beiden Konzepte relativ sind; ein Fehler, der in einem Satz global ist, kann zu einem lokalen Fehler werden, wenn dieser Satz in einen größeren Satz eingebettet wird (zitiert nach Lennon, 1991, s. 183). Diese Unterscheidung bietet einen wichtigen pädagogischen Leitfaden für Lehrkräfte, welche Fehler priorisiert werden sollten: Die Korrektur von globalen Fehlern, die die Kommunikation vollständig stören, hat in der Regel Vorrang vor lokalen Fehlern, die das Verständnis der Botschaft meist nicht verhindern.

Fehlerquellen

Die Analyse der im Zweitspracherwerbsprozess gemachten Fehler zielt nicht nur darauf ab, die Art der Fehler zu bestimmen, sondern auch die zugrunde liegenden Ursachen und Quellen zu verstehen. Laut Brown (2007, s. 232) bildet die Entschlüsselung dieser komplexen Beziehung zwischen den kognitiven und affektiven Prozessen des Lernenden und dem linguistischen System die Grundlage für ein ganzheitliches Verständnis des Zweitspracherwerbsprozesses. In diesem Zusammenhang untersucht die Fehleranalyse-Literatur die Fehlerquellen hauptsächlich unter zwei Hauptüberschriften: interlingualer Transfer und intralingualer Transfer (Erdoğan, 2005, s. 265).

Bevor jedoch auf diese beiden Hauptquellen eingegangen wird, sollte erwähnt werden, dass Fehlerquellen auch aus einer breiteren Perspektive betrachtet werden können. Taylor hat vier verschiedene Kategorien definiert, um die Mehrdimensionalität von Fehlerquellen zu verdeutlichen (zitiert nach Ellis, 1994, s. 57f):

1. Psycholinguistische Quellen: Beziehen sich auf die Natur des Zweitsprachen-Wissenssystems im Gehirn des Lernenden und die mentalen

Verarbeitungsschwierigkeiten, die bei der Anwendung dieses Wissens im Sprechen oder Schreiben auftreten.

2. Soziolinguistische Quellen: Resultieren aus den Mängeln des Lernenden, die Sprache an den jeweiligen sozialen Kontext (z. B. Formalität, Vertrautheit) anzupassen.
3. Epistemische Quellen: Fehler, die nicht aus der Sprache selbst, sondern aus dem mangelnden Weltwissen des Lernenden über das Thema, das er auszudrücken versucht, entstehen.
4. Diskursive Quellen: Fehler, die aus Problemen bei der Organisation von Informationen zu einem kohärenten und zusammenhängenden Text resultieren, auch wenn die Sätze auf Satzebene korrekt sind.

Ellis (1994, s. 57f) stellt fest, dass sich die Zweitspracherwerbsforschung im Allgemeinen auf die erste dieser Quellen konzentriert, nämlich auf die psycholinguistischen Prozesse, was uns zum Konzept des Transfers führt.

Transfer ist ein allgemeiner Begriff, der die Übertragung von zuvor gelerntem Wissen oder einer zuvor erlernten Leistung auf einen späteren Lernprozess bezeichnet (Brown, 2007, s. 102). Im Kontext des Sprachenlernens kann dieser Transfer den Lernprozess entweder erleichtern oder erschweren. Dementsprechend wird der Transfer in positiven und negativen Transfer unterteilt:

Positiver Transfer tritt auf, wenn früheres Wissen (in der Regel aus der Muttersprache) dem Erlernen der neuen Sprache zugutekommt. Die Ähnlichkeit von Regeln oder Strukturen zwischen zwei Sprachen ermöglicht es dem Lernenden, korrekte Verbindungen herzustellen und beschleunigt das Lernen (Brown, 2007; Jabeen et al., 2015).

Negativer Transfer (Interferenz) ist der Fall, in dem früheres Wissen die Leistung bei einer neuen Aufgabe behindert. Brown (2007, s. 102) zufolge wird dieser Zustand auch als

Interferenz bezeichnet, da ein Element aus der Muttersprache fälschlicherweise auf die Zielsprache übertragen wird. Schwierigkeiten und Fehler des Lernenden aufgrund unterschiedlicher Regeln zwischen den beiden Sprachen sind ein Ergebnis des negativen Transfers (Jabeen et al., 2015, s. 55).

Der interlinguale Transfer ist die häufigste und offensichtlichste Form des negativen Transfers. Erdoğan (2005, s. 265) definiert diese Fehlerart als Fehler, die durch die Interferenz von Gewohnheiten und Regeln aus der Muttersprache des Lernenden in die Zweitsprache entstehen. Diese Interferenz kann auf allen Ebenen der Sprache auftreten, wie der phonologischen (Aussprache), morphologischen (Affixe), grammatikalischen (Satzstruktur) und lexikalisch-semantischen (Wort und Bedeutung).

Brown (2007, s. 232) betont, dass der interlinguale Transfer eine bedeutende Fehlerquelle ist, insbesondere für Lernende in den Anfangsstadien des Spracherwerbs. In dieser Phase ist das Wissenssystem des Lernenden über die Zielsprache noch sehr schwach, weshalb seine eigene Muttersprache die einzige verlässliche linguistische Quelle ist, auf die er zurückgreifen kann. Daher neigt der Lernende bewusst oder unbewusst dazu, die Muster seiner Muttersprache auf die Zielsprache zu übertragen. An diesem Punkt ist es ein großer pädagogischer Vorteil, wenn der Lehrer die Muttersprache des Lernenden kennt, um solche Fehler zu diagnostizieren und ihre Quelle zu erklären.

Mit fortschreitendem Spracherwerb treten neben den durch den Einfluss der Muttersprache verursachten Fehlern auch Fehler auf, die direkt aus dem System der Zielsprache selbst resultieren. Erdoğan (2005, s. 265) definiert *intralinguale Transferfehler* als Fehler, die durch falsches Verständnis oder Übergeneralisierung der Regeln der Zweitsprache entstehen. Diese Fehler treten unabhängig von der Muttersprache als Ergebnis fehlerhafter Hypothesen und Verallgemeinerungen auf, die der Lernende beim Erlernen der Regeln der Zielsprache bildet (Erdoğan, 2005, s. 266). Wenn beispielsweise ein Lernender, der die

Vergangenheitsendung *-ed* für regelmäßige Verben im Englischen gelernt hat, diese Regel auch auf das Verb *go* anwendet und *goed* sagt, ist dies ein typischer intralingualer Fehler.

Im Sprachlernprozess besteht eine dynamische Beziehung zwischen diesen beiden Fehlerquellen. Brown (2007, s. 232) stellt fest, dass in den frühen Phasen des Lernens interlinguale Fehler (Einfluss der Muttersprache) dominieren, während mit zunehmender Kompetenz des Lernenden in der Zielsprache intralinguale Fehler (Verallgemeinerungen innerhalb der Zielsprache) immer mehr in den Vordergrund treten. In ähnlicher Weise beobachtete auch Ellis (1994, s. 61f), dass mit zunehmender Sprachkompetenz des Lernenden die interlingualen Transferfehler abnehmen, während der Anteil der intralingualen Fehler zunimmt. Dies ist ein positives Entwicklungszeichen, das darauf hindeutet, dass der Lernende nun weniger von seiner Muttersprache abhängig ist und aktiv Hypothesen innerhalb des Systems der Zielsprache aufstellt und testet.

Zusätzlich zum inter- und intralingualen Transfer weist Brown (2007, s. 234) auf den Lernkontext als dritte Fehlerquelle hin. Der Lernkontext umfasst das Klassenzimmer, den Lehrer, die verwendeten Lehrmaterialien oder das soziale Umfeld. Manchmal resultiert ein Fehler nicht aus dem Lernenden selbst, sondern aus irreführenden Hinweisen der Lernumgebung, in der er sich befindet. Beispielsweise kann eine unvollständige oder irreführende Grammatikerklärung des Lehrers, ein fehlerhaftes Beispiel im Lehrbuch oder die Verallgemeinerung eines auswendig gelernten Dialogs auf falsche Kontexte dazu führen, dass der Lernende fehlerhafte Hypothesen über die Sprache bildet. Daher sollte bei der Fehleranalyse auch der Lehrprozess berücksichtigt werden, dem der Lernende ausgesetzt ist.

Feedback

Feedback, das als eine der stärksten Einflussquellen auf das Lernen gilt, wird in seiner allgemeinsten Form als Information definiert, die von einem Vermittler (z. B. einem Lehrer, einem Mitschüler, einem Buch, den Eltern, der Person selbst oder einer Erfahrung) über

Aspekte der Leistung oder des Verständnisses eines Lernenden bereitgestellt wird (Hattie & Timperley, 2007, s. 81). Diese Information zielt darauf ab, den Lernenden zu helfen, die Lücke zwischen ihrer aktuellen Leistung und der angestrebten Leistung zu schließen (Nicol & Macfarlane-Dick, 2006, zitiert nach Lee, 2017, s. 56).

Der Begriff Feedback wird überwiegend im Kontext des korrektiven Feedbacks behandelt und als ein Phänomen definiert, das:

- eine Antwort auf die Äußerung eines Lernenden ist, die einen sprachlichen Fehler enthält (Ellis, 2009; Ellis, Loewen & Erlam, 2006; Lyster, Saito & Sato, 2013)
- aus Elementen besteht wie dem Hinweis auf einen Fehler, der Bereitstellung der korrekten Form oder der Vermittlung grammatischer Informationen über den Fehler (Ellis, 2009; Ellis et al., 2006),
- je nach Explizitheit des gegebenen Hinweises als implizit oder explizit klassifiziert wird (Ellis, 2009; Ellis et al., 2006),
- je nachdem, ob die korrekte Form bereitgestellt wird oder nicht, direkte oder indirekte Strategien umfasst (Bitchener, 2008; Bitchener, Young & Cameron, 2005; Ellis, 2008; Ellis, Sheen, Murakami & Takashima, 2008),
- in zwei Hauptkategorien unterteilt werden kann: Neuformulierungen, bei denen die korrekte Form angeboten wird, und Prompts, die den Lernenden zur Selbstkorrektur anregen (Ellis, 2009; Lyster et al., 2013).

Im speziellen Kontext der Schreibdidaktik kann Feedback als eine pädagogische Gattung betrachtet werden, die Kommentare sowohl zum Inhalt als auch zur Form eines Textes liefert, um die Studierenden zur Verbesserung ihrer Schreibfähigkeiten und zur Festigung ihres Lernens zu ermutigen. Dieser Prozess geht über die reine Bewertung eines fertigen Textes hinaus; er gibt auch Aufschluss für zukünftige Texte der Studierenden, hilft ihnen,

den Schreibkontext zu verstehen, und bietet ihnen eine Leserperspektive (Hyland & Hyland, 2019, s. 165).

Nach dem Modell von Hattie und Timperley (2007) beantwortet ein effektiver Feedbackprozess drei grundlegende Fragen:

1. *Wohin gehe ich?*[13]:Diese Phase bezieht sich auf die Klärung der Lernziele und Erfolgskriterien. Die Lernenden verstehen, was sie erreichen sollen.
2. *Wie komme ich voran?*[14]: Diese Phase liefert Informationen über die Lücke zwischen der aktuellen Leistung des Lernenden und den Zielen. Das Feedback sollte darauf abzielen, diese Lücke zu schließen.
3. *Was kommt als Nächstes?*[15]: Diese Phase konzentriert sich auf die Schritte, die der Lernende unternehmen muss, um seine zukünftige Leistung zu verbessern. Das Feedback enthält Anleitungen, die den Lernenden voranbringen (zitiert nach Lee, 2017, s. 4f).

In diesem Rahmen ist Feedback nicht nur ein Akt der Benotung eines Textes, sondern auch ein fortlaufender Dialog zwischen Lehrkraft und Lernendem und ein dynamischer Prozess, der das Lernen unterstützt (Hyland & Hyland, 2019, s. 171).

Arten von Feedback

Feedback wird je nach dem Medium, in dem es gegeben wird, grundsätzlich in mündliches und schriftliches Feedback unterteilt (Lee, 2017, s. 58). Beide Arten haben ihre eigenen spezifischen Vorteile und Einschränkungen.

[13] Feed Up

[14] Feed Back

[15] Feed Forward

Mündliches korrektives Feedback

Mündliches korrektives Feedback wird in der Regel unmittelbar während Interaktionen im Klassenzimmer oder in Einzelgesprächen gegeben (Lee, 2017, s. 71). Der offensichtlichste Vorteil dieser Art von Feedback ist die Möglichkeit, sofort auf Fehler der Lernenden einzugreifen und durch wechselseitige Interaktion die Bedeutung auszuhandeln (Goldstein, 2004, s. 65). Die Forschung hat gezeigt, dass mündliches Feedback in verschiedenen Formen auftreten kann. Einige davon wurden als Umformung, Klärungsbitte, metalinguistisches Feedback, Elizitierung und Wiederholung kategorisiert (Lightbown & Spada, 2013, s. 140f).

Diese Feedbackarten können grundsätzlich in zwei Hauptgruppen unterteilt werden: Neuformulierungen und Prompts. Neuformulierungen sind Feedbackformen, die die korrekte Form der fehlerhaften Äußerung des Lernenden präsentieren. Zu dieser Kategorie gehören Recast und explizite Korrektur. Prompts hingegen sind verschiedene Signale, die den Lernenden dazu anregen, seinen eigenen Fehler zu korrigieren, anstatt die korrekte Form direkt zu liefern. Zu dieser Kategorie gehören Klärungsbitten, metalinguistisches Feedback, Elizitierung und Wiederholung (Brown, 2014; Lyster et al., 2013).

Darüber hinaus kann dieses Feedback auch als implizit oder explizit klassifiziert werden, je nachdem, wie offen die Aufmerksamkeit des Lernenden auf den Fehler gelenkt wird (Li & Vuono, 2019; Sheen, 2010).

Umformung

Eine Umformung *(Recast)* ist die Neuformulierung einer fehlerhaften Äußerung eines Lernenden durch die Lehrkraft, bei der die gesamte Äußerung oder ein Teil davon korrekt wiedergegeben wird, ohne die Bedeutung zu verändern und den Fehler zu eliminieren (Lyster & Ranta, 1997; Lyster et al., 2013; Nicholas, Lightbown & Spada, 2001; Rassaei,

2015). Er wird im Allgemeinen als eine implizite Form des Feedbacks betrachtet, da die Korrekturabsicht der Lehrkraft nicht explizit geäußert wird. Studien zeigen, dass dies die von Lehrkräften am häufigsten verwendete Feedbackart ist (Brown, 2014; Li & Vuono, 2019).

Klärungsbitte

Bei einer Klärungsbitte verwendet die Lehrkraft Ausdrücke wie *Wie bitte?, Ich habe das nicht verstanden* oder *Könntest du das wiederholen?,* um anzuzeigen, dass sie die Äußerung des Lernenden nicht verstanden hat oder dass die Äußerung grammatikalisch problematisch ist. Mit dieser Methode wird der Fehler indirekt signalisiert und vom Lernenden erwartet, seine Äußerung verständlicher oder korrekter zu wiederholen (Lochtman, 2002, s. 277). Dies ist eine Art von Prompt, der den Lernenden zur Selbstkorrektur anregt.

Metalinguistisches Feedback

Metalinguistisches Feedback liegt vor, wenn die Lehrkraft Kommentare, Informationen oder Fragen zur Richtigkeit der Äußerung des Lernenden liefert, ohne die korrekte Form direkt zu nennen (Lochtman, 2002; Lyster & Ranta, 1997). Dieses Feedback lenkt die Aufmerksamkeit auf die Natur des Fehlers, indem es beispielsweise an eine grammatische Regel erinnert oder einen Fachbegriff verwendet (Li & Vuono, 2019, s. 7). Es ist eine explizite Form des Feedbacks und ermutigt den Lernenden zur Selbstkorrektur (Rassaei, 2015, s. 99).

Elizitierung

Bei der Elizitierung fordert die Lehrkraft die korrekte Form direkt vom Lernenden ein (Lyster & Ranta, 1997, s. 48). Dies kann geschehen, indem die Lehrkraft den Satz bis zur

fehlerhaften Stelle wiederholt und pausiert, eine Frage stellt oder den Lernenden bittet, den Satz neu zu formulieren. Diese Methode ermutigt den Lernenden, die Lücke zu füllen oder den Fehler zu finden und zu korrigieren (Lochtman, 2002, s. 277).

Wiederholung

Bei der Wiederholung wiederholt die Lehrkraft die fehlerhafte Äußerung des Lernenden isoliert, oft mit einer steigenden Intonation, um den Fehler zu betonen (Lochtman, 2002; Lyster & Ranta, 1997). Diese Methode lenkt die Aufmerksamkeit des Lernenden auf den Fehler und bietet eine Gelegenheit zur Selbstkorrektur (Lyster et al., 2013, s. 21).

Schriftliches korrektives Feedback

Das schriftliche korrektive Feedback hingegen besteht aus Kommentaren, Codes oder Korrekturen, die von der Lehrkraft auf dem Text des Lernenden angebracht werden. Einer der wichtigsten Vorteile des schriftlichen Feedbacks ist, dass es dem Lernenden Zeit gibt, das Feedback zu prüfen, zu verstehen und umzusetzen (Bruton, 2010, s. 493). Im Gegensatz zum mündlichen Feedback ist das schriftliche Feedback dauerhaft und kann vom Lernenden wiederholt überprüft werden (Bruton, 2010, s. 495). Diese Eigenschaft ist besonders wichtig für das Verständnis und die Internalisierung komplexer grammatischer Themen.

Das schriftliche Feedback, und genauer gesagt das schriftliche korrektive Feedback, sind schriftliche Markierungen, Codes oder Kommentare, die von einer Lehrkraft auf einem von einem Lernenden verfassten Text angebracht werden, um morphologische, syntaktische und lexikalische Formen zu behandeln, die von den Regeln der Zielsprache abweichen (Ferris, 2011; Ferris & Kurzer, 2019). Diese pädagogische Praxis beschränkt sich nicht nur auf die Korrektur von Fehlern in einem Text; sie erfüllt auch eine informative, pädagogische und

interpersonale Funktion, die den Lernenden hilft, ihre Schreibfähigkeiten zu entwickeln und ihr Lernen zu festigen (Hyland & Hyland, 2019, s. 165).

Das schriftliche korrektive Feedback sollte nicht nur als eine objektive Interaktion zwischen einer Lehrkraft und einem Text betrachtet werden, sondern auch als eine soziale Handlung, die mit dem Autor, d.h. dem Lernenden, eingegangen wird (Hyland & Hyland, 2019; Lee, 2017). Die Art und Weise, wie die Lehrkraft das Feedback formuliert, enthält Annahmen über die Beziehung zwischen Lehrendem und Lernendem und kann die Reaktion des Lernenden auf dieses Feedback sowie das Ausmaß, in dem er es in seinen Revisionen verwendet, direkt beeinflussen (Hyland & Hyland, 2019, s. 165). Daher ist das schriftliche korrektive Feedback ein komplexer und mehrdimensionaler Prozess, der nicht nur darauf abzielt, die grammatische Korrektheit zu verbessern, sondern auch den Lernenden als Autor zu fördern und ihn im Lernprozess zu begleiten.

Historische Entwicklung

Die Praxis des schriftlichen korrektiven Feedbacks ist eng mit der Evolution der pädagogischen Ansätze im L2-Schreibunterricht verbunden. Dieser Prozess hat sich von einem strengen Präskriptivismus hin zu lernerzentrierten Ansätzen und schließlich zu einer ausgewogeneren, auf empirischen Belegen basierenden Perspektive entwickelt.

Bis in die 1970er Jahre wurde das L2-Schreiben hauptsächlich als eine Sprachübung zum Einüben grammatischer Formen und zur Verwendung neuer Vokabeln betrachtet. In dieser Zeit korrigierten Lehrkräfte, beeinflusst von Theorien der behavioristischen Psychologie und der strukturellen Linguistik, kontinuierlich alle Fehler der Lernenden, um die Bildung schlechter Gewohnheiten zu verhindern (Ferris, 2011, s. 7f).

In den 1970er und 1980er Jahren kam es jedoch zu einem großen Paradigmenwechsel, der als prozessorientierter Ansatz bekannt wurde. Dieser Ansatz legte den Schwerpunkt nicht auf die Korrektheit des Endprodukts, sondern auf die Prozesse der Autoren wie

Ideenfindung, Entwerfen und Überarbeiten. Themen wie Grammatik und Lektorat wurden an das Ende des Schreibprozesses verlegt. Dies führte zu einer Phase der wohlwollenden Vernachlässigung von Fehlern durch Lehrende und Lernende. Bald darauf äußerten einige Akademiker ihre Besorgnis über diese Vernachlässigung und stellten fest, dass die sprachlichen Probleme von L2-Autoren nicht auf magische Weise durch den Prozess verschwanden und dass die Lernenden Interventionen benötigten (Ferris, 2011, s. 8f).

Der wichtigste Wendepunkt in diesem Bereich war der 1996 von John Truscott veröffentlichte Artikel mit dem Titel *The Case Against Grammar Correction in L2 Writing Classes*. Truscott argumentierte, dass Grammatikkorrekturen unwirksam, ja sogar schädlich seien und vollständig aufgegeben werden sollten. Seine Argumente waren wie folgt:

- Die Forschung zeigt, dass Korrekturen nicht funktionieren (Truscott, 1996, s. 328).
- Die Spracherwerbstheorie und praktische Probleme legen nahe, dass wir ein Scheitern der Korrektur erwarten sollten. Die Korrektur ignoriert die schrittweise Natur des Lernens, Entwicklungssequenzen und das Risiko eines oberflächlichen Pseudolernens (Truscott, 1996, s. 342, 344f).
- Korrekturen haben schädliche Auswirkungen auf die Lernenden, wie z. B. die Erzeugung von Stress, die Verringerung der Motivation und die Förderung von Vermeidungsstrategien (Truscott, 1996, s. 354).
- Die Argumente für die Beibehaltung der Korrektur (z. B. dass sie die Fossilisierung verhindere oder dass die Lernenden sie wünschten) sind ungültig (Truscott, 1996, s. 357).

Die stärkste Antwort auf diese Behauptungen kam von Dana Ferris (1999). Ferris erklärte, dass Truscotts Schlussfolgerungen voreilig und überzogen seien (S. 2). Hauptkritikpunkte von Ferris waren:

- Truscott verwendete den Begriff der Grammatikkorrektur zu allgemein und ignorierte wichtige Unterschiede zwischen verschiedenen Arten von Feedback. Laut Ferris kann jedoch eine effektive Korrektur, die selektiv, priorisiert und klar ist, den Studierenden helfen (S. 4).
- Truscotts Forschungsüberblick war fehlerhaft; er umfasste unvergleichbar unterschiedliche Lernergruppen und Lehrmethoden und ignorierte Studien, die seine eigene These nicht stützten (S. 5).

Diese Debatte führte zur Durchführung rigoroserer und kontrollierterer Studien in dem Feld (Ferris, 2004, s. 56f, zitiert nach Ferris & Kurzer, 2019, s. 107). Ab den 2000er Jahren zeigten kontrollierte Studien mit einem Prätest-Posttest-verzögerten Posttest-Design, die von Forschern wie Bitchener, Sheen und Ellis durchgeführt wurden, konsistent, dass ein auf spezifische grammatische Merkmale (z. B. Artikel im Englischen) ausgerichtetes schriftliches korrektives Feedback die Korrektheit der Lernenden im Laufe der Zeit verbessert (Ferris & Kurzer, 2019, s. 106f).

Während diese neue Forschungswelle starke empirische Belege gegen Truscotts Behauptungen lieferte, wiesen Kritiker wie Anthony Bruton (2009) darauf hin, dass diesen Studien die ökologische Validität fehle, da sie künstliche Aufgaben verwendeten und nicht die reale Unterrichtsumgebung widerspiegelten (S. 601). Diese Kritik verdeutlichte den Unterschied zwischen den Zielen von Zweitsprachenerwerb-Forschern (Erwerb spezifischer linguistischer Strukturen) und den Zielen von L2-Schreibpraktikern (Verbesserung der Gesamtwirksamkeit von Texten und Entwicklung von Selbstkorrekturstrategien der Lernenden) (Ferris, 2010, s. 181).

Heute hat sich die Debatte von der Frage *Funktioniert Korrektur?* zu der Frage *Wie wird sie am besten durchgeführt und für wen funktioniert sie?* entwickelt (Ferris & Kurzer, 2019, s. 109ff). Der allgemeine Konsens ist nun, dass schriftliches korrektives Feedback unter den richtigen Bedingungen wirksam sein kann. Forscher konzentrieren sich jetzt auf individuelle

Unterschiede (Kompetenzniveau, Motivation, Überzeugungen), kontextuelle Faktoren und neue pädagogische Ansätze wie das dynamische schriftliche korrektive Feedback[16] (Ferris & Kurzer, 2019, s. 106).

Direktes vs. Indirektes Feedback

Dies ist eine der grundlegendsten Unterscheidungen in der Literatur des schriftlichen korrektiven Feedbacks und bezieht sich auf den Grad der Explizitheit des Feedbacks (Ferris & Kurzer, 2019, s. 109).

- Direktes korrektives Feedback: Die Art von Feedback, bei der die Lehrkraft oder der Forscher die korrekte sprachliche Form über oder in der Nähe des Fehlers des Lernenden bereitstellt (Ferris, 2011; Lee, 2017). Dies kann in Form des Durchstreichens eines unnötigen Wortes, des Einfügens eines fehlenden Wortes oder des Schreibens der korrekten Form über die falsche Form geschehen (Lee, 2017, s. 68). Bei diesem Ansatz wird vom Lernenden erwartet, dass er bei der Überarbeitung lediglich die Korrekturen der Lehrkraft in den Text übernimmt (Ferris, 2011, s. 31f).
- Indirektes korrektes Feedback: Die Art von Feedback, bei der die Lehrkraft anzeigt, dass ein Fehler gemacht wurde, aber nicht die korrekte Form liefert, wodurch die Aufgabe, das Problem zu lösen und den Fehler zu korrigieren, dem Lernenden überlassen wird (Ferris, 2011; Lee, 2017). Es wird argumentiert, dass dieser Ansatz ein höheres kognitives Engagement fördert, indem er den Lernenden in einen Prozess des angeleiteten Lernens und der Problemlösung einbezieht (Lalande, 1982, s. 140).

Die Forschung zur Wirksamkeit dieser beiden Ansätze hat unterschiedliche Ergebnisse erbracht. Zweitsprachenerwerbsorientierte Studien haben im Allgemeinen ergeben, dass direktes korrektives Feedback effektiver ist, um eine bestimmte Struktur zu lehren

[16] Eng. *Dynamic Written Corrective Feedback*

(Bitchener & Knoch, 2009, zitiert nach Ferris & Kurzer, 2019; Ellis et al., 2008; Sheen, 2007), während klassenzimmerbasierte Studien im Bereich des L2-Schreibens die Überlegenheit des indirekten korrektiven Feedbacks zur langfristigen Förderung der Lernerautonomie und Korrektheit gezeigt haben (Lalande, 1982; Ferris, Hyland & Hyland, 2006).

Fokussiertes vs. Nicht-fokussiertes Feedback

Diese Unterscheidung bezieht sich darauf, auf wie viele und welche Arten von Fehlern sich die Lehrkraft konzentriert.

- Fokussiertes/Selektives Feedback: Der Ansatz, bei dem das Feedback auf eine begrenzte, vorab festgelegte Anzahl von Fehlertypen abzielt (z. B. nur Fehler bei Artikeln oder Vergangenheitsformen) (Ellis et al., 2008; Lee, 2017). Die grundlegende Logik dieses Ansatzes besteht darin, das Lernen zu erleichtern und die kognitive Belastung zu reduzieren, indem die Aufmerksamkeit der Lernenden auf weniger und klarer definierte Fehlertypen gelenkt wird (Ferris, 2010, s. 182).
- Nicht-fokussiertes/Umfassendes Feedback: Der traditionelle Ansatz, bei dem die Lehrkraft alle oder eine sehr breite Palette von Fehlern korrigiert, die sie im Text beobachtet (Ellis et al., 2008, s. 356). Befürworter dieses Ansatzes argumentieren, dass die Erwartungen der realen Welt ein hohes Maß an Korrektheit erfordern und die Lernenden lernen müssen, ihre gesamten Texte zu überarbeiten, nicht nur einige wenige Fehlermuster (Evans, Hartshorn, McCollom, & Wolfersberger, 2010, zitiert nach Ferris, 2011; Hartshorn et al., 2010).

Obwohl die meisten Forschungsarbeiten darauf hindeuten, dass fokussiertes Feedback effektiver ist (Bitchener & Ferris, 2012, zitiert nach Lee, 2017, s. 67), fand eine der wenigen Studien, die beide Typen direkt verglich, nämlich die von Ellis et al. (2008), keinen signifikanten Unterschied zwischen den beiden Gruppen (S. 366).

Kodiertes vs. Nicht-kodiertes Feedback

Bei der Anwendung von indirektem Feedback ist auch die Deutlichkeit, mit der die Lehrkraft den Fehler kennzeichnet, eine Variable.

- Nicht-kodiertes Feedback: Der Ansatz, bei dem die Lehrkraft die Stelle des Fehlers lediglich markiert (z. B. durch Unterstreichen oder Einkreisen), aber keine Kennzeichnung über die Art des Fehlers vornimmt (Ferris, 2011, s. 34).
- Kodiertes Feedback: Der Ansatz, bei dem die Lehrkraft nicht nur die Stelle des Fehlers markiert, sondern auch einen Code verwendet, der die Art des Fehlers angibt (z. B. VT für Verb-Tempus, S für Substantivendungen) (Ferris, 2011; Lee, 2017).

Obwohl Umfragen unter Lernenden und Lehrenden eine Tendenz zeigen, dass kodiertes Feedback als nützlicher empfunden wird (Ferris & Roberts, 2001, s. 162), haben Textanalysen gezeigt, dass kodiertes Feedback im Hinblick auf den Überarbeitungserfolg keine signifikante Überlegenheit gegenüber nicht-kodiertem Feedback aufweist (Robb, Ross & Shortreed, 1986; Ferris & Roberts, 2001).

Metalinguistisches Feedback

Diese Art von Feedback kann mit anderen Arten kombiniert werden und beinhaltet die Bereitstellung expliziter grammatischer Informationen oder Hinweise auf Regeln über die Natur des Fehlers des Lernenden (Ellis et al., 2008; Lee, 2017). Beispielsweise kann die Lehrkraft neben einer direkten Korrektur eine kurze Regelerklärung hinzufügen.

Die Faktoren, die die Wirksamkeit von Feedback beeinflussen

Die Wirksamkeit von schriftlichem korrektivem Feedback hängt nicht nur von der Art des verwendeten Feedbacks ab. Dieser Prozess besitzt eine komplexe Struktur, in der eine Reihe von lerner-, lehrer- und kontextbezogenen Faktoren zusammenspielen. Es sollte nicht vergessen werden, dass Feedback weniger ein mechanischer Korrekturakt als vielmehr eine soziale Handlung ist und innerhalb eines Aktivitätssystems stattfindet (Lee, 2017, s. 57).

Lernerfaktoren

Die individuellen Unterschiede der Lernenden haben einen großen Einfluss darauf, wie sie Feedback wahrnehmen und nutzen.

- Sprachkompetenz und Lernhintergrund: Das L2-Kompetenzniveau der Lernenden und ihre bisherigen Sprachlernerfahrungen sind von entscheidender Bedeutung. Beispielsweise benötigen visuelle Lerner, die eher eine formale Ausbildung genossen haben, und auditive Lerner, die die Sprache eher durch Hören gelernt haben, möglicherweise unterschiedliche Arten von Feedback (Reid, 1998, zitiert nach Ferris, 2011, s. 4f). Lernende mit geringerem Kompetenzniveau haben möglicherweise Schwierigkeiten, indirektes Feedback zu verstehen, während fortgeschrittene Lernende von dieser Art, die mehr kognitiven Aufwand erfordert, stärker profitieren können (Zheng & Yu, 2018, s. 21f).
- Überzeugungen, Einstellungen und Motivation: Die Überzeugungen der Lernenden bezüglich Feedback und Grammatiklernen bestimmen, wie sehr sie sich auf den Prozess einlassen (Han, 2017, s. 134f). Nahezu alle Lernenden geben an, dass sie Feedback wünschen und der Meinung sind, dass es ihnen hilft (Ferris, 2011, s. 13). Lernende mit geringer Motivation oder hohem Angstniveau können durch kritisches Feedback negativ beeinflusst werden (Hyland & Hyland, 2019, s. 168).

- Kognitives und affektives Engagement: Wie sehr die Lernenden das Feedback beachten, es verstehen und anwenden, bestimmt, ob ein Lernen stattfindet oder nicht (Han & Hyland, 2019, s. 248f). Affektive Faktoren beeinflussen ebenfalls die Bereitschaft der Lernenden, Feedback zu akzeptieren, und die Wahrscheinlichkeit, dass sie es im Gedächtnis behalten (Storch & Wigglesworth, 2010, s. 328).

Lehrerfaktoren

Die Rolle und der Ansatz der Lehrkraft stehen im Zentrum des Feedback-Prozesses.

- Wissen und Fähigkeiten: Das Wissen der Lehrkräfte über die L2-Grammatik und ihre Fähigkeit, Fehler konsistent zu diagnostizieren, beeinflussen direkt die Qualität des Feedbacks (Ferris, 2011, s. 57f.).
- Philosophie und Überzeugungen: Die Philosophie der Lehrkräfte in Bezug auf Feedback bestimmt, auf welche Fehler sie sich konzentrieren und wie sie das Feedback präsentieren. Viele Lehrkräfte haben aufgrund von institutionellem Druck oder ihren eigenen Lernerfahrungen bestimmte Feedback-Gewohnheiten (Lee, 2010; Lee, 2017).
- Interpersonale Dimension: Feedback ist nicht nur eine informative, sondern auch eine interpersonale Interaktion (Hyland & Hyland, 2019, s. 165). Die Art und Weise, wie Lehrkräfte Lob, Kritik und Vorschläge äußern (z. B. durch den Einsatz von Abschwächungsstrategien), baut eine Beziehung zum Lernenden auf (Hyland & Hyland, 2001, s. 201). Die Kombination von Kritik mit Lob, die Verwendung von Frageformen oder die Formulierung in einer persönlichen Sprache die negative Auswirkung des Feedbacks verringern (Hyland & Hyland, 2019, s. 169).

Kontextfaktoren

Feedback findet nicht in einem Vakuum statt; es wird zutiefst von dem Kontext beeinflusst, in dem es sich befindet.

- Institutioneller und kultureller Kontext: Faktoren wie die Bewertungsrichtlinien der Schule oder Institution, die Klassengrößen und die Prüfungskultur prägen die Feedback-Praktiken der Lehrkräfte (Lee, 2008, s. 69). Beispielsweise können sich Lehrkräfte in einem prüfungsorientierten System unter Druck gesetzt fühlen, sich eher auf die grammatische Korrektheit als auf den Inhalt zu konzentrieren (Lee, 2017, s. 73).
- Pädagogischer Kontext: Es ist wichtig, inwieweit das Feedback in andere Unterrichtsaktivitäten (z. B. Grammatik-Minilektionen, Peer-Feedback) integriert ist. Feedback ist effektiver, wenn es nicht eine isolierte Handlung ist, sondern Teil einer umfassenderen pädagogischen Strategie (Hyland & Hyland, 2019, s. 170). Das dynamische schriftliche korrektive Feedback zielt beispielsweise darauf ab, diese Integration zu gewährleisten, indem es kurze und regelmäßige Schreibaufgaben, kodiertes Feedback und eine obligatorische Überarbeitung kombiniert (Evans et al., 2010, s. 454f).

Die Bevorzugung von schriftlichem Feedback in der Arbeit

Da sich diese Arbeit mit KI-basierten Feedback befasst, liegt der Fokus auf schriftlichem korrektivem Feedback. Der Hauptgrund für diese Wahl liegt in der Natur der bestehenden technologischen Systeme, die von Haus aus die Fähigkeit zur textbasierten Analyse und zur Generierung von Feedback besitzen. Aspekte wie die unmittelbare Interaktion, Intonation und Körpersprache, die mündliches Feedback erfordert, sind für KI-Systeme noch nicht vollständig gelöst. Schriftliches Feedback hingegen bietet für solche Systeme eine passendere und funktionalere Grundlage.

Darüber hinaus bietet schriftliches korrektives Feedback von Natur aus einige einzigartige pädagogische Vorteile. Laut Bruton (2010) hat schriftliches Feedback das Potenzial, kontextualisiert, individualisiert, personalisiert und in einen bedeutungsvollen, in der Regel vom Lernenden selbst produzierten Diskurs eingebettet zu sein (S. 493). Die Lernenden haben ausreichend Zeit, das Feedback zu überprüfen und darüber nachzudenken. Im Gegensatz zu Informationen, die in einer mündlichen Umgebung sofort verloren gehen, ist schriftliches Feedback dauerhaft und ermöglicht es dem Lernenden, seinen Originaltext immer wieder mit den Korrekturen zu vergleichen (Bruton, 2010, s. 495). Diese Eigenschaften fördern eine tiefere Reflexion des Lernenden über seine Fehler und unterstützen das langfristige Lernen. Daher bildet die etablierte und strukturierte Natur des schriftlichen Feedbacks eine solide Grundlage für den Fokus der Arbeit, während das Potenzial und die Grenzen von KI-Systemen untersucht werden.

Künstliche Intelligenz

Da es keine einheitliche Definition des Begriffs der *Intelligenz* gibt, ist es nicht einfach, eine allgemein gültige Definition des Begriffs der *KI* zu finden. Es ist jedoch festzustellen, dass es einige Gemeinsamkeiten gibt, auf die Forscher bei der Definition des Begriffs der KI stoßen. Um diese Gemeinsamkeiten zu ermitteln, werden im Folgenden die Definitionen des Begriffs *KI* aufgeführt, die von Forschern aus verschiedenen Bereichen vorgenommen wurden.

Wie von Arslan erklärt, können die menschenähnlichen Fähigkeiten eines Computers, wie zum Beispiel logisches Denken, Problemlösung, Inferenz und Generalisierung von Verhaltensweisen, also die Nutzung hochrangiger kognitiver Fähigkeiten, als KI definiert werden (2020, S.76). Laut Elmas kann KI als die Fähigkeit definiert werden, geistige Prozesse wie Entscheidungsfindung, Bedeutungsentnahme, Verallgemeinerung, Lernen und Nutzen aus Erfahrungen von Menschen oder einigen Lebewesen mithilfe von Computern,

Software und integrierten Schaltkreisen umzusetzen (2021, S.26). KI wird grundlegend als die Fähigkeit eines Computers oder computergesteuerten Geräts von Nabiyev definiert, Aufgaben durchzuführen, die normalerweise als hochrangige geistige Prozesse des Menschen gelten, wie logisches Denken, Bedeutungsentnahme, Verallgemeinerung und Lernen aus vergangenen Erfahrungen (2021. S.27). McCharty zufolge ist KI die Wissenschaft und Technik der Entwicklung intelligenter Maschinen, insbesondere intelligenter Computerprogramme (2007, S.2). Hamet und Tremblay denken, dass KI ein allgemeiner Begriff ist, der die Verwendung eines Computers zur Modellierung intelligenten Verhaltens mit minimalem menschlichem Eingriff bezeichnet (2017, S.36). Laut Nilsson befasst sich KI nach einer weit gefassten (und etwas zirkulären) Definition mit dem intelligenten Verhalten von Artefakten. Intelligentes Verhalten wiederum umfasst Wahrnehmung, logisches Denken, Lernen, Kommunikation und Handeln in komplexen Umgebungen. Eines der langfristigen Ziele der KI ist die Entwicklung von Maschinen, die diese Dinge genauso gut wie Menschen oder möglicherweise sogar besser können (2009, S.1). Yılmaz ist der Ansicht, dass KI die Fähigkeit ist, einem Computer oder einer Maschine die Fähigkeit zu verleihen, Schlussfolgerungen zu ziehen, zu verallgemeinern und aus früheren Erfahrungen zu lernen, um die menschliche Intelligenz zu modellieren (2017, S.1). KI ist die Anwendung von Denk-, Verstehens-, Verständnis-, Interpretations- und Lernprozessen, die von der menschlichen Intelligenz ausgeführt werden, auf die Lösung von Problemen durch Computerprogramme (Aydemir, 2019, S.28). Sebetci erklärte KI als den Einsatz von Computern zur Modellierung intelligenten Verhaltens mit minimalem menschlichem Eingriff (2023, S.36).

In diesen Zitaten werden einige Wörter in der Definition von KI wiederholt. Diese Wörter können in einem Metabegriff zusammengefasst werden: Nachahmung komplexer Prozesse im menschlichen Geist und deren Umsetzung in Aktionen. Auf der Grundlage dieser Zitate und Gemeinsamkeiten kann daher die allgemeinste Definition von KI gegeben werden: KI

ist die Fähigkeit einer Maschine, die komplexen Prozesse im menschlichen Geist zu imitieren, um zu einem Urteil zu gelangen und ein menschenähnliches Verhalten zu zeigen. Britannica, eine der größten Enzyklopädien der Welt, erklärt den Begriff *KI* auf ähnliche Weise:

> „KI, die Fähigkeit eines digitalen Computers oder eines computergesteuerten Roboters, Aufgaben auszuführen, die üblicherweise mit intelligenten Wesen in Verbindung gebracht werden. Der Begriff wird häufig auf das Projekt der Entwicklung von Systemen angewandt, die mit den für den Menschen charakteristischen intellektuellen Prozessen ausgestattet sind, wie z. B. der Fähigkeit zu denken, Bedeutungen zu erkennen, zu verallgemeinern oder aus früheren Erfahrungen zu lernen“ (2024)

In den Zitaten über KI und in dem aus diesen Zitaten erstellten Definitionsvorschlag werden vor allem die Ausdrücke *Nachahmung menschlicher Intelligenz* und *Durchführung menschenähnlicher Handlungen* stärker hervorgehoben, weil sie die Studie unterstützen. Denn in der Studie wird von der KI erwartet, dass sie sich wie ein Lehrer verhält, der sich auf die Verbesserung der Schreibfähigkeiten von Lernenden in einem FSU konzentriert. Es ist auch bekannt, dass es drei verschiedene Arten von KI gibt: Enge KI, Generelle KI und Super KI.

> „Enge KI wird für KI-Anwendungen verwendet, die in einem bestimmten Bereich trainiert und entwickelt werden. […] Der Unterschied zwischen allgemeiner KI und enger KI besteht darin, dass sie über soziale und emotionale Intelligenz verfügt, Fähigkeiten zum Denken in der Vergangenheit und in der Zukunft besitzt und gleichzeitig verschiedene Eigenschaften wie Kreativität und Originalität, die der Mensch besitzt, einbeziehen kann. […] Super KI ist ein Begriff, der sich auf den Zeitraum bezieht, der als Endphase der KI-Entwicklung angesehen wird. Dieses Konzept beschreibt die Existenz eines eigenständigen Phänomens, das über Entdeckungsfähigkeiten verfügt, die die menschliche Intelligenz übertreffen“ (Eryılmaz, 2023, 11f).

Nabiyev erklärt dagegen, dass es drei Arten der KI gibt, und diese sind: Schwache KI, Starke KI und Superkünstliche Intelligenz. Schwache KI ist ein künstliches System, das einen Menschen imitieren kann oder in einem bestimmten Bereich ähnliche Leistungen wie ein Mensch erbringt. Starke künstliche Intelligenzsysteme werden als Systeme betrachtet, die als allgemeine Problemlöser fungieren und aufgrund ihrer Fähigkeit zur Herstellung von

Ursache-Wirkungs-Beziehungen als näher an der menschlichen Intelligenz liegend betrachtet werden. Superkünstliche Intelligenzsysteme sollten ein menschliches Gehirn in einem realen Sinne simulieren, neue Theoreme mit Hilfe von Denkfähigkeiten beweisen und volle Kreativität zeigen (2021. S.64f).

Hodgins ist der Meinung, dass enge künstliche Intelligenzeinheiten spezifische Aufgaben lernen, insbesondere durch maschinelles Lernen und tiefes Lernen. Während maschinelles Lernen lernt, eine Aufgabe mit nur einer Methode zu erledigen, verfügt tiefes Lernen über komplexere Datenverarbeitungsfähigkeiten. Maschinelles Lernen benötigt Daten in einem bestimmten Format und auf vordefinierte Weise, aber tiefes Lernen kann Daten auf flexiblere Weise verarbeiten (2022, s. 76).

Obwohl einige Forscher behaupten, dass Computerpower und menschliche Intelligenz grundlegend unterschiedlich funktionieren und dass Computerpower die menschliche Intelligenz nicht ersetzen kann (Fjelland, 2020, s. 2), wird die zukünftige Technologie als auf dem Niveau einer generellen KI stehend betrachtet.

Während Attila das Ziel der KI darin sieht, das Wesen der Intelligenz zu verstehen und eine neue Art von intelligenten Maschinen oder intelligenten Robotern zu schaffen, die auf dieser Grundlage ähnlich wie die menschliche Intelligenz reagieren können (2022, S.27), erklärt Yılmaz das Ziel der KI darin, Maschinen herzustellen, die Aufgaben ausführen können, die menschliche Intelligenz erfordern (2017, S.5). Sebetci ist der Meinung, dass der Hauptzweck von Projekten im Bereich der KI darin besteht, den Arbeits- und Energieaufwand zu verringern, indem das Verarbeitungsvolumen und die Verarbeitungsgeschwindigkeit an den Stellen erhöht werden, an denen Arbeitskräfte benötigt werden, und den Menschen durch eine hohe Genauigkeit Komfort zu bieten (2023, S.315).

Der historische Prozess des KI-gestützten Fremdsprachenlernens

Die KI-gestützte Sprachausbildung hat eine Entwicklungslinie parallel zur Evolution der Technologie durchlaufen und so ihre heutige komplexe und interaktive Form erreicht. Dieser

Prozess umfasst eine tiefgreifende Transformation, die von einfachen regelbasierten Systemen bis hin zu generativen KI-Modellen reicht, die menschenähnliche Texte erzeugen können.

Die Grundlagen der KI-Forschung reichen bis in die Mitte des 20. Jahrhunderts zurück. Der Begriff *KI* wurde erstmals 1956 von McCarthy geprägt, wobei seine Ursprünge auf Turings bahnbrechende Arbeiten zu intelligenten Maschinen zurückgehen (Cristianini, 2016, zitiert nach Crompton & Burke, 2023, s. 3). In dieser frühen Phase wurde KI als Maschinen und Algorithmen definiert, die die menschliche Intelligenz nachahmen können (Ng et al., 2021, zitiert nach Gayed, 2025, s. 1).

Die ersten Auswirkungen im Bereich der Sprachausbildung zeigten sich mit den ersten Schritten in der *Verarbeitung natürlicher Sprache*. Einer der wichtigsten Meilensteine in diesem Bereich ist das Dialogsystem ELIZA, das zwischen 1964 und 1966 von Joseph Weizenbaum am *MIT Artificial Intelligence Laboratory* entwickelt wurde (Zhai & Wibowo, 2023, s. 2). ELIZA gilt als die erste Software zur Verarbeitung natürlicher Sprache und wurde entwickelt, um Gespräche mit einem Psychotherapeuten zu simulieren (Sharma, 2023, zitiert nach Liu et al., 2023, s. 3). Dieses System basierte auf einer regelbasierten Struktur und zielte darauf ab, durch die Generierung von Antworten auf eingegebene Texte die Oberflächlichkeit der Mensch-Maschine-Interaktion hervorzuheben. Die revolutionäre Natur von ELIZA ebnete den Weg für spätere Chatbot-Projekte (Zhai & Wibowo, 2023, s. 2).

Eine weitere wichtige technologische Entwicklung dieser Zeit waren traditionelle Textverarbeitungsfunktionen wie die Rechtschreib- und Grammatikprüfung, die in den 1970er Jahren mit den ersten Beispielen des digitalen Schreibens aufkamen (Peterson, 1980, zitiert nach Gayed, Carlon, Oriola & Cross, 2022, s. 2). Diese Werkzeuge können als die ersten praktischen Anwendungen der KI in der Sprachausbildung betrachtet werden. In diesen frühen Jahren verfolgte die KI eine machtorientierte Theorie der Intelligenz, indem

sie sich auf Techniken wie das Beweisen von Theoremen und die Suche konzentrierte (Goldstein & Papert, 1977, s. 89).

In den 1980er Jahren wurde die Computertechnologie durch ihre rasanten Fortschritte zu einem wichtigen Bestandteil der Zweitsprachenpädagogik (Lai & Kritsonis, 2006, s. 2). Auch die Verbreitung des digitalen Schreibens als Forschungsthema fällt in diese Zeit (Kirschenbaum, 2017, zitiert nach Gayed et al., 2022, s. 2). Diese Periode markiert die Geburt des computergestützten Sprachenlernens und die Formung seiner ersten Ausprägungen.

Nach der Klassifikation von Warschauer (2000, zitiert nach Kannan & Munday, 2018, s. 14f) kann die erste Phase dieser Ära als *behavioristisches computergestütztes Sprachenlernen* bezeichnet werden. In diesem Ansatz übernahm der Computer die Rolle eines mechanischen Lehrers, der sich auf Drill-and-Practice-Übungen konzentrierte, wobei das Hauptziel des Sprachenlernens die Korrektheit war. Bax (2003) hingegen bezeichnete diese Periode als *restriktives computergestütztes Sprachenlernen*, da die Computerprogramme eher auf Auswendiglernen als auf Interaktion basierten (zitiert nach Kannan & Munday, 2018, s. 15).

Die Rolle der KI in der Sprachausbildung wurde in den 1980er Jahren mit dem Aufkommen von intelligenten Tutor-Systemen deutlicher (Kannan & Munday, 2018, s. 21). Intelligente Tutor-Systeme wurden als computerbasierte Lernsysteme definiert, die sich an die Bedürfnisse der Lernenden anzupassen versuchen, und galten als großes Versprechen für eine personalisierte Bildung (Self, 1998, zitiert nach Kannan & Munday, 2018, s. 21). Diese ersten Systeme wurden als wichtig für das individualisierte Lernen angesehen, da sie im Vergleich zu einem menschlichen Lehrer unendliche Wiederholungs- und Übungsmöglichkeiten boten (Sumakul, Hamied & Sukyadi, 2022, s. 235).

Die 1990er Jahre waren eine Übergangsphase für KI-Anwendungen in der Sprachausbildung. Laut Liang, Hwang, Chen und Darmawansah (2021) kann die Periode von 1990-1999 in Anlehnung an die Klassifikation von Bax (2011) als die

Computergestütztes-Sprachenlernen-Ära bezeichnet werden (S. 4275). Die Forschung in dieser Zeit konzentrierte sich hauptsächlich auf die Konzeption von KI und die Systementwicklung, jedoch blieben empirische Daten zur Wirksamkeit dieser Systeme unzureichend (Liang et al., 2021, s. 4278). In dieser Phase übernahm die KI die Rolle eines intelligenten Tutor-Systems und erfüllte Funktionen wie die Organisation von Lernmaterialien und die Bereitstellung von Feedback (Liang et al., 2021, s. 4282). Die zunehmende Verbreitung der KI führte zur Entstehung des Bereichs *intelligentes computergestütztes Sprachenlernen* als Teilbereich der computergestützten Sprachenlernen-Forschung (Lu, 2018, zitiert nach Sumakul et al., 2022, s. 235). Die erste anerkannte Publikation zu intelligentem computergestütztem Sprachenlernen war die Arbeit von Weischedel et al. (1978) über ein KI-System zum Deutschlernen (zitiert nach Sumakul et al., 2022, s. 235).

Der Beginn des 21. Jahrhunderts leitete mit der Verbreitung des Internets und mobiler Technologien eine neue Ära in der Sprachausbildung ein. Diese Periode ist durch die Paradigmen des *kommunikativen computergestützten Sprachenlernens* und des *integrativen computergestützten Sprachenlernens* gekennzeichnet, bei denen der Zugang zu Multimedia-Tools und zum Internet zunahm (Warschauer, 2000, zitiert nach Kannan & Munday, 2018, s. 15).

Die entscheidendste Neuerung dieser Zeit war das Aufkommen des Mobilen Sprachenlernens. Das Akronym *mobilgestütztes Sprachenlernen* fand Anfang der 2000er Jahre mit dem Aufkommen mobiler Technologien Eingang in die Literatur (O'Malley et al., 2003, zitiert nach Kannan & Munday, 2018, s. 16). Diese Periode wird von Liang et al. (2021) in Anlehnung an Bax (2011) als Mobilgestütztes Sprachenlernen-Ära bezeichnet (S. 4275). Mit diesem neuen Ansatz war das Lernen nicht mehr auf einen festen Ort beschränkt, sondern konnte durch die Möglichkeiten mobiler Technologien jederzeit und überall stattfinden.

Der Anstieg empirischer Studien zur KI-gestützten Sprachausbildung fällt ebenfalls in diese Zeit. Die erste empirische KI-Studie zu den Lernergebnissen von Studierenden wurde 2004 von Lo et al. veröffentlicht (zitiert nach Liang et al., 2021, s. 4278). Die dominierenden KI-Technologien in dieser Zeit waren die Verarbeitung natürlicher Sprache und Intelligente Agenten; die Rolle der KI in der Bildung konzentrierte sich weitgehend auf die Diagnose der Stärken von Lernenden und die Bereitstellung von automatischem Feedback (Liang et al., 2021, s. 4281f).

In den letzten zehn Jahren gab es einen beispiellosen und exponentiellen Anstieg der Fähigkeiten von Technologie und KI (Zhai & Wibowo, 2023, s. 4). Diese Ära wird von Liang et al. (2021) in Anlehnung an die Arbeiten von Bax (2011) als die Intelligentes-Computergestütztes-Sprachenlernen-Ära definiert (S. 4275). In dieser Zeit begann die intelligente Funktionalität von KI-Systemen eine wichtige Rolle im Sprachunterricht zu spielen (Liang et al., 2021, s. 4278).

Der wichtigste technologische Durchbruch dieser Periode war die Ankündigung der Transformer-Architektur im Jahr 2017 (Godwin-Jones, 2024; Vaswani et al., 2017, zitiert nach Law, 2024). Diese Architektur führte zu einer radikalen Veränderung im Sprachlernansatz der KI. Es erfolgte ein Übergang von regelbasierten Systemen zu einem statistischen Ansatz, bei dem die KI durch die Analyse riesiger Mengen an Textdaten selbstständig entdeckt, wie Sprache funktioniert (Godwin-Jones, 2024, s. 9). Für diese neue Generation von Großen Sprachmodellen wurde Sprache zu einem statistischen Problem, das auf der Häufigkeit der Wortverwendung und dem kontextuellen Gebrauch basiert (Godwin-Jones, 2024, s. 9).

Das populärste Ergebnis dieser technologischen Revolution war ChatGPT, das im November 2022 auf den Markt kam. Das Erscheinen von ChatGPT sorgte dafür, dass generative KI im Bereich der Sprachausbildung breite Aufmerksamkeit erlangte (Law, 2024, s. 3). Diese Entwicklung führte auch zu einem massiven Anstieg akademischer Veröffentlichungen; die Zahl der in den Jahren 2021 und 2022 publizierten Artikel hat sich

im Vergleich zu den Vorjahren verdoppelt bis verdreifacht (Crompton & Burke, 2023, s. 5). Die KI ist nun nicht mehr nur ein Werkzeug, das Feedback gibt oder Materialien bereitstellt, sondern hat sich zu einem Produzenten entwickelt, der eine Vielzahl von Aufgaben wie das Zusammenfassen von Texten, das Erstellen personalisierter Inhalte und Simulationen virtueller Patienten durchführen kann (Alier, García-Peñalvo, & Camba, 2024; Eysenbach, 2023). Diese neue Generation von KI-Werkzeugen stellt den Beginn eines weiteren Paradigmenwechsels in der Sprachausbildung und dem Spracherwerb dar (Law, 2024, s. 1). Der historische Prozess der KI-gestützten Fremdsprachenausbildung offenbart eine dynamische Evolution, in der Technologie und pädagogische Ansätze in ständiger Wechselwirkung stehen. Der Übergang von der regelbasierten Struktur des ersten Dialogsystems ELIZA in den 1960er Jahren zum computergestützten Sprachenlernen und den intelligenten Tutor-Systemen der 1980er und 1990er Jahre hat die Rolle der Technologie in der Bildung schrittweise erweitert. Die Ära des mobilgestützten Sprachenlernens, die in den 2000er Jahren mit dem Aufstieg mobiler Technologien begann, hat das Lernen über die Klassenzimmerwände hinausgetragen. Ab den 2010er Jahren haben *Big Data, Learning Analytics* und schließlich *generative KI-Modelle* auf Basis der Transformer-Architektur das Feld vollständig transformiert. Insbesondere mit dem Aufkommen von Werkzeugen wie ChatGPT ist die KI nicht mehr nur ein Hilfsmittel, sondern auch ein Partner, der Inhalte erstellt, personalisierte Lernerfahrungen bietet und die Grenzen der Mensch-Maschine-Kollaboration neu definiert. Diese historische Reise zeigt, dass sich die Rolle der KI im FSU von einer einfachen Nachahmung eines *Lehrers* zu der eines *intelligenten Partners* entwickelt hat, der ein integraler Bestandteil des Lernökosystems ist.

Verarbeitung natürlicher Sprache

Die Verarbeitung natürlicher Sprache ist ein interdisziplinäres Feld an der Schnittstelle von KI und Linguistik, das sich auf die Interaktion zwischen Computern und menschlichen Sprachen konzentriert (Juhn & Liu, 2020; Schmidt & Strasser, 2022). Diese Technologie

wird als die Fähigkeit von Computern definiert, menschliche Sprache sowohl in mündlicher als auch in schriftlicher Form zu analysieren, zu verstehen und zu erzeugen (Yang & Kyun, 2022, s. 186). Im Wesentlichen befasst sich Verarbeitung natürlicher Sprache mit dem Prozess der Umwandlung von Text in strukturierte Daten, und dieser Prozess umfasst auch Fortschritte im Bereich der Verarbeitung natürlicher Sprache, was die Fähigkeit von KI-Systemen beschreibt, aus Texten Bedeutung zu extrahieren (Godwin-Jones, 2022, s. 5f). Als angewandter Aspekt der Computerlinguistik, einem interdisziplinären Forschungsfeld, zielt Verarbeitung natürlicher Sprache darauf ab, durch die computergestützte Nutzung menschlicher Sprachen relevante Konzepte zu identifizieren (Juhn & Liu, 2020; Schmidt & Strasser, 2022). Dieser Prozess beinhaltet die computergestützte Syntax, die für die automatische Zerlegung von Satzstrukturen in ihre Wortarten verantwortlich ist, und die Programme, die diese Analyse durchführen, werden als *Parser* bezeichnet (Dodigovic, 2007, s. 100).

Die Entwicklung von der Verarbeitung natürlicher Sprache hat sich insbesondere durch Fortschritte im tiefen und netzwerkbasierten Lernen sowie durch wachsende technologische Kapazitäten für die Big-Data-Analyse beschleunigt. In den letzten Jahren haben großes Sprachmodell und generative künstliche Intelligenz (GenKI) die Bereiche der natürlichen Sprachverarbeitung, Inhaltserstellung und des Verständnisses transformiert (Bouschery, Blazevic, & Piller, 2023). GenKI ist im Wesentlichen eine KI-gestützte Großes-Sprachmodell-Technologie, die in Form von Verarbeitung natürlicher Sprache-Software-Tools existiert (Law, 2024). Systeme wie ChatGPT zeigen beispielsweise zahlreiche Verarbeitung natürlicher Sprache-Fähigkeiten wie das Lösen von Anfragen, das Erstellen von Erzählungen, das Treffen logischer Bewertungen und die maschinelle Übersetzung (Karataş, Abedi, Gunyel, Karadeniz & Kuzgun, 2024, s. 19344). Die spezifischen Techniken und Algorithmen, die in solchen transformatorbasierten Modellen verwendet werden, können Gegenstand eines KI-Lehrplans sein (Eysenbach, 2023, s. 8).

Die Verarbeitung natürlicher Sprache-Technologie ermöglicht es KI-Dialogsystemen und Chatbots, natürliche menschliche Dialoge durch Text- oder Sprach-zu-Text-Funktionen zu simulieren; dieser Prozess wird in der Regel durch maschinelles Lernen oder tiefes Lernen realisiert (Zhai & Wibowo, 2023). Im Bildungsbereich, insbesondere in der Sprachausbildung, setzt die KI in der Bildung in hohem Maße auf Verarbeitung natürlicher Sprache-Technologien, und Verarbeitung natürlicher Sprache hat sich in diesem Bereich zu einem der am häufigsten verwendeten Schlüsselwörter entwickelt (Liang et al. 2021, s. 4277). Anwendungen wie intelligente Tutor-Systeme nutzen Verarbeitung natürlicher Sprache; eine der zentralen Herausforderungen bei diesen Systemen besteht jedoch darin, mithilfe von Verarbeitung natürlicher Sprache eine valide automatische Analyse der Schülersprache zu erstellen (Chounta, Bardone, Raudsep, & Pedaste, 2022, s. 734).

Im Kontext des Sprachenlernens macht Verarbeitung natürlicher Sprache die KI zu einem wertvollen Werkzeug, indem es Maschinen ermöglicht, menschliche Sprache zu verstehen (Son, Ružić & Philpott, 2023, s. 95). Anwendungen wie Duolingo nutzen die Verarbeitung natürlicher Sprache-Technologie, um Fähigkeiten wie Sprechen, Hören, Lesen und Schreiben zu unterstützen (Qiao & Zhao, 2023, s. 6). KI-gestützte Verarbeitung natürlicher Sprache-Tools können Sprachlernern Feedback zu Grammatik, Rechtschreibung und Zeichensetzung geben, um ihre Schreibfähigkeiten zu verbessern (Owan, Abang, Idika & Bassey, 2023, s. 3). Einer der Hauptvorteile von Verarbeitung natürlicher Sprache ist die Möglichkeit, den Lernenden personalisierte Lernmaterialien anzubieten und so eine Präzisionsbildung zu ermöglichen (Liu, 2023, s. 3; Wang, Pang, Wallace, Wang, & Chen, 2024, s. 816). Dennoch gibt es auch Herausforderungen bei Verarbeitung natürlicher Sprache-Prozessen, wie zum Beispiel algorithmischen Bias. Da die meisten Sprachmodelle auf standardmäßigem, geschriebenem, professionellem Englisch trainiert werden, wurde beobachtet, dass die Leistung von Verarbeitung natürlicher Sprache bei unterrepräsentierten Gruppen nachlässt (Godwin-Jones, 2021, s. 9).

Maschinelles Lernen und Tiefes Lernen

Als Teilbereich der KI bezeichnet maschinelles Lernen die Fähigkeit von Computern oder Systemen, ohne explizite Programmierung durch die Nutzung von Daten und vergangenen Erfahrungen zu lernen und ihre Leistung zu verbessern (Juhn & Liu, 2020; Schmidt & Strasser, 2022). Im Grunde handelt es sich um eine datengesteuerte Technologie, die entwickelt wurde, um Aufgaben zu erfüllen, die typischerweise menschliche Intelligenz erfordern (Tapalova & Zhiyenbayeva, 2022, s. 642). Maschinelles Lernen basiert auf statistischen Lernmethoden. Algorithmen analysieren vorhandene Daten, um darin enthaltene Muster und Regelmäßigkeiten zu erkennen. Ausgehend von diesen Mustern erstellen sie ein prädiktives Modell, mit dem sie Schlussfolgerungen über zukünftige Daten ziehen (Godwin-Jones, 2024; Tapalova & Zhiyenbayeva, 2022). Je mehr Daten ein System ausgesetzt ist, desto mehr verbessert es seine Leistung durch Erfahrung.

Ansätze des maschinellen Lernens werden im Allgemeinen in zwei Kategorien unterteilt. Wenn das System von vorab durch menschliche Anleitung etikettierten Daten lernt, wird dies als *überwachtes Lernen* bezeichnet. Entdeckt es hingegen ohne menschliches Eingreifen verborgene Muster in den Daten selbstständig, nennt man dies *unüberwachtes Lernen.* Dank dieser Lernfähigkeit bietet die Technologie Lösungen für viele Probleme in Bereichen wie Sehen, Sprechen, Erkennung und Robotik (Schmidt & Strasser, 2022, s. 167). Sie wird insbesondere in Technologien der Automatischen Spracherkennung, die Funktionen wie Spracherkennung und Sprache-zu-Text-Umwandlung bieten, und im Bereich der Verarbeitung natürlicher Sprache weit verbreitet eingesetzt (Juhn & Liu, 2020; Son et al., 2023). Auch GenKI-Systeme nutzen maschinelles Lernen-Techniken, die auf der Transformer-Architektur basieren, um die Fähigkeit zu erlangen, das nächste Wort in einer Textsequenz vorherzusagen und sinnvolle Texte zu erstellen (Godwin-Jones, 2024, s. 9).

Tiefes Lernen, das als eine Teilmenge oder Erweiterung des maschinellen Lernens definiert wird, ist ein Bereich der KI, in dem Entscheidungen und Vorhersagen ebenfalls von einer datengesteuerten Technologie getroffen werden (Lameras & Arnab, 2022; Tapalova &

Zhiyenbayeva, 2022). Die Grundlage bilden mehrschichtige künstliche neuronale Netze, die ähnlich wie die neuronalen Netze im menschlichen Gehirn konzipiert sind (Schmidt & Strasser, 2022, s. 167). Tiefes Lernen verdankt seinen Namen dieser mehrschichtigen Struktur des künstlichen neuronalen Netzes. Es arbeitet mit komplexen Algorithmen und nutzt große, leistungsstarke Computer-Arrays, um in den gleichzeitig zugänglichen, multiplen Schichten eines künstlichen neuronale Netzes zu suchen und zu filtern (Godwin-Jones, 2024, s. 9). In diesem Prozess werden neuronale Netze und iterative Clustering-Algorithmen verwendet, um Verbindungen zwischen Objekten zu bestimmen, und dieser Vorgang wird fortgesetzt, bis das Objekt erkannt wird (Lameras & Arnab, 2022, s. 10).

Aufgrund dieser komplexen Strukturen gelten tiefes Lernen-Modelle für viele Anwendungen als überlegen gegenüber flachen Modellen des maschinellen Lernens und traditionellen Datenanalyseansätzen (Tapalova & Zhiyenbayeva, 2022, s. 643). Obwohl es sich anfangs auf sichtbasierte Aufgaben wie die Bilderkennung konzentrierte, wird es auch intensiv für Zwecke der Verarbeitung natürlicher Sprache eingesetzt (Schmidt & Strasser, 2022, s. 167). Heutige große Sprachmodelle verwenden tiefes Lernen-Architekturen, die aus mehreren parallel arbeitenden Schichten mathematischer Berechnungen bestehen (Godwin-Jones, 2022, s. 12). Ebenso können große, vorab trainierte Modelle mit etwa 175 Milliarden Parametern dank des tiefen Lernens Informationen aus sowohl etikettierten als auch nicht etikettierten Datensätzen extrahieren und so menschenähnliche Texte erzeugen (Zhai & Wibowo, 2023, s. 5).

Generative KI und Große Sprachmodelle

GenKI hat sich in den letzten Jahren als ein Bereich herauskristallisiert, der zu revolutionären Entwicklungen in der Technologie- und Bildungswelt geführt hat. Im Wesentlichen bezeichnet GenKI KI-Systeme, die unter Verwendung von Mustern und Beziehungen, die sie aus einem vorhandenen Datensatz gelernt haben, völlig neue und originelle Inhalte wie Text, Bilder, Musik oder Videos erstellen (Law, 2024, s. 1). Diese

Technologie arbeitet über ein Modell des maschinellen Lernens, das große Mengen von menschenerstellten Daten analysiert und die daraus gewonnenen Informationen nutzt, um neue Ergebnisse zu schaffen. Das grundlegendste Merkmal, das GenKI von früheren KI-Technologien unterscheidet, ist ihr Fokus auf die Erstellung neuer textueller und multimodaler Inhalte, insbesondere durch große Sprachmodelle sowie kunst- und videobasierte Modelle (Law, 2024, s. 1).

Eines der bekanntesten Beispiele für diese Technologie ist ChatGPT, dessen Veröffentlichung durch OpenAI im November 2022 das weltweite Interesse an GenKI auf einen Höhepunkt brachte. Es basiert auf einem großen Sprachmodell, das als Generative Pre-trained Transformer (GPT) bezeichnet wird (Law, 2024, s. 1). Solche Systeme verarbeiten riesige Mengen an Sprachdaten mithilfe von Techniken des maschinellen Lernens, die auf der Transformer-Architektur basieren, und lernen so selbstständig die Struktur und Funktionsweise der menschlichen Sprache (Godwin-Jones, 2024, s. 9). Die aus diesem Prozess hervorgehenden großen Sprachmodelle behandeln Sprache als ein statistisches Problem; sie analysieren die Häufigkeit und den Kontext von Wörtern, um vorherzusagen, welches Wort in einer Textsequenz als Nächstes folgen sollte. Dank dieser Fähigkeit können sie lange und zusammenhängende Texte produzieren, die grammatikalisch korrekt und kontextuell kohärent sind (Godwin-Jones, 2024, s. 9). Diese Modelle sind auch als große vorab trainierte Modelle wie *Bidirectional Encoder Representations from Transformers* und GPT bekannt und stellen äußerst komplexe autoregressive Sprachmodelle dar, die bis zu 175 Milliarden Parameter aufweisen können (Zhai & Wibowo, 2023, s. 7).

Das Potenzial, das GenKI-Tools insbesondere im Bildungsbereich bieten, ist beträchtlich. Studien zeigen, dass Tools wie ChatGPT das Potenzial haben, die Schreibfähigkeiten von Lernenden in einer Zweitsprache, ihre Lernmotivation, ihr Interesse am Thema, ihre Unterrichtsbeteiligung und die Kreativität in ihren Texten zu steigern (Karataş et al., 2024; Law, 2024; Song & Song, 2023). Eine Untersuchung ergab sogar, dass KI-gestützter Unterricht im Vergleich zu traditionellen Lehrmethoden bei Studierenden, die EFL lernen,

zu signifikant größeren Fortschritten in Bereichen wie Schreibkompetenz, Motivation, Textorganisation, Kohärenz, Grammatik und Wortschatz führte (Karataş et al., 2024, s. 19343, 19356).

Neben dem Potenzial, das diese Technologie bietet, gibt es jedoch auch erhebliche Einschränkungen und Risiken. Eines der gravierendsten Probleme im Zusammenhang mit großen Sprachmodellen ist die als Halluzination bekannte Tendenz, falsche oder erfundene Informationen zu generieren, die vom KI-System nicht durch Trainingsdaten verifiziert sind, aber dennoch selbstbewusst präsentiert werden (Eysenbach, 2023, s. 12). Obwohl die erzeugten Texte grammatikalisch kohärent und flüssig sind, stimmen sie nicht immer mit der realen menschlichen Erfahrung überein und sind möglicherweise nicht sinnvoll (Godwin-Jones, 2021; Godwin-Jones, 2024). Dies kann zur Produktion falscher, irreführender und sogar beleidigender Aussagen führen. Darüber hinaus sind diese Systeme unzureichend darin, soziale Normen, kulturelle Praktiken und die pragmatischen Feinheiten der Sprache zu verstehen (Godwin-Jones, 2024, s. 9).

Über technische Einschränkungen hinaus bestehen auch ernsthafte ethische Bedenken hinsichtlich der Entwicklung und Nutzung von großen Sprachmodellen. Zu diesen Bedenken gehören das unbefugte Scannen digitaler Textquellen für das Training der Modelle unter Missachtung von Urheberrechten, der immense Energieverbrauch, den diese Prozesse erfordern, und der Einsatz oft niedrig entlohnter Arbeitskräfte für das Training der Modelle (Godwin-Jones, 2024, s. 18f). Zudem wird kritisiert, dass diese Systeme kulturelle Vorurteile enthalten, die die Perspektive von *westlichen, gebildeten, industrialisierten, reichen und demokratischen Gesellschaften* widerspiegeln und daher die Sichtweisen von Minderheitengruppen, Frauen und anderen nicht privilegierten Gemeinschaften ausschließen (Godwin-Jones, 2024, s. 18f). Der böswillige Einsatz von GenKI stellt ebenfalls eine erhebliche Bedrohung dar. Anwendungen wie die Erstellung von Deepfake-Videos, die Schaffung synthetischer Identitäten für bösartige Cyberkampagnen, die Verbreitung gezielter Desinformation und die Produktion fortschrittlicher Malware zeigen

die dunkle Seite dieser Technologie. Insbesondere wurde berichtet, dass KI-generierte synthetische Identitäten, die von echten Personen nicht zu unterscheiden sind, bereits für Aktivitäten wie Spionage und Online-Belästigung eingesetzt werden.

Eine der offensichtlichsten Sorgen für die akademische Welt betrifft Plagiat und akademische Integrität. Texte, die von GenKI-Tools erstellt werden, können von herkömmlicher Plagiatserkennungssoftware nicht aufgespürt werden, da sie bei jeder Anfrage einzigartig generiert werden (Godwin-Jones, 2022; Sumakul et al., 2022). Dies birgt das Risiko, dass Studierende ihre Hausarbeiten und Abschlussarbeiten von diesen Tools schreiben lassen, was die Grundsätze der akademischen Integrität fundamental untergräbt (Karataş et al., 2024; Law, 2024; Malik et al., 2023). Aus diesem Grund betonen Experten und Akademiker die Notwendigkeit, GenKI mit Bedacht in die Bildung zu integrieren. Es ist von entscheidender Bedeutung, Missbrauch in Lehre und Forschung zu verhindern, KI-Kompetenz in die Lehrpläne aufzunehmen und den Studierenden klare Richtlinien über die Grenzen und den korrekten Einsatz dieser Werkzeuge an die Hand zu geben (Liu et al., 2023, s. 5).

Obwohl die GenKI ein mächtiges Werkzeug ist, das das Potenzial hat, Lern- und Lehrprozesse zu transformieren, bringt sie ernsthafte technische, ethische und akademische Herausforderungen mit sich. Um die Vorteile dieser Technologie zu maximieren und gleichzeitig ihre Risiken zu minimieren, ist die Annahme eines ausgewogenen, bewussten und kritischen Ansatzes unerlässlich.

KI-gestützte Werkzeuge und Systeme im FSU

KI transformiert die Sprachausbildung durch intelligente Tutor-Systeme, indem sie personalisierte und interaktive Lernerfahrungen bietet, die an die individuellen Bedürfnisse der Lernenden angepasst sind (Chassignol, Khoroshavin, Klimova, & Bilyatdinova, 2018; Liang et al., 2021). Diese Systeme erfüllen eine breite Palette von Funktionen, die von der Diagnose und Korrektur von Fehlern der Lernenden (Dodigovic, 2007) bis hin zur

Bereitstellung von Aussprache- und Sprechübungen durch automatische Spracherkennung und intelligente persönliche Assistenten reichen (Dizon, 2017; Son et al., 2023). Schreibfähigkeiten werden wiederum durch GenKI-Assistenten wie ChatGPT und Tools zur automatischen Schreibbewertung wie *Grammarly* mit sofortigem Feedback unterstützt (Law, 2024; Malik et al., 2023). Heutzutage reichen diese Technologien von vielseitigen mobilen Plattformen wie Duolingo über ausspracheorientierte Anwendungen wie *Elsa Speak* bis hin zu immersiven virtuellen Lernumgebungen wie CILLE, wodurch das Sprachenlernen autonomer, effektiver und motivierender wird (Divekar et al., 2022; Godwin-Jones, 2024; Karataş et al., 2024).

Intelligente Tutor-Systeme

Intelligente Tutor-Systeme sind Computersysteme, die entwickelt wurden, um durch die Kombination von Techniken der KI und pädagogischen Methoden den Studierenden eine personalisierte und interaktive Lernerfahrung zu bieten (Crompton & Burke, 2023; Son et al., 2023). Die grundlegende Idee dieser Systeme, deren Ursprünge in den 1970er Jahren liegen, ist der Gedanke, dass ein Computer nicht nur als Werkzeug, sondern auch als Tutor agieren kann (Chassignol et al., 2018, s. 19). Intelligente Tutor-Systeme gelten als ein durch KI weiterentwickelter Zweig der Lernmanagementsysteme, und ihr Hauptzweck ist es, durch das Sammeln von Daten aus den Antworten der Studierenden „das Wissen, die Motivation oder die Emotionen der Studierenden zu modellieren und den Unterricht an individuelle Bedürfnisse anzupassen“ (Chassignol et al., 2018, s. 19). Diese Systeme übernehmen eine der häufigsten Rollen der KI in der Bildung und werden als eine der grundlegenden Methoden für den Einsatz von KI in der Hochschulbildung angesehen (Crompton & Burke, 2023; Liang et al., 2021).

Die Architektur von intelligenten Tutor-Systemen basiert in der Regel auf drei grundlegenden Modellen: dem Domänenmodell, das das Wissen über das behandelte Thema enthält, dem Bewertungsmodell, das die Leistung des Studierenden kontinuierlich evaluiert,

und dem Studierendenmodell, das Daten über die Beherrschung des Themas durch den Studierenden sammelt und aktualisiert (Schmidt & Strasser, 2022, s. 174-175). Diese Studierendenmodelle sind ein zentrales Element von intelligenten Tutor-Systemen (Godwin-Jones, 2021, s. 8). Mithilfe dieser Modelle überwacht das System die Leistung des Studierenden, identifiziert Lerndefizite und bietet die am besten geeigneten Lernressourcen an (Chassignol et al., 2018; Lameras & Arnab, 2022). Das grundlegendste Merkmal, das intelligente Tutor-Systeme von traditionellen computerbasierten Lehrsystemen unterscheidet, ist, dass sie nicht erst nach Abschluss einer Aufgabe Feedback geben, sondern in jedem Schritt des Prozesses mit dem Studierenden interagieren (Chassignol et al., 2018, s. 20).

Die Sprachausbildung ist einer der häufigsten Anwendungsbereiche von intelligenten Tutor-Systemen und wird in diesem Feld auch als Intelligente Sprachlernsysteme oder intelligentes computergestütztes Sprachenlernen bezeichnet (Dodigovic, 2007; Schmidt & Strasser, 2022). Diese Systeme bieten personalisierte Lernerfahrungen, indem sie sich an den einzigartigen Lernpfad, die Stärken und die Entwicklungsbereiche jedes Studierenden anpassen (Malik et al., 2023, s. 1). Beispielsweise können sie Fehler von Studierenden erkennen und korrigieren, gezielte Aktivitäten in bestimmten Bereichen wie Aussprache, Wortschatz oder Grammatik anbieten und sofortiges, personalisiertes Feedback geben (Son et al., 2023; Song & Song, 2023). In diesem Prozess übernimmt ein intelligentes Tutor-System zwei grundlegende Rollen: eine pädagogische Rolle als *intelligenter Tutor*, der die Studierenden anleitet, und eine kognitive Rolle als *intelligenter Bewerter*, der Lernprobleme diagnostiziert und Empfehlungen gibt (Liang et al., 2021, s. 4282).

Intelligente Tutor-Systeme sind auch ein wichtiges Unterstützungswerkzeug für Lehrende. Im Rahmen der geteilten Verantwortlichkeiten zwischen Lehrkraft und KI entwirft die Lehrkraft als Inhaltsexperte die Lerninteraktionen, während der intelligente Tutor Aufgaben wie die „Steuerung des Lernflusses, die Bereitstellung personalisierter Inhalte, das Geben von sofortigem Feedback und die Überwachung der Leistung“ übernimmt (Godwin-Jones,

2024, s. 15f). Dadurch können Lehrende den Unterricht auch bei steigenden Klassengrößen an die individuellen Bedürfnisse der Studierenden anpassen (Kannan & Munday, 2018, s. 23).

Allerdings gibt es bei der Entwicklung und Implementierung von intelligenten Tutor-Systemen einige Herausforderungen. Insbesondere die Erstellung einer validen automatischen Analyse der Schülersprache mittels Verarbeitung natürlicher Sprache stellt eine grundlegende Schwierigkeit dar (Schmidt & Strasser, 2022, s. 175). Zudem wird berichtet, dass die ersten Versionen dieser seit den 1980er Jahren existierenden Systeme aufgrund ihres *begrenzten pädagogischen Designs* nur eine moderate positive Auswirkung auf das akademische Lernen von Hochschulstudierenden zeigten (Kannan & Munday, 2018, s. 21). Jedoch beleben die heutigen Fortschritte in der KI-Technologie das Potenzial dieser Systeme für personalisiertes Lernen neu (Kannan & Munday, 2018, s. 21).

Automatische Schreibbewertung

Automatisierte Schreibbewertung ist eine Technologie, die innerhalb des breiten Spektrums der Fähigkeiten KI angesiedelt ist, welches Bereiche wie die Verarbeitung natürlicher Sprache und die Textgenerierung umfasst (Rad et al., 2024, s. 5032). Diese Systeme, die über traditionelle Rechtschreib- und Grammatikprüfungstools hinausgehen, sind speziell für die Bewertung von Aufsätzen konzipiert und bieten umfassende Unterstützung, indem sie Schreibprobleme erkennen und auf Bereiche hinweisen, die einer Überarbeitung bedürfen (Godwin-Jones, 2022, s. 5). Automatisierte Schreibbewertungstechnologien werden auf allen Bildungsstufen, von der Grundschule bis zur Universität, sowohl im muttersprachlichen als auch im Zweitsprachenunterricht weit verbreitet eingesetzt (Godwin-Jones, 2022, s. 5).

Die Hauptfunktion von automatisierten Schreibbewertungssystemen liegt in ihrem Potenzial, Studierenden schnelles und konsistentes korrektives Feedback zu ihren

schriftlichen Arbeiten zu geben; dies wird als wesentlicher Vorteil gegenüber menschlichen Bewertern angesehen (Godwin-Jones, 2022, s. 7).

Diese Werkzeuge ermöglichen es den Studierenden, wertvolle Einblicke in die Arten der von ihnen gemachten Fehler zu gewinnen. Die Möglichkeit zur Selbstkorrektur wird als positiver Aspekt angesehen, der autonomes Lernen für Studierende fördert, die ohne die Unterstützung eines Lehrers arbeiten. In diesem Zusammenhang wird den automatisierten Schreibbewertungsentwicklern empfohlen, ihre Ressourcen zu erweitern und eine Option für kommentiertes Feedback in der Muttersprache anzubieten (Son et al., 2023, s. 97).

Automatisierte Schreibbewertung ist ein intensiv erforschtes Feld, und KI-gestützte automatisierte Schreibbewertungssysteme stellen einen aufstrebenden Teilbereich dar. Der Fokus akademischer Studien hat sich von der anfänglichen Überprüfung der Wirksamkeit KI-gestützter automatisierter Schreibbewertungssysteme im Laufe der Zeit darauf verlagert zu verstehen, wie Studierende mit dem von diesen Systemen angebotenen Feedback interagieren (Yang & Kyun, 2022, s. 192ff). Es wird jedoch angemerkt, dass diese Technologien zwar häufig zu Bewertungszwecken eingesetzt werden, es aber nur begrenzte Studien darüber gibt, wie KI-basierte Werkzeuge über die Bewertung hinaus als Lerninstrument genutzt werden können (Gayed et al., 2022, s. 1).

Heutzutage können neben spezifischen automatisierten Schreibbewertungssystemen wie *Criterion* oder *MY Access!* auch allgemeine KI-Werkzeuge für diesen Zweck verwendet werden. Studententexte können der KI zusammen mit detaillierten Prompts oder einer Bewertungsrubrik zur Analyse vorgelegt werden. Die Verwendung eines allgemeinen KI-Werkzeugs kann mehr Flexibilität für die Personalisierung bieten, indem es ermöglicht, sich auf spezifische Aspekte des Schreibens zu konzentrieren (Godwin-Jones, 2024, s. 12).

Die Verbreitung des Themas in der akademischen Literatur zeigt sich auch in Studien, in denen Systeme wie *Pigai* untersucht werden (Kohnke, Moorhouse & Zou, 2023, s. 2) oder das Thema im Kontext der Hochschulbildung behandelt wird (Malik et al., 2023, s. 9).

Automatische Spracherkennung

Automatische Spracherkennung wird als eine Technologie definiert, die es Maschinen ermöglicht, Sprachsignale in Text oder Befehle umzuwandeln (Zhai & Wibowo, 2023, s. 7f). Diese Technologie nutzt Techniken der KI und des maschinellen Lernens, um gesprochenen und geschriebenen Text zu verstehen und zu produzieren (Son et al., 2023, s. 100). Im Grunde ermöglicht automatische Spracherkennung den Benutzern, durch Sprechen mit Software wie intelligente persönliche Assistenten zu interagieren; indem sie Sprache in Text umwandelt, befähigt sie diese Systeme, dem Benutzer angemessene Antworten zu geben (Dizon, 2017, s. 812). In einem breiteren Rahmen wird die Spracherkennung als eine Fähigkeit der KI betrachtet und ist Teil von „Computersystemen, die in der Lage sind, Aufgaben auszuführen, die normalerweise menschliche Intelligenz erfordern“ (Chassignol et al., 2018; Tapalova & Zhiyenbayeva, 2022).

Die automatische Spracherkennungstechnologie wird häufig in Software eingesetzt, die Spracherkennungs- und Sprache-zu-Text-Funktionen nutzt, wie z. B. in intelligenten persönlichen Assistenten, automatischen Transkriptionsprogrammen und Notizanwendungen (Son et al., 2023, s. 100). Die Anwendung von automatischer Spracherkennung im Kontext des Fremdsprachenlernens erfolgt durch das Design dialogbasierter Inhalte, die akustische, lexikalische und sprachliche Modelle umfassen. In diesem Bereich treten zwei Hauptkomponenten der Technologie hervor: (1) die Fähigkeit, eine akzentuierte Aussprache von einer fehlerhaften Aussprache zu unterscheiden, und (2) die Fähigkeit, eine Bewertung der Aussprachequalität zu liefern (Zhai & Wibowo, 2023, s. 7f).

Die Integration von automatischer Spracherkennung in den Zweitspracherwerb bietet den Lernenden erhebliche Vorteile. Sie unterstützt die Verbesserung der Zweitsprachenaussprache, da die Benutzer sofortiges, personalisiertes und autonomes Feedback erhalten können (Son et al., 2023, s. 100). Automatische Spracherkennungssysteme steigern die interaktionale Kompetenz der Lernenden, indem sie

Möglichkeiten zur Produktion von Output in der Zielsprache bieten und das Memorieren und Behalten von Vokabeln verbessern (Zhai & Wibowo, 2023, s. 7f). Darüber hinaus können sie eine angstreduzierte Lernumgebung schaffen, insbesondere für Lernende, die zögern, vor anderen zu sprechen, was sich positiv auf ihre Kommunikationsbereitschaft und Motivation auswirken kann (Zhai & Wibowo, 2023; Son et al., 2023).

Chatbots und Intelligente Persönliche Assistenten

Die Entwicklung von KI-Technologien hat zur Entstehung neuer Werkzeuge in vielen Bereichen, einschließlich in der Bildung, geführt. Unter diesen Werkzeugen nehmen Chatbots und intelligente persönliche Assistenten, die sich durch ihre Fähigkeit zu menschenähnlichen Dialogen auszeichnen, einen wichtigen Platz ein. Chatbots sind Softwareanwendungen, die darauf ausgelegt sind, menschliche Konversation durch text- oder sprachbasierte Dialoge mit Benutzern zu simulieren (Zhai & Wibowo, 2023, s. 1). In der Literatur werden diese Technologien auch mit verschiedenen anderen Namen wie Bot, Chatterbot, Dialogsystem, Konversationsagent oder virtueller Assistent bezeichnet (Son et al., 2023, s. 102). Das grundlegende Funktionsprinzip dieser Programme basiert darauf, die Eingabe des Benutzers mit der Technologie der Verarbeitung natürlicher Sprache zu analysieren und eine Antwort auf der Grundlage der in der Systemdatenbank enthaltenen Informationen zu generieren (Haristiani, 2019, s. 2). Insbesondere im Bereich des Sprachenlernens haben Chatbots und virtuelle Gesprächspartner das Potenzial, Lernende in realistische Gesprächssituationen einzubeziehen, indem sie interaktive und authentische Dialogmöglichkeiten bieten (Malik et al., 2023, s. 1). Eines der populärsten Beispiele in diesem Bereich ist ChatGPT, ein von OpenAI entwickeltes, auf Transformationen basierendes großes Sprachmodell, das mit einem umfangreichen Textdatensatz trainiert wurde (Liu et al., 2023, s. 72). Es wird prognostiziert, dass ChatGPT sowohl für Lernende als auch für Lehrende ein wertvolles Werkzeug sein kann, da es das Potenzial hat, die

Lernerfahrungen der Studierenden zu bereichern und die Arbeitslast der Lehrkräfte zu reduzieren (Liu et al., 2023, s. 89).

Ein weiteres wichtiges Anwendungsfeld der KI-Technologien sind intelligente persönliche Assistenten, die weit verbreitet eingesetzt werden, um Benutzer bei der Erledigung alltäglicher Aufgaben zu unterstützen und ihre Erfahrungen zu personalisieren (Kohnke et al., 2023; Wang et al., 2024). Sprachbasierte KI-Anwendungen, die in der Branche weithin bekannt sind, wie Alexa, Siri und Google Assistant, fallen in diese Kategorie (Wang et al., 2024, s. 814). Diese Assistenten nutzen die Technologie der automatischen Spracherkennung, um die Sprachbefehle der Benutzer zu erkennen und in Text umzuwandeln (Son et al., 2023, s. 100). Ihre Fähigkeit, realistische Interaktionen zu führen, macht intelligente persönliche Assistenten für das Sprachenlernen äußerst geeignet (Wang et al., 2024, s. 814). Tatsächlich zeigt sich, dass sie von Zweitsprachenlernern besonders im Kontext von Aussprache- und Sprechübungen positiv aufgenommen werden (Dizon, 2017, s. 824). Sie bergen das Potenzial, autonomes Lernen zu fördern, indem sie Lernenden außerhalb des traditionellen Klassenzimmers sinnvolle Interaktionsmöglichkeiten schaffen (Dizon, 2017, s. 825). Darüber hinaus steigert die Möglichkeit, den Lernenden ein angstfreies Übungsumfeld zu bieten, ihr Potenzial im Zweit- und Fremdsprachenerwerb (Son et al., 2023, s. 101). Eine diesbezügliche Studie hat gezeigt, dass die Verwendung von Amazon Echo Show (Alexa) einen positiven Einfluss auf die Hör- und Sprechfähigkeiten von Studierenden hatte, die EaF lernen. In der Studie wurde festgestellt, dass Studierende, die regelmäßig mit dem intelligenten persönlichen Assistenten interagierten, eine signifikante Verbesserung ihrer Sprechfähigkeiten zeigten und dass die durch solche Werkzeuge gebotenen mündlichen Interaktionsmöglichkeiten die Sprechangst reduzierten (Qiao & Zhao, 2023, s. 4). Interessanterweise wird berichtet, dass Lehrkräfte sich möglicherweise nicht bewusst waren, dass Assistenten wie Siri oder Google Assistant, die sie zur Erleichterung ihrer Aufgaben im Alltag nutzten, KI-Anwendungen sind, bis generative KI-Tools wie ChatGPT populär wurden (Kohnke et al., 2023, s. 3f).

ChatGPT

ChatGPT, das im November 2022 von OpenAI der Öffentlichkeit vorgestellt wurde, ist ein auf KI basierender Chatbot (Adeshola & Adepoju, 2023; Alkaissi & McFarlane, 2023; Baidoo-Anu & Owusu Ansah, 2023; Lo, 2023; Roumeliotis & Tselikas, 2023). Innerhalb von zwei Monaten nach seiner Einführung erreichte es über 100 Millionen Nutzer und wurde damit zur schnellst wachsenden Verbraucheranwendung in der Geschichte (Eysenbach, 2023). Seine technische Infrastruktur nutzt einen vortrainierten Transfer-Learning-Ansatz auf der Grundlage riesiger Textdatenmengen aus dem Internet wie Büchern und Artikeln (Liu et al., 2023) sowie die Transformer-Architektur zum Verständnis kontextueller Beziehungen innerhalb von Texten (Vaswani et al., 2017, zitiert nach Cao et al., 2023; Rahman & Watanobe, 2023; Roumeliotis & Tselikas, 2023). Das Modell wurde speziell mit der Technik des bestärkenden Lernens durch menschliches Feedback trainiert; dieser Prozess bringt die Antworten des Modells stärker in Einklang mit menschlichen Präferenzen und Werten (Liu et al., 2023; Ouyang et al., 2022, zitiert nach Qin et al., 2023).

Zu seinen Kernfähigkeiten gehören das Erzeugen (Generative) neuer Inhalte wie Texte, Code und Gedichte (Haleem, Javaid & Singh, 2023; Javaid, Haleem & Singh, 2023), die Beantwortung komplexer Fragen (Haleem et al., 2023; Liu et al., 2023), das Schreiben und Debuggen von Code (Rahman & Watanobe, 2023) sowie das Zusammenfassen länger Texte und die Übersetzung zwischen Sprachen (Javaid, Haleem, Singh, Khan & Khan, 2023; Liu et al., 2023). ChatGPT, eine Variante eines Modells der GPT-3.5-Serie (Antaki, Touma, Milad, El-Khoury, & Duval, 2023), wurde so konzipiert, dass Fehler und unangemessene Ausdrücke früherer Sprachmodelle minimiert werden (Rahman & Watanobe, 2023; Taecharungroj, 2023). Das Gemini-Modell von Google (früher bekannt als Bard) gilt als einer der wichtigsten Konkurrenten von ChatGPT (Carlà et al., 2024; Lee et al., 2024; Gemini Team, 2023), und der Wettbewerb zwischen diesen beiden Modellen beschleunigt

die Innovationen im Bereich der GenKI (Rudolph, Tan & Tan, 2023). ChatGPT stellt einen wichtigen Meilenstein in der Geschichte der KI dar und birgt sowohl durch seine technischen Fähigkeiten als auch durch seine einer breiten Öffentlichkeit zugängliche Benutzeroberfläche ein transformatives Potenzial für viele Bereiche der Gesellschaft (Adeshola & Adepoju, 2023; Atlas, 2023; Roumeliotis & Tselikas, 2023).

DeepSeek

DeepSeek, ein in China ansässiges Unternehmen für KI, wurde von dem Unternehmer Liang Wenfeng gegründet und fordert die Dominanz westlicher KI-Giganten durch das Angebot von Open-Source-, hochleistungsfähigen und kostengünstigen Modellen heraus (Sallam, Al-Mahzoum, Sallam & Mijwil, 2025). Der strategische Erfolg des Unternehmens basiert auf dem Erwerb von über 10.000 Nvidia A100 GPUs vor den US-amerikanischen Chip-Exportbeschränkungen (Sallam et al., 2025). DeepSeek bietet eine breite Palette von Modellen an, die von effizienten Mixture-of-Experts-Modellen wie DeepSeek-V2 und V3 (DeepSeek-AI, 2024c, 2024d) über den auf Codierung spezialisierten DeepSeek-Coder (Zhu et al., 2024) bis hin zum auf logisches Schlussfolgern ausgerichteten DeepSeek-R1 (DeepSeek-AI, 2025) reicht. Dieser Ansatz birgt das Potenzial, die globale Innovation zu fördern, indem er die Demokratisierung der KI ermöglicht (Sallam et al., 2025).

Im Zentrum der technologischen Innovationen von DeepSeek steht die Mixture-of-Experts-Architektur, die trotz einer enormen Anzahl von Parametern für jeden Token nur einen kleinen Teil davon aktiviert (DeepSeek-AI, 2024c, 2024d; Sallam et al., 2025). Darüber hinaus steigert der Multi-head-Latent-Attention-Mechanismus die Effizienz, indem er den Key-Value-Cache komprimiert, was die Speichernutzung während der Inferenz erheblich reduziert (DeepSeek-AI, 2024c, 2024d). Eines der bemerkenswertesten Modelle des Unternehmens, DeepSeek-R1, wurde direkt durch bestärkendes Lernen trainiert, wobei die Phase des überwachten Feinabstimmungs übersprungen wurde. In diesem Prozess

entwickelte es menschenähnliche Fähigkeiten zum logischen Schlussfolgern, wie beispielsweise die schrittweise Generierung von Antworten (DeepSeek-AI, 2025; Sallam et al., 2025). Diese Offenheit birgt jedoch auch ernsthafte existenzielle Risiken (Sallam et al., 2025). Es besteht die Sorge, dass böswillige Akteure diese Modelle für Desinformation, Cyberangriffe oder den Entwurf biologischer Waffen nutzen könnten (Sallam et al., 2025). Tatsächlich wurde in einer Sicherheitsbewertung berichtet, dass DeepSeek-R1 eine Angriffs-Erfolgsrate von 100% gegenüber schädlichen Anfragen aufwies (Zhang et al., 2025).

Gemini

Gemini, von Google entwickelt, ist eine Familie hochmoderner, multimodaler Modelle für KI, die darauf ausgelegt ist, verschiedene Datentypen wie Text, Code, Bilder, Audio und Video integriert zu verarbeiten (Gemini Team, 2023; McIntosh et al., 2025; Perera & Lankathilaka, 2023). Die Gemini-Familie, ursprünglich als Nachfolger von Google Bard angekündigt (Carlà et al., 2024) und als Konkurrent zu OpenAIs ChatGPT-Modell positioniert (McIntosh et al., 2025), wird in drei verschiedenen Größen angeboten: Ultra für die komplexesten Aufgaben, Pro für hochskalierbare Leistung und Nano, das für den effizienten Einsatz auf Geräten wie Smartphones konzipiert wurde (Gemini Team, 2023; McIntosh et al., 2025). Das grundlegende Unterscheidungsmerkmal von Gemini ist, dass es von Grund auf multimodal trainiert wurde; dadurch kann es Daten in verschiedenen Formaten wie Text, Bilder, Audio und Video auf natürliche Weise gemeinsam verstehen und verarbeiten (Gemini Team, 2023; Imran & Almusharraf, 2024).

Die Architektur von Gemini basiert auf Transformer-Decodern und verwendet in fortschrittlichen Versionen wie Gemini 1.5 Pro eine effizientere Mixture-of-Experts-Architektur (Gemini Team, 2023; Gemini Team, 2024). Eine der bemerkenswertesten Fähigkeiten dieser Modellfamilie ist die Verarbeitung von extrem langen Kontextfenstern,

die Millionen von Tokens umfassen können; dies stellt einen Generationssprung gegenüber bestehenden Modellen wie Claude 3.0 (200k) und GPT-4 Turbo (128k) dar (Gemini Team, 2024, s. 1). Dank dieser Fähigkeit kann Gemini 1.5 Pro stundenlange Video- oder Audioaufnahmen oder eine gesamte Codebasis in einer einzigen Eingabe analysieren (Gemini Team, 2024, s. 1). Gemini Ultra zeichnet sich auch dadurch aus, dass es das erste Modell ist, das in anspruchsvollen akademischen Benchmarks wie MMLU die Leistung menschlicher Experten übertrifft (Gemini Team, 2023; McIntosh et al., 2025).

Die Leistung von Gemini hat in verschiedenen Vergleichsstudien zu unterschiedlichen Ergebnissen geführt. So zeigten beispielsweise sowohl ChatGPT als auch Gemini eine hohe Genauigkeit bei Fragen zur Patientenschulung über Bluthochdruck, wobei jedoch festgestellt wurde, dass die Antworten von Gemini ein niedrigeres Leseniveau aufwiesen (also zugänglicher waren) (Lee et al., 2024, s. 4). In Analysen zur politischen Ausrichtung wurde festgestellt, dass Gemini eine eher zentristische Haltung einnimmt, während ChatGPT eine liberale Tendenz aufweist (Choudhary, 2025, s. 1, 11). Einige Bewertungen haben gezeigt, dass die faktische Genauigkeit von Gemini im Vergleich zu ChatGPT-4 geringer sein kann (Kalluri, Kokala & Parvatha, 2024, s. 4675), die Halluzinationsrate jedoch niedriger ist als bei anderen Modellen wie Kimi (Shan, Ming, Hong & Wu, 2024, s. 4). Dies deutet darauf hin, dass Gemini ein leistungsfähiges KI-Werkzeug ist, dessen Leistung jedoch je nach Anwendungsfall und verglichenem Modell variieren kann.

TEIL III

METHODE

In diesem Kapitel werden Angaben zum Forschungsmodell, zum Untersuchungsmaterial, zum Prozess und den Instrumenten der Datenerhebung sowie zur Analyse und Interpretation der Daten dargelegt.

Forschungsdesign

Da diese Studie darauf abzielt, die Leistung von KI-Chatbots bei der Fehlererkennung, Korrektur und Feedback in deutschen Texten vergleichend zu bewerten, wurde grundsätzlich eine quantitative Forschungsstrategie verfolgt (Bryman, 2015; Creswell & Creswell, 2023; Johnson & Christensen, 2020). Quantitative Forschung zielt darauf ab, vorab festgelegte Hypothesen zu testen, Beziehungen zwischen Variablen zu untersuchen und Ergebnisse zu generalisieren (Creswell & Creswell, 2023, s. 5, 18). Das Design der Studie basiert auf einer experimentellen Logik, da es darauf abzielt, die Wirkung der unabhängigen Variablen (Art des KI-Chatbots und Textniveau) auf die abhängigen Variablen (Leistungsmaße) zu untersuchen (Bryman, 2015; Creswell & Creswell, 2023; Johnson & Christensen, 2020). Genauer gesagt, wurde in dieser Untersuchung ein quasi-experimentelles Design verwendet. Während bei einem echten Experiment die Teilnehmenden den Gruppen zufällig zugewiesen

werden müssen (Bryman, 2015; Johnson & Christensen, 2020), handelt es sich bei den in dieser Studie als Teilnehmende fungierenden KI-Modellen (DeepSeek, ChatGPT, Gemini) um bereits existierende und strukturell voneinander abweichende Gruppen. Da eine zufällige Zuweisung zu diesen Gruppen nicht möglich ist, trägt die Studie einen quasi-experimentellen Charakter (Bryman, 2015; Creswell & Creswell, 2023; Johnson & Christensen, 2020). Die drei verschiedenen KI-Modelle fungieren füreinander als Vergleichsgruppen (Bryman, 2015, s. 51). Die Untersuchung weist zugleich Merkmale einer strukturierten Beobachtung oder einer Feldstimulation auf, bei der eine spezielle Situation (fehlerhafte Texte) geschaffen wurde, um eine bestimmte Reaktion (das Fehlerkorrekturverhalten der KI) hervorzurufen (Bryman, 2015, s. 277). Indem die Texte absichtlich Fehler enthalten gestaltet wurden, wurde ein Stimulus erzeugt, und die Reaktionen der KI-Modelle auf diesen Stimulus wurden systematisch beobachtet (Bryman, 2015; Johnson & Christensen, 2020).

Studienmaterial und Datensatz

Das Studienmaterial besteht aus jeweils 20 von KI erstellten Texten auf den Niveaustufen A1, A2 und B1. Jeder Text wurde anschließend so überarbeitet, dass er Fehler enthielt. Bei diesem Vorgehen handelt es sich um eine Form der gezielten Stichprobenziehung, bei der Fälle oder Dokumente mit spezifischen Merkmalen, die für die Bearbeitung der Forschungsfragen erforderlich sind, strategisch ausgewählt werden (Bryman, 2015; Creswell & Plano Clark, 2018; Johnson & Christensen, 2020; Patton, 2015).

Diese ausgewählten Texte wurden so manipuliert, dass sie die unabhängige Variable der Untersuchung darstellen. Auf der Grundlage von Corders Klassifikation der Fehleranalyse wurde in jeden Text gezielt jeweils ein Fehler bestimmter Fehlertypen integriert. Auf diese Weise wurde jeder Text auf jeder Niveaustufe (A1, A2, B1) zu einem standardisierten Stimulus-Set für die Messung der Leistung der KI-Modelle umgewandelt.

Datenerhebungsprozess und -instrumente

Zur Datenerhebung wurde in dieser Studie der KI-Chatbot Gemini eingesetzt. Innerhalb von Gemini wurde ein spezielles GEM erstellt, und bei der Erstellung dieses GEM wurden diesem die folgenden Anweisungen bzw. Prompts erteilt:

- Rolle: Du bist eine erfahrene Lehrkraft, die Materialien für das Erlernen der deutschen Sprache vorbereitet.
- Aufgabe: Du wirst unter Verwendung der von mir bereitgestellten deutschen Wortlisten für die Niveaustufen A1, A2 und B1 sowie der von mir angestrebten Sprachniveaustufe einen originellen Lesetext erstellen.
- Regeln:
    - Der zu erstellende Text muss eine Länge von 250–300 Wörtern haben.
    - Der Text muss der von mir angegebenen [ZIELSTUFE] entsprechen.
    - Im Text sollen überwiegend Wörter aus den Wortlisten der [ZIELSTUFE] und niedrigerer Stufen verwendet werden. (Beispiel: Ist das Ziel A2, sollen überwiegend Wörter aus den A1- und A2-Listen verwendet werden).
    - Es muss eine sinnvolle und kohärente Geschichte oder ein informativer Text erstellt werden.
    - Der Text muss einen deutschen Titel erhalten.
- Bereitgestellte Daten: Zielstufe [Wird angegeben]
- Erwünschte Ausgabe:
    - Titel: (Deutsch)
    - Text: (Lesetext)

Der KI wurden die vom Goethe-Institut veröffentlichten Wortlisten für A1, A2 und B1 zur Verfügung gestellt, und sie wurde angewiesen, die Wörter für die zu erstellenden Texte speziell aus diesen Listen auszuwählen. Auf der Grundlage dieser Anweisungen wurden von dem texterstellenden GEM jeweils 20 Texte für die Niveaustufen A1, A2 und B1 generiert. Anschließend mussten in diese Texte Fehler eingefügt werden. Für diesen Vorgang wurde ebenfalls ein GEM erstellt, und der Prozess der Fehlereinfügung wurde ebenfalls von der KI durchgeführt. Bei der Einfügung der Fehler erhielt die KI die folgenden Anweisungen:

- Aufgabendefinition Und Identität: Du bist ein KI-Assistent, der auf die Bereiche Zweitspracherwerb und deutsche Linguistik spezialisiert ist und die Fehleranalysetheorie von S. Pit Corder beherrscht. Deine Hauptaufgabe ist es, die dir vorgegebenen deutschen Lesetexte auf den Niveaustufen A1, A2 oder B1 gemäß den unten aufgeführten strengen Regeln absichtlich mit Fehlern zu versehen und neu zu gestalten. Bei der Ausführung dieser Aufgabe wirst du dich auf Corders Arbeit von 1982 stützen, die dir als Referenz dient, jedoch wirst du ausschließlich die von mir angegebenen Fehlerkategorien und -typen verwenden.
- Anzuwendende Regeln: Für jeden dir vorgegebenen Text musst du exakt 8 Fehler erstellen. Diese Fehler müssen sich buchstabengetreu an die folgende Struktur halten:
    - Zu ignorierende Fehlerebene: Die Fehlerebene "Graphological/Phonological" von Corder wirst du AUF KEINEN FALL verwenden.
    - Zu verwendende Fehlerebenen: Die Fehler werden nur auf diesen beiden Ebenen liegen:
        - Grammatical (Grammatisch)
        - Lexico-semantic (Lexikalisch-semantisch)
- Fehlerverteilung: Die insgesamt 8 Fehler müssen im Text wie folgt verteilt werden:

- 4 Fehler auf der Grammatical-Ebene: Bei diesem Fehlertyp müssen alle Fehler die Grammatik betreffen.
    - 1 Omission (Auslassung): Dieser Fehlertyp bedeutet, dass ein grammatisch notwendiges Element im Satz fehlt. Zum Beispiel das Fehlen grammatischer Elemente wie eines Verbs oder einer Präposition im Satz.
    - 1 Addition (Hinzufügung): Dieser Fehlertyp bedeutet, dass ein grammatisch nicht notwendiges Element im Satz vorhanden ist. Zum Beispiel werden dem Satz zusätzlich grammatische Elemente wie ein Verb oder eine Präposition hinzugefügt.
    - 1 Selection (Falsche Auswahl): Dieser Fehlertyp bedeutet, dass anstelle eines grammatisch korrekten Elements ein anderes Element verwendet wird. Zum Beispiel kann das Verb im Satz falsch konjugiert sein, der Artikel kann aufgrund einer Präposition im falschen Kasus stehen oder der Artikel kann in einem falschen Kasus stehen.
    - 1 Ordering (Falsche Reihenfolge): Dieser Fehlertyp bezieht sich auf die falsche Anordnung grammatischer Elemente im Satz. Zum Beispiel wird bei der Adjektivdeklination im Satz eine Reihenfolge wie Adjektiv+Artikel+Nomen statt Artikel+Adjektiv+Nomen verwendet. Oder im Perfekt steht das Partizip 2 an 3. Position, obwohl es am Satzende stehen müsste.
- 4 Fehler auf der Lexico-semantic-Ebene: Bei diesem Fehlertyp müssen alle Fehler Wortbedeutung betreffen.
    - 1 Omission (Auslassung): Dieser Fehlertyp bedeutet, dass ein für die Satzbedeutung notwendiges Element im Satz fehlt. Zum Beispiel das

Fehlen lexikalisch-semantischer Elemente wie Substantiv, Adjektiv, Pronomen oder Subjekt im Satz.

- 1 Addition (Hinzufügung): Dieser Fehlertyp bedeutet, dass ein für die Satzbedeutung nicht notwendiges Element im Satz vorhanden ist. Zum Beispiel, dass zusätzlich zu lexikalisch-semantischen Elementen wie Substantiv, Adjektiv, Pronomen oder Subjekt ein zusätzliches Element vorhanden ist.
- 1 Selection (Falsche Auswahl): Dieser Fehlertyp bedeutet, dass anstelle eines für die Satzbedeutung notwendigen Elements ein anderes Element verwendet wird. Zum Beispiel, dass anstelle von lexikalisch-semantischen Elementen wie Substantiv, Adjektiv, Pronomen oder Subjekt ein Element verwendet wird, das die semantische Kohärenz des Satzes stört.
- 1 Ordering (Falsche Reihenfolge): Dieser Fehlertyp bedeutet, dass Fehler in der Satzstruktur vorliegen, die für die Satzbedeutung notwendig ist. Zum Beispiel, dass lexikalisch-semantische Elemente wie Substantiv, Adjektiv, Pronomen oder Subjekt im Satz an einer anderen als der erforderlichen Position stehen.

- Du sollst nicht für jeden Fehlertyp auf die gleiche Weise Fehler machen. Zum Beispiel solltest du beim Fehlertyp Lexiko-Semantic Addition nicht immer dasselbe Wort an derselben Stelle hinzufügen. Du musst einen Fehler machen, der kontextuell zum Satz passt.

- Erläuterungen Zu Den Fehlerarten Und Beispiele:
  - Berücksichtige bei der Erstellung der Fehler die folgenden Definitionen und Beispiele:

- Grammatical (Grammatische) Fehler:
    - Omission: Das Auslassen eines für die grammatische Struktur des Satzes notwendigen Elements (z. B. Artikel, Präposition, Hilfsverb). Beispiel: "Ich gehe in Schule" (korrekt: "in die Schule").
    - Addition: Das Hinzufügen eines in der grammatischen Struktur des Satzes unnötigen Elements (z. B. ein überflüssiger Artikel, eine falsche Pluralendung). Beispiel: "Er hat ein Geld" (korrekt: "Er hat Geld").
    - Selection: Die Auswahl eines falschen grammatischen Elements (z. B. falsches Tempus, falscher Artikel, falscher Kasus – Dativ statt Akkusativ). Beispiel: "Ich helfe dich" (korrekt: "dir").
    - Ordering: Die Anordnung von Wörtern oder Satzgliedern entgegen den grammatischen Regeln (z. B. das Verb im Nebensatz steht nicht am Ende). Beispiel: "Ich weiß, dass er kommt heute" (korrekt: "...dass er heute kommt").
- Lexico-semantic (Lexikalisch-semantische) Fehler:
    - Omission: Das Auslassen eines Wortes (Verb, Substantiv usw.), das zur Vervollständigung der Satzbedeutung notwendig ist. Beispiel: "Der Mann ist sehr" (Adjektiv fehlt, z. B. "glücklich").
    - Addition: Das Hinzufügen eines semantisch unnötigen oder kontextfremden Wortes. Hier ist zu beachten, dass das hinzugefügte Wort dem Kontext des betreffenden Satzes

ähneln, aber dennoch falsch sein sollte. Der Ausdruck "Auto" im Satz "Ich trinke Kaffee Auto am Morgen." ist ein offensichtlicher Fehler im Kontext. Stattdessen muss ein Wort eingefügt werden, das kontextuell sinnvoll ist, aber die semantische Kohärenz im Satz stört und unnötig ist. Beispiel: "Ich trinke Kaffee Brot am Morgen." (Das Wort Brot passt als Lebensmittel in den Kontext. Es ist im Satz jedoch semantisch irrelevant).

- Selection: Die Auswahl eines semantisch falschen Wortes (z. B. ein ähnlich klingendes, aber anders bedeutsames Wort, "falsche Freunde" (false friends), ein falsches Wort aus derselben Kategorie). Beispiel: "Ich schreibe mit einem Löffel" (korrekt: "Stift"). (Falsches Objekt für die Tätigkeit des Schreibens).
- Ordering: Die falsche Anordnung der Teile eines semantisch zusammengehörigen Ausdrucks oder eines zusammengesetzten Wortes. Mache dabei keinen Fehler, indem du die Teile zusammengesetzter Substantive vertauschst. Wenn du beispielsweise das Wort Tomatensoße in Soßentomate änderst, kann dies nicht als Lexico-semantic Ordering-Fehler gewertet werden. Bitte achte darauf.
- Beispiel: "Ich stehe um 7 Uhr auf" (falsche Position des trennbaren Teils des Verbs). Dieser Fehler kann sich manchmal mit der grammatischen Wortstellung überschneiden, achte also darauf, dass es sich um eine die semantische Kohärenz störende Reihenfolge handelt.

- Ausgabeformat: Du musst deine Antwort in dem folgenden Format unter zwei Hauptüberschriften präsentieren:
    - Fehlerhafter Text: Die endgültige Version des Textes, in die du die 8 Fehler eingefügt hast.
    - Fehleranalyse-Liste: Liste jeden von dir erstellten Fehler nummeriert mit den folgenden Details auf: Fehler 1: [Ebene] - [Typ]
    - Position: Schreibe hier den gesamten fehlerhaften Satz aus dem Text.
    - Fehlerdetail: Zeige den Unterschied zwischen der ursprünglichen/korrekten Version des Satzes und der fehlerhaften Version. (Bsp.: Korrekt: "...in die Schule." -> Fehlerhaft: "...in Schule.")
    - Erläuterung: Erkläre kurz, warum dieser Fehler der angegebenen Ebene und dem angegebenen Typ zugeordnet wird. Fehler 2: [Ebene] - [Typ] Position:... Fehlerdetail: ... Erläuterung: ...
    - (Diese Struktur wird für alle 8 Fehler wiederholt)
- Finale Checkliste
    - Bleibt die ursprüngliche Bedeutung des Textes weitgehend erhalten?
    - Ähneln die erstellten Fehler natürlichen Fehlern, die ein Lernender auf dem Niveau des Textes machen könnte?
    - Gibt es insgesamt exakt 8 Fehler?
    - Gibt es von jedem Fehlertyp und jeder Ebene genau die angegebene Anzahl?
    - Entspricht das Ausgabeformat exakt den Vorgaben?
    - Ist die Erläuterung für den 8. Fehler vollständig angegeben?
    - Sind alle Fehler vollständig wie gewünscht bearbeitet und mitgeteilt worden?

- BEGINNE MIT DER AUFGABE.

Nachdem die KI jeden Text im gewünschten Format bearbeitet, d. h. die Texte so umgeschrieben hatte, dass sie Fehler enthielten, wurde dieser Zustand vom Autor kontrolliert. Nachdem überprüft worden war, ob jede Fehlerebene und jeder Fehler in jedem Text vorhanden waren, wurde zur nächsten Phase übergegangen.

In der nächsten Phase wurden die 20 auf A1-Niveau, 20 auf A2-Niveau und 20 auf B1-Niveau verfassten fehlerhaften Texte an Gemini, DeepSeek und ChatGPT gesendet. Der folgende Prompt wurde in dieser Phase allen KI-Chatbots vor allen Texten wiederholt gegeben:

- Rolle: Sie sind ein erfahrener Deutschlehrer und Korrekturleser. Ich lege Ihnen einen deutschen Text auf [A1/A2/B1] vor. Dieser Text enthält einige Fehler, die Teil des Lernprozesses sind. Ihre Aufgabe ist es, die folgenden drei Schritte der Reihe nach und vollständig zu bearbeiten:
- Fehlererkennung: Identifizieren Sie alle grammatikalischen und lexikalisch-semantischen Fehler im Text und listen Sie sie in Stichpunkten auf.
- Textkorrektur: Korrigieren Sie die Fehler und schreiben Sie den Text in eine vollständig korrekte und flüssige Version um.
- Feedback: Erklären Sie für jeden gefundenen Fehler klar und prägnant, warum der Fehler falsch ist und welche deutsche Regel richtig ist.
- Ausgabenformat: Bitte gliedern Sie Ihre Antwort in die Überschriften „1. Erkannte Fehler“, „2. Korrigierter Text“ und „3. Feedback“. Text:

Nach diesem Vorgang wurde überprüft, ob die KI-Chatbots die in jedem Text vorhandenen insgesamt 8 Fehler unterschiedlicher Ebenen und Arten erkannt, korrekt korrigiert und zutreffendes Feedback gegeben haben; zudem wurden die Daten einzeln gezählt. Zur Gewährleistung der Reliabilität dieses Vorgangs wurde dieser Prozess nach der Berechnung

durch den Autor auch von den anderen KI-Chatbots durchgeführt. Die als Ergebnis dieses Vorgangs gewonnenen Daten wurden in einer Excel-Datei aufgelistet und dort nach verschiedenen Fehlertypenebenen, Sprachniveaus, Chatbots und anderen Faktoren aufgeschlüsselt. Durch all diese Vorgänge erfolgte die Bewertung der gegebenen Antworten aus verschiedenen Perspektiven. Für jeden erfolgreichen Vorgang wurde eine Häufigkeit ermittelt, und diese Häufigkeit wurde anschließend prozentual ausgedrückt. Darüber hinaus wurde auch bei der Interpretation der im Abschnitt *Befunde und Kommentare* vorhandenen Tabellen auf KI-Chatbots zurückgegriffen. Zusätzlich zu den Kommentaren des Autors wurden auch Kommentare von der KI angefordert, und deren Kommentare wurden ebenfalls berücksichtigt. Insbesondere konnte die KI einige Daten, die der Autor übersehen hatte, mühelos erkennen und lieferte Informationen dazu.

Datenanalyse

Die Analyse der erhobenen Daten wurde unter Verwendung quantitativer Datenanalysetechniken durchgeführt (Creswell & Creswell, 2023; Hair, Black, Babin & Anderson, 2018). Das Hauptziel der Analyse besteht darin, die Leistung jedes KI-Chatbots auf den verschiedenen Sprachniveaus vergleichend darzustellen.

Im Datenanalyseprozess wurden zunächst, ausgehend von den kodierten Leistungsdaten für jedes KI-Modell und jedes Sprachniveau (A1, A2, B1), im Rahmen der univariaten Analyse, die die Verteilung einer einzelnen Variablen untersucht, Häufigkeitsverteilungen erstellt (Bryman, 2015, s. 333). Diese Häufigkeiten wurden anschließend in prozentuale Werte umgerechnet, um die Erfolgsquoten der Modelle bei der Fehlererkennung, Korrektur und Rückmeldung zu berechnen.

Validität und Reliabilität

Um die wissenschaftliche Qualität dieser als quantitative Studie angelegten Arbeit zu gewährleisten, wurden verschiedene Maße der Validität und Reliabilität berücksichtigt.

- Reliabilität: Reliabilität bezieht sich auf die Konsistenz oder Stabilität einer Messung (Bryman, 2015; Field, 2024; Hair et al., 2018; Johnson & Christensen, 2020). In dieser Studie ist die Konsistenz des Kodierungsprozesses, bei dem die KI-Outputs bewertet wurden, von entscheidender Bedeutung. Zu diesem Zweck wurde auf die Gewährleistung der Inter-Rater-Reliabilität geachtet (Bryman, 2015; Johnson & Christensen, 2020). Da alle Schritte der Untersuchung detailliert beschrieben wurden, ist auch die Replizierbarkeit der Studie hoch (Bryman, 2015, s. 40, 164).
- Validität: Validität bezieht sich auf die Frage, ob ein Instrument tatsächlich das misst, was es zu messen beabsichtigt (Bryman, 2015; Field, 2024; Hair et al., 2018; Johnson & Christensen, 2020).
    - Interne Validität: Die interne Validität befasst sich mit der Frage, ob ein Ergebnis eine kausale Beziehung beinhaltet und ob diese Beziehung robust ist (Bryman, 2015; Johnson & Christensen, 2020). Obwohl in diesem quasi-experimentellen Design die fehlende randomisierte Zuweisung zu den Gruppen eine Bedrohung für die interne Validität darstellt (Bryman, 2015, s. 51), wurden externe Faktoren wie die Aufgabenschwierigkeit kontrolliert, indem allen KI-Modellen exakt die gleichen manipulierten Texte vorgelegt wurden.
    - Messvalidität: Dies bezieht sich darauf, ob die erstellten Kriterien zur Messung der Fehlererkennung, Korrektur und Feedbackqualität tatsächlich die Sprachkompetenzleistung der KI messen (Bryman, 2015; Hair et al., 2018). Die Verwendung von Fehlern, die auf einem etablierten theoretischen Rahmen im Bereich des Sprachunterrichts (Corders Fehleranalyse) basieren,

trägt in dieser Studie zur Konstruktvalidität bei, da das gemessene Konzept eine theoretische Grundlage hat (Bryman, 2015; Field, 2024; Hair et al., 2018; Johnson & Christensen, 2020). Dass die Leistungskriterien die entsprechende Fähigkeit direkt widerspiegeln, unterstützt die Augenscheinvalidität, die bedeutet, dass ein Indikator den Inhalt des betreffenden Konzepts widerzuspiegeln scheint (Bryman, 2015; Field, 2024; Hair et al., 2018).

- Externe Validität: Die externe Validität befasst sich mit der Frage, ob die Ergebnisse einer Studie über den Forschungskontext hinaus generalisierbar sind (Bryman, 2015; Johnson & Christensen, 2020). Die Ergebnisse dieser Studie sind auf die verwendeten Texte und die drei untersuchten spezifischen KI-Modelle beschränkt. Ähnlich wie in der qualitativen Forschung wird angestrebt, die Ergebnisse dieser Studie nicht auf Populationen, sondern auf die Theorie zu generalisieren (Bryman, 2015; Patton, 2015).

TEIL IV

BEFUNDE UND INTERPRETATIONEN

In diesem Abschnitt der Studie wird versucht, die Ergebnisse, die aus den im Einklang mit der angewandten Methodik gewonnenen Daten hervorgehen, mithilfe von Tabellen zu erläutern.

Der in den Tabellen verwendete Begriff *Phase (P)* bezieht sich auf die Phase im Datenerhebungsprozess. Im Datenerhebungsprozess wurden 3 Phasen berechnet, jedoch ergab sich basierend auf den Antworten der KI-Chatbots eine weitere Phase. Die erste der 3 geplanten Phasen, d.h. *Phase 1 (P1)*, bezieht sich auf die Fehlererkennung. In dieser Phase versuchten die KI-Chatbots, die Fehler zu erkennen. Die zweite Phase, d.h. *Phase 2 (P2)*, ist die Phase, in der die erkannten Fehler von den KI-Chatbots korrigiert werden. Fehler, die in P1 nicht erkannt wurden, konnten hier naturgemäß nicht korrigiert werden. Die letzte Phase, d.h. *Phase 3 (P3)*, bezieht sich auf die Abgabe von korrigierendem Feedback durch die KI-Chatbots zu den erkannten und korrekt korrigierten Fehlern. Ein Fehler, der in P1 nicht erkannt wurde, konnte in P2 nicht korrigiert werden, und folglich konnte in P3 kein Feedback gegeben werden. Darüber hinaus konnte auch für Fehler, die in P1 zwar erkannt, aber in P2 nicht korrekt korrigiert wurden, in P3 kein Feedback gegeben werden. Die nicht geplante, sondern nachträglich hinzugefügte *Phase 4 (P4)* bezieht sich auf die Abgabe von nicht-korrigierendem Feedback durch die KI-Chatbots. Die KI-Chatbots machten manchmal

Vorschläge aus unterschiedlichen Gründen (wie Stil, Ausdrucksweise, Gebrauchshäufigkeit), obwohl ein Satz oder Wort korrekt war. Die Einbeziehung auch dieser Vorschläge in die Studie ist aus wissenschaftlicher Sicht produktiver. Zudem muss erwähnt werden, dass alle genannten Phasen innerhalb der von der KI gegebenen Antwort stattfinden. In jeder Antwort der KI-Chatbots wurden alle Phasen gesondert berücksichtigt; die Trennung der Phasen wurde vom Autor vorgenommen.

Der in den Tabellen vorhandene Begriff *Status (S)* hilft dabei, die Leistung der KI in der jeweiligen Phase zu erkennen. Der Begriff *Erfolgreich* drückt aus, dass der KI-Chatbot in der betreffenden Phase eine korrekte Handlung ausgeführt hat. Zum Beispiel, dass er in P1 den Fehler erkannt, in P2 den erkannten Fehler korrekt korrigiert und in P3 zu dem korrekt korrigierten Fehler ein korrektes Feedback gegeben hat. Der Begriff *Fehlgeschlagen* bezeichnet das genaue Gegenteil von *Erfolgreich*. Der im S enthaltene Begriff *Bedingt Fehlgeschlagen* gilt nur für die zweite und dritte Phase. Der Ausdruck *Bedingt Fehlgeschlagen* bedeutet, dass die KI automatisch als nicht erfolgreich gilt, wenn sie in der/den vorherigen Phase(n) nicht erfolgreich war. Wenn beispielsweise in der ersten Phase der Fehler nicht erkannt wurde, galten die zweite und dritte Phase automatisch als nicht erfolgreich; dieser Zustand wurde als *Bedingt Fehlgeschlagen* bezeichnet. Da in jeder Phase 20 Texte und 8 Fehler analysiert wurden, musste die Leistung hier summiert angegeben werden: *Gesamt Erfolgreich (GE)*, *Gesamt Fehlgeschlagen (GF)*, *Gesamt Bedingt Fehlgeschlagen(GBF)*. Der Begriff *Gesamt (GES)* für Phase 4 zeigt die Gesamtanzahl der vom KI-Chatbot gemachten Vorschläge an.

Des Weiteren sind in jeder Zeile prozentuale und nicht-prozentuale Daten vorhanden. Die nicht-prozentualen Daten messen die Leistung bei den jeweiligen Fehlern als Stückzahl, während die prozentual angegebenen Daten die Leistung in der jeweiligen Phase und Fehlerart prozentual darstellen.

Wie im Methodikteil erläutert, gibt es für jedes Fremdsprachenniveau 20 Texte, und in diesen 20 Texten befinden sich 8 Fehler unterschiedlicher Art. Diese Fehlerarten wurden in den Tabellen auf *grammatikalischer Ebene (GR)* und *lexikalisch-semantischer Ebene (LS)* als *Auslassung (AL), Hinzufügung (HF), Auswahl (AW)* und *Umstellung (UM)* abgekürzt.

Die Leistung von ChatGPT

Tabelle 2.
Die Leistung von ChatGPT für das Sprachniveau A1

P	S	Fehlerebene und -typ							
		GR				LS			
		AL	HF	AW	UM	AL	HF	AW	UM
P1	GE	20	20	20	20	19	20	14	19
		100%	100%	100%	100%	95%	100%	70%	95%
	GF	0	0	0	0	1	0	6	1
		0%	0%	0%	0%	5%	0%	30%	5%
P2	GE	20	20	19	20	19	20	14	19
		100%	100%	95%	100%	95%	100%	70%	95%
	GF	0	0	1	0	0	0	0	0
		0%	0%	5%	0%	0%	0%	0%	0%
	GBF	0	0	0	0	1	0	6	1
		0%	0%	0%	0%	5%	0%	30%	5%
P3	GE	19	20	18	19	17	20	12	19
		95%	100%	90%	95%	85%	100%	60%	95%
	GF	1	0	1	1	2	0	2	0
		5%	0%	5%	5%	10%	0%	10%	0%
	GBF	0	0	1	0	1	0	6	1
		0%	0%	5%	0%	5%	0%	30%	5%
P4	GES	0	0	5	1	0	1	1	2

In dieser Tabelle wird die Leistung des ChatGPT-Modells anhand des deutschen A1-Datensatzes im Kontext von vier verschiedenen Analysephasen und acht spezifischen Fehlerkategorien detailliert untersucht. Der Datensatz misst die Kompetenzen des Modells in den Prozessen der Fehlererkennung (Phase 1), Fehlerkorrektur (Phase 2), Feedback-Generierung (P3) und des nicht-korrektiven Feedbacks (Phase 4). Die Fehlerkategorien sind unter zwei Hauptüberschriften klassifiziert: Grammatik und Lexikalisch/Semantisch (GR-AL, GR-HF, GR-AW, GR-UM; LS-AL, LS-HF, LS-AW, LS-UM). Die gewonnenen Daten

zeigen, dass das Modell in den Fehlererkennungs- und Korrekturprozessen generell eine hohe Kompetenz aufweist, die Leistung jedoch je nach Komplexität der Aufgabe und dem spezifischen Fehlertyp variiert.

In der ersten Analysephase, der Fehlererkennung, wurde beobachtet, dass das Modell insbesondere bei Grammatikfehlern eine fehlerfreie Leistung erbringt. Das Modell erreichte in allen Kategorien GR-AL, GR-HF, GR-AW und GR-UM eine Erfolgsquote von 100 % (20/20). Auch in der Kategorie Lexikalisch/Semantisch sind die Erfolgsquoten bei den Fehlertypen LS-HF (100 %), LS-AL (95 %) und LS-UM (95 %) sehr hoch. Der deutlichste Leistungsabfall in dieser Phase wurde jedoch in der Kategorie LS-AW verzeichnet, in der die Erfolgsquote auf 70 % (14/20) sank; 6 Fälle (30 %) in dieser Kategorie konnten vom Modell nicht erkannt werden. Die Befunde der Fehlerkorrektur (P2) spiegeln diese Tendenz aus der Erkennungsphase weitgehend wider. Die Metrik GF blieb auf nur 1 Fall (5 %) in der Kategorie GR-AW beschränkt. Im Gegensatz dazu ist die Metrik GBF eine direkte Widerspiegelung der in Phase 1 nicht erkannten Fehler; die 6 Erkennungsfehler (30 %) in der Kategorie LS-AW sowie jeweils ein Fehler (5 %) in den Kategorien LS-AL und LS-UM wurden in dieser Phase als bedingt fehlgeschlagen registriert.

In der Phase der Feedback-Generierung (Phase 3), in der die Aufgabenkomplexität zunimmt, war ein merklicher Rückgang der Erfolgsquoten im Vergleich zu den vorherigen Phasen zu verzeichnen. Obwohl es dem Modell gelang, in den Kategorien GR-HF und LS-HF eine 100%ige Erfolgsquote beizubehalten, wurde die niedrigste Leistung auch in dieser Phase mit einer Erfolgsquote von 60 % (12/20) in der Kategorie LS-AW festgestellt. Die Erfolgsquoten in den anderen Kategorien variierten zwischen 85 % und 95 % (z. B. GR- AL: 95 %, GR-AW: 90 %, LS-AL: 85 %). Es fällt auf, dass in dieser Phase die Fälle von GF zugenommen haben; in den Kategorien GR-AL (5 %), GR-AW (5 %), GR-UM (5 %), LS-AL (10 %) und LS-AW (10 %) wurden direkte Feedback-Fehler beobachtet. Die Daten zu GBF bestätigen hingegen den aus Phase 1 übertragenen Zustand der Nichterkennung in den Kategorien LS-AW (30 %) und LS-UM (5 %). Schließlich ist das nicht-korrektive Feedback (P4) keine

Leistungsmetrik, sondern die Frequenz eines bestimmten Feedback-Typs. Es wurde festgestellt, dass diese Art von Feedback insgesamt 10 Mal generiert wurde, sich am häufigsten auf die Kategorie GR-AW (5 Fälle) konzentrierte und in den Kategorien GR-AL, GR-HF und LS-AL überhaupt nicht (0 Fälle) verwendet wurde.

Zusammenfassend zeigt diese umfassende Analyse des A1-Datensatzes, dass das ChatGPT-Modell eine außerordentlich hohe Erkennungs- und Korrekturkompetenz aufweist, insbesondere in allen Grammatik-Kategorien und der LS-HF-Kategorie. Dennoch legen die gewonnenen Ergebnisse drei kritische Tendenzen offen: Erstens ist mit zunehmender Aufgabenkomplexität (von Phase 1 zu Phase 3) ein leichter Leistungsabfall des Modells zu verzeichnen. Zweitens stellt die Kategorie LS-AW mit durchweg den niedrigsten Erfolgsquoten in allen Leistungsphasen (Erkennung: 70 %, Korrektur: 70 %, Feedback: 60 %) einen spezifischen Problembereich für das Modell dar. Drittens hat die Analyse das Phänomen der Fehlerfortpflanzung deutlich gemacht; es wurde nachgewiesen, dass Fehler in der Erkennungsphase (P1) die Leistung in den nachfolgenden Korrektur- und Feedback-Phasen (Phase 2 und 3) als bedingtes Scheitern direkt beeinflussen.

Tabelle 3.

Die Leistung von ChatGPT für das Sprachniveau A2

P	S	Fehlerebene und -typ							
		GR				LS			
		AL	HF	AW	UM	AL	HF	AW	UM
P1	GE	20	20	20	20	17	20	17	19
		100%	100%	100%	100%	85%	100%	85%	95%
	GF	0	0	0	0	3	0	3	1
		0%	0%	0%	0%	15%	0%	15%	5%
P2	GE	20	19	20	20	17	19	16	19
		100%	95%	100%	100%	85%	95%	80%	95%
	GF	0	1	0	0	0	1	1	0
		0%	5%	0%	0%	0%	5%	5%	0%
	GBF	0	0	0	0	3	0	3	1
		0%	0%	0%	0%	15%	0%	15%	5%
P3	GE	17	17	16	19	16	17	15	18
		85%	85%	80%	95%	80%	85%	75%	90%
	GF	3	2	4	1	1	2	1	1
		15%	10%	20%	5%	5%	10%	5%	5%
	GBF	0	1	0	0	3	1	4	1
		0%	5%	0%	0%	15%	5%	20%	5%
P4	GES	0	1	4	5	0	0	3	1

Diese Tabelle untersucht die Leistung des ChatGPT-Modells auf dem A2-Niveau-Datensatz in einem vierstufigen Analyserahmen (Fehlererkennung, Korrektur, Feedback-Generierung und nicht-korrektives Feedback). Die analysierten Daten zeigen, dass das Modell auf dem A2-Niveau sowohl bemerkenswerte Erfolge als auch deutliche Problembereiche aufweist. Der auffälligste Befund ist, dass das Modell bei der Erkennung (P1) von Fehlern in allen Grammatik-Kategorien (AL, HF, AW, UM) eine fehlerfreie Erfolgsquote von 100 % aufweist. Es wurde jedoch festgestellt, dass die deutlichste Schwäche des Modells im A2-Datensatz in den Lexikalisch/Semantischen Kategorien liegt, insbesondere bei den Fehlertypen LS-AL und LS-AW.

Bei einer spezifischen Betrachtung der Daten für Phase 1 (Fehlererkennung) wird eine klare interkategoriale Unterscheidung in der Fehlererkennungsfähigkeit des Modells beobachtet. Wie bereits erwähnt, hat das Modell in allen vier Grammatikkategorien (GR-AL, GR-HF, GR-AW, GR-UM) alle 20 von 20 Fällen korrekt erkannt. Im lexikalisch/semantischen

Bereich wurden ebenfalls hohe Erfolgsquoten von 100 % (20/20) in der Kategorie LS-HF und 95 % (19/20) in der Kategorie LS-UM verzeichnet. Im Gegensatz dazu wurde die niedrigste Leistung in dieser Phase mit einer Erfolgsquote von 85 % (17/20) sowohl in den Kategorien LS-AL als auch LS-AW beobachtet; in jeder dieser Kategorien konnten 3 Fälle (15 %) vom Modell nicht erkannt werden.

Die Untersuchung von Phase 2 (Fehlerkorrektur) und Phase 3 (Feedback-Generierung) bestätigt den Trend, dass die Erfolgsquoten mit zunehmender Aufgabenkomplexität sinken. In der Korrekturphase (P2) wurden die niedrigsten Erfolgsquoten mit 80 % in der Kategorie LS-AW und 85 % in der Kategorie LS-AL verzeichnet. In der Feedback-Phase (P3) sank die Leistung weiter; die Kategorie LS-AW verzeichnete mit 75 % (15/20) die niedrigste Erfolgsquote der gesamten Analyse, gefolgt von den Kategorien GR-AW und LS-AL mit 80 % Erfolg. Es wird deutlich, dass das Phänomen der Fehlerfortpflanzung bei diesem Leistungsabfall eine entscheidende Rolle spielt. Die 3 in Phase 1 nicht erkannten Fälle von LS-AL (15 %) und 3 Fälle von LS-AW (15 %) spiegelten sich in den folgenden Phasen als GBF wider und senkten die Gesamterfolgsquoten in diesen Kategorien direkt. Darüber hinaus wurde beobachtet, dass die Rate der GF des Modells in Phase 3 ebenfalls anstieg; insbesondere 4 Fälle (20 %) in der Kategorie GR-AW wurden als die höchste Rate an direkten Feedback-Fehlern registriert.

Die vierte Phase der Analyse konzentrierte sich auf die Frequenzverteilung des nicht-korrektiven Feedbacks. Es wurde festgestellt, dass diese Art von Feedback insgesamt 14 Mal generiert wurde und die Konzentration in den Kategorien GR-UM (5 Fälle) und GR-AW (4 Fälle) lag. Darauf folgten die Kategorien LS-AW (3 Fälle), GR-HF (1 Fall) und LS-UM (1 Fall). Als bemerkenswerter Befund wurde festgestellt, dass in den Kategorien GR-AL, LS-AL und LS-HF diese Art von Feedback überhaupt nicht auftrat (0 Fälle).

Zusammenfassend lässt sich sagen, dass die Leistung des ChatGPT-Modells auf dem A2-Datensatz eine duale Struktur aufweist. Während das Modell bei der Erkennung (P1) von

grammatikbasierten Fehlern eine fehlerfreie Kompetenz zeigt, weist es eine deutliche Schwäche bei der Erkennung lexikalischer und semantischer Nuancen (insbesondere LS-AL und LS-AW) auf. Im Lichte der gewonnenen Daten ist deutlich ersichtlich, dass die Fehler in dieser ersten Erkennungsphase (bedingtes Scheitern) eine Fehlerfortpflanzung erzeugen, die die Leistung der nachfolgenden Korrektur- und Feedback-Phasen direkt negativ beeinflusst.

Tabelle 4.

Die Leistung von ChatGPT für das Sprachniveau B1

P	S	Fehlerebene und -typ							
		GR				LS			
		AL	HF	AW	UM	AL	HF	AW	UM
P1	GE	16	20	20	19	18	20	11	17
		80%	100%	100%	95%	90%	100%	55%	85%
	GF	4	0	0	1	2	0	9	3
		20%	0%	0%	5%	10%	0%	45%	15%
P2	GE	15	20	19	18	13	20	11	16
		75%	100%	95%	90%	65%	100%	55%	80%
	GF	1	0	1	1	5	0	0	1
		5%	0%	5%	5%	25%	0%	0%	5%
	GBF	4	0	0	1	2	0	9	3
		20%	0%	0%	5%	10%	0%	45%	15%
P3	GE	12	18	18	18	12	20	9	15
		60%	90%	90%	90%	60%	100%	45%	75%
	GF	3	2	1	0	1	0	2	1
		15%	10%	5%	0%	5%	0%	10%	5%
	GBF	5	0	1	2	7	0	9	4
		25%	0%	5%	10%	35%	0%	45%	20%
P4	GES	2	0	4	0	2	0	1	1

Diese Tabelle untersucht die Leistung des ChatGPT-Modells auf dem B1-Niveau-Datensatz anhand einer vierstufigen Methodik. Im Vergleich zu den vorherigen A1- und A2-Niveau-Datensätzen wurde festgestellt, dass der B1-Datensatz für das Modell deutlich anspruchsvoller ist. Die gewonnenen Erkenntnisse zeigen, dass die Erfolgsquoten im B1-Set in fast allen Kategorien niedriger und die Raten sowohl des direkten als auch des bedingten Scheiterns höher sind. Es wurde beobachtet, dass die deutlichste Schwäche des Modells auf

diesem fortgeschrittenen Niveau in der Kategorie LS-AW liegt, gefolgt von der Kategorie GR-AL, wobei jedoch in bestimmten Kategorien eine hohe Leistung beibehalten wurde.

Der erste Schritt der vierstufigen Analyse, die Fehlererkennung (Phase 1), legt die Schwierigkeit des B1-Sets deutlich dar. Im Gegensatz zu den vorherigen Niveaus konnte das Modell selbst in den Grammatikkategorien keinen konsistenten Erfolg vorweisen. Die niedrigste Grammatikerkennung wurde in der Kategorie GR-AL mit einer Erfolgsquote von 80 % bei 4 fehlgeschlagenen Fällen (20 %) verzeichnet. Es wird jedoch deutlich, dass die eigentliche Schwierigkeit des Modells im B1-Set im LS-Bereich liegt; die Kategorie LS-AW wies mit 55 % Erfolgsquote den deutlichsten Misserfolg in dieser Phase auf, da 9 von 20 Fällen (45 %) nicht erkannt wurden. Trotz dieser Schwächen ist es bemerkenswert, dass das Modell selbst im anspruchsvollen B1-Set in den Kategorien GR-HF (100 %), GR-AW (100 %) und LS-HF (100 %) eine fehlerfreie Erkennungsleistung zeigte.

Mit zunehmender Aufgabenkomplexität, d.h. beim Übergang zu Phase 2 (Korrektur) und Phase 3 (Feedback-Generierung), sinkt die Leistung des Modells weiter. Fehler, die in Phase 1 nicht erkannt wurden, spiegelten sich in den folgenden Phasen als GBF wider. Die Kategorie, in der dieses Phänomen der Fehlerfortpflanzung am deutlichsten zu beobachten war, war LS-AW; die 9 Erkennungsfehler (45 %) aus Phase 1 wurden bis zu Phase 3 übertragen und senkten die Gesamterfolgsquote dieser Kategorie auf bis zu 45 % (9/20). Ein ähnlicher Rückgang wurde in der Kategorie GR-AL beobachtet, die bei der Erkennung eine Erfolgsquote von 80 % aufwies; die Erfolgsquote in dieser Kategorie sank in Phase 3 auf 60 %. Im B1-Set nahmen nicht nur die bedingten Fehlschläge zu, sondern auch die direkt gemachten Fehler. Ein bemerkenswerter Befund ist, dass die Kategorie LS-AL in Phase 2 (Korrektur) mit 5 Fällen (25 %) die höchste Rate an direkten Korrekturfehlern aufwies. Dies zeigt, dass das Modell Schwierigkeiten hat, die korrekte Korrektur bereitzustellen, selbst wenn es den Fehler erkennt. In Phase 3 (Feedback-Generierung) wurden die meisten direkten Fehler mit 3 Fällen (15 %) in der Kategorie GR-AL gemacht.

Die vierte Phase der Analyse konzentrierte sich auf die Frequenzverteilung des nicht-korrektiven Feedbacks. Es wurde festgestellt, dass diese Art von Feedback im B1-Set insgesamt 10 Mal generiert wurde. Diese Feedbacks konzentrierten sich am häufigsten mit 4 Fällen auf die Kategorie GR-AW. Darauf folgten die Kategorien GR-AL und LS-AL, die jeweils 2 Fälle produzierten.

Zusammenfassend lässt sich sagen, dass der B1-Datensatz für ChatGPT im Vergleich zu den A1- und A2-Sets ein wesentlich komplexeres Herausforderungsprofil darstellt. Es wurde bestätigt, dass die deutlichste Schwäche des Modells in der Kategorie LS-AW liegt, wobei der hohe Erkennungsfehler von 45 % (P1) in dieser Kategorie die Leistung der nachfolgenden Korrektur- und Feedback-Phasen ernsthaft beeinträchtigt. Darüber hinaus sticht die Kategorie GR-AL durch ihre geringe Leistung sowohl in der Erkennungs- als auch in der Feedback-Phase hervor. Die Fähigkeit des Modells, selbst auf B1-Niveau in spezifischen Kategorien wie GR-HF und LS-HF hohe Erfolgsquoten beizubehalten, zeigt jedoch, wie sich die Kompetenzen des Modells je nach Fehlertyp unterscheiden.

Tabelle 5.

Die Gesamtleistung von ChatGPT

P	S	Fehlerebene und -typ							
		GR				LS			
		AL	HF	AW	UM	AL	HF	AW	UM
P1	GE	56	60	60	59	54	60	42	55
		93%	100%	100%	98%	90%	100%	70%	92%
	GF	4	0	0	1	6	0	18	5
		7%	0%	0%	2%	10%	0%	30%	8%
P2	GE	55	59	58	58	49	59	41	54
		92%	98%	97%	97%	82%	98%	68%	90%
	GF	1	1	2	1	5	1	1	1
		2%	2%	3%	2%	8%	2%	2%	2%
	GBF	4	0	0	1	6	0	18	5
		7%	0%	0%	2%	10%	0%	30%	8%
P3	GE	48	55	52	56	45	57	36	52
		80%	92%	87%	93%	75%	95%	60%	87%
	GF	7	4	6	2	4	2	5	2
		12%	7%	10%	3%	7%	3%	8%	3%
	GBF	5	1	2	2	11	1	19	6
		8%	2%	3%	3%	18%	2%	32%	10%
P4	GES	2	1	13	6	2	1	5	4

Diese Tabelle bewertet die kumulative Leistung des ChatGPT-Modells auf den A1-, A2- und B1-Datensätzen im Rahmen einer vierstufigen Methodik (Fehlererkennung, Korrektur, Feedback-Generierung und nicht-korrektives Feedback). Die Gesamtdaten zeigen, dass das Modell eine hohe Gesamtkompetenz aufweist, diese Leistung jedoch je nach Fehlerkategorie deutliche Unterschiede zeigt. Der bemerkenswerteste Befund ist, dass das Modell in den Kategorien GR-HF, GR-AW und LS-HF eine außerordentlich starke Leistung zwischen 97 % und 100 % erbringt. Im Gegensatz dazu sticht die Kategorie LS-AW als konsistentester und deutlichster Schwachstellenbereich des Modells hervor, da sie in allen Phasen die niedrigsten Erfolgsquoten aufweist.

Die Daten der Phase 1 (Fehlererkennung), die die allgemeine Fehlererkennungsfähigkeit des Modells zusammenfassen, bestätigen diese duale Struktur. Das Modell zeigte kumulativ eine fehlerfreie Erkennungsleistung in den Kategorien GR-HF (100 %), GR-AW (100 %) und LS-HF (100 %). Auch in anderen Kategorien liegen die Erfolgsquoten auf sehr hohem

Niveau, wie GR-AL (93 %) und GR-UM (98 %). Die deutlichste Abweichung von diesem positiven Bild wurde jedoch in der Kategorie LS-AW mit einer Erfolgsquote von 70 % (42/60) verzeichnet. Die Tatsache, dass in dieser Kategorie insgesamt 18 Fälle (30 %) nicht erkannt wurden, stellt das schwächste Glied in der kumulativen Leistung des Modells dar.

Die Forschungsergebnisse bestätigen den allgemeinen Trend, dass die Erfolgsquoten mit zunehmender Aufgabenkomplexität (von Phase 1 zu Phase 3) sinken. Der Hauptgrund für diesen Rückgang ist das Phänomen der Fehlerfortpflanzung. Die 18 in Phase 1 nicht erkannten LS-AW-Fälle (30 %) spiegelten sich in Phase 2 (Korrektur) als GBF wider und senkten den Erfolg in dieser Kategorie auf 68 %. Dieser Effekt wurde in Phase 3 (Feedback-Generierung) noch deutlicher; die bedingte Misserfolgsquote in der Kategorie LS-AW stieg auf 32 % (19 Fälle) und die Gesamterfolgsquote sank auf 60 % (36/60), das niedrigste Niveau der gesamten Analyse. Ein ähnlicher Trend wurde in der Kategorie LS-AL mit einer bedingten Misserfolgsquote von bis zu 18 % beobachtet. Der Leistungsabfall in Phase 3 ist jedoch nicht nur auf bedingte Fehler zurückzuführen. Es wurde festgestellt, dass das Modell auch direkte Feedback-Fehler (GF) machte; diese Situation, die insbesondere in den Kategorien GR-AL (7 Fälle, 12 %) und GR-AW (6 Fälle, 10 %) auftrat, zeigt, dass das Modell eine zusätzliche Schwierigkeit hat, diese Korrektur zu erklären, selbst wenn es den Fehler erkennt und korrigiert.

Die kumulative Frequenzerhebung des nicht-korrektiven Feedbacks, die vierte Phase der Analyse, ergab, dass diese Art von Feedback insgesamt 34 Mal generiert wurde. Diese Feedbacks konzentrierten sich mit 13 Fällen am deutlichsten auf die Kategorie GR-AW Darauf folgten die Kategorien GR-UM (6 Fälle), LS-AW (5 Fälle) und LS-UM (4 Fälle).

Zusammenfassend lässt sich sagen, dass die kumulative Analyse der A1-, A2- und B1-Datensätze die Stärken und Schwächen des ChatGPT-Modells bei der Fehleranalyse und -korrektur klar aufzeigt. Während das Modell bei spezifischen Fehlertypen wie GR-HF, GR-AW und LS-HF eine außerordentliche Kompetenz aufweist, wurde die Kategorie LS-AW

als die konsistenteste und strukturellste Schwäche des Modells identifiziert. Die gewonnenen Daten belegen, dass die niedrige Erkennungsleistung (70 %) in dieser Kategorie die nachfolgenden Korrektur- (68 %) und Feedback-Phasen (60 %) durch Fehlerfortpflanzung direkt negativ beeinflusst. Darüber hinaus deuten die direkten Misserfolgsquoten in der Feedback-Phase darauf hin, dass die Generierung von Erklärungen einen separaten, von der Erkennung und Korrektur unabhängigen, Schwierigkeitsbereich darstellt.

Die Leistung von DeepSeek

Tabelle 6.
Die Leistung von DeepSeek für das Sprachniveau A1

P	S	Fehlerebene und -typ							
		GR				LS			
		AL	HF	AW	UM	AL	HF	AW	UM
P1	GE	20	20	20	20	20	20	16	20
		100%	100%	100%	100%	100%	100%	80%	100%
	GF	0	0	0	0	0	0	4	0
		0%	0%	0%	0%	0%	0%	20%	0%
P2	GE	20	20	20	20	20	20	16	20
		100%	100%	100%	100%	100%	100%	80%	100%
	GF	0	0	0	0	0	0	0	0
		0%	0%	0%	0%	0%	0%	0%	0%
	GBF	0	0	0	0	0	0	4	0
		0%	0%	0%	0%	0%	0%	20%	0%
P3	GE	20	20	20	20	19	20	16	19
		100%	100%	100%	100%	95%	100%	80%	95%
	GF	0	0	0	0	1	0	0	1
		0%	0%	0%	0%	5%	0%	0%	5%
	GBF	0	0	0	0	0	0	4	0
		0%	0%	0%	0%	0%	0%	20%	0%
P4	GES	0	0	5	1	0	1	1	2

Diese Tabelle untersucht die Leistung des DeepSeek-Modells auf dem A1-Niveau-Datensatz im Rahmen einer vierstufigen Methodik (Fehlererkennung, Korrektur, Feedback-Generierung und nicht-korrektives Feedback). Die gewonnenen Erkenntnisse zeigen, dass das Modell im A1-Set eine außerordentliche, nahezu fehlerfreie Leistung erbringt. Diese hohe Kompetenz wird dadurch belegt, dass das Modell in sieben der acht Fehlerkategorien

sowohl in Phase 1 (Erkennung) als auch in Phase 2 (Korrektur) eine Erfolgsquote von 100 % erreichte.

Bei der Untersuchung der Fehlererkennungs- (P1) und Korrekturfähigkeiten (P2) des Modells wird dieses hohe Erfolgsprofil deutlich sichtbar. Das Modell zeigte in sieben der acht Fehlerkategorien (GR-AL, GR-HF, GR-AW, GR-UM, LS-AL, LS-HF und LS-UM) eine fehlerfreie Erfolgsquote von 100 % (20/20). Der einzige und konsistente Schwachstellenbereich des Modells im A1-Set ist die Kategorie LS-AW; in dieser Kategorie blieb die Erfolgsquote bei 80 % (16/20), und 4 Fälle (20 %) konnten nicht erkannt werden. Einer der beeindruckendsten Befunde des Modells trat in Phase 2 (Korrektur) zutage: Das Modell korrigierte mit einer Rate von 0 % GBF in allen Kategorien jeden Fehler, den es im A1-Set erkennen konnte, fehlerfrei. Daher resultiert die 80%ige Erfolgsquote in der Kategorie LS-AW ausschließlich aus der Metrik GBF von 4 Fällen (20 %), die aus Phase 1 übertragen wurden.

Die Phase der Feedback-Generierung (P3) ist die einzige Phase, in der die Leistung des Modells einen sehr geringfügigen Rückgang erlitt. Während die meisten Kategorien eine 100%ige Erfolgsquote beibehielten, sank der Erfolg in den Kategorien LS-AL und LS-UM auf 95 % (19/20). Dieser leichte Rückgang ist auf jeweils 1 Fall (5 %) eines direkten Feedback-Fehlers zurückzuführen, der in diesen beiden Kategorien verzeichnet wurde. Die Kategorie LS-AW wies hingegen aufgrund der 4 bedingt fehlgeschlagenen Fälle (20 %) aus Phase 1 weiterhin die niedrigste Leistung mit einer Erfolgsquote von 80 % auf. Bei der Untersuchung der Frequenzverteilung des nicht-korrektiven Feedbacks, der vierten Phase der Analyse, wurde festgestellt, dass diese Art von Feedback insgesamt 10 Mal generiert wurde. Die Konzentration wurde mit 5 Fällen am häufigsten in der Kategorie GR-AW beobachtet; darauf folgten die Kategorien LS-UM (2 Fälle) sowie GR-UM, LS-HF und LS-AW (jeweils 1 Fall).

Zusammenfassend lässt sich sagen, dass das DeepSeek-Modell im A1-Datensatz eine außerordentlich hohe Leistung erbringt. Die Kompetenz des Modells wird durch die 100%ige Erfolgsquote in sieben Kategorien in der Erkennungs- und Korrekturphase belegt. Im Lichte der gewonnenen Daten wurde festgestellt, dass die einzige deutliche Schwäche des Modells in diesem Datensatz die Kategorie LS-AW ist; der Erkennungsfehler von 20 % (4 Fälle) in dieser Kategorie hat die Leistung der nachfolgenden Phasen durch Fehlerfortpflanzung bedingt gesenkt. Die bemerkenswerteste Stärke des Modells ist jedoch die direkte Fehlerrate von 0 % in Phase 2 (Korrektur); dieser Befund zeigt, dass das Modell alle erkannten Fehler ausnahmslos korrekt korrigiert hat.

Tabelle 7.

Die Leistung von DeepSeek für das Sprachniveau A2

P	S	Fehlerebene und -typ							
		GR				LS			
		AL	HF	AW	UM	AL	HF	AW	UM
P1	GE	20	20	19	20	18	20	17	20
		100%	100%	95%	100%	90%	100%	85%	100%
	GF	0	0	1	0	2	0	3	0
		0%	0%	5%	0%	10%	0%	15%	0%
P2	GE	20	20	18	20	17	19	16	20
		100%	100%	90%	100%	85%	95%	80%	100%
	GF	0	0	1	0	1	1	1	0
		0%	0%	5%	0%	5%	5%	5%	0%
	GBF	0	0	1	0	2	0	3	0
		0%	0%	5%	0%	10%	0,0%	15%	0%
P3	GE	20	19	17	20	14	19	16	20
		100%	95%	85%	100%	70%	95%	80%	100%
	GF	0	1	1	0	3	0	0	0
		0%	5%	5%	0%	15%	0%	0%	0%
	GBF	0	0	2	0	3	1	4	0
		0%	0%	10%	0%	15%	5%	20%	0%
P4	GES	1	1	5	1	1	0	4	1

Diese Tabelle untersucht die Leistung des DeepSeek-Modells auf dem A2-Niveau-Datensatz unter Verwendung einer vierstufigen Methodik (Fehlererkennung, Korrektur, Feedback-Generierung und nicht-korrektives Feedback). Die gewonnenen Erkenntnisse zeigen, dass der A2-Datensatz im Vergleich zum A1-Set für das DeepSeek-Modell anspruchsvoller ist;

die im A1-Set beobachteten 100%-Erfolgsquoten sind im A2-Set gesunken. Es wurde festgestellt, dass sich die deutlichsten Schwächen des Modells im A2-Set auf die Kategorien LS-AW und LS-AL konzentrieren. Der deutlichste Unterschied zum A1-Set besteht darin, dass das Modell im A2-Set begonnen hat, direkte Korrekturfehler (Phase 2, GF) zu machen.

Bei der Untersuchung der Fehlererkennungsfähigkeit des Modells (Phase 1, Fehlererkennung) zeigt sich, dass die Grammatikerkennung nach wie vor sehr stark ist; in den Kategorien GR-AL, GR-HF und GR-UM wurde eine Erfolgsquote von 100 % erzielt. Im Gegensatz zur fehlerfreien Leistung im A1-Set wurde jedoch in der Kategorie GR-AW 1 Fall (5 %) nicht erkannt (95 % Erfolg). Im lexikalisch/semantischen Bereich wurde die 100%ige Erfolgsquote in den Kategorien LS-HF und LS-UM ebenfalls beibehalten. Der deutlichste Misserfolg in dieser Phase wurde jedoch in der Kategorie LS-AW verzeichnet, in der 3 Fälle (15 %) nicht erkannt wurden (85 % Erfolg). Darauf folgte die Kategorie LS-AL, in der 2 Fälle (10 %) nicht erkannt wurden.

Mit zunehmender Aufgabenkomplexität (Phase 2 und 3) wurden Leistungsabfälle beobachtet, die im A1-Set nicht zu verzeichnen waren. In Phase 2 (Korrektur) wurden die niedrigsten Erfolgsquoten in den Kategorien LS-AW (80 %) und LS-AL (85 %) verzeichnet. Einer der Hauptgründe für diesen Rückgang ist die Tatsache, dass sich die Erkennungsfehler aus Phase 1 (3 LS-AW und 2 LS-AL Fälle) in dieser Phase als GBF niederschlugen. Anders als im A1-Set machte das Modell im A2-Set jedoch in 4 verschiedenen Kategorien (GR-AW [5 %], LS-AL [5 %], LS-HF [5 %] und LS-AW [5 %]) jeweils 1 Fall, also insgesamt 4 GF. Der Leistungsabfall wurde in Phase 3 (Feedback-Generierung) noch deutlicher; die Kategorie LS-AL verzeichnete mit 70 % (14/20) die niedrigste Erfolgsquote in diesem Datensatz. Bemerkenswert sind die 3 Fälle (15 %) direkter Feedback-Fehler in dieser Kategorie. Die Metriken für bedingtes Scheitern in Phase 3 (4 Fälle für LS-AW, 3 Fälle für LS-AL) bestätigen die anhaltenden Auswirkungen der anfänglichen Erkennungsfehler.

Die vierte Phase der Analyse konzentrierte sich auf die Häufigkeitszählung des nicht-korrektiven Feedbacks. Es wurde festgestellt, dass diese Art von Feedback insgesamt 14 Mal generiert wurde. Die Konzentration wurde am häufigsten in den Kategorien GR-AW (5 Fälle) und LS-AW (4 Fälle) beobachtet. Mit Ausnahme der Kategorie LS-HF (0 Fälle) wurde in allen anderen Kategorien 1 Fall verzeichnet.

Zusammenfassend lässt sich sagen, dass der A2-Datensatz für das DeepSeek-Modell im Vergleich zu A1 ein anspruchsvolleres Profil aufweist. Trotz des gestiegenen Schwierigkeitsgrads hat das Modell seine hohe Erkennungsleistung in Kategorien wie GR-AL, GR-HF und GR-UM weitgehend beibehalten. Die deutlichsten Schwächen des Modells liegen jedoch in den Kategorien LS-AW (mit einer Fehlerrate von 15 % in der Erkennungsphase) und LS-AL (mit 15 % direkten Fehlern in der Feedback-Phase und 70 % Gesamterfolg). Der wichtigste Befund beim Übergang von A1 zu A2 ist, dass die Rate der direkten Korrekturfehler (Phase 2), die bei A1 noch 0 % betrug, im A2-Set in 4 verschiedenen Kategorien auftrat.

Tabelle 8.

Die Leistung von DeepSeek für das Sprachniveau B1

P	S	Fehlerebene und -typ							
		GR				LS			
		AL	HF	AW	UM	AL	HF	AW	UM
P1	GE	16	20	20	19	18	20	16	18
		80%	100%	100%	95%	90%	100%	80%	90%
	GF	4	0	0	1	2	0	4	2
		20%	0%	0%	5%	10%	0%	20%	10%
P2	GE	16	18	20	19	18	14	16	18
		80%	90%	100%	95%	90%	70%	80%	90%
	GF	0	2	0	0	0	6	0	0
		0%	10%	0%	0%	0%	30%	0%	0%
	GBF	4	0	0	1	2	0	4	2
		20%	0%	0%	5%	10%	0%	20%	10%
P3	GE	14	15	18	18	17	14	16	17
		70%	75%	90%	90%	85%	70%	80%	85%
	GF	2	3	2	1	1	0	0	1
		10%	15%	10%	5%	5%	0%	0%	5%
	GBF	4	2	0	1	2	6	4	2
		20%	10%	0%	5%	10%	30%	20%	10%
P4	GES	2	1	9	6	0	1	3	0

Diese Tabelle untersucht die Leistung des DeepSeek-Modells auf dem B1-Niveau-Datensatz anhand einer vierstufigen Methodik (Fehlererkennung, Korrektur, Feedback-Generierung und nicht-korrektives Feedback). Im Vergleich zu den A1- und A2-Datensätzen ist ersichtlich, dass der B1-Datensatz für das Modell deutlich anspruchsvoller ist und die Erkennungsraten gesunken sind. Der auffälligste Befund in diesem Datensatz ist der Leistungseinbruch des Modells in der Kategorie LS-HF, in der es in den A1- und A2-Datensätzen eine Erfolgsquote von nahezu 100 % aufwies. Diese Kategorie hat sich im B1-Set zur deutlichsten Schwäche des Modells entwickelt und wies insbesondere in Phase 2 (Korrektur) eine hohe direkte Fehlerrate von 30 % (6 Fälle) auf.

Die gestiegene Schwierigkeit des B1-Sets ist bereits in Phase 1 (Fehlererkennung) zu beobachten. Selbst in diesem anspruchsvollen Set hat das Modell die fehlerfreie Erkennungsrate von 100 % in den Kategorien GR-HF, GR-AW und LS-HF beibehalten. Jedoch zeigten die Kategorien GR-AL und LS-AW, die sich als konsistente Schwächen des

Modells herauskristallisierten, beide eine Erfolgsquote von 80 %, da 4 Fälle (20 %) nicht erkannt wurden. Diese nicht erkannten Fehler haben die Leistung der nachfolgenden Phasen als bedingtes Scheitern direkt beeinflusst.

Die kritischsten Befunde der Analyse traten in Phase 2 (Korrektur) zutage; in dieser Phase wurde ein im Vergleich zu den vorherigen Datensätzen abweichendes und ausgeprägtes Schwächeprofil beobachtet. Die Kategorie LS-HF, in der das Modell in den vorherigen Sets (A1/A2) eine Erfolgsquote von nahezu 100 % aufwies, verzeichnete in dieser Phase mit 70 % (14/20) die niedrigste Erfolgsquote. Der Grund für diesen dramatischen Rückgang ist nicht die Fehlerfortpflanzung (bedingtes Scheitern), sondern 6 Fälle (30 %) direkter Korrekturfehler (GF). In ähnlicher Weise wurden auch in der Kategorie GR-HF 2 Fälle (10 %) direkter Korrekturfehler gemacht; in allen anderen Kategorien wurde dieser Fehlertyp mit 0 % verzeichnet. Zusätzlich zu diesen neuen Schwächen senkten die 4 aus Phase 1 übertragenen Fälle von GR-AL (20 %) und 4 Fälle von LS-AW (20 %) die Erfolgsquote auch in diesen Kategorien auf 80 %. Diese Fehler spiegelten sich weiterhin in Phase 3 (Feedback-Generierung) wider; die 6 Korrekturfehler (30 %) in der Kategorie LS-HF schlugen sich in dieser Phase als bedingtes Scheitern nieder und zogen den Erfolg dieser Kategorie zusammen mit GR-AL (70 %) auf das niedrigste Niveau. Der direkte Feedback-Fehler in Phase 3 trat hingegen am häufigsten mit 3 Fällen (15 %) in der Kategorie GR-HF auf.

Die vierte Phase der Analyse untersuchte die Frequenzverteilung des nicht-korrektiven Feedbacks. Es wurde festgestellt, dass diese Art von Feedback insgesamt 23 Mal generiert wurde und sich am häufigsten auf die Kategorien GR-AW (9 Fälle) und GR-UM (6 Fälle) konzentrierte. Als bemerkenswerter Befund wurde beobachtet, dass es dem Modell trotz der allgemeinen Schwierigkeit des B1-Sets gelungen ist, seine fehlerfreie Leistung von 100 % in der Kategorie GR-AW in den Phasen der Erkennung (P1) und Korrektur (P2) beizubehalten.

Zusammenfassend lässt sich sagen, dass der B1-Datensatz für das DeepSeek-Modell im Vergleich zu den vorherigen Niveaus deutlich anspruchsvoller ist. Das Leistungsprofil des Modells hat sich im B1-Set verändert; während die Kategorien GR-AL und LS-AW mit 20 % Erkennungsfehlern (P1) konsistente Schwächen bleiben, trat die deutlichste Schwäche in der Kategorie LS-HF auf, die in den A1/A2-Sets stark war. Der in dieser Kategorie beobachtete direkte Korrekturfehler von 30 % (P2) belegt, dass der B1-Datensatz nicht nur in der Erkennungsphase, sondern auch in der komplexeren Korrekturphase neue und deutliche Herausforderungen für das Modell bietet.

Tabelle 9.

Die Gesamtleistung von DeepSeek

P	S	Fehlerebene und -typ							
		GR				LS			
		AL	HF	AW	UM	AL	HF	AW	UM
P1	GE	56	60	59	59	56	60	49	58
		93%	100%	98%	98%	93%	100%	82%	97%
	GF	4	0	1	1	4	0	11	2
		7%	0%	2%	2%	7%	0%	18%	3%
P2	GE	56	58	58	59	55	53	48	58
		93%	97%	97%	98%	92%	88%	80%	97%
	GF	0	2	1	0	1	7	1	0
		0%	3%	2%	0%	2%	12%	2%	0%
	GBF	4	0	1	1	4	0	11	2
		7%	0%	2%	2%	7%	0%	18%	3%
P3	GE	54	54	55	58	50	53	48	56
		90%	90%	92%	97%	83%	88%	80%	93%
	GF	2	4	3	1	5	0	0	2
		3%	7%	5%	2%	8%	0%	0%	3%
	GBF	4	2	2	1	5	7	12	2
		7%	3%	3%	2%	8%	12%	20%	3%
P4	GES	3	2	15	7	1	1	14	1

Diese Tabelle bewertet die kumulative Leistung des DeepSeek-Modells auf den A1-, A2- und B1-Datensätzen im Rahmen einer vierstufigen Methodik (Fehlererkennung, Korrektur, Feedback-Generierung und nicht-korrektives Feedback). Die Gesamtdaten zeigen, dass das Modell eine hohe allgemeine Erkennungskompetenz aufweist, dieses Leistungsprofil jedoch von zwei spezifischen und qualitativ unterschiedlichen Schwachstellenbereichen

überschattet wird. Die deutlichste und auffälligste Schwäche des Modells ist der in der Kategorie LS-HF beobachtete direkte Korrekturfehler (P2) von 12 % (7 Fälle). Darüber hinaus wurde festgestellt, dass die konsistenteste Schwäche des Modells mit einer Erkennungsfehlerrate von 18 % (11 Fälle) die Kategorie LS-AW ist.

Die Daten der Phase 1 (Fehlererkennung), die die kumulative Fehlererkennungsfähigkeit des Modells zusammenfassen, bestätigen dieses zwiespältige Profil. Das Modell zeigte in den Kategorien GR-HF (100 %) und LS-HF (100 %) eine fehlerfreie Gesamtleistung. Auch in anderen Grammatikkategorien (GR-AW 98 %, GR-UM 98 %) und lexikalischen Kategorien (LS-UM 97 %, LS-AL 93 %) ist der Erfolg sehr hoch. Der deutlichste Misserfolg in dieser Phase wurde jedoch in der Kategorie LS-AW mit einer Erfolgsquote von 82 % (49/60) verzeichnet; die Nichterkennung von insgesamt 11 Fällen (18 %) in dieser Kategorie stellt die größte Erkennungsschwäche des Modells dar.

Mit zunehmender Aufgabenkomplexität (Phase 2 und 3) werden die Auswirkungen dieser beiden grundlegenden Schwächen deutlich. In Phase 2 (Korrektur) liegt die niedrigste Erfolgsquote mit 80 % in der Kategorie LS-AW; dies ist größtenteils auf die aus Phase 1 übertragene Metrik GBF von 11 Fällen (18 %) zurückzuführen. Der analytischste Befund in Phase 2 trat jedoch in der Kategorie LS-HF zutage, deren Erkennungserfolg bei 100 % lag. Diese Kategorie wies mit 7 Fällen (12 %) die höchste Rate an direkten Korrekturfehlern (GF) auf (88 % Erfolg), was zeigt, dass das Modell eine von der Erkennungsfähigkeit unabhängige Korrekturschwierigkeit aufweist. Diese 7 Korrekturfehler spiegelten sich als bedingtes Scheitern in Phase 3 (Feedback-Generierung) wider und senkten die Erfolgsquote der LS-HF-Kategorie auch in dieser Phase. In Phase 3 wurde zudem mit 5 Fällen (8 %) ein deutlicher direkter Feedback-Fehler in der Kategorie LS-AL beobachtet.

Die vierte Phase der Analyse konzentrierte sich auf die kumulative Häufigkeitszählung des nicht-korrektiven Feedbacks. Es wurde festgestellt, dass das DeepSeek-Modell diese Art von Feedback insgesamt 44 Mal generiert hat; dies ist eine der höchsten Frequenzen unter den

untersuchten Modellen. Diese hohe Frequenz konzentrierte sich am deutlichsten auf die Kategorie GR-AW mit 15 Fällen und die Kategorie LS-AW mit 14 Fällen.

Zusammenfassend lässt sich sagen, dass die kumulative Leistung des DeepSeek-Modells eine Spannung zwischen starken Erkennungsfähigkeiten (GR-HF und LS-HF) und zwei deutlichen, jedoch qualitativ unterschiedlichen Schwachstellenbereichen zeigt. Die konsistenteste Schwäche des Modells ist der Erkennungsfehler von 18 % in der Kategorie LS-AW und dessen Widerspiegelung in den nachfolgenden Phasen durch Fehlerfortpflanzung. Die auffälligste Schwäche ist jedoch der direkte Korrekturfehler von 12 % in der Kategorie LS-HF, obwohl der Erkennungserfolg fehlerfrei (100 %) war. Dieser Befund belegt, dass die Fähigkeiten des Modells zur Fehlererkennung und Fehlerkorrektur voneinander unabhängige Herausforderungen beinhalten. Darüber hinaus fällt die Häufigkeit des nicht-korrektiven Feedbacks mit 44 Fällen, konzentriert in den Kategorien GR-AW und LS-AW, als ein Verhaltensmerkmal des Modells auf.

Die Leistung von Gemini

Tabelle 10.

Die Leistung von Gemini für das Sprachniveau A1

P	S	Fehlerebene und -typ							
		GR				LS			
		AL	HF	AW	UM	AL	HF	AW	UM
P1	GE	20	20	20	20	20	20	16	19
		100%	100%	100%	100%	100%	100%	80%	95%
	GF	0	0	0	0	0	0	4	1
		0%	0%	0%	0%	0%	0%	20%	5%
P2	GE	20	19	20	20	20	20	16	20
		100%	95%	100%	100%	100%	100%	80%	100%
	GF	0	1	0	0	0	0	0	0
		0%	5%	0%	0%	0%	0%	0%	0%
	GBF	0	0	0	0	0	0	4	0
		0%	0%	0%	0%	0%	0%	20%	0%
P3	GE	18	19	19	20	20	20	16	20
		90%	95%	95%	100%	100%	100%	80%	100%
	GF	2	0	1	0	0	0	0	0
		10%	0%	5%	0%	0%	0%	0%	0%
	GBF	0	1	0	0	0	0	4	0
		0%	5%	0%	0%	0%	0%	20%	0%
P4	GES	0	0	4	0	0	1	3	2

Diese Tabelle untersucht die Leistung des Gemini-Modells auf dem A1-Niveau-Datensatz im Rahmen einer vierstufigen Methodik (Fehlererkennung, Korrektur, Feedback-Generierung und nicht-korrektives Feedback). Die gewonnenen Erkenntnisse zeigen, dass das Modell im A1-Set eine außerordentlich starke Leistung erbringt; in den meisten Kategorien erreichen die Erkennungs- und Korrekturraten 100 %. Die Leistung des Modells in diesem Datensatz ist durch zwei unterschiedliche Arten von Fehlermechanismen gekennzeichnet: eine konsistente Schwäche in der Erkennungsphase und eine spezifische Anomalie, die in der Korrekturphase auftritt.

Die Fehlererkennungsleistung (P1) des Modells bestätigt diese hohe Kompetenz. In sechs der acht Fehlerkategorien (GR-AL, GR-HF, GR-AW, GR-UM, LS-AL und LS-HF) wurde ein fehlerfreier Erfolg von 100 % (20/20) erzielt. Die deutlichste und konsistenteste Schwäche des Modells im A1-Set wurde in der Kategorie LS-AW (80 % Erfolg) verzeichnet,

in der 4 Fälle (20 %) nicht erkannt werden konnten. Dieser Erkennungsfehler spiegelte sich durch Fehlerfortpflanzung als GBF in Phase 2 (Korrektur) und Phase 3 (Feedback-Generierung) wider und senkte die Erfolgsquote dieser Kategorie in beiden Phasen auf 80 %.

Der zweite und methodisch andersartige Fehlertyp, der im A1-Set beobachtet wurde, trat in der Kategorie GR-HF auf. Obwohl der Erkennungserfolg (P1) in dieser Kategorie 100 % betrug, wurde in Phase 2 (Korrektur) 1 Fall (5 %) eines direkten Korrekturfehlers (GF) gemacht. Abgesehen von diesem Einzelfall ist die direkte Korrekturleistung des Modells hervorragend. Dieser direkte Korrekturfehler in Phase 2 spiegelte sich als 1 Fall (5 %) bedingten Scheiterns in Phase 3 (Feedback-Generierung) wider. In der Feedback-Phase (P3) machte das Modell zusätzlich zu diesen beiden Quellen bedingten Scheiterns (4 LS-AW und 1 GR-HF) direkte Feedback-Fehler (GF) in den Kategorien GR-AL (2 Fälle, 10 %) und GR-AW (1 Fall, 5 %).

Die vierte Phase der Analyse untersuchte die Frequenzverteilung des nicht-korrektiven Feedbacks. Es wurde festgestellt, dass diese Art von Feedback insgesamt 10 Mal generiert wurde. Diese Feedbacks konzentrierten sich am häufigsten auf die Kategorie GR-AW (4 Fälle) und die Kategorie LS-AW (3 Fälle).

Zusammenfassend lässt sich sagen, dass das Gemini-Modell im A1-Datensatz eine außerordentlich hohe Leistung erbringt; die 100%ige Erkennung des Modells in sechs Kategorien und seine nahezu fehlerfreie Korrekturfähigkeit belegen dies. Die konsistenteste Schwäche des Modells in diesem Set ist die Kategorie LS-AW, in der sich der Erkennungsfehler von 20 % in den nachfolgenden Phasen widerspiegelte. Die in der Kategorie GR-HF beobachtete Fehlerfortpflanzung (Erkennung 100 %, Korrektur 95 %) ist jedoch ein wichtiger methodischer Befund, der zeigt, dass Erkennungs- und Korrekturprozesse voneinander unabhängige Fehlerquellen sein können.

Tabelle 11.

Die Leistung von Gemini für das Sprachniveau A2

P	S	Fehlerebene und -typ							
		GR				LS			
		AL	HF	AW	UM	AL	HF	AW	UM
P1	GE	19	20	20	20	17	20	15	19
		95%	100%	100%	100%	85%	100%	75%	95%
	GF	1	0	0	0	3	0	5	1
		5%	0%	0%	0%	15%	0%	25%	5%
P2	GE	19	20	19	20	17	20	15	19
		95%	100%	95%	100%	85%	100%	75%	95%
	GF	0	0	1	0	0	0	0	0
		0%	0%	5%	0%	0%	0%	0%	0%
	GBF	1	0	0	0	3	0	5	1
		5%	0%	0%	0%	15%	0%	25%	5%
P3	GE	18	20	19	20	17	20	14	19
		90%	100%	95%	100%	85%	100%	70%	95%
	GF	1	0	0	0	0	0	1	0
		5%	0%	0%	0%	0%	0%	5%	0%
	GBF	1	0	1	0	3	0	5	1
		5%	0%	5%	0%	15%	0%	25%	5%
P4	GES	0	1	1	2	0	0	0	0

Diese Tabelle untersucht die Leistung des Gemini-Modells auf dem A2-Niveau-Datensatz im Rahmen einer vierstufigen Methodik (Fehlererkennung, Korrektur, Feedback-Generierung und nicht-korrektives Feedback). Die Befunde des A2-Datensatzes legen einen bemerkenswerten Aspekt des Leistungsprofils des Modells offen: Die direkte Fehlerrate des Modells (Phase 2 und 3 - GF) ist außerordentlich niedrig. Im Gegensatz dazu wurde festgestellt, dass der Hauptfaktor, der den Gesamterfolg des Modells einschränkt, spezifische Schwächen in der Erkennungsphase (P1) und die Übertragung dieser Fehler in die nachfolgenden Phasen durch Fehlerfortpflanzung ist.

Die Daten der Phase 1 (Fehlererkennung), welche die Fehlererkennungsfähigkeit des Modells messen, zeigen diese duale Struktur deutlich. Das Modell zeigte in den Grammatik-Kategorien eine sehr starke Leistung; in den Kategorien GR-HF, GR-AW und GR-UM wurde eine Erfolgsquote von 100 % erzielt. In der Kategorie GR-AL konnte hingegen nur 1 Fall (5 %) nicht erkannt werden (95 % Erfolg). Auch im lexikalisch/semantischen Bereich

wurde in der Kategorie LS-HF ein fehlerfreier Erfolg von 100 % erzielt. Die deutlichste Schwäche des Modells im A2-Set wurde jedoch in der Kategorie LS-AW (75 % Erfolg) verzeichnet, in der 5 Fälle (25 %) nicht erkannt werden konnten. Darauf folgte die Kategorie LS-AL (85 % Erfolg), in der 3 Fälle (15 %) nicht erkannt werden konnten.

Die kritischsten Befunde der Analyse ergaben sich aus der vergleichenden Untersuchung der Daten von Phase 2 (Korrektur) und Phase 3 (Feedback-Generierung). Die direkte Fehlerrate des Modells ist in diesen beiden Phasen minimal. In Phase 2 (Korrektur) liegt die Rate der direkten Korrekturfehler (GF) in sieben der acht Fehlerkategorien bei 0 %; nur in der Kategorie GR-AW wurde 1 Fall (5 %) als direkter Fehler gemacht. In ähnlicher Weise wurden in Phase 3 (Feedback-Generierung) insgesamt nur 2 direkte Feedback-Fehler (GR-AL 5 % und LS-AW 5 %) beobachtet. Im Lichte dieser Befunde wird deutlich, dass der Leistungsabfall in den Kategorien LS-AW (75 % in Phase 2, 70 % in Phase 3) und LS-AL (85 % in Phase 2), in denen die niedrigsten Erfolgsquoten verzeichnet wurden, nicht auf eine direkte Unzulänglichkeit zurückzuführen ist. Vielmehr ist der Hauptgrund für diesen Rückgang die Tatsache, dass sich die 5 in Phase 1 nicht erkannten Fälle von LS-AW (25 %) und 3 Fälle von LS-AL (15 %) als GBF in den nachfolgenden Phasen niederschlugen.

Die vierte Phase der Analyse konzentrierte sich auf die Frequenzverteilung des nicht-korrektiven Feedbacks. Es wurde festgestellt, dass das Gemini-Modell diese Art von Feedback im A2-Set mit insgesamt 4 Fällen sehr selten generierte. Diese niedrigfrequente Nutzung wurde am häufigsten in der Kategorie GR-UM mit 2 Fällen beobachtet.

Zusammenfassend lässt sich sagen, dass das Gemini-Modell im A2-Datensatz ein sehr starkes Leistungsprofil aufweist, insbesondere in den Grammatik-Kategorien und der Kategorie LS-HF Die Fähigkeiten des Modells zur direkten Korrektur (P2) und zum direkten Feedback (P3) sind nahezu fehlerfrei. Die gewonnenen Daten belegen, dass der Hauptfaktor, der den Gesamterfolg des Modells in diesem Set einschränkt, die Erkennungsfehler (P1) in

den Kategorien LS-AW (25 %) und LS-AL (15 %) sowie die Übertragung dieser initialen Fehler in die nachfolgenden Prozesse durch Fehlerfortpflanzung ist.

Tabelle 12.
Die Leistung von Gemini für das Sprachniveau B1

P	S	Fehlerebene und -typ							
		GR				LS			
		AL	HF	AW	UM	AL	HF	AW	UM
P1	GE	16	18	20	20	16	19	17	20
		80%	90%	100%	100%	80%	95%	85%	100%
	GF	4	2	0	0	4	1	3	0
		20%	10%	0%	0%	20%	5%	15%	0%
P2	GE	16	18	20	20	16	19	17	20
		80%	90%	100%	100%	80%	95%	85%	100%
	GF	0	0	0	0	0	0	0	0
		0%	0%	0%	0%	0%	0%	0%	0%
	GBF	4	2	0	0	4	1	3	0
		20%	10%	0%	0%	20%	5%	15%	0%
P3	GE	14	17	17	20	16	19	16	19
		70%	85%	85%	100%	80%	95%	80%	95%
	GF	2	1	3	0	0	0	1	1
		10%	5%	15%	0%	0%	0%	5%	5%
	GBF	4	2	0	0	4	1	3	0
		20%	10%	0%	0%	20%	5%	15%	0%
P4	GES	0	1	3	2	1	3	1	0

Diese Tabelle untersucht die Leistung des Gemini-Modells auf dem B1-Niveau-Datensatz anhand einer vierstufigen Methodik (Fehlererkennung, Korrektur, Feedback-Generierung und nicht-korrektives Feedback). Im Vergleich zu den A1- und A2-Datensätzen wurde festgestellt, dass der B1-Datensatz für das Modell insbesondere in der Erkennungsphase anspruchsvoller ist. Der auffälligste und methodisch wichtigste Befund in diesem Datensatz ist jedoch die Leistung des Modells in Phase 2 (Korrektur): Das Modell hat jeden Fehler, den es erkennen konnte, unabhängig von der Kategorie, mit einer direkten Fehlerrate von 0 % fehlerfrei korrigiert. Dieser Befund zeigt, dass die Ursache für den Leistungsabfall im B1-Set nicht die Korrekturfähigkeit ist, sondern die Erkennungsfehler in Phase 1 und die Schwierigkeiten bei der Feedback-Generierung in Phase 3.

Der erste Bereich, in dem das Modell Schwierigkeiten hatte, ist Phase 1 (Fehlererkennung). Obwohl das Modell in den Kategorien GR-AW, GR-UM und LS-UM eine fehlerfreie Erfolgsquote von 100 % aufwies, wurde die niedrigste Leistung in diesem Set mit einer Erfolgsquote von 80 % (16/20) sowohl in den Kategorien GR-AL als auch LS-AL verzeichnet. In jeder dieser Kategorien konnten 4 Fälle (20 %) vom Modell nicht erkannt werden. Darüber hinaus wurden auch in den Kategorien GR-HF (10 %) und LS-AW (15 %) Erkennungsfehler beobachtet.

Die zweite Phase der Analyse (Fehlerkorrektur) offenbart die größte Stärke des Modells im B1-Set. In allen acht Fehlerkategorien beträgt die Rate der direkten Korrekturfehler (GF) 0 %. Dies belegt, dass das Modell alle Fehler, die es erkennen konnte, selbst auf B1-Niveau erfolgreich korrigiert hat. Folglich ist der Grund für die niedrigsten beobachteten Erfolgsquoten in dieser Phase (GR-AL 80 % und LS-AL 80 %) nicht eine Korrekturunzulänglichkeit des Modells, sondern ausschließlich der Widerspiegelung von jeweils 4 nicht erkannten Fällen (20 %) aus Phase 1 als GBF durch Fehlerfortpflanzung.

Der zweite Bereich, in dem das Modell Schwierigkeiten hatte, ist Phase 3 (Feedback-Generierung). In dieser Phase traten zusätzlich zu den aus Phase 1 übertragenen bedingten Fehlern neue direkte Feedback-Fehler auf. Der deutlichste Beleg hierfür wurde in der Kategorie GR-AW beobachtet: Obwohl das Modell die Fehler in dieser Kategorie mit 100 % Erfolg erkannt (P1) und mit 100 % Erfolg korrigiert (P2) hatte, machte es in Phase 3 3 Fälle (15 %) direkte Feedback-Fehler. Dieser Befund zeigt, dass der Mechanismus zur Feedback-Generierung eine von der Korrektur unabhängige Schwierigkeit birgt. Die niedrigste Gesamterfolgsquote in dieser Phase wurde mit 70 % (14/20) in der Kategorie GR-AL verzeichnet. In der vierten Phase der Analyse (Nicht-korrektives Feedback) wurde festgestellt, dass diese Art von Feedback insgesamt 11 Mal generiert wurde und am häufigsten mit jeweils 3 Fällen in den Kategorien GR-AW und LS-HF beobachtet wurde.

Zusammenfassend lässt sich sagen, dass die Leistung des Gemini-Modells im B1-Datensatz eine klare Unterscheidung hinsichtlich der Fähigkeiten des Modells offenbart. Die größte Stärke des Modells ist seine fehlerfreie Korrekturfähigkeit mit einer direkten Fehlerrate von 0 % (P2). Im Lichte der gewonnenen Daten stammen die Misserfolge in diesem Set aus zwei Hauptquellen: Erstens die Erkennungsschwäche (P1), die sich auf die Kategorien GR-AL und LS-AL konzentriert (20 % Fehlerrate). Zweitens die Tatsache, dass der Prozess der Feedback-Generierung (P3) eine unabhängige Fehlerquelle darstellt, selbst wenn die Erkennungs- und Korrekturprozesse erfolgreich sind, wie in der Kategorie GR-AW zu sehen ist.

Tabelle 13.
Die Gesamtleistung von Gemini

P	S	Fehlerebene und -typ							
		GR				LS			
		AL	HF	AW	UM	AL	HF	AW	UM
P1	GE	55	58	60	60	53	59	48	58
		92%	97%	100%	100%	88%	98%	80%	97%
	GF	5	2	0	0	7	1	12	2
		8%	3%	0%	0%	12%	2%	20%	3%
P2	GE	55	57	59	60	53	59	48	59
		92%	95%	98%	100%	88%	98%	80%	98%
	GF	0	1	1	0	0	0	0	0
		0%	2%	2%	0%	0%	0%	0%	0%
	GBF	5	2	0	0	7	1	12	1
		8%	3%	0%	0%	12%	2%	20%	2%
P3	GE	50	56	55	60	53	59	46	58
		83%	93%	92%	100%	88%	98%	77%	97%
	GF	5	1	4	0	0	0	2	1
		8%	2%	7%	0%	0%	0%	3%	2%
	GBF	5	3	1	0	7	1	12	1
		8%	5%	2%	0%	12%	2%	20%	2%
P4	GES	0	2	8	4	1	4	4	2

Diese Tabelle bewertet die kumulative Leistung des Gemini-Modells auf den A1-, A2- und B1-Datensätzen anhand einer vierstufigen Methodik (Fehlererkennung, Korrektur, Feedback-Generierung und nicht-korrektives Feedback). Die kumulativen Befunde legen eine methodisch wichtige Unterscheidung im Leistungsprofil des Modells offen. Die größte

Stärke des Modells ist der außerordentlich hohe direkte Korrekturerfolg, den es in Phase 2 (Korrektur) zeigt. Die kumulativen Daten zeigen, dass das Modell die erkannten Fehler nahezu fehlerfrei korrigiert; in allen acht Kategorien mit 60 Fällen (insgesamt 480 Szenarien) wurden insgesamt nur 2 direkte Korrekturfehler verzeichnet. Im Gegensatz dazu gibt es zwei Hauptfaktoren, die den Gesamterfolg des Modells einschränken: spezifische Schwächen in Phase 1 (Erkennung) und die direkten Fehlerraten in Phase 3 (Feedback-Generierung).

Der größte negative Einfluss auf die Gesamtleistung des Modells stammt von den Fehlschlägen in Phase 1 (Erkennung). Obwohl das Modell in den Kategorien GR-AW (100 %) und GR-UM (100 %) eine kumulativ fehlerfreie Erkennungsleistung zeigte, konzentrierte sich die deutlichste Schwäche auf den lexikalisch-semantischen Bereich. Die niedrigste Erkennungsleistung des Modells in der Gesamtsumme wurde in der Kategorie LS-AW (80 % Erfolg) verzeichnet, in der 12 Fälle (20 %) nicht erkannt werden konnten. Darauf folgte die Kategorie LS-AL (88 % Erfolg), in der 7 Fälle (12 %) nicht erkannt werden konnten.

Die Analyse der Phase 2 (Korrektur) legt die größte Stärke des Modells offen. Die Rate der direkten Korrekturfehler (GF) des Modells ist außerordentlich niedrig; während diese Rate in sechs Kategorien 0 % betrug, wurde in den Kategorien GR-HF (2 %) und GR-AW (2 %) nur jeweils 1 Fall verzeichnet. Dieser Befund belegt, dass der Grund für die niedrigen Gesamterfolgsquoten in dieser Phase (z. B. LS-AW 80 %, LS-AL 88 %) nicht eine Korrekturunzulänglichkeit des Modells darstellt. Vielmehr sind diese niedrigen Raten ausschließlich eines Ergebnisses der Widerspiegelung der 12 in Phase 1 nicht erkannten LS-AW-Fälle (20 %) und der 7 LS-AL-Fälle (12 %) als GBF durch Fehlerfortpflanzung.

Obwohl das Modell die Fehler nahezu fehlerfrei korrigierte (P2), wies es unabhängige Schwierigkeiten bei der Generierung von Feedback zu diesen Korrekturen (P3) auf. In dieser Phase traten zusätzlich zu den aus Phase 1 übertragenen bedingten Fehlschlägen (LS-AW 20 %, LS-AL 12 %) neue direkte Feedback-Fehler auf. Direkte Fehler traten am häufigsten

mit 5 Fällen (8 %) in der Kategorie GR-AL und 4 Fällen (7 %) in der Kategorie GR-AW auf. Diese Situation deutet darauf hin, dass die Feedback-Generierung einen von der Erkennung und Korrektur unabhängigen Schwierigkeitsbereich darstellt. Die vierte Phase der Analyse (Nicht-korrektives Feedback) zeigte, dass das Modell diesen Typ insgesamt 25 Mal generierte und sich am häufigsten mit 8 Fällen auf die Kategorie GR-AW konzentrierte. Diese Gesamtfrequenz positioniert Gemini als das Modell, das diesen Feedback-Typ am seltensten generiert.

Zusammenfassend lässt sich sagen, dass die kumulative Leistung des Gemini-Modells eine methodisch klare Unterscheidung aufweist. Das stärkste und entscheidendste Merkmal des Modells ist die in Phase 2 gezeigte, nahezu fehlerfreie direkte Korrekturfähigkeit (insgesamt 2 Fehler). Die gewonnenen Daten belegen, dass der die Gesamtleistung des Modells einschränkende Faktor nicht eine Korrekturunzulänglichkeit ist, sondern vielmehr aus zwei unterschiedlichen Quellen stammt: Erstens die Erkennungsschwäche (P1), die sich auf die Kategorien LS-AW (20 %) und LS-AL (12 %) konzentriert, was zu Fehlerfortpflanzung führt. Zweitens die unabhängigen Schwierigkeiten, die im Prozess der Feedback-Generierung (P3) trotz erfolgreicher Korrekturen auftreten, wie in den Kategorien GR-AL (8 %) und GR-AW (7 %) zu sehen ist.

Vergleichende Leistungen der Chatbots nach Fehlertypen

In diesem Abschnitt werden die Namen der KI-Chatbots aus Gründen der formalen Eignung der Tabellen als *ChatGPT (C), DeepSeek (D)* und *Gemini (G)* abgekürzt. Unter der Überschrift *Durchschnitt* sind hingegen die Durchschnittswerte der KI-Chatbots unter Berücksichtigung ihrer Leistungen bei den jeweiligen Fehlerarten auf allen Sprachniveaus angegeben.

Tabelle 14.

Vergleich der GR-AL Fehler nach Niveau, Phase und Chatbot

P	S	GR-AL											
		A1			A2			B1			Durchschnitt		
		C	D	G	C	D	G	C	D	G	C	D	G
P1	GE	100%	100%	100%	100%	100%	95%	80%	80%	80%	93%	93%	92%
	GF	0%	0%	0%	0%	0%	5%	20%	20%	20%	7%	7%	8%
P2	GE	100%	100%	100%	100%	100%	95%	75%	80%	80%	92%	93%	92%
	GF	0%	0%	0%	0%	0%	0%	5%	0%	0%	2%	0%	0%
	GBF	0%	0%	0%	0%	0%	5%	20%	20%	20%	7%	7%	8%
P3	GE	95%	100%	90%	85%	100%	90%	60%	70%	70%	80%	90%	83%
	GF	5%	0%	10%	15%	0%	5%	15%	10%	10%	12%	3%	8%
	GBF	0%	0%	0%	0%	0%	5%	25%	20%	20%	8%	7%	8%

Diese Tabelle untersucht vergleichend die Leistung von drei KI-Modellen (ChatGPT, DeepSeek und Gemini) speziell für den Fehlertyp GR-AL (Grammatik - Auslassung) anhand der Datensätze A1, A2 und B1. Die kumulativen Befunde zeigen, dass das DeepSeek-Modell über alle Phasen und Datensätze hinweg die höchste und konsistenteste Leistung erbringt. Die Analyse bestätigt zudem, dass der B1-Datensatz für alle drei Modelle das anspruchsvollste Niveau darstellt und Phase 3 (Feedback-Generierung) die Phase ist, in der die deutlichsten Leistungsunterschiede zwischen den Modellen auftreten und die die größten Schwierigkeiten bereitet.

Bei der Untersuchung der Daten aus Phase 1, welche die Fehlererkennungsfähigkeit misst, zeigt sich, dass im A1-Datensatz alle drei Modelle (ChatGPT, DeepSeek, Gemini) eine Erfolgsquote von 100 % aufweisen. Als der Schwierigkeitsgrad auf das A2-Set anstieg, behielten ChatGPT und DeepSeek ihre 100%ige Erfolgsquote bei, während Gemini mit 95 % Erfolg (5 % Fehlschlag) einen leichten Rückgang verzeichnete. Im B1-Datensatz, dem

anspruchsvollsten Niveau, fielen alle drei Modelle auf eine gleiche Leistung von 80 % Erfolg (20 % Fehlschlag) zurück. Im kumulativen Durchschnitt weisen DeepSeek (93 % Erfolg) und ChatGPT (93 % Erfolg) die höchste Erkennungsrate auf, während Gemini (92 % Erfolg) diesen Modellen sehr dicht folgt.

Die Analyse der Phase 2 (Fehlerkorrektur) offenbart wichtige Unterschiede in den Fehlerkorrekturstrategien der Modelle. Im A1-Set sind alle drei Modelle zu 100 % erfolgreich. Im B1-Set hingegen zeigen DeepSeek und Gemini 80 % Erfolg (20 % GBF), während ChatGPT auf 75 % Erfolg abfiel. Dieser Rückgang bei ChatGPT ist darauf zurückzuführen, dass es zusätzlich zu den 20 % bedingtem Scheitern einen direkten Korrekturfehler (GF) von 5 % machte. Dieser Befund spiegelt sich auch in den kumulativen Durchschnittswerten wider: Während DeepSeek und Gemini in dieser Fehlerkategorie durchschnittlich 0 % direkte Korrekturfehler machten, lag dieser Wert bei ChatGPT bei 2 %. Die gewonnenen Daten zeigen, dass der Hauptfaktor für den Leistungsabfall in dieser Phase weniger die direkte Korrekturunzulänglichkeit ist, sondern vielmehr die Widerspiegelung der Erkennungsfehler aus Phase 1 im Bereich von 7 % bis 8 % als bedingtes Scheitern in den nachfolgenden Phasen.

Die Phase, in der die Leistungsunterschiede zwischen den Modellen am deutlichsten sind und die niedrigsten Erfolgsquoten verzeichnet wurden, ist Phase 3 (Feedback-Generierung). Selbst im A1-Set erreichte DeepSeek 100 % Erfolg, während Gemini bei 90 % und ChatGPT bei 95 % Erfolg blieben. Im A2-Set behielt DeepSeek 100 % Erfolg bei, während ChatGPT auf 85 % und im B1-Set sogar auf 60 % zurückfiel. Im B1-Set fielen auch DeepSeek und Gemini auf 70 % Erfolg ab. Im kumulativen Durchschnitt ist DeepSeek mit einer Erfolgsquote von 90 % in dieser Phase das mit Abstand erfolgreichste Modell; ihm folgen Gemini (83 %) und ChatGPT (80 %). Der Hauptindikator für die Schwierigkeit in dieser Phase sind die Raten der direkten Feedback-Fehler (GF). Während DeepSeek mit 3 % das Modell mit den wenigsten direkten Fehlern ist, wurde diese Rate für Gemini mit 8 % und für ChatGPT mit 12 % gemessen.

Zusammenfassend lässt sich sagen, dass die vergleichende Analyse speziell für den Fehlertyp GR-AL zeigt, dass das DeepSeek-Modell die konsistenteste und erfolgreichste Leistung erbringt. Der B1-Datensatz sticht als das anspruchsvollste Niveau für alle drei Modelle hervor, das die Erkennungs- (80 %) und Feedback-Raten (60 %-70 %) stark absinken lässt. Obwohl die Korrekturfähigkeiten (P2) der Modelle (0 % direkte Fehler bei DeepSeek und Gemini) hoch bleiben, wurde festgestellt, dass zwei Hauptfaktoren die Gesamtleistung in der Kategorie GR-AL bestimmen: Erstens die Widerspiegelung der Erkennungsfehler aus Phase 1 in den nachfolgenden Phasen durch Fehlerfortpflanzung (bedingtes Scheitern); zweitens die direkten Fehlerraten in der Feedback-Phase (P3), mit denen insbesondere ChatGPT (12 %) Schwierigkeiten hatte.

Tabelle 15.
Vergleich der GR-HF Fehler nach Niveau, Phase und Chatbot

P	S	GR-HF											
		A1			A2			B1			Durchschnitt		
		C	D	G	C	D	G	C	D	G	C	D	G
P1	GE	100%	100%	100%	100%	100%	100%	100%	100%	90%	100%	100%	97%
	GF	0%	0%	0%	0%	0%	0%	0%	0%	10%	0%	0%	3%
P2	GE	100%	100%	95%	95%	100%	100%	100%	90%	90%	98%	97%	95%
	GF	0%	0%	5%	5%	0%	0%	0%	10%	0%	2%	3%	2%
	GBF	0%	0%	0%	0%	0%	0%	0%	0%	10%	0%	0%	3%
P3	GE	100%	100%	95%	85%	95%	100%	90%	75%	85%	92%	90%	93%
	GF	0%	0%	0%	10%	5%	0%	10%	15%	5%	7%	7%	2%
	GBF	0%	0%	5%	5%	0%	0%	0%	10%	10%	2%	3%	5%

Diese Tabelle untersucht vergleichend die Leistung von drei KI-Modellen (ChatGPT, DeepSeek und Gemini) speziell für den Fehlertyp GR-HF (Grammatik - Hinzufügung) anhand der Datensätze A1, A2 und B1. Die gewonnenen Erkenntnisse zeigen, dass die Kategorie GR-HF ein Fehlertyp ist, bei dem alle Modelle generell hohe Erfolgsquoten

aufweisen. Der bemerkenswerteste Befund in dieser Kategorie ist jedoch, dass sich die Leistungsprofile der Modelle je nach Phase unterscheiden; kein Modell zeigt in allen Phasen eine konsistente Führung, das erfolgreichste Modell wechselt beim Übergang von der Erkennung (P1) zur Korrektur (P2) und zum Feedback (P3).

Die Daten der Phase 1, welche die Fehlererkennungsfähigkeit messen, bestätigen, dass dieser Fehlertyp für fast alle Modelle sehr leicht zu erkennen ist. ChatGPT und DeepSeek zeigten in allen drei Datensätzen (A1, A2, B1) eine fehlerfreie Erfolgsquote von 100 % und erreichten somit auch im Durchschnitt 100 %. Gemini zeigte mit einem durchschnittlichen Erfolg von 97 % eine sehr nahe Leistung an diesen Modellen; seinen einzigen Erkennungsfehler verzeichnete es im B1-Datensatz (90 % Erfolg, 10 % Fehlschlag).

Im Gegensatz zur Erkennungsphase begannen die Modelle in Phase 2 (Fehlerkorrektur), direkte Korrekturfehler (GF) zu machen. Die Hauptquelle für Leistungsverluste in dieser Kategorie sind weniger die Erkennungsfehler aus Phase 1 (nur 10 % bei Gemini), sondern vielmehr diese in Phase 2 gemachten direkten Fehler. In dieser Phase sticht ChatGPT mit einer kumulativen Erfolgsquote von 98 % als das erfolgreichste Modell hervor; ihm folgen DeepSeek (97 %) und Gemini (95 %). Die Leistung der Modelle variierte je nach Datensatz; während Gemini bei A1 5 % direkte Fehler machte, machte ChatGPT bei A2 5 % direkte Fehler, und im B1-Set verzeichnete DeepSeek 10 % direkte Fehler.

Bei der Untersuchung der Phase 3 (Feedback-Generierung) ging die Führung mit einem durchschnittlichen Erfolg von 93 % an das Gemini-Modell über. Der Erfolg von Gemini in dieser Phase basiert darauf, dass es mit durchschnittlich 2 % die niedrigste Rate an direkten Feedback-Fehlern (GF) aufweist. ChatGPT (92 % Durchschnitt) und DeepSeek (90 % Durchschnitt) hatten in dieser Phase größere Schwierigkeiten; beide Modelle machten durchschnittlich 7 % direkte Feedback-Fehler. Das anspruchsvollste Szenario in dieser Phase war die Leistung des DeepSeek-Modells im B1-Datensatz; der Erfolg des Modells sank auf 75 %, und der Hauptgrund für diesen Rückgang waren 15 % direkte Feedback-Fehler. Die

in dieser Phase beobachteten bedingten Fehlschläge (z. B. 5 % bei Gemini A1, 5 % bei ChatGPT A2) zeigen, dass sich die in Phase 2 gemachten direkten Korrekturfehler in dieser Phase widerspiegelten.

Zusammenfassend lässt sich sagen, dass die Kategorie GR-HF zwar für alle Modelle einen Bereich hoher Kompetenz darstellt, die Stärken der Modelle jedoch in unterschiedlichen Phasen zutage treten. Während ChatGPT und DeepSeek bei der Erkennung (P1) eine fehlerfreie Leistung zeigten, erreichte ChatGPT bei der Korrektur (P2) den höchsten Durchschnitt, und Gemini übernahm die Führung in der Feedback-Phase (P3), indem es die niedrigste Rate an direkten Fehlern aufwies. Es wurde festgestellt, dass die Leistungsverluste in dieser Kategorie weniger auf Fehlerfortpflanzung (bedingtes Scheitern) aufgrund von Erkennungsfehlern in Phase 1 zurückzuführen sind, sondern vielmehr auf die direkten Fehler, welche die Modelle in Phase 2 und Phase 3 zeigten.

Tabelle 16.
Vergleich der GR-AW Fehler nach Niveau, Phase und Chatbot

P	S	GR-AW											
		A1			A2			B1			Durchschnitt		
		C	D	G	C	D	G	C	D	G	C	D	G
P1	GE	100%	100%	100%	100%	95%	100%	100%	100%	100%	100%	98%	100%
	GF	0%	0%	0%	0%	5%	0%	0%	0%	0%	0%	2%	0%
P2	GE	95%	100%	100%	100%	90%	95%	95%	100%	100%	97%	97%	98%
	GF	5%	0%	0%	0%	5%	5%	5%	0%	0%	3%	2%	2%
	GBF	0%	0%	0%	0%	5%	0%	0%	0%	0%	0%	2%	0%
P3	GE	90%	100%	95%	80%	85%	95%	90%	90%	85%	87%	92%	92%
	GF	5%	0%	5%	20%	5%	0%	5%	10%	15%	10%	5%	7%
	GBF	5%	0%	0%	0%	10%	5%	5%	0%	0%	3%	3%	2%

Diese Tabelle untersucht vergleichend die Leistung von drei KI-Modellen (ChatGPT, DeepSeek und Gemini) speziell für den Fehlertyp GR-AW (Grammatik - Auswahl). Die Kategorie GR-AW zeigt ein Profil, bei dem sich die Stärken und Schwächen der Modelle je nach Phase deutlich unterscheiden. Während die Erkennungsphase (P1) für alle Modelle äußerst einfach ist (98 %-100 % Erfolg), wurde festgestellt, dass die Hauptquelle für Leistungsverluste nicht Erkennungsfehler sind, sondern vielmehr direkte Fehler (GF), die in Phase 2 (Korrektur) und insbesondere in Phase 3 (Feedback-Generierung) gemacht werden. Das Führungsprofil wechselt zwischen den Phasen: ChatGPT und Gemini erreichen bei der Erkennung, Gemini bei der Korrektur und DeepSeek sowie Gemini beim Feedback die höchsten Durchschnittswerte.

Die Daten der Phase 1, welche die Fehlererkennungsfähigkeit messen, bestätigen, dass die Erkennung dieses Fehlertyps für alle Modelle äußerst einfach ist. ChatGPT und Gemini zeigten in allen drei Datensätzen (A1, A2, B1) eine fehlerfreie Leistung und erreichten im Durchschnitt 100 % Erfolg. DeepSeek hingegen verzeichnete lediglich im A2-Datensatz einen Fehler (95 % Erfolg) und erzielte einen sehr hohen Erfolgsdurchschnitt von 98 %.

Im Gegensatz zum Erkennungserfolg machten alle Modelle in Phase 2 (Fehlerkorrektur) direkte Korrekturfehler (GF). In dieser Phase ist Gemini mit einem durchschnittlichen Erfolg von 98 % führend; ChatGPT und DeepSeek folgen mit einem Durchschnitt von 97 %. Die Leistung variierte je nach Datensatz: Während ChatGPT in den A1- und B1-Sets 5 % direkte Fehler machte, machte Gemini im A2-Set 5 % direkte Fehler. DeepSeek fiel hingegen im A2-Set auf 90 % Erfolg ab; dieser Rückgang resultierte sowohl aus 5 % direkten Fehlern als auch aus 5 % bedingtem Scheitern, das aus Phase 1 übertragen wurde.

Die Phase, in der die Leistungsunterschiede zwischen den Modellen am deutlichsten sind und die Erfolgsquoten am stärksten sinken, ist Phase 3 (Feedback-Generierung). In dieser Phase teilen sich DeepSeek und Gemini (92 % durchschnittlicher Erfolg) die Führung, während ChatGPT (87 % Durchschnitt) zurückbleibt. Der Grund für diesen niedrigen

Durchschnitt bei ChatGPT ist, dass es mit 10 % die höchste Rate an direkten Feedback-Fehlern (GF) aufweist. Das anspruchsvollste Szenario in dieser Phase trat im A2-Datensatz auf; ChatGPT fiel auf 80 % Erfolg ab und verzeichnete eine Rate von 20 % an direkten Feedback-Fehlern. Auch DeepSeek (85 % Erfolg) und Gemini (95 % Erfolg) hatten im A2-Set Schwierigkeiten.

Zusammenfassend lässt sich sagen, dass die Kategorie GR-AW zeigt, dass sich die Leistungsprofile der Modelle je nach Phase deutlich trennen. Während die Führung bei der Erkennung (P1) bei ChatGPT und Gemini liegt, geht sie bei der Korrektur (P2) an Gemini und beim Feedback (P3) an DeepSeek und Gemini über. Es wurde festgestellt, dass die Leistungsverluste in dieser Kategorie weniger auf Fehlerfortpflanzung durch Erkennungsfehler (0 %-2 %) zurückzuführen sind, sondern vielmehr auf die direkten Unzulänglichkeiten, welche die Modelle in Phase 2 (Korrektur) und insbesondere in Phase 3 (Feedback-Generierung) zeigten. Die Feedback-Phase im A2-Datensatz sticht, insbesondere für ChatGPT (20 % direkte Fehler), als das anspruchsvollste Szenario der gesamten Analyse hervor.

Tabelle 17.

Vergleich der GR-UM Fehler nach Niveau, Phase und Chatbot

P	S	GR-UM											
		A1			A2			B1			Durchschnitt		
		C	D	G	C	D	G	C	D	G	C	D	G
P1	GE	100 %	100 %	100 %	100 %	100 %	100 %	95 %	95 %	100 %	98 %	98 %	100 %
	GF	0%	0%	0%	0%	0%	0%	5%	5%	0%	2%	2%	0%
P2	GE	100 %	100 %	100 %	100 %	100 %	100 %	90 %	95 %	100 %	97 %	98 %	100 %
	GF	0%	0%	0%	0%	0%	0%	5%	0%	0%	2%	0%	0%
	GBF	0%	0%	0%	0%	0%	0%	5%	5%	0%	2%	2%	0%
P3	GE	95%	100 %	100 %	95%	100 %	100 %	90 %	90 %	100 %	93 %	97 %	100 %
	GF	5%	0%	0%	5%	0%	0%	0%	5%	0%	3%	2%	0%
	GBF	0%	0%	0%	0%	0%	0%	10 %	5%	0%	3%	2%	0%

Diese Tabelle untersucht vergleichend die Leistung von drei KI-Modellen (ChatGPT, DeepSeek und Gemini) speziell für den Fehlertyp GR-UM (Grammatik - Umstellung) anhand der Datensätze A1, A2 und B1. Die gewonnenen Erkenntnisse zeigen, dass diese Fehlerkategorie im Vergleich zu den anderen untersuchten Kategorien für alle Modelle eine der am wenigsten anspruchsvollen ist. Der auffälligste und entscheidendste Befund in dieser Kategorie ist, dass das Gemini-Modell mit einer fehlerfreien durchschnittlichen Erfolgsquote von 100 % über alle Datensätze und alle Phasen hinweg das mit Abstand erfolgreichste Modell ist.

Die Phasen der Fehlererkennung (P1) und Fehlerkorrektur (P2) bestätigen die allgemeine Einfachheit dieses Fehlertyps und die Überlegenheit des Gemini-Modells. In der Erkennungsphase erreichte Gemini in allen drei Datensätzen 100 % Erfolg und erzielte somit einen fehlerfreien Durchschnitt. ChatGPT und DeepSeek hingegen verzeichneten lediglich im B1-Datensatz 95 % Erfolg (5 % Fehlschlag) und erreichten einen Durchschnitt von 98 %. Die Korrekturphase (P2) verdeutlicht diese Tendenz noch weiter; Gemini zeigte auch in

dieser Phase einen fehlerfreien Durchschnitt von 100 %. Für DeepSeek (98 % Durchschnitt) und ChatGPT (97 % Durchschnitt) war B1 der einzige Datensatz, in dem Leistungsabfälle auftraten.

Bei der Untersuchung der Ursache für die Leistungsunterschiede zwischen den Modellen zeigt sich, dass die Leistungsverluste von DeepSeek und ChatGPT im B1-Set auf unterschiedliche Gründe zurückzuführen sind. Der 95%ige Erfolg von DeepSeek in der B1-Korrekturphase (P2) resultierte aus einem bedingten Scheitern von 5 %, das die Widerspiegelung des 5%igen Erkennungsfehlers aus Phase 1 darstellt. Im Gegensatz dazu resultierte der 90%ige Erfolg von ChatGPT sowohl aus 5 % bedingtem Scheitern als auch aus 5 % direkten Korrekturfehler (Gesamt fehlgeschlagen). Die Feedback-Phase (P3) zeigte eine ähnliche Tendenz; während Gemini seinen fehlerfreien Durchschnitt von 100 % beibehielt, verzeichnete DeepSeek 97 % und ChatGPT 93 % Erfolgsdurchschnitt. Die relativ geringe Leistung von ChatGPT in dieser Phase ist darauf zurückzuführen, dass es zusätzlich zu B1 auch in den A1- und A2-Sets jeweils 5 % direkte Feedback-Fehler machte.

Zusammenfassend lässt sich sagen, dass die Fehlerkategorie GR-UM als ein Bereich hoher Kompetenz für alle Modelle (durchschnittlich 97 %-98 % Erfolg) hervorsticht. Gemini ist mit einer fehlerfreien Erfolgsquote von 100 % in allen Phasen und allen Datensätzen das mit Abstand erfolgreichste Modell in dieser Kategorie. Die Hauptursachen für die geringfügigen Leistungsverluste, die ChatGPT und DeepSeek in dieser Kategorie (nur im B1-Set deutlich werdend) erlitten, sind unterschiedlich: während die Fehler von DeepSeek fast ausschließlich auf Fehlerfortpflanzung (bedingtes Scheitern) durch Erkennungsfehler in Phase 1 zurückzuführen sind, resultieren die Fehler von ChatGPT sowohl aus Erkennungsfehlern als auch aus direkten Korrektur-/Feedback-Fehlern (GF).

Tabelle 18.

Vergleich der LS-AL Fehler nach Niveau, Phase und Chatbot

P	S	LS-AL											
		A1			A2			B1			Durchschnitt		
		C	D	G	C	D	G	C	D	G	C	D	G
P1	GE	95%	100%	100%	85%	90%	85%	90%	90%	80%	90%	93%	88%
	GF	5%	0%	0%	15%	10%	15%	10%	10%	20%	10%	7%	12%
P2	GE	95%	100%	100%	85%	85%	85%	65%	90%	80%	82%	92%	88%
	GF	0%	0%	0%	0%	5%	0%	25%	0%	0%	8%	2%	0%
	GBF	5%	0%	0%	15%	10%	15%	10%	10%	20%	10%	7%	12%
P3	GE	85%	95%	100%	80%	70%	85%	60%	85%	80%	75%	83%	88%
	GF	10%	5%	0%	5%	15%	0%	5%	5%	0%	7%	8%	0%
	GBF	5%	0%	0%	15%	15%	15%	35%	10%	20%	18%	8%	12%

Diese Tabelle untersucht vergleichend die Leistung von drei KI-Modellen (ChatGPT, DeepSeek und Gemini) anhand der A1-, A2- und B1-Datensätze, spezifisch für den Fehlertyp LS-AL (Lexikalisch/Semantisch - Auslassung). Die gewonnenen Erkenntnisse zeigen, dass diese lexikalisch-semantische Kategorie im Vergleich zu den in früheren Analysen untersuchten Grammatiktypen für die Modelle deutlich anspruchsvoller ist. Die Leistungsprofile offenbaren, dass die Führung je nach Phase (Erkennung, Korrektur, Feedback-Generierung) wechselt und insbesondere ChatGPT bei diesem Fehlertyp, speziell im B1-Datensatz, erhebliche Schwierigkeiten aufwies.

Die Daten der Phase 1, welche die Fehlererkennungsfähigkeit messen, legen die Unterschiede zwischen den Modellen deutlich dar. Im kumulativen Durchschnitt ist DeepSeek mit einer Erfolgsquote von 93 % das Modell, das in dieser Kategorie die beste Erkennungsleistung erbringt. Ihm folgen ChatGPT (90 %) und Gemini (88 %); Gemini ist mit 12 % das Modell mit der höchsten durchschnittlichen Erkennungsfehlerrate. Diese Schwäche wird dadurch bestätigt, dass Gemini im B1-Datensatz mit 80 % die niedrigste

Erkennungsleistung aufwies. DeepSeek hingegen wies mit Erfolgsquoten von 100 % in A1, 90 % in A2 und 90 % in B1 das konsistenteste Erkennungsprofil auf.

Die Analyse der Phase 2 (Fehlerkorrektur) zeigt die qualitativen Unterschiede in den Fehlerkorrekturstrategien der Modelle und beleuchtet die Herausforderung durch die Kategorie LS-AL Obwohl DeepSeek im kumulativen Durchschnitt (92 %) den höchsten Erfolg aufweist, ist der methodisch wichtigste Befund, dass Gemini eine Rate von 0 % an direkten Korrekturfehlern (GF) hat; dieser Umstand zeigt, dass Gemini alle LS-AL-Fehler, die es erkennen konnte, fehlerfrei korrigierte. Im Gegensatz dazu wies ChatGPT in dieser Phase eine erhebliche Schwäche auf; das Modell fiel im B1-Datensatz auf eine Erfolgsquote von 65 % ab, und der Hauptgrund für diesen Rückgang war mit 25 % ein sehr hoher direkter Korrekturfehler. Kumulativ ist ChatGPT mit 8 % ebenfalls das Modell, das die meisten direkten Korrekturfehler macht.

Phase 3, welche die Fähigkeit zur Feedback-Generierung misst, wies für alle drei Modelle niedrige Erfolgsquoten auf, und die Führung ging an Gemini über. Gemini ist mit einem Erfolgsdurchschnitt von 88 % (einschließlich 100 % Erfolg bei A1) das erfolgreichste Modell in dieser Phase. Ihm folgen DeepSeek (83 %) und ChatGPT (75 %). Die niedrigsten Leistungen in dieser Phase wurden mit 60 % Erfolg von ChatGPT im B1-Set und 70 % Erfolg von DeepSeek im A2-Set verzeichnet.

Zusammenfassend lässt sich sagen, dass die Kategorie LS-AL ein Fehlertyp ist, bei dem die Modelle im Vergleich zu den Grammatiktypen deutlich größere Schwierigkeiten haben. Die Leistungsprofile trennen sich klar: Während DeepSeek bei der Erkennung (P1) das erfolgreichste Modell ist (93 %), wies Gemini im Korrekturprozess (P2) mit einer direkten Fehlerrate von 0 % die sauberste Leistung und in der Feedback-Phase (P3) den höchsten Gesamterfolg (88 %) auf. ChatGPT hingegen wies in dieser Kategorie in allen Phasen den niedrigsten kumulativen Erfolg auf; insbesondere der im B1-Datensatz gezeigte direkte

Korrekturfehler von 25 % deutet darauf hin, dass das Modell eine strukturelle Schwierigkeit bei der Korrektur dieses spezifischen lexikalisch-semantischen Fehlertyps aufweist.

Tabelle 19.
Vergleich der LS-HF Fehler nach Niveau, Phase und Chatbot

P	S	LS-HF											
		A1			A2			B1			Durchschnitt		
		C	D	G	C	D	G	C	D	G	C	D	G
P1	GE	100%	100%	100%	100%	100%	100%	100%	100%	95%	100%	100%	98%
	GF	0%	0%	0%	0%	0%	0%	0%	0%	5%	0%	0%	2%
P2	GE	100%	100%	100%	95%	95%	100%	100%	70%	95%	98%	88%	98%
	GF	0%	0%	0%	5%	5%	0%	0%	30%	0%	2%	12%	0%
	GBF	0%	0%	0%	0%	0%	0%	0%	0%	5%	0%	0%	2%
P3	GE	100%	100%	100%	85%	95%	100%	100%	70%	95%	95%	88%	98%
	GF	0%	0%	0%	10%	0%	0%	0%	0%	0%	3%	0%	0%
	GBF	0%	0%	0%	5%	5%	0%	0%	30%	5%	2%	12%	2%

Diese Tabelle untersucht vergleichend die Leistung von drei KI-Modellen (ChatGPT, DeepSeek und Gemini) anhand der Datensätze A1, A2 und B1, spezifisch für den Fehlertyp LS-HF (Lexikalisch/Semantisch - Hinzufügung). Die gewonnenen Erkenntnisse zeigen, dass diese Kategorie für Gemini und ChatGPT einen Bereich hohen Erfolgs darstellt, jedoch für DeepSeek im B1-Datensatz eine deutliche und spezifische Schwäche birgt. Der kritischste Befund der Analyse ist, dass DeepSeek trotz einer 100%igen Erkennungsleistung (P1) im B1-Set einen hohen direkten Korrekturfehler (P2) von 30 % aufwies. Im Gegensatz dazu zeigte Gemini mit einem durchschnittlichen Erfolg von 98 % in allen Phasen und einer Rate von 0 % an direkten Korrektur- oder Feedback-Fehlern die konsistenteste Leistung.

Die Daten der Phase 1, welche die Fehlererkennungsfähigkeit messen, bestätigen, dass dieser Fehlertyp von allen Modellen mit hohem Erfolg erkannt wird. ChatGPT und

DeepSeek zeigten in allen drei Datensätzen (A1, A2, B1) eine fehlerfreie Leistung und erreichten im Durchschnitt 100 % Erfolg. Gemini hingegen war in den A1- und A2-Sets zu 100 % erfolgreich, verzeichnete jedoch im B1-Set einen Erfolg von 95 % (5 % Fehlschlag) und erreichte einen Durchschnitt von 98 %.

Der eigentliche Leistungsunterschied zwischen den Modellen tritt in Phase 2 (Korrektur) zutage. Gemini zeigte in dieser Phase mit einer Rate von 0 % an direkten Korrekturfehlern (GF) und einem Gesamterfolgsdurchschnitt von 98 % die sauberste und stabilste Leistung. Auch ChatGPT (%98 Durchschnitt, 2 % direkte Fehler) wies eine sehr hohe Kompetenz auf. Im Gegensatz dazu erlitt die Leistung von DeepSeek im B1-Datensatz einen dramatischen Rückgang; das Modell fiel in diesem Set auf 70 % Erfolg zurück, und dieser Rückgang war ausschließlich auf 30 % direkte Korrekturfehler (GF) zurückzuführen. Dieser spezifische Fehlschlag zog den kumulativen Durchschnitt von DeepSeek mit 12 % direkten Fehlern auf 88 %.

Die Auswirkung dieses in Phase 2 beobachteten direkten Korrekturfehlers spiegelte sich durch Fehlerfortpflanzung in Phase 3 (Feedback-Generierung) wider. Der 30%ige direkte Fehler von DeepSeek im B1-Set wandelte sich in Phase 3 zu 30 % bedingtem Scheitern um und senkte den Erfolg des Modells auch in dieser Phase auf 70 %. Diese Situation führte dazu, dass der kumulative Durchschnitt von DeepSeek bei 88 % blieb und es mit 12 % die höchste Rate an GBF aufwies. ChatGPT (95 % Durchschnitt) und Gemini (98 % Durchschnitt) behielten hingegen auch in dieser Phase ihre hohe Leistung bei.

Zusammenfassend lässt sich sagen, dass die Kategorie LS-HF die Leistungsprofile der Modelle klar voneinander trennt. Gemini sticht mit einer direkten Fehlerrate von 0 % (P2) in allen Phasen als das erfolgreichste und konsistenteste Modell hervor; auch ChatGPT zeigte eine sehr starke Leistung. Für DeepSeek wurde hingegen nachgewiesen, dass das Hauptproblem nicht bei der Erkennung (100 % Erfolg), sondern in der Korrekturphase (P2) lag. Der im B1-Set beobachtete direkte Korrekturfehler von 30 % deutet auf eine strukturelle

Schwäche hin, die dieses Modell in dieser spezifischen Kategorie aufweist, und es wurde festgestellt, dass sich dieser Fehler als bedingtes Scheitern in Phase 3 widerspiegelte.

Tabelle 20.
Vergleich der LS-AW Fehler nach Niveau, Phase und Chatbot

P	S	LS-AW											
		A1			A2			B1			Durchschnitt		
		C	D	G	C	D	G	C	D	G	C	D	G
P1	GE	70%	80%	80%	85%	85%	75%	55%	80%	85%	70%	82%	80%
	GF	30%	20%	20%	15%	15%	25%	45%	20%	15%	30%	18%	20%
P2	GE	70%	80%	80%	80%	80%	75%	55%	80%	85%	68%	80%	80%
	GF	0%	0%	0%	5%	5%	0%	0%	0%	0%	2%	2%	0%
	GBF	30%	20%	20%	15%	15%	25%	45%	20%	15%	30%	18%	20%
P3	GE	60%	80%	80%	75%	80%	70%	45%	80%	80%	60%	80%	77%
	GF	10%	0%	0%	5%	0%	5%	10%	0%	5%	8%	0%	3%
	GBF	30%	20%	20%	20%	20%	25%	45%	20%	15%	32%	20%	20%

Diese Tabelle untersucht vergleichend die Leistung von drei KI-Modellen (ChatGPT, DeepSeek und Gemini) anhand der A1-, A2- und B1-Datensätze, spezifisch für den Fehlertyp LS-AW (Lexikalisch/Semantisch - Auswahl). Die gewonnenen Erkenntnisse legen kategorisch dar, dass die Kategorie LS-AW unter allen analysierten Kategorien der Bereich ist, in dem alle drei Modelle die größten Schwierigkeiten hatten und die niedrigsten Erfolgsquoten aufwiesen. In dieser anspruchsvollen Kategorie zeigten DeepSeek und Gemini eine relativ konsistente Leistung von etwa 80 %, während ChatGPT einen deutlich geringeren Erfolg als die beiden anderen Modelle aufwies und dieser Fehlertyp sich als der ausgeprägteste Schwachstellenbereich von ChatGPT herauskristallisierte.

Der Hauptgrund für die geringe Leistung der Modelle in dieser Kategorie sind die hohen Fehlerraten in Phase 1 (Fehlererkennung). Im kumulativen Durchschnitt ist DeepSeek mit

82 % Erfolg (18 % Fehlschlag) das Modell mit der besten Erkennungsleistung; ihm folgt Gemini (80 % Erfolg, 20 % Fehlschlag). ChatGPT hingegen wies mit durchschnittlich 70 % Erfolg (30 % Fehlschlag) die niedrigste Erkennungsleistung auf. Der Hauptgrund für diesen niedrigen Durchschnitt von ChatGPT ist seine Leistung im B1-Datensatz; das Modell konnte in diesem Set 45 % der LS-AW-Fehler nicht erkennen und erreichte lediglich eine Erfolgsquote von 55 %. Auch im A1-Set blieb ChatGPT (70 %) hinter den beiden anderen Modellen (80 %) zurück.

Diese hohen Erkennungsfehler in Phase 1 haben die Leistung der nachfolgenden Phasen durch Fehlerfortpflanzung direkt bestimmt. Die Analyse der Phase 2 (Fehlerkorrektur) belegt dieses Phänomen eindeutig. Während DeepSeek und Gemini (80 %) im kumulativen Durchschnitt ihre Führung behielten, weist ChatGPT (68 %) den niedrigsten Durchschnitt auf. Der methodisch wichtigste Befund ist, dass der Durchschnitt der direkten Korrekturfehler (GF) bei den Modellen ChatGPT und DeepSeek mit 2 % gleich hoch ist. Dieser Befund bestätigt, dass der Leistungsunterschied der Modelle in dieser Kategorie nicht auf die Korrekturfähigkeit zurückzuführen ist; er bestätigt, dass der Unterschied aus der sehr hohen Rate von 30 % an GBF resultiert, die ChatGPT aus Phase 1 übernommen hat (DeepSeek 18 %).

Eine ähnliche Tendenz setzte sich auch in Phase 3 (Feedback-Generierung) fort. Während DeepSeek (80 % Durchschnitt) und Gemini (77 % Durchschnitt) ihre Konsistenz im 80 %-Bereich beibehielten, fiel der Durchschnitt von ChatGPT auf bis zu 60 %. Die niedrigste Leistung in dieser Phase war erneut die von ChatGPT im B1-Datensatz verzeichnete Erfolgsquote von 45 %. Der Hauptgrund auch für diesen Rückgang ist die kumulative Rate von 32 % an GBF von ChatGPT in dieser Phase.

Zusammenfassend lässt sich sagen, dass die Kategorie LS-AW als der anspruchsvollste Fehlertyp für alle Modelle identifiziert wurde. DeepSeek und Gemini zeigten selbst in dieser anspruchsvollen Kategorie eine relativ konsistente Leistung, indem sie einen

Erfolgsdurchschnitt von nahezu 80 % erreichten. ChatGPT hingegen wies mit Erfolgsquoten, die insbesondere im B1-Set auf 45 %-55 % fielen (60 %-70 % Gesamtdurchschnitt), eine deutliche Schwäche in dieser Kategorie auf. Die Analyse belegt, dass der Hauptgrund für die geringe Leistung in dieser Kategorie nicht ein direktes Versagen der Modelle bei den Korrektur- (P2) oder Feedback-Fähigkeiten (P3) ist; sie belegt, dass das Problem grundlegend auf die hohen Fehlerraten in Phase 1 (Erkennung) (insbesondere 30 % bei ChatGPT) und die Widerspiegelung dieser nicht erkannten Fälle als bedingtes Scheitern in den nachfolgenden Phasen zurückzuführen ist.

Tabelle 21.
Vergleich der LS-UM Fehler nach Niveau, Phase und Chatbot

P	S	LS-UM											
		A1			A2			B1			Durchschnitt		
		C	D	G	C	D	G	C	D	G	C	D	G
P1	GE	95%	100%	95%	95%	100%	95%	85%	90%	100%	92%	97%	97%
	GF	5%	0%	5%	5%	0%	5%	15%	10%	0%	8%	3%	3%
P2	GE	95%	100%	100%	95%	100%	95%	80%	90%	100%	90%	97%	98%
	GF	0%	0%	0%	0%	0%	0%	5%	0%	0%	2%	0%	0%
	GBF	5%	0%	0%	5%	0%	5%	15%	10%	0%	8%	3%	2%
P3	GE	95%	95%	100%	90%	100%	95%	75%	85%	95%	87%	93%	97%
	GF	0%	5%	0%	5%	0%	0%	5%	5%	5%	3%	3%	2%
	GBF	5%	0%	0%	5%	0%	5%	20%	10%	0%	10%	3%	2%

Diese Tabelle untersucht vergleichend die Leistung von drei KI-Modellen (ChatGPT, DeepSeek und Gemini) anhand der A1-, A2- und B1-Datensätze, spezifisch für den Fehlertyp LS-UM (Lexikalisch/Semantisch - Umstellung). Die gewonnenen Erkenntnisse zeigen, dass in dieser Kategorie eine klare Leistungsdifferenzierung zwischen den Modellen besteht und Gemini als das mit Abstand erfolgreichste Modell hervorsticht. Gemini erreichte in allen Phasen (Erkennung: 97 %, Korrektur: 98 %, Feedback: 97 %) die höchsten

durchschnittlichen Erfolgsquoten. Im Gegensatz dazu wies ChatGPT in dieser Kategorie in allen Phasen die niedrigste kumulative Leistung (im Bereich von 87 %-92 %) auf.

Der Hauptfaktor, der den Leistungsunterschied zwischen den Modellen verursacht, ist die Kompetenz in Phase 1 (Fehlererkennung). Im kumulativen Durchschnitt sind DeepSeek und Gemini mit einer Erfolgsquote von 97 % die Modelle mit der besten Erkennung, während ChatGPT bei 92 % (8 % Fehlschlag) blieb. Dieser 8%ige Erkennungsfehler von ChatGPT wird besonders im B1-Set (85 % Erfolg) deutlich und bildet eine Grundlage für die nachfolgenden Phasen. Die Auswirkungen dieses anfänglichen Erkennungsfehlers spiegeln sich durch Fehlerfortpflanzung in Phase 2 (Korrektur) wider. Während Gemini (98 % Durchschnitt) und DeepSeek (97 % Durchschnitt) in der Korrekturphase ihre hohe Leistung beibehalten, sinkt der Durchschnitt von ChatGPT auf 90 %. Der methodisch wichtigste Befund ist, dass Gemini in dieser Phase eine Rate von 0 % an direkten Korrekturfehlern (GF) aufweist. Der niedrige Erfolg von ChatGPT von 80 % im B1-Set hingegen resultiert sowohl aus 5 % direkten Korrekturfehlern als auch aus 15 % bedingtem Scheitern, das sich aus Phase 1 widerspiegelte.

Phase 3, welche die Fähigkeit zur Feedback-Generierung misst, bestätigt diese Tendenz. Während Gemini (97 % Durchschnitt) und DeepSeek (93 % Durchschnitt) die Führung beibehalten, ist der Durchschnitt von ChatGPT auf bis zu 87 % gesunken. Die niedrigste Leistung von ChatGPT ist der im B1-Set verzeichnete Erfolg von 75 %. Die kumulativen Daten zeigen, dass die Hauptfaktoren, die die Leistung von ChatGPT senken, die auf 10 % ansteigende Rate des GBF und die Rate von 3 % GF (direkter Feedback-Fehler) sind.

Zusammenfassend lässt sich sagen, dass die Fehlerkategorie LS-UM die Leistung der Modelle deutlich trennt. Gemini hat seine Überlegenheit in diesem Bereich bewiesen, indem es in allen Phasen die höchsten Durchschnittswerte erreichte (97 %-98 %) und 0 % direkte Korrekturfehler aufwies. Es wurde festgestellt, dass der Hauptgrund dafür, dass ChatGPT in allen Phasen die niedrigste Leistung (%87-%92) zeigte, nicht eine direkte Schwäche bei der

Korrektur- (P2) oder Feedback-Fähigkeit (P3) ist, sondern vielmehr die Widerspiegelung des 8%igen Erkennungsfehlers aus Phase 1 als hohe Raten an bedingtem Scheitern (%8-%10) in den nachfolgenden Phasen.

Vergleichende Leistungen der Chatbots nach ihrem nicht-korrigierenden Feedback

Tabelle 22.
Vergleich der Chatbots für das A1-Niveau hinsichtlich P4

Chatbot	Fehlerebene und -typ							
	GR				LS			
	AL	HF	AW	UM	AL	HF	AW	UM
ChatGPT	0	0	5	1	0	1	1	2
DeepSeek	0	0	5	1	0	1	1	2
Gemini	0	0	4	0	0	1	3	2
Gesamt	0	0	14	2	0	3	5	6

Diese Tabelle untersucht ein Verhaltensprofil von drei KI-Modellen (ChatGPT, DeepSeek und Gemini), spezifisch für den A1-Datensatz. Bei dieser Analyse handelt es sich nicht um eine Messung von Erfolg oder Misserfolg, sondern um einen Vergleich der Häufigkeit (Frequenz) der Generierung von nicht-korrektivem Feedback, das als vierte Phase definiert ist. Der auffälligste Befund bezüglich des A1-Datensatzes ist, dass die Modelle ChatGPT und DeepSeek hinsichtlich der Generierung dieser Feedback-Art exakt dasselbe Verhaltensprofil aufweisen und sich dieses Verhalten am stärksten auf die Kategorie GR-AW konzentriert.

Im A1-Datensatz wurde eine vollständige Übereinstimmung zwischen den Frequenzverteilungen der Modelle ChatGPT und DeepSeek beobachtet. Beide Modelle generierten diese Art von Feedback insgesamt 10 Mal, und 5 dieser Fälle wurden als höchste Frequenz in der Kategorie GR-AW (Grammatik - Auswahl) verzeichnet. Obwohl das Gemini-Modell ebenfalls insgesamt 10 Fälle generierte, unterschied sich die Verteilung dieser Fälle von den beiden anderen Modellen. Die Frequenzen von Gemini zeigten eine

ausgewogenere Verteilung zwischen der Kategorie GR-AW (4 Fälle) und der Kategorie LS-AW (Lexikalisch/Semantisch - Auswahl) (3 Fälle).

Bei der Untersuchung der kumulativen Frequenzen (Zeile *Gesamt*) der drei Modelle im A1-Datensatz wird deutlich, dass die Kategorie GR-AW mit insgesamt 14 Fällen die Kategorie ist, in der diese Art von Feedback am häufigsten ausgelöst wurde. Auf diese Kategorie folgen die Kategorien LS-UM (Lexikalisch/Semantisch - Umstellung) mit 6 Fällen und LS-AW mit 5 Fällen. Als bemerkenswerter Befund gilt, dass die Kategorien GR-AL (Grammatik - Auslassung), GR-HF (Grammatik - Hinzufügung) und LS-AL (Lexikalisch/Semantisch - Auslassung) im A1-Set bei keinem der drei Modelle (insgesamt 0 Fälle) diese Art von Feedback ausgelöst haben.

Zusammenfassend zeigt die Analyse der Phase 4 im A1-Datensatz, dass die Modelle diese spezifische Feedback-Art zwar insgesamt mit gleicher Häufigkeit (10 Fälle) generierten, ihre Verhaltensprofile sich jedoch unterschieden. Während ChatGPT und DeepSeek exakt dieselbe Verteilung aufwiesen, zeigten die Frequenzen von Gemini ein ausgewogeneres Erscheinungsbild zwischen GR-AW und LS-AW Während GR-AW (insgesamt 14 Fälle) der stärkste Auslöser für dieses Verhalten war, wurden die Kategorien GR-AL, GR-HF und LS-AL als Kategorien identifiziert, in denen diese Art von Feedback im A1-Set nicht beobachtet wurde.

Tabelle 23.
Vergleich der Chatbots für das A2-Niveau hinsichtlich P4

Chatbot	Fehlerebene und -typ							
	GR				LS			
	AL	HF	AW	UM	AL	HF	AW	UM
ChatGPT	0	1	4	5	0	0	3	1
DeepSeek	1	1	5	1	1	0	4	1
Gemini	0	1	1	2	0	0	0	0
Gesamt	1	3	10	8	1	0	7	2

Diese Tabelle vergleicht die Häufigkeit (Frequenz) der Generierung von nicht-korrektivem Feedback (P4) durch drei KI-Modelle (ChatGPT, DeepSeek und Gemini), spezifisch für den A2-Datensatz. Hierbei handelt es sich nicht um eine Erfolgsmessung, sondern um eine

Verhaltensprofilanalyse. Der auffälligste Befund bezüglich des A2-Datensatzes ist, dass ein deutlicher Unterschied in der Nutzungshäufigkeit dieses Feedbacks zwischen den Modellen besteht: Während ChatGPT und DeepSeek dieses Verhalten jeweils 14 Mal zeigten, setzte Gemini diese Art von Feedback mit nur 4 generierten Fällen deutlich seltener ein.

Bei der Untersuchung der individuellen Verhaltensprofile der Modelle ist ersichtlich, dass sich ChatGPT und DeepSeek trotz ähnlicher Gesamtfrequenzen (14 Fälle) auf unterschiedliche Kategorien konzentrierten. ChatGPT wies die höchste Frequenz mit 5 Fällen in der Kategorie GR-UM (Grammatik - Umstellung) und 4 Fällen in der Kategorie GR-AW (Grammatik - Auswahl) auf; in den Kategorien GR-AL, LS-AL und LS-HF erfolgte hingegen keine (0 Fälle) Generierung. DeepSeek hingegen zeigte die höchste Frequenz mit 5 Fällen in GR-AW und 4 Fällen in LS-AW (Lexikalisch/Semantisch - Auswahl); nur in der Kategorie LS-HF wurden 0 Fälle verzeichnet. Im Gegensatz dazu verteilte sich die Gesamtfrequenz von Gemini (4 Fälle) auf die Kategorien GR-UM (2 Fälle), GR-HF (1 Fall) und GR-AW (1 Fall); in fünf verschiedenen Kategorien (GR-AL, LS-AL, LS-HF, LS-AW, LS-UM) wurde diese Art von Feedback nicht generiert.

Bei der Betrachtung der kumulativen Frequenzen (Zeile *Gesamt*) der drei Modelle im A2-Datensatz wird deutlich, dass die Kategorie GR-AW mit insgesamt 10 Fällen die Kategorie ist, in der diese Art von Feedback am häufigsten ausgelöst wurde. Auf diese Kategorie folgen GR-UM mit 8 Fällen und LS-AW mit 7 Fällen. Im Unterschied zum A1-Set war LS-HF (Lexikalisch/Semantisch - Hinzufügung) die einzige Kategorie im A2-Set, die diese Art von Feedback überhaupt nicht (insgesamt 0 Fälle) auslöste.

Zusammenfassend lässt sich sagen, dass die Analyse des A2-Datensatzes offenbart, dass die Modelle in ihren Phase-4-Verhaltensprofilen deutlichere Unterschiede aufweisen als im A1-Set. Während ChatGPT und DeepSeek dieses Feedback häufig (14 Fälle) verwendeten, bevorzugte Gemini eine sehr seltene (4 Fälle) Nutzung. Kategorisch gesehen waren GR-AW (insgesamt 10 Fälle) und GR-UM (insgesamt 8 Fälle) die herausragendsten Auslöser für

dieses Verhalten im A2-Set, während die Kategorie LS-HF dieses Verhalten bei keinem Modell auslöste.

Tabelle 24.
Vergleich der Chatbots für das B1-Niveau hinsichtlich P4

Chatbot	Fehlerebene und -typ							
	GR				LS			
	AL	HF	AW	UM	AL	HF	AW	UM
ChatGPT	2	0	4	0	2	0	1	1
DeepSeek	2	1	9	6	0	1	3	0
Gemini	0	1	3	2	1	3	1	0
Gesamt	4	2	16	8	3	4	5	1

Diese Tabelle vergleicht die Häufigkeit (Frequenz) der Generierung von nicht-korrektivem Feedback (P4) durch drei KI-Modelle (ChatGPT, DeepSeek und Gemini) im B1-Datensatz, dem anspruchsvollsten Niveau. Der deutlichste Befund bezüglich des B1-Datensatzes ist, dass ein scharfer Unterschied hinsichtlich der Häufigkeit dieses Verhaltens zwischen den Modellen besteht. Während DeepSeek mit insgesamt 22 Fällen als das Modell hervorsticht, das diese Feedback-Art mit Abstand am häufigsten generiert, weisen ChatGPT (10 Fälle) und Gemini (11 Fälle) ähnliche und deutlich niedrigere Frequenzen auf.

Bei der Untersuchung der individuellen Verhaltensprofile der Modelle ist ersichtlich, dass sich die hohe Frequenz von 22 Fällen bei DeepSeek hauptsächlich auf zwei Grammatikkategorien konzentriert: GR-AW (Grammatik - Auswahl) (9 Fälle) und GR-UM (Grammatik - Umstellung) (6 Fälle). Obwohl die Gesamtfrequenzen der beiden anderen Modelle (ChatGPT 10, Gemini 11) nahe beieinander liegen, unterscheiden sich ihre Verteilungen. Während die höchste Frequenz von ChatGPT mit 4 Fällen in der Kategorie GR-AW beobachtet wurde (0 Fälle in den Kategorien GR-HF, LS-HF und LS-UM), verteilte sich die höchste Frequenz von Gemini mit jeweils 3 Fällen gleichmäßig auf die Kategorien GR-AW und LS-HF (Lexikalisch/Semantisch - Hinzufügung) (0 Fälle in den Kategorien GR-AL und LS-UM).

Bei der Betrachtung der kumulativen Frequenzen (Zeile *Gesamt*) der drei Modelle im B1-Datensatz wird bestätigt, dass die Kategorie GR-AW mit insgesamt 16 Fällen die Kategorie

ist, in der diese Art von Feedback am häufigsten ausgelöst wurde. Auf diese Kategorie folgen GR-UM mit 8 Fällen und LS-AW (Lexikalisch/Semantisch - Auswahl) mit 5 Fällen. Einer der bemerkenswertesten Befunde im B1-Set ist, dass die Kategorie LS-UM (insgesamt 1 Fall) diese Art von Feedback fast nie ausgelöst hat; diese Kategorie wurde bei den Modellen ChatGPT und Gemini mit 0 Fällen verzeichnet.

Zusammenfassend lässt sich sagen, dass die Analyse des B1-Datensatzes (P4) eine schärfere Trennung in den Verhaltensprofilen der Modelle im Vergleich zu den A1- und A2-Sets zeigt. Obwohl kumulativ GR-AW (insgesamt 16 Fälle) der Hauptauslöser für dieses Verhalten ist, ist der entscheidendste Befund der modellbasierte Frequenzunterschied. DeepSeek hat mit 22 Fällen (konzentriert insbesondere auf die Kategorien GR-AW und GR-UM) diese Feedback-Art im Vergleich zu den anderen Modellen weitaus häufiger verwendet. ChatGPT (10 Fälle) und Gemini (11 Fälle) hingegen wiesen ein ähnliches und konservativeres Frequenzprofil auf.

Tabelle 25.
Gesamtvergleich der Chatbots hinsichtlich P4

Chatbot	Fehlerebene und -typ							
	GR				LS			
	AL	HF	AW	UM	AL	HF	AW	UM
ChatGPT	2	1	13	6	2	1	5	4
DeepSeek	3	2	15	7	1	1	14	1
Gemini	0	2	8	4	1	4	4	2
Gesamt	5	5	36	17	4	6	23	7

Diese Tabelle vergleicht die kumulative (Gesamtsumme) Leistung von drei KI-Modellen (ChatGPT, DeepSeek und Gemini) in den A1-, A2- und B1-Datensätzen hinsichtlich der Häufigkeit der Generierung von nicht-korrektivem Feedback (P4). Es handelt sich hierbei nicht um eine Messung von Erfolg oder Misserfolg, sondern um eine Frequenzanalyse, die die Verhaltensprofile der Modelle offenlegt. Die gewonnenen kumulativen Erkenntnisse zeigen, dass hinsichtlich der Nutzungshäufigkeit dieser Feedback-Art eine deutliche Hierarchie zwischen den Modellen besteht: Während DeepSeek mit insgesamt 44 Fällen das Modell ist, das dieses Verhalten am häufigsten zeigt, ist Gemini mit insgesamt 25 Fällen das

Modell, das dieses Verhalten am seltensten zeigt. ChatGPT liegt mit 34 Fällen zwischen diesen beiden Modellen.

Bei der Untersuchung der individuellen Verhaltensprofile der Modelle wird deutlich, dass sich neben diesem Frequenzunterschied auch die Konzentrationsbereiche unterscheiden. DeepSeek (insgesamt 44 Fälle), das die höchste Frequenz aufweist, konzentrierte diese Nutzung auf die Kategorien GR-AW (Grammatik - Auswahl) (15 Fälle) und LS-AW (Lexikalisch/Semantisch - Auswahl) (14 Fälle). ChatGPT (insgesamt 34 Fälle) zeigte dieses Verhalten mit 13 Fällen am häufigsten in der Kategorie GR-AW Gemini (insgesamt 25 Fälle), das diese Art von Feedback am seltensten verwendete, verzeichnete seine höchste Frequenz ebenfalls in der Kategorie GR-AW mit 8 Fällen.

Werden die kumulativen Frequenzen (Zeile *Gesamt*) aller drei Modelle über alle Datensätze hinweg untersucht, treten die Kategorien, die dieses Verhalten auslösen, deutlich hervor. Die Kategorie GR-AW wurde mit insgesamt 36 Fällen als die Kategorie identifiziert, in der diese Art von Feedback mit Abstand am häufigsten ausgelöst wird. Auf diese Kategorie folgen LS-AW mit 23 Fällen und GR-UM (Grammatik - Umstellung) mit 17 Fällen. Im Gegensatz dazu waren die Kategorien LS-AL (Lexikalisch/Semantisch - Auslassung) (insgesamt 4 Fälle), GR-AL (Grammatik - Auslassung) (insgesamt 5 Fälle) und GR-HF (Grammatik - Hinzufügung) (insgesamt 5 Fälle) die Bereiche, in denen diese Feedback-Art kumulativ am seltensten beobachtet wurde.

Zusammenfassend bestätigen die kumulativen Daten der Phase 4, dass zwischen den Modellen ein klarer Unterschied hinsichtlich der Nutzungshäufigkeit dieser Feedback-Art besteht. Während DeepSeek (44 Fälle) dieses Verhalten kumulativ am häufigsten zeigt, ist Gemini (25 Fälle) das Modell, das es am seltensten zeigt. Auf kategorialer Ebene stachen GR-AW (insgesamt 36 Fälle) und LS-AW (insgesamt 23 Fälle) als die stärksten Auslöser dieses Verhaltensprofils hervor, während die Kategorien LS-AL, GR-AL und GR-HF als die

Bereiche identifiziert wurden, die kumulativ am wenigsten mit dieser Feedback-Art assoziiert sind.

TEIL V

SCHLUSSFOLGERUNG

Dieses abschließende Kapitel, das eine umfassende Synthese der im Abschnitt *Befunde und Interpretationen* der Untersuchung vorgestellten und mittels Tabellen detaillierten Daten anstrebt, sucht Antworten auf die zentralen Forschungsfragen der Dissertation. Diese Studie, die sich auf die Schreibfähigkeit im Kontext des Unterrichts von DaF konzentriert, hat die Fehlermanagement-Kompetenzen von drei führenden KI-Chatbots (ChatGPT, DeepSeek und Gemini) mittels einer methodischen und mehrstufigen Analyse bewertet. Die Leistungen der Modelle in den Prozessen des Erkennens (Phase 1: Erkennung), Lösens (Phase 2: Korrektur) und Erklärens (Phase 3: Feedback-Generierung) von acht verschiedenen Fehlertypen (in zwei Fehlerebenen (Grammatik und Lexikalisch/Semantisch)) in Texten der Niveaustufen A1, A2 und B1 wurden sowohl individuell als auch vergleichend untersucht.

Die kumulative Auswertung der gewonnenen Befunde hat eindeutig ergeben, dass keines der untersuchten KI-Modelle eine konsistente und absolute Überlegenheit über alle Fehlertypen, Schwierigkeitsgrade und Aufgabenphasen hinweg aufweist. Stattdessen wurde festgestellt, dass jedes Modell individuelle Leistungsmerkmale aufweist, die sich aus spezifischen, methodisch differenzierten Stärken und Schwächen zusammensetzen.

Diese qualitative Unterscheidung zwischen den Modellprofilen manifestiert sich wie folgt: Das Gemini-Modell sticht durch seine außerordentliche Kompetenz im Prozess der Phase 2 (Korrektur) hervor. Die kumulativen Daten zeigen, dass die Rate der direkten Korrekturfehler (GF) des Modells selbst in komplexen Kategorien wie LS-HF, LS-AL und LS-UM 0 % beträgt und im Gesamtdurchschnitt im Bereich von 0 %-2 % liegt, was eine nahezu fehlerfreie Korrekturleistung darstellt. Es wurde ersichtlich, dass der Hauptfaktor, der den Gesamterfolg dieses Modells einschränkt, weniger ein Mangel an Korrekturfähigkeit als vielmehr seine Schwäche in Phase 1 (Erkennung) ist. Insbesondere die Erkennungsfehler in den Kategorien LS-AW (20 %) und LS-AL (12 %) spiegelten sich als *GBF* in den nachfolgenden Phasen wider, wobei das Phänomen der Fehlerfortpflanzung den grundlegenden Schwächemechanismus dieses Modells darstellt.

DeepSeek fällt als das Profil auf, das unter den Modellen die ausgewogenste Gesamtleistung erbringt. DeepSeek zeigte sowohl bei der Erkennung von LS-AL (93 %) den höchsten Erfolg als auch in der Kategorie LS-AW (im Bereich von 80 %-82 %), die allen Modellen die größten Schwierigkeiten bereitete, eine im Vergleich zu ChatGPT deutlich stärkere Leistung. Dieses ausgewogene Profil birgt jedoch eine spezifische und unerwartete Schwäche, die durch den B1-Datensatz ausgelöst wurde. Die Schwachstelle des Modells trat in Phase 2 (Korrektur) der Kategorie LS-HF zutage; trotz einer 100%igen Erkennungsleistung wurde in dieser Kategorie im B1-Set ein hoher direkter Korrekturfehler von 30 % und in der kumulativen Gesamtsumme von 12 % verzeichnet.

ChatGPT weist hingegen ein deutliches Dilemma auf: Das Modell zeigt eine unbestreitbare Expertise in den Grammatik-Kategorien (GR); in den Kategorien GR-HF und GR-AW erzielte es fehlerfreie Erkennungsraten von 100 %. Im Gegensatz dazu trat die konsistenteste und deutlichste Schwäche des Modells bei Lexikalisch/Semantischen (LS) Aufgaben zutage, insbesondere in der Kategorie LS-AW ChatGPT blieb in dieser Kategorie mit kumulativen Erfolgsquoten zwischen 60 % (Feedback) und 70 % (Erkennung) stabil 10 bis 20 Prozentpunkte hinter den beiden anderen Modellen zurück. Darüber hinaus wies das

Fehlerprofil von ChatGPT das vielfältigste Fehlermuster auf, da es sowohl hohe bedingte Fehlschläge (Erkennungsfehler) in Kategorien wie LS-AW (30 %-32 %) als auch hohe direkte Feedback-Fehler (Erklärungsfehler) in Kategorien wie GR-AL (12 %) beinhaltete.

Die kumulativen Analysen bestätigten den Trend, dass die Aufgabenschwierigkeit bei allen Modellen parallel zur kognitiven Komplexität zunimmt (Phase 1 < Phase 2 < Phase 3). Phase 3 (Feedback-Generierung) belegte, dass die korrekte Korrektur eines Fehlers nicht garantiert, dass dieser Fehler auch korrekt erklärt werden kann (z. B. ChatGPT, GR-AL 12 % direkte Fehler), und dass diese Phase daher einen unabhängigen Schwierigkeitsbereich darstellt. Auch die kategoriale Schwierigkeitshierarchie trat deutlich zutage; während grammatikbasierte, regelbasierte Kategorien wie GR-UM und GR-HF die höchsten Erfolgsquoten aufwiesen, waren die Kategorien LS-AW und LS-AL, die kontextuellen Nuancen erfordern, die Bereiche, in denen die niedrigsten Erfolgsquoten verzeichnet wurden. Im Kontext der Datensätze fungierte das B1-Niveau als die effektivste Testfunktion, um diese oben definierten strukturellen Schwächen der Modelle (z. B. DeepSeeks LS-HF-Korrekturfehler, ChatGPTs LS-AL-Korrekturfehler) auszulösen.

Schließlich zeigten die Daten der Phase 4 (Nicht-korrektives Feedback), die eine Verhaltensprofilanalyse bieten, dass sich die Modelle auch hinsichtlich der Nutzungshäufigkeit dieser Art von Feedback unterschieden. Während DeepSeek (insgesamt 44 Fälle) diesen Typ am häufigsten verwendete, war Gemini (insgesamt 25 Fälle) das Modell, das ihn am seltensten nutzte. Kumulativ wurden GR-AW (36 Fälle) und LS-AW (23 Fälle) als die Kategorien identifiziert, in denen dieses Verhalten am häufigsten ausgelöst wurde.

Im Einleitungsteil der Arbeit werden unter der Problemstellung die Forschungsfragen dargelegt. Im Schlussteil wird versucht, diese Fragen einzeln zu beantworten.

Wie viele Fehler in schriftlichen Produkten im DaF-Unterricht können die KI-gestützten Chatbots ChatGPT, DeepSeek und Gemini identifizieren?

Die Antwort auf die Frage weist auf der Grundlage der in dieser Untersuchung gewonnenen kumulativen (Gesamtsumme) und vergleichenden (A1, A2, B1) Daten eine Struktur auf, die zu vielschichtig und komplex ist, um sie mit einem einzigen Prozentsatz auszudrücken. Die Fehlererkennungskompetenzen (P1) der Modelle variieren stark in Abhängigkeit von

1. der kategorialen Struktur des Fehlers (Grammatik [GR] oder Lexikalisch/Semantisch [LS]),
2. dem sprachlichen Schwierigkeitsniveau des Textes (deutlicher Einfluss des B1-Datensatzes) und
3. dem spezifisch verwendeten KI-Modell (ChatGPT, DeepSeek oder Gemini).

Die gewonnenen Befunde zeigen, dass es eine klare Schwierigkeitshierarchie bei der Fehlererkennungsleistung der Modelle gibt. An der Spitze dieser Hierarchie stehen Kategorien, in denen die Modelle eine nahezu fehlerfreie Kompetenz aufweisen, während sich am unteren Ende spezifische Fehlertypen befinden, bei denen alle Modelle erhebliche Schwierigkeiten haben.

Die kumulativen Daten zeigen, dass alle drei untersuchten KI-Modelle eine außerordentlich hohe Erfolgsquote bei der Erkennung von Fehlern in den Grammatik-Kategorien (GR) aufweisen. Die Kompetenz in diesem Bereich erreicht in einigen Kategorien Fehlerfreiheit.

In der Kategorie GR-HF (Grammatik - Hinzufügung) wiesen ChatGPT und DeepSeek einen kumulativen Erfolg von 100 % auf, während auch Gemini mit 97 % eine sehr hohe Rate erreichte. In der Kategorie GR-AW (Grammatik - Auswahl) zeigten ChatGPT und Gemini einen kumulativen Erfolg von 100 %, dicht gefolgt von DeepSeek mit 98 %. In der Kategorie GR-UM (Grammatik - Umstellung) erzielte Gemini einen fehlerfreien Durchschnitt von 100 %. Diese Befunde deuten darauf hin, dass die Erkennung von regelbasierten und strukturellen Grammatikfehlern für die aktuellen KI-Modelle ein weitgehend gelöstes Problem darstellt.

Selbst innerhalb der Grammatikkategorien besteht ein Schwierigkeitsunterschied. Die Kategorie GR-AL (Grammatik - Auslassung) sticht als der anspruchsvollste Fehlertyp unter der Überschrift Grammatik hervor. Obwohl die Modelle in den kumulativen Durchschnitten eine Leistung im Bereich von 92 %-93 % zeigten, legte der B1-Datensatz die Schwierigkeit dieser Kategorie deutlich offen: Auf dem B1-Niveau fiel der Erkennungserfolg aller drei Modelle (ChatGPT, DeepSeek und Gemini) auf 80 % (20 % Fehlschlag) ab.

Im Gegensatz zum hohen Erfolg der Modelle in den Grammatikkategorien stellen die Lexikalisch/Semantischen (LS) Kategorien den Bereich dar, in dem die Schwierigkeiten bei der Erkennung (P1) beginnen und die Leistungsunterschiede zwischen den Modellen deutlich werden.

Die Kategorie LS-HF (Lexikalisch/Semantisch - Hinzufügung) wies mit kumulativen Erkennungsraten von 100 % (ChatGPT, DeepSeek) und 98 % (Gemini) eine ähnliche Einfachheit wie die Grammatikkategorien auf. In ähnlicher Weise weist auch die Kategorie LS-UM (Lexikalisch/Semantisch - Umstellung) mit 97 % (DeepSeek, Gemini) und 92 % (ChatGPT) hohe Erfolgsquoten auf.

Das eigentliche Erkennungsproblem beginnt bei den Kategorien LS-AL (Lexikalisch/Semantisch - Auslassung) und LS-AW (Lexikalisch/Semantisch - Auswahl). Die Kategorie LS-AL differenzierte die Modelle im kumulativen Durchschnitt: Während DeepSeek mit 93 % Erfolg führend war, folgten ChatGPT (90 %) und Gemini (88 %). Der Abfall von Gemini auf 80 % im B1-Set in dieser Kategorie zeigt, dass die Erkennungsschwierigkeit hier niveauabhängig zunimmt.

Einer der deutlichsten Befunde der Untersuchung ist, dass die Kategorie LS-AW für alle drei Modelle bei weitem die anspruchsvollste Kategorie ist und die niedrigsten Erkennungserfolge verzeichnet wurden. Diese Kategorie legt die Grenzen der Modelle bei der Bewertung von kontextuellen Nuancen und der Angemessenheit der Wortwahl offen.

Selbst in den kumulativen Durchschnitten sind die Erfolgsquoten deutlich niedrig. DeepSeek erwies sich mit 82 % Erfolg (18 % Erkennungsfehler) als das leistungsstärkste Modell in dieser anspruchsvollen Kategorie. Ihm folgte Gemini mit 80 % Erfolg (20 % Erkennungsfehler). ChatGPT hingegen zeigte mit 70 % kumulativem Erfolg (30 % Erkennungsfehler) die niedrigste Leistung; dieser Befund zeigt, dass ChatGPT fast ein Drittel dieser lexikalisch-semantischen Fehler nicht erkennen konnte.

Die Schwierigkeit in der Kategorie LS-AW erreichte im B1-Datensatz ihren Höhepunkt. Während DeepSeek (80 %) und Gemini (85 %) selbst in diesem anspruchsvollen Set den 80 %-Bereich halten konnten, fiel die Erkennungsleistung von ChatGPT auf bis zu 55 % (45 % Fehlschlag) ab. Das bedeutet, dass ChatGPT fast die Hälfte der komplexen LS-AW-Fehler auf B1-Niveau nicht erkennen konnte.

Die Antwort auf die Frage, *wie viele der Fehler sie erkennen können*, ändert sich auch je nach dem spezifischen Profil des Modells:

- DeepSeek: Nach den kumulativen Daten weist DeepSeek das stärkste Erkennungsprofil (P1) auf. Es zeigte nicht nur in einfachen Kategorien wie GR-HF (100 %) und LS-HF (100 %) den höchsten kumulativen Erkennungserfolg, sondern auch in den beiden anspruchsvollsten lexikalisch-semantischen Kategorien LS-AL (93 %) und LS-AW (82 %).
- ChatGPT: Das Modell zeichnet mit seiner fehlerfreien Leistung in den Kategorien GR-HF (100 %) und GR-AW (100 %) das Profil eines Grammatikexperten. Diese Expertise lässt sich jedoch nicht auf den lexikalisch-semantischen Bereich übertragen. ChatGPT wies in der Kategorie LS-AW (70 %) die niedrigste kumulative Erkennung auf und zeigte im B1-Set (55 %) eine deutliche Schwäche in dieser Kategorie.
- Gemini: Gemini zeigte eine fehlerfreie Grammatikerkennung in den Kategorien GR-AW (100 %) und GR-UM (100 %). Die Gesamtleistung des Modells wird jedoch

von seiner Schwäche in der lexikalisch-semantischen Erkennungsphase (P1) beeinflusst. Die Erkennungsraten in den Kategorien LS-AL (88 %) und LS-AW (80 %) zeigen, dass dieses Modell eher Schwierigkeiten bei der Erkennungsfähigkeit (P1) als bei der Korrekturfähigkeit (P2) hat.

Zusammenfassend lässt sich sagen, dass sich die Fehlererkennungskompetenz von KI-Modellen auf einem Spektrum verteilt. An einem Ende dieses Spektrums befinden sich Kategorien wie GR-HF, GR-AW und LS-HF, in denen die Modelle diese Fehler mit Raten zwischen 97 % und 100 %, also nahezu fehlerfrei, erkennen können. Am anderen Ende des Spektrums befindet sich die Kategorie LS-AW, in der die Modelle Schwierigkeiten haben, kontextuelle und semantische Nuancen zu verstehen; in dieser Kategorie sinkt der kumulative Erkennungserfolg auf den Bereich von 70 %-82 % und im anspruchsvollsten B1-Szenario (für ChatGPT) auf bis zu 55 %.

Bei einer Gesamtbewertung bietet DeepSeek die ausgewogenste und höchste Erkennungsleistung, während ChatGPT sich auf Grammatik spezialisiert hat, jedoch die schwächste Leistung bei der Erkennung komplexer lexikalisch-semantischer Fehler zeigte. Gemini hingegen stach als ein Modell hervor, bei dem trotz starker Grammatikerkennung die Erkennungsdefizite im lexikalisch-semantischen Bereich (insbesondere LS-AL und LS-AW) das Gesamtprofil bestimmten.

Wie viele der identifizierten Fehler in schriftlichen Produkten im DaF-Unterricht können die KI-gestützten Chatbots ChatGPT, DeepSeek und Gemini korrekt korrigieren?

Die Frage bildet den methodischen Kern dieser Untersuchung und ist eine der Fragen, die die Kompetenzprofile der Modelle am deutlichsten voneinander abgrenzt. Die Antwort auf diese Frage basiert auf der korrekten Interpretation der beiden zentralen Metriken, die für Phase 2 (Fehlerkorrektur) verwendet werden: *GBF* und *GF* (direkter Korrekturfehler).

Wenn gemessen wird, *wie viele der erkannten Fehler korrigiert wurden*, wird der Einfluss der Erkennungsfehler aus Phase 1 (d. h. gesamt bedingt fehlgeschlagen) methodisch aus der

Gleichung herausgenommen. Der Fokus verlagert sich in diesem Fall vollständig auf die Metrik GF (direkter Korrekturfehler), die ausschließlich die Fähigkeit des Modells misst, einen erkannten Fehler korrekt zu lösen. Bei der Untersuchung der kumulativen (Gesamtsumme) und der B1-Niveau-Daten zeigt sich, dass die Modelle hinsichtlich dieser spezifischen Kompetenz deutliche und strukturelle Unterschiede aufweisen.

Der deutlichste Befund ist die außerordentliche Leistung des Gemini-Modells in diesem Bereich. Gemini wies ein nahezu fehlerfreies Profil bei der Korrektur der Fehler auf, die es erkennen konnte (Phase 2 direkter Fehler). Laut der kumulativen Gesamttabelle machte das Modell während aller Korrekturvorgänge in den acht Fehlerkategorien (insgesamt 480 Szenarien) nur 2 direkte Korrekturfehler (0,4 %) (jeweils 1 Fall in den Kategorien GR-HF und GR-AW). Noch wichtiger ist, dass die kumulative direkte Korrekturfehlerrate in allen als anspruchsvoll geltenden Lexikalisch/Semantischen Kategorien wie LS-AL, LS-HF und LS-UM 0 % beträgt. Diese Befunde belegen, dass der die Leistung von Gemini einschränkende Faktor nicht die Korrekturfähigkeit ist; sie belegen, dass das Modell, sobald es einen Fehler erkennt, diesen (unabhängig von der Kategorie) mit einer Genauigkeit von nahezu 100 % korrigiert. Die niedrigen Gesamterfolgsquoten des Modells in Phase 2 (z. B. LS-AW 80 %, LS-AL 88 %) sind daher ausschließlich ein Ergebnis der Fehlerfortpflanzung (bedingtes Scheitern), die aus den mangelnden Erkennungsleistungen in Phase 1 resultiert.

Im Gegensatz dazu sticht das ChatGPT-Modell als das Modell hervor, das die deutlichsten Schwierigkeiten bei der Korrektur erkannter Fehler aufweist. Im Gegensatz zu Gemini ist die Leistung von ChatGPT nicht nur durch die Erkennung (P1) begrenzt; das Modell macht auch in Phase 2 signifikante direkte Korrekturfehler. Die Kategorie, in der diese Schwäche am deutlichsten zutage trat, war LS-AL (Lexikalisch/Semantisch - Auslassung). ChatGPT wies in dieser Kategorie kumulativ 8 % (5 Fälle) direkte Korrekturfehler auf. Der Ursprung dieser Schwäche liegt im B1-Datensatz; das Modell erreichte bei der Korrektur von LS-AL-Fehlern im B1-Set eine sehr hohe direkte Fehlerrate von 25 % (5 Fälle). Dieser Umstand

zeigt, dass ChatGPT eine strukturelle Schwierigkeit aufweist, die richtige Lösung für diesen spezifischen lexikalisch-semantischen Fehler bereitzustellen, selbst wenn es ihn erkannt hat.

Das DeepSeek-Modell zeigt zwischen diesen beiden Extremprofilen ein ausgewogenes Profil, das im Allgemeinen hoch, jedoch mit einem spezifischen Makel behaftet ist. DeepSeek zeigte in Kategorien wie GR-AL mit 0 % direkten Korrekturfehlern eine ähnliche Fehlerfreiheit wie Gemini. Der Punkt, an dem das Modell jedoch Schwierigkeiten hatte, trat in der Kategorie LS-HF (Lexikalisch/Semantisch - Hinzufügung) zutage. DeepSeek wies in dieser Kategorie kumulativ eine direkte Korrekturfehlerrate von 12 % (7 Fälle) auf und schnitt damit in dieser spezifischen Kategorie sogar schlechter ab als ChatGPT (2 %). Diese hohe Fehlerrate ist ausschließlich auf den B1-Datensatz zurückzuführen; DeepSeek scheiterte bei der Korrektur von LS-HF-Fehlern im B1-Set in 30 % (6 Fälle) der Fälle direkt. Dieser Befund legt offen, dass die Korrekturfähigkeit von DeepSeek im Allgemeinen sehr hoch ist, es jedoch bei diesem spezifischen und komplexen Fehlertyp (auf B1-Niveau) eine deutliche Schwäche aufweist.

Innerhalb der Grammatik-Kategorien (GR) wurde hingegen beobachtet, dass direkte Korrekturfehler (Phase 2 GF) bei allen Modellen wesentlich seltener auftraten. Kumulativ blieben die direkten Fehlerraten in der Kategorie GR-HF mit 2 %-3 %, in der Kategorie GR-AW mit 2 % und in der Kategorie GR-AL mit 0 %-2 % auf sehr niedrigem Niveau. Dies zeigt, dass die Modelle eine hohe Kompetenz bei der Korrektur von regelbasierten Grammatikfehlern besitzen, die sie erkennen können.

Zusammenfassend lässt sich sagen, dass die Antwort auf die Frage, *wie viele der erkannten Fehler korrigiert werden*, je nach Modell stark variiert. Gemini kann nahezu alle Fehler, die es erkennt (99,6 %), korrekt korrigieren; für dieses Modell ist die einzige Hürde die Erkennungsphase. DeepSeek korrigiert die große Mehrheit der erkannten Fehler korrekt, zeigt jedoch eine spezifische direkte Korrekturschwäche von 30 % in der Kategorie LS-HF auf B1-Niveau. ChatGPT hingegen ist das Modell mit der geringsten Leistung bei der

Korrektur erkannter Fehler; die direkte Fehlerrate von 25 %, die es in der Kategorie LS-AL auf B1-Niveau aufwies, belegt, dass sich die Schwierigkeiten des Modells im lexikalisch-semantischen Bereich nicht nur auf die Erkennung beschränken, sondern sich auch im Korrekturprozess widerspiegeln.

Wie viele der korrekt korrigierten Fehler in schriftlichen Produkten im DaF-Unterricht können die KI-gestützten Chatbots ChatGPT, DeepSeek und Gemini mit angemessenem Feedback versehen?

Die Frage stellt die komplexeste kognitive Aufgabe (P3) dieser Untersuchung in den Mittelpunkt und bewertet die pädagogischen Kompetenzen der Modelle. Die Antwort auf diese Frage erfordert es, über die Metrik GBF der Phase 3 (d. h. die Fehlerfortpflanzung aus Erkennungsfehlern der Phase 1 oder Korrekturfehlern der Phase 2) hinauszugehen und sich ausschließlich auf die erst in Phase 3 neu generierten Fehler zu konzentrieren, d. h. auf die Metrik *GF* (direkter Feedback-Fehler).

Die gewonnenen kumulativen (Gesamtsumme) und vergleichenden Daten belegen eindeutig, dass diese Phase ein methodisch von Phase 1 (Erkennung) und Phase 2 (Korrektur) unabhängiger Bereich ist, der seine eigenen spezifischen Herausforderungen birgt. Die Tatsache, dass ein Modell einen Fehler mit 100 % Erfolg erkennt und mit 100 % Erfolg korrigiert, garantiert nicht, dass es auch 100 % korrektes Feedback zu diesem Fehler geben wird. Der deutlichste Beleg für dieses Phänomen ist die Leistung des Gemini-Modells in der Kategorie GR-AW (Grammatik - Auswahl) im B1-Datensatz: Obwohl das Modell die Fehler in dieser Kategorie zu 100 % erkannte und zu 100 % korrigierte, machte es in Phase 3 (Feedback-Generierung) 15 % (3 Fälle) direkte Feedback-Fehler. Dieser Befund zeigt, dass die Fähigkeit, Erklärungen zu generieren, ein vom Lösungsfindungsprozess getrennter kognitiver Vorgang ist.

Bei der Untersuchung der Raten, mit denen die Modelle zu erkannten und korrekt korrigierten Fehlern direkt falsches Feedback geben, zeigt sich, dass das ChatGPT-Modell

in diesem Bereich die deutlichsten Schwierigkeiten aufweist. Den kumulativen Daten zufolge konzentrieren sich die direkten Feedback-Fehlerraten von ChatGPT insbesondere auf die Grammatik-Kategorien (GR). Das Modell weist hohe direkte Fehlerraten in der Kategorie GR-AL (Grammatik - Auslassung) mit 12 % (7 Fälle) und in der Kategorie GR-AW mit 10 % (6 Fälle) auf. Diese Schwäche wird durch den direkten Feedback-Fehler von 20 % (4 Fälle) in der Kategorie GR-AW des A2-Datensatzes (Quelle 15) bestätigt. Dies deutet darauf hin, dass ChatGPT zwar kompetent bei der Korrektur von Grammatikfehlern ist, jedoch eine strukturelle Schwierigkeit aufweist, den Grund für diese Korrekturen zu erklären.

Die Modelle DeepSeek und Gemini zeigen bei direkten Feedback-Fehlern zwar eine bessere Leistung als ChatGPT, weisen aber ebenfalls spezifische Schwachstellen auf. Die kumulativen direkten Feedback-Fehler des DeepSeek-Modells konzentrieren sich auf die Kategorien LS-AL (8 %, 5 Fälle) und GR-HF (7 %, 4 Fälle). Die größte Stärke von DeepSeek in diesem Bereich ist die beste Leistung in der Kategorie GR-AL, in der es nur 3 % direkte Fehler machte, verglichen mit 12 % bei ChatGPT. Das Gemini-Modell erlebt im Vergleich zu seinem nahezu fehlerfreien Profil bei direkten Fehlern (nahe 0 %) in Phase 2 (Korrektur) einen sichtbaren Leistungsabfall in Phase 3 (Feedback-Generierung). Die kumulativen direkten Feedback-Fehler des Modells konzentrieren sich auf die Kategorien GR-AL (8 %, 5 Fälle) und GR-AW (7 %, 4 Fälle). Dieser Umstand zeigt deutlich, dass die Leistungsfähigkeit der Korrektur- und Erklärungsmechanismen bei Gemini auseinanderfällt.

Zusammenfassend lautet die Antwort auf die Frage, *zu wie vielen erkannten und korrekt korrigierten Fehlern korrektes Feedback gegeben werden kann*, nicht zu allen, und diese Rate variiert je nach Fehlerkategorie und Modell. Alle drei Modelle machen direkte Feedback-Fehler, indem sie in Phase 3 falsche Erklärungen zu Problemen generieren, die sie in Phase 2 erfolgreich gelöst haben. Bei dieser anspruchsvollen Aufgabe ist ChatGPT das Modell, das mit Raten von 10 % bis 12 % die größten Schwierigkeiten bei Grammatikerklärungen hat. DeepSeek und Gemini sind zwar insgesamt erfolgreicher,

belegen aber durch direkte Feedback-Fehler von 7 %-8 % in spezifischen Kategorien wie GR-AL, GR-AW oder LS-AL, dass der Akt der Erklärung für die aktuellen KI-Modelle einen komplexeren und vom Akt der Korrektur unabhängigen Schwierigkeitsbereich darstellt.

Inwieweit können die KI-gestützten Chatbots ChatGPT, DeepSeek und Gemini erfolgreich eingesetzt werden, um die Schreibfertigkeit im DaF-Unterricht zu fördern?

Die Antwort auf die Frage basiert auf einer kumulativen und vergleichenden Datensynthese der Tabellen, die im Abschnitt *Befunde und Interpretationen* dieser Untersuchung vorgestellt wurden. Die gewonnenen Daten zeigen, dass diese Unterstützung nicht eindimensional oder absolut ist; im Gegenteil, sie offenbaren, dass sich die als Unterstützung definierte Handlung radikal verändert, je nachdem, ob es sich um

1. Bewusstseinsbildung (Phase 1: Erkennung),
2. Bereitstellung des korrekten Modells (Phase 2: Korrektur) oder
3. Bereitstellung pädagogischer Erklärungen (Phase 3: Feedback-Generierung) handelt.

Bei einer Gesamtbewertung wurde festgestellt, dass alle drei Modelle eine starke Unterstützung bei der Bewältigung von regelbasierten, strukturellen Grammatikfehlern bieten; jedoch in komplexen lexikalisch-semantischen Aufgaben, die kontextuelle Nuancen, Wortwahl und pädagogische Erklärungen erfordern, ein begrenztes, variables und bisweilen riskantes Unterstützungsprofil aufweisen. Das Potenzial der Modelle zur Unterstützung der DaF-Schreibkompetenz wird in der detaillierten Analyse dieser drei Phasen deutlich.

Der erste Schritt bei der Entwicklung der Schreibkompetenz eines Lernenden ist das Erkennen seines Fehlers. Diese Unterstützung der Modelle bei der Bewusstseinsbildung variiert stark je nach Fehlertyp.

Die Modelle bieten eine außerordentliche Unterstützung bei der Erkennung von Grammatikfehlern. In Kategorien wie GR-HF (Grammatik - Hinzufügung) und GR-AW

(Grammatik - Auswahl) liegen die kumulativen Erkennungsraten im Bereich von 98 %-100 %. Dies zeigt, dass sich Lernende in hohem Maße auf die Modelle verlassen können, um regelbasierte Grammatikfehler (z. B. falsche Präposition, überflüssiges Wort) zu erkennen.

Das Unterstützungsniveau sinkt im lexikalisch-semantischen Bereich deutlich. Das größte Versagen der Modelle bei der Bewusstseinsbildung tritt in der Kategorie LS-AW (Lexikalisch/Semantisch - Auswahl) auf, der anspruchsvollsten Kategorie für alle Modelle. Kumulativ kann ChatGPT 30 % dieser Fehler nicht erkennen, Gemini 20 % und DeepSeek 18 %. Dieser Befund bedeutet, dass das Risiko, dass die KI einen Fehler übersieht und somit verhindert, dass der Lernende seinen Fehler erkennt, wenn er eine kontextuell falsche Wortwahl trifft (z.B. die Verwendung von machen statt tun), zwischen 18 % und 30 % liegt. In ähnlicher Weise bot DeepSeek (93 %) auch in der Kategorie LS-AL (Lexikalisch/Semantisch - Auslassung) die höchste Unterstützung bei der Bewusstseinsbildung, während Gemini (88 %) in diesem Bereich die schwächste Erkennung aufwies.

Bei der Entwicklung der Schreibkompetenz ist die kritischste Unterstützung nach der Fehlererkennung die Bereitstellung der korrekten Form oder des korrekten Modells für den Lernenden. Diese Phase misst die Zuverlässigkeit der Modelle und offenbart die pädagogisch wichtigsten Unterschiede. Diese Analyse geht davon aus, dass das Anbieten einer falschen Korrektur für einen erkannten Fehler (direkter Korrekturfehler) der gefährlichste pädagogische Fehler ist, da er das Risiko einer negativen Verstärkung birgt.

Die Daten belegen eindeutig, dass das Gemini-Modell die zuverlässigste Unterstützung bei der Bereitstellung des korrekten Modells (P2) bietet. Die kumulative direkte Korrekturfehlerrate des Modells liegt bei insgesamt nur 2 Fällen (0,4 %) und ist selbst in anspruchsvollen lexikalisch-semantischen Kategorien wie LS-AL (0 %), LS-HF (0 %) und LS-UM (0 %) fehlerfrei. Dies bedeutet, dass die Unterstützung durch Gemini auf die

Erkennung begrenzt ist; sobald es einen Fehler erkennen kann, ist die dem Lernenden angebotene Korrektur mit nahezu 100%iger Genauigkeit (und somit pädagogisch sicher).

Im Gegensatz dazu ist die Korrekturunterstützung von ChatGPT und DeepSeek riskant und unausgewogen. ChatGPT wies insbesondere bei der Korrektur von LS-AL-Fehlern auf B1-Niveau eine direkte Fehlerrate von 25 % (5 Fälle) auf. DeepSeek zeigte bei der Korrektur von LS-HF-Fehlern auf B1-Niveau eine direkte Fehlerrate von 30 % (6 Fälle). Diese Befunde deuten auf ein kritisches pädagogisches Risiko hin: In ihren spezifischen Schwachstellenbereichen bergen diese Modelle das Risiko, den Fehler des Lernenden zu verfestigen, indem sie zu dem erkannten Fehler eine falsche Korrektur anbieten (d.h. falsch korrigieren).

Die Entwicklung der Schreibkompetenz erfordert nicht nur die Korrektur des Fehlers, sondern auch das Verständnis, warum dieser Fehler gemacht wurde. Diese pädagogische Unterstützung (P3) ist der Bereich, in dem alle drei Modelle die größten Schwierigkeiten haben und unabhängig von ihren Kompetenzen in Phase 2 (Korrektur) scheitern.

Die Daten belegen, dass eine 100%ig korrekte Erkennung und Korrektur eines Fehlers nicht garantieren, dass auch eine 100%ig korrekte Erklärung für diesen Fehler geliefert werden kann. (z. B. Gemini, B1 GR-AW: 100 % Erkennung, 100 % Korrektur, aber 15 % direkte Feedback-Fehler).

Bei dieser pädagogischen Erklärungsaufgabe bietet ChatGPT die schwächste Unterstützung. Das Modell machte selbst in grundlegenden Grammatikkategorien wie GR-AL (12 %) und GR-AW (10 %) hohe Raten an direkten Feedback-Fehlern (Quelle 13, 15). Paradoxerweise zeigt dies, dass ChatGPT zwar ein Experte bei der Erkennung von Grammatikfehlern ist, jedoch das unzuverlässigste Modell bei der Erklärung dieser Fehler ist. Obwohl DeepSeek und Gemini in diesem Bereich eine bessere Leistung zeigten, machten sie dennoch direkte Feedback-Fehler in Kategorien wie GR-AL (bis zu 8 %) oder GR-HF (7 %) (Quelle 13, 14), was zeigt, dass diese Unterstützung noch nicht vollständig zuverlässig ist.

Schließlich umfasst der Begriff der Unterstützung auch das Verhaltensprofil des Modells; das heißt, wie oft es unnötige oder nicht-korrektive (z. B. stilistische) Eingriffe vornimmt.

DeepSeek ist mit insgesamt 44 Fällen das Modell, das kumulativ am häufigsten diese Art von Feedback generiert. Dies zeigt, dass DeepSeek die interventionistischste Unterstützung für Lernendentexte bietet, wobei die Neigung zu stilistischen oder alternativen Vorschlägen insbesondere in den Kategorien GR-AW (15 Fälle) und LS-AW (14 Fälle) am höchsten ist.

Gemini ist mit insgesamt 25 Fällen das Modell, das diese Art von Feedback am seltensten generiert. Dies deutet darauf hin, dass Gemini eine konservativere Unterstützung bietet und dazu neigt, seine Eingriffe weitgehend auf eindeutige Fehler zu beschränken.

Zusammenfassend lässt sich sagen, dass die Antwort auf die Frage: *Inwieweit unterstützen KI-Modelle die DaF-Schreibkompetenz?* von den Fragen *Welche Art von Unterstützung?* und *Für welchen Fehlertyp?* abhängt.

Im Bereich der Strukturellen/Regelbasierten Unterstützung (Grammatik) ist die Unterstützung stark. Die Modelle schaffen Bewusstsein (P1), indem sie Grammatikfehler mit einer hohen Rate (93 %-100 %) erkennen. Gemini bietet für diese Fehler ein korrektes Modell (P2) mit nahezu 100%iger Genauigkeit.

Im Bereich der Kontextuellen/Semantischen Unterstützung (Lexikalisch/Semantisch) ist die Unterstützung begrenzt und riskant. Die Unterstützung bei der Bewusstseinsbildung (P1) sinkt in der Kategorie LS-AW auf 70 %-82 %. Die Unterstützung bei der Bereitstellung des korrekten Modells (P2) birgt für ChatGPT (LS-AL 25 % Fehler) und DeepSeek (LS-HF 30 % Fehler) auf B1-Niveau das Risiko, aktiv irreführend zu sein.

Die Pädagogische/Erklärende Unterstützung (P3) ist der schwächste Unterstützungsbereich der Modelle. Alle drei Modelle neigen dazu (insbesondere ChatGPT), direkte Feedback-Fehler zu machen, während sie Fehler erklären, die sie korrekt korrigiert haben.

Im Lichte der kumulativen Daten kann interpretiert werden, dass diese Modelle derzeit eine starke Unterstützung als Grammatikprüfer oder Korrekturleser bieten; jedoch als pädagogische Tutoren oder Erklärer eine schwache und entwicklungsbedürftige Unterstützung leisten. Für einen Benutzer, der die zuverlässigste Korrekturunterstützung (P2) sucht, ist Gemini die beste Option; für einen Benutzer, der jedoch die umfassendste Erkennungsunterstützung (P1) sucht, insbesondere im LS-Bereich, bietet DeepSeek eine bessere Leistung.

Gesamtbewertung

Die kumulative Synthese der Daten aus allen in der Untersuchung verwendeten Tabellen liefert eine ganzheitliche Bewertung der Fähigkeiten der drei untersuchten KI-Modelle (ChatGPT, DeepSeek und Gemini) im Kontext der Fehlererkennung, Korrektur und Feedback-Generierung. Die umfassende Analyse zeigt, dass kein Modell in allen Kategorien eine absolute Überlegenheit aufweist; stattdessen offenbart jedes Modell eigene, methodisch differenzierte Stärken und Schwächen.

Den gewonnenen Daten zufolge werden die grundlegenden Leistungsprofile der Modelle wie folgt zusammengefasst: Die deutlichste Stärke des Gemini-Modells liegt in seiner Fähigkeit zur Phase 2 (Korrektur). Die kumulativen Daten zeigen, dass die direkte Fehlerrate des Modells bei der Korrektur erkannter Fehler zwischen 0 % und 2 %, also auf einem Niveau von nahezu null, liegt. Der Leistungsverlust bei diesem Modell ist fast ausschließlich auf Fehlschläge in Phase 1 (Erkennung) und die Widerspiegelung dieser Fehler in den nachfolgenden Phasen durch Fehlerfortpflanzung zurückzuführen. DeepSeek sticht als das Modell mit der ausgewogensten Gesamtleistung hervor. Insbesondere in anspruchsvollen lexikalisch-semantischen Kategorien (z. B. LS-AW) zeigte es eine deutlich höhere Erfolgsquote als ChatGPT. Es weist jedoch mit einer hohen direkten Fehlerrate von 12 % in der Phase der Korrektur (P2) der Kategorie LS-HF eine spezifische und deutliche Schwäche

auf. ChatGPT hingegen zeigt eine außerordentliche Expertise bei Grammatik-Aufgaben (GR); das Modell erreicht in den Kategorien GR-HF und GR-AW eine fehlerfreie Erkennungsrate von 100 %. Die deutlichste Schwäche des Modells tritt im Bereich LS-AW (Lexikalisch/Semantisch - Auswahl) zutage, der anspruchsvollsten Kategorie für alle Modelle; ChatGPT blieb in dieser Kategorie sowohl in der Erkennungs- (70 %) als auch in der Korrektur- (68 %) und Feedback-Phase (60 %) deutlich hinter den beiden anderen Modellen zurück.

Allgemein wurde bestätigt, dass die Leistung mit zunehmender Aufgabenkomplexität sinkt und sich die Aufgabenschwierigkeit wie folgt staffelt: Phase 1 (Erkennung) < Phase 2 (Korrektur) < Phase 3 (Feedback-Generierung). Während LS-AW als die anspruchsvollste Fehlerkategorie für alle Modelle identifiziert wurde, wurde deutlich, dass das B1-Niveau der Datensatz ist, der die spezifischen Schwächen der Modelle am deutlichsten offenlegt.

Die kumulative Leistung der Modelle zeigt einen konsistenten Rückgang, während die kognitive Komplexität der Aufgabe zunimmt (von der Erkennung zur Korrektur und zum Feedback).

Phase 1: Fehlererkennung Diese Phase ist der grundlegende Kompetenzbereich, in dem die höchsten Erfolgsquoten erzielt wurden. Bei der Erkennung von Grammatikfehlern (GR) variieren die kumulativen Erfolgsdurchschnitte zwischen 93 % und 100 %. Die eigentliche Herausforderung für die Modelle beginnt bei den Lexikalisch/Semantischen (LS) Kategorien. In diesem Kontext sticht die Erkennung von LS-AW mit Erfolgsquoten zwischen 70 % (ChatGPT) und 82 % (DeepSeek) als die anspruchsvollste Erkennungsaufgabe hervor.

Phase 2: Fehlerkorrektur Die Erfolgsquoten in dieser Phase sind aufgrund des methodischen Einflusses der Metrik *GBF* zwangsläufig niedriger als in Phase 1. Jeder in Phase 1 nicht erkannte Fehler (z. B. 18 Fälle in der Kategorie LS-AW bei ChatGPT) spiegelt sich in dieser Phase direkt als Fehlerfortpflanzung (bedingtes Scheitern) wider. Der grundlegende

qualitative Unterschied zwischen den Modellen tritt in den Raten des *GF* (direkter Korrekturfehler) zutage, die sie in dieser Phase aufweisen.

Phase 3: Feedback-Generierung Diese Phase wurde für alle drei Modelle als die anspruchsvollste Aufgabe identifiziert, und hier wurden die niedrigsten Gesamterfolgsquoten verzeichnet. Die Daten zeigen, dass die Modelle, selbst wenn sie einen Fehler erfolgreich erkennen und korrigieren (in Phase 2 erfolgreich sind), bei der Erklärung dieser Korrektur (P3) direkt scheitern können (gesamt fehlgeschlagen). Dass ChatGPT beispielsweise in der Kategorie GR-AL 12 % und in der Kategorie GR-AW 10 % direkte Feedback-Fehler machte, belegt, dass diese Phase einen unabhängigen Schwierigkeitsbereich darstellt.

Die Fehlerkategorien weisen eine klare Hierarchie hinsichtlich des Schwierigkeitsgrads auf:

- Anspruchsvollste Kategorie: LS-AW (Lexikalisch/Semantisch - Auswahl) Diese Kategorie ist für alle Modelle mit Abstand der Bereich, in dem die größte Herausforderung und die niedrigste Leistung verzeichnet wurden.
    - ChatGPT: Erkennung 70 %, Korrektur 68 %, Feedback 60 %.
    - DeepSeek: Erkennung 82 %, Korrektur 80 %, Feedback 80 %.
    - Gemini: Erkennung 80 %, Korrektur 80 %, Feedback 77 %.
- Kategorie mit hoher Schwierigkeit: LS-AL (Lexikalisch/Semantisch - Auslassung) Diese Kategorie stellt insbesondere für ChatGPT ein anspruchsvolles Profil dar. ChatGPT machte bei der Korrektur (P2) dieser Kategorie im B1-Datensatz 25 % (5 Fälle) sehr hohe direkte Korrekturfehler (Gesamt fehlgeschlagen). Im kumulativen Durchschnitt wies ChatGPT (in der Feedback-Phase) mit 75 % die niedrigste Leistung auf.
- Spezifische Schwierigkeit: LS-HF (Lexikalisch/Semantisch - Hinzufügung) Diese Kategorie sticht als eine der am leichtesten zu erkennenden LS-Kategorien hervor

(98 %-100 % Erkennungserfolg). Gleichzeitig stellt diese Kategorie jedoch für DeepSeek einen spezifischen und methodisch bedeutsamen Schwachstellenbereich dar. DeepSeek wies in dieser Kategorie einen Durchschnitt von 12 % an direkten Korrekturfehlern (GF) auf.

- Einfachste Kategorien: Grammatik (GR) Die Grammatikkategorien sind die Bereiche, in denen die Modelle die höchste Kompetenz zeigen. Die Kategorien GR-UM, GR-HF und GR-AW wurden mit kumulativen Erfolgsdurchschnitten im Bereich von 97 %-100 % als die einfachsten Aufgaben identifiziert. GR-AL (92 %-93 % Korrekturerfolg) ist die relativ anspruchsvollste unter den Grammatikkategorien.

ChatGPT: Die deutlichste Stärke des Modells ChatGPT ist seine Meisterschaft bei Grammatikaufgaben (GR). Es zeigte in den Kategorien GR-HF und GR-AW einen kumulativen Erkennungserfolg von 100 % und in der Kategorie GR-UM von 98 %. Auch in der Kategorie LS-HF blieb es mit 98 % Korrekturerfolg (P2) stark. Die Schwäche des Modells ist seine konsistent niedrige Leistung bei Lexikalisch/Semantischen (LS) Aufgaben. Im Bereich LS-AW, der anspruchsvollsten Kategorie, blieb es im Bereich von 60 %-70 % und lag damit 10-20 Prozentpunkte hinter seinen Konkurrenten. Darüber hinaus weist es in der Kategorie LS-AL in allen Phasen den niedrigsten kumulativen Durchschnitt auf. Das Fehlerprofil ist vielfältig. Das Modell weist sowohl hohe Raten an *GBF*(Erkennungsfehler) (z. B. LS-AW 30 %-32 %) als auch hohe Raten an *gesamt fehlgeschlagen* (direkter Feedback-Fehler) auf (z. B. GR-AL 12 %).

DeepSeek: Das Modell zeigt generell eine hohe und ausgewogene Leistung. In den anspruchsvollen Kategorien LS-AW (80 %-82 %) und LS-AL (92 %-93 %) hat es eine deutlich bessere Leistung als ChatGPT gezeigt. Es weist die besten kumulativen Durchschnittswerte beim GR-AL-Feedback (90 %) und bei der LS-AL-Erkennung (93 %) auf. Die Kategorie LS-HF ist die methodisch deutlichste Schwäche des Modells. Obwohl es bei der Erkennung zu 100 % erfolgreich war, wies es in Phase 2 (Korrektur) eine hohe direkte

Fehlerrate von 12 % (7 Fälle) auf. Es wurde festgestellt, dass diese Schwäche auf den spezifischen Fehler von 30 % (6 Fälle) im B1-Set zurückzuführen ist. Die Fehler des Modells resultieren in der Regel entweder aus der Erkennungsphase (Gesamt bedingt fehlgeschlagen) oder aus diesem spezifischen LS-HF-Korrekturfehler (Gesamt fehlgeschlagen).

Gemini: Die stärkste und entscheidendste Seite des Modells ist seine Leistung in Phase 2 (Korrektur). Die kumulative Rate an *gesamt fehlgeschlagen* (direkter Korrekturfehler) liegt bei nahezu null. In den Kategorien LS-HF (0 %), LS-AL (0 %) und LS-UM (0 %) machte es im Durchschnitt keinerlei direkte Korrekturfehler. Darüber hinaus wies es in der Kategorie GR-UM in allen Phasen einen fehlerfreien kumulativen Durchschnitt von 100 % auf. Die einzige Schwachstelle des Modells ist die Phase 1 (Erkennung). Der Hauptfaktor, der seine Leistung mindert, ist die Unfähigkeit, den Fehler von vornherein in Kategorien wie LS-AW (20 % Erkennungsfehler), LS-AL (12 % Erkennungsfehler) und GR-AL (8 % Erkennungsfehler) zu erkennen. Das Fehlerprofil des Modells wird fast ausschließlich durch die Metrik *GBF* erklärt. Dies zeigt, dass Gemini einen erkannten Fehler fast sicher korrigiert, die Wahrscheinlichkeit, den Fehler nicht zu erkennen, jedoch im Vergleich zu den anderen Modellen relativ höher sein kann.

Während die Datensätze A1 und A2 für die Modelle ähnliche und relativ hohe Erfolgsquoten boten, sticht der B1-Datensatz als der mit Abstand anspruchsvollste Set für alle Modelle hervor. Der B1-Datensatz war das Niveau, das die oben definierten spezifischen und strukturellen Schwächen der Modelle auslöste:

- ChatGPT fiel im B1-Set bei der LS-AL-Korrektur auf 65 % Erfolg (25 % direkte Fehler) ab und fiel bei der LS-AW-Erkennung auf 55 % zurück (45 % Fehlschlag).
- DeepSeek fiel im B1-Set bei der LS-HF-Korrektur auf 70 % Erfolg ab (30 % direkte Fehler).
- Gemini verzeichnete im B1-Set in den Kategorien GR-AL (80 %) und LS-AL (80 %) die niedrigsten Erkennungsraten.

Analyse der Phase 4 (Nicht-korrektives Feedback): Auch die Nutzungshäufigkeit (Frequenz) dieser Feedback-Art zeigt deutliche Verhaltensunterschiede zwischen den Modellen. DeepSeek ist mit insgesamt 44 Fällen das Modell, das diese Feedback-Art kumulativ am häufigsten generiert. Gemini ist mit insgesamt 25 Fällen das Modell, das diese Art von Feedback kumulativ am seltensten generiert. GR-AW ist mit insgesamt 36 Fällen die Kategorie, die über alle drei Modelle hinweg am häufigsten mit diesem Feedback in Verbindung gebracht wird. An zweiter Stelle folgt die anspruchsvollste Kategorie LS-AW mit 23 Fällen. LS-AL (4 Fälle), GR-AL (5 Fälle) und GR-HF (5 Fälle) sind die Bereiche, in denen diese Art von Feedback kumulativ am seltensten auftrat.

Die umfassende vergleichende Analyse legt eindeutig dar, dass die drei untersuchten KI-Modelle unterschiedliche Kompetenzprofile im Fehlermanagement aufweisen. Kein Modell ist in allen Bereichen dominant; stattdessen weist jedes Modell ein methodisch unterscheidbares Merkmal auf.

Gemini sticht als Korrekturexperte hervor; die Leistung des Modells ist fast ausschließlich auf seine Erkennungsfähigkeit (P1) beschränkt. Sobald es einen Fehler erkennt, ist sein Korrekturerfolg (P2) nahezu fehlerfrei (0 % direkte Fehler). ChatGPT zeigt ein deutliches Dilemma: Während das Modell eine hohe Expertise in Grammatik-Kategorien (GR) aufweist, zeigt es eine ernste und konsistente Schwäche in lexikalisch-semantischen (LS) Kategorien (insbesondere LS-AW). DeepSeek bietet zwischen diesen beiden Modellen das ausgewogenste Profil; es zeigte in den meisten anspruchsvollen Kategorien eine hohe Leistung, hat aber mit einer direkten Fehlerrate von 30 %, ausgelöst durch das B1-Set bei der LS-HF-Korrektur, gezeigt, dass es einen spezifischen und unvorhersehbaren Schwachstellenbereich aufweist.

Im Lichte der gewonnenen Daten wurde bestätigt, dass die Aufgabenschwierigkeit sowohl zwischen den Phasen (Erkennung < Korrektur < Feedback) als auch zwischen den

Kategorien (Grammatik < Lexikalisch/Semantisch) zunimmt. Es wurde festgestellt, dass die Leistungsverluste aus zwei Hauptquellen stammen:

1. Die Widerspiegelung von Erkennungsfehlern aus Phase 1 in den nachfolgenden Phasen (Fehlerfortpflanzung, das Hauptproblem von Gemini) und
2. Direktes Scheitern in Phase 2 oder 3, selbst bei erfolgreicher Erkennung (die spezifische Schwäche von ChatGPT und DeepSeek).

Abschließend fungierte der B1-Datensatz als der effektivste Stresstest, der diese strukturellen Schwächen der Modelle offenlegte.

Die umfassende Analyse der Untersuchung ergab, dass keines der drei untersuchten KI-Modelle (ChatGPT, DeepSeek und Gemini) eine absolute Überlegenheit im Fehlermanagement aufweist, sondern jedes methodologisch differenzierte Stärken und Schwächen besitzt. Gemini erwies sich zwar als *Korrekturspezialist* bei der nahezu fehlerfreien Korrektur (P2) der von ihm erkannten Fehler (0–2 % direkte Fehlerquote), seine grundlegende Schwäche lag jedoch in der initialen Fehlererkennung (P1), insbesondere in den lexikalisch-semantischen Kategorien (LS-AW 20 %, LS-AL 12 %). Diese Schwäche wirkte sich als *Fehlerpropagation (bedingter Fehlschlag)* auf die nachfolgenden Phasen aus. DeepSeek bot das ausgewogenste Profil und zeigte die stärkste Erkennungsfähigkeit (P1) insbesondere in den anspruchsvollen lexikalisch-semantischen Kategorien (LS-AL 93 %, LS-AW 82 %), wies jedoch auf dem B1-Niveau in der Kategorie LS-HF eine spezifische und hochriskante Korrekturschwäche (P2) auf (30 % direkte Korrekturfehler). ChatGPT wiederum zeigte als *Grammatikspezialist* (100 % Erkennung bei GR-HF/GR-AW) ein deutliches Dilemma: Obwohl es grammatikalische Aufgaben meisterhaft bewältigte, erbrachte es im lexikalisch-semantischen Bereich die schwächste Leistung und blieb in der für alle Modelle anspruchsvollsten Kategorie LS-AW (60–70 % Erfolgsquote) sowie bei LS-AL konsistent hinter den Konkurrenten zurück. Darüber hinaus zeigte ChatGPT insbesondere bei Grammatikerklärungen (bis zu 12 % direkte Fehler) die geringste Fähigkeit

zu pädagogischem Feedback (P3). Folglich nahm die Leistung aller Modelle mit steigender Aufgabenkomplexität (Erkennung < Korrektur < Feedback) und Fehlerkomplexität (Grammatik < Lexikalisch-Semantisch) ab, wobei das B1-Niveau sich als der effektivste Stresstest erwies, der diese strukturellen Schwächen auslöste. Diese Befunde bestätigen, dass KI-Modelle aktuell als starke *Grammatikprüfer* Unterstützung bieten können, jedoch noch keine zuverlässigen *pädagogischen Tutoren* sind.

LITERATURVERZICHNIS

Abbott, G. (1980). Towards a more rigorous analysis of foreign language errors. *International Review of Applied Linguistics in Language Teaching, 18*, 121-134. https://doi.org/10.1515/iral.1980.18.1-4.121

Adeshola, I., & Adepoju, A. P. (2023). The opportunities and challenges of ChatGPT in education. *Interactive Learning Environments, 32*(10), 6159-6172. https://doi.org/10.1080/10494820.2023.2253858

Alier, M., García-Peñalvo, F. J., & Camba, J. D. (2024). Generative artificial intelligence in education: From deceptive to disruptive. *International Journal of Interactive Multimedia and Artificial Intelligence*, *9*(2), 7–16. https://doi.org/10.9781/ijimai.2024.02.011

Alkaissi, H., & McFarlane, S. I. (2023). Artificial hallucinations in chatgpt: Implications in scientific writing. *Cureus, 15*(2), e35179. DOI: 10.7759/cureus.35179

Altıparmak Yılmaz, H. M., & Demir, N. (2020). Error analysis: Approaches to written texts of turks living in the sydney. *International Education Studies,13* (2), 104-114. DOI:10.5539/ies.v13n2p104

Antaki, F., Touma, S., Milad, D., El-Khoury, J., & Duval, R. (2023). Evaluating the performance of chatgpt in ophthalmology: An analysis of its successes and shortcomings. *Ophthalmology Science, 3*(3), 100324. https://doi.org/10.1016/j.xops.2023.100324

Arslan, K. (2020). Eğitimde yapay zeka ve uygulamaları. *Batı Anadolu Eğitim Bilimleri Dergisi, 11*(1), 71-88.

Artificial Intelligence. (2024, April 10). In Encyclopedia Britannica online. Abgerufen von https://www.britannica.com/technology/artificial-intelligence

Atlas, S. (2023). *ChatGPT for higher education and professional development: A guide to conversational AI*. DigitalCommons@URI.

Attila, A. Ş. (2022). *Uygulamalı örneklerle yapay zekâ algoritmalari ve programlama*. Ankara: Seçkin.

Aydemir, E. (2019). *Weka ile yapay zeka*. Ankara: Seçkin.

Baidoo-Anu, D., & Owusu Ansah, L. (2023). Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of chatgpt in promoting teaching and learning. *Journal of AI, 7*(1), 52-62. https://doi.org/10.61969/jai.1337500

Bitchener, J. (2008). Evidence in support of written corrective feedback. *Journal of Second Language Writing*, *17*(2), 102–118. https://doi.org/10.1016/j.jslw.2007.11.004

Bitchener, J., Young, S., & Cameron, D. (2005). The effect of different types of corrective feedback on ESL student writing. *Journal of Second Language Writing*, *14*(3), 191–205. https://doi.org/10.1016/j.jslw.2005.08.001

Botley, S. P. (2015). Errors versus mistakes: A false dichotomy*? Malaysian Journal of ELT Research*, *11*(1), 81–94.

Bouschery, S. G., Blazevic, V., & Piller, F. T. (2023). Augmenting human innovation teams with artificial intelligence: Exploring transformer-based language models. *Journal of Product Innovation Management*, *40*(2), 139-160. https://doi.org/10.1111/jpim.12656

Brown, H. D. (2014). The type and linguistic foci of oral corrective feedback in the L2 classroom: A meta-analysis. *Language Teaching Research, 20*(4), 436-458. https://doi.org/10.1177/1362168814563200

Brown, H. D. (2007). *Principles of language learning and teaching*. New York: Pearson

Bruton, A. (2009). Improving accuracy is not the only reason for writing, and even if it were. . . *System, 37*(4), 600–613. https://doi.org/10.1016/j.system.2009.09.005

Bruton, A. (2010). Another reply to Truscott on error correction: Improved situated designs over statistics. *System, 38*(3), 491–498. https://doi.org/10.1016/j.system.2010.07.001

Burt, M. K. (1975). Error analysis in the adult EFL classroom. *TESOL Quarterly*, *9*(1), 53-63. https://doi.org/10.2307/3586012

Bryman, A. (2015). *Social research methods*. Oxford: Oxford University.

Cao, Y., Li, S., Liu, Y., Yan, Z., Dai, Y., Yu, P. S., & Sun, L. (2023). A comprehensive survey of AI-generated content (AIGC): A history of generative AI from GAN to ChatGPT. *ArXiv, abs/2303.04226.* https://doi.org/10.48550/arXiv.2303.04226

Carlà, M. M., Gambini, G., Baldascino, A., Boselli, F., Giannuzzi, F., Margollicci, F., & Rizzo, S. (2024). Large language models as assistance for glaucoma surgical cases: a ChatGPT vs. Google Gemini comparison. *Graefe's Archive for Clinical and Experimental Ophthalmology, 262*, 2945–2959. https://doi.org/10.1007/s00417-024-06470-5

Chassignol, M., Khoroshavin, A., Klimova, A., & Bilyatdinova, A. (2018). Artificial intelligence trends in education: A narrative overview. *Procedia Computer Science, 136*, 16–24. https://doi.org/10.1016/j.procs.2018.08.233

Chomsky, N. (1965). *Aspects of the theory of syntax.* Cambridge, MA: MIT.

Choudhary, T. (2025). Political bias in large language models: A comparative analysis of ChatGPT-4, Perplexity, Google Gemini, and Claude. *IEEE Access, 13*, 11341-11379, 2025. doi: 10.1109/ACCESS.2024.3523764.

Chounta, I.-A., Bardone, E., Raudsep, A., & Pedaste, M. (2022). Exploring teachers' perceptions of artificial intelligence as a tool to support their practice in Estonian K-12 education. *International Journal of Artificial Intelligence in Education*, *32*, 725–755. https://doi.org/10.1007/s40593-021-00243-5

Corder, S. P. (1967). The significance of learner's errors. *International Review of Applied Linguistics in Language Teaching, 5*(4), 161-170. https://doi.org/10.1515/iral.1967.5.1-4.161

Corder, S. P. (1975). Error analysis, interlanguage and second language acquisition. *Language Teaching, 8*, 201-218. http://dx.doi.org/10.1017/S0261444800002822

Corder, S. P. (1982). *Error analysis and interlanguage*. Oxford: Oxford University. https://doi.org/10.2307/326720

Creswell, J. W., & Creswell, J. D. (2023). *Research design: Qualitative, quantitative, and mixed methods approaches*. SAGE.

Creswell, J. W., & Plano Clark, V. L. (2018). *Designing and conducting mixed methods research*. SAGE.

Crompton, H., & Burke, D. (2023). Artificial intelligence in higher education: The state of the field. *International Journal of Educational Technology in Higher Education*, *20*(1), 22. https://doi.org/10.1186/s41239-023-00392-8

DeepSeek-AI. (2024a). DeepSeek-V2: A strong, economical, and efficient mixture-of-experts language model. *ArXiv, abs/2405.04434*. https://doi.org/10.48550/arXiv.2405.04434

DeepSeek-AI. (2024b). DeepSeek-V3 technical report. *ArXiv, abs/2412.19437*. https://doi.org/10.48550/arXiv.2412.19437

DeepSeek-AI. (2025). DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *ArXiv, abs/2501.12948*. https://doi.org/10.48550/arXiv.2501.12948

Divekar, R. R., Drozdal, J., Chabot, S., Zhou, Y., Su, H., Chen, Y., … Braasch, J. (2022). Foreign language acquisition via artificial intelligence and extended reality: design and evaluation. *Computer Assisted Language Learning*, *35*(9), 2332–2360. https://doi.org/10.1080/09588221.2021.1879162

Dizon, G. (2017). Using intelligent personal assistants for second language learning: A case study of Alexa. *TESOL Journal, 8*(4), 811–830. https://doi.org/10.1002/tesj.353

Dodigovic, M. (2007). Artificial intelligence and second language learning: An efficient approach to error remediation. *Language Awareness, 16*(2), 99-113. https://doi.org/10.2167/la416.0

Dulay, H., Burt, M., & Krashen, S. (1982). *Language two*. Oxford: Oxford University. https://doi.org/10.2307/327086

Ellis, R. (1994). *The study of second language acquisition*. Oxford: Oxford University.

Ellis, R. (2008). A typology of written corrective feedback types. *ELT Journal*, *63*(2), 97–107. https://doi.org/10.1093/elt/ccn023

Ellis, R. (2009). Corrective feedback and teacher development. *L2 Journal*, *1*(1), 3–18. https://doi.org/10.5070/l2.v1i1.9054

Ellis, R., Loewen, S., & Erlam, R. (2006). Implicit and explicit corrective feedback and the acquisition of L2 grammar. *Studies in Second Language Acquisition*, *28*(2), 339–368. https://doi.org/10.1017/S0272263106060141

Ellis, R., Sheen, Y., Murakami, M., & Takashima, H. (2008). The effects of focused and unfocused written corrective feedback in an English as a foreign language context. *System, 36*(3), 353–371. https://doi.org/10.1016/j.system.2008.02.001

Elmas, Ç. (2021). *Yapay zeka uygulamaları*, Ankara: Seçkin.

Erdoğan, V. (2005). Contribution of error analysis to foreign language teaching. *Mersin University Journal of the Faculty of Education, 1*(2), 261-270. https://doi.org/10.17860/efd.22900

Eryılmaz, H. E. (2023). Yapay zekâ çağında kişisel veri mahremiyeti. *UMAY Sanat Ve Sosyal Bilimler Dergisi, 1*(2), 6-25.

Evans, N. W., Hartshorn, K. J., McCollom, R. M., & Wolfersberger, M. (2010). Contextualizing corrective feedback in second language writing pedagogy. *Language Teaching Research, 14*(4), 445–463. https://doi.org/10.1177/1362168810375367

Eysenbach, G. (2023). The role of ChatGPT, generative language models, and artificial intelligence in medical education: A conversation with ChatGPT and a call for papers. *JMIR Medical Education*, *9*, e46885. doi:10.2196/46885

Field, A. (2024). *Discovering statistics using IBM SPSS statistics*. SAGE.

Ferris, D. (1999). The case for grammar correction in L2 writing classes: A response to Truscott (1996). *Journal of Second Language Writing, 8*(1), 1–11. https://doi.org/10.1016/S1060-3743(99)80110-6

Ferris, D., Hyland, K., & Hyland, F. (2006). Does error feedback help student writers? New evidence on the short- and long-term effects of written error correction. In *Feedback*

in Second Language Writing: Contexts and Issues (pp. 81–104). Cambridge: Cambridge University. https://doi.org/10.1017/CBO9781139524742.007

Ferris, D. R. (2010). Second language writing research and written corrective feedback in SLA: Intersections and practical applications. *Studies in Second Language Acquisition, 32*(2), 181–201. https://doi.org/10.1017/S0272263109990490

Ferris, D. R. (2011). *Treatment of error in second language student writing*. Michigan: The University of Michigan. https://doi.org/10.3998/mpub.2173290

Ferris, D., & Kurzer, K. (2019). Does Error Feedback Help L2 Writers? Latest Evidence on the Efficacy of Written Corrective Feedback. In K. Hyland & F. Hyland (Eds.), *Feedback in Second Language Writing (Contexts and Issues)* (pp. 106-124). Cambridge: Cambridge University. https://doi.org/10.1017/9781108635547.008

Ferris, D. R. & Roberts, B. (2001). Error feedback in L2 writing classes: How explicit does it need to be? *Journal of Second Language Writing, 10*(3), 161–184. https://doi.org/10.1016/S1060-3743(01)00039-X

Fjelland, R. (2020). Why general artificial intelligence will not be realized. *Humanit Soc Sci Commun 7*(10), 1-9. https://doi.org/10.1057/s41599-020-0494-4

Gass, S. M., & Selinker, L. (2008). *Second language acquisition: An introductory course*. New York and London: Routledge. https://doi.org/10.2307/416225

Gayed, J. M. (2025). Educators' perspective on artificial intelligence: Equity, preparedness, and development. *Cogent Education*, *12*(1), 2447169. https://doi.org/10.1080/2331186X.2024.2447169

Gayed, J. M., Carlon, M. K. J., Oriola, A. M., & Cross, J. S. (2022). Exploring an AI-based writing assistant's impact on English language learners. *Computers and Education: Artificial Intelligence*, *3*, 100055. https://doi.org/10.1016/j.caeai.2022.100055

Gemini Team, Google. (2023). Gemini: A family of highly capable multimodal models. *ArXiv, abs/2312.11805*. https://doi.org/10.48550/arXiv.2312.11805

Gemini Team, Google. (2024). Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *ArXiv, abs/2403.05530*. https://doi.org/10.48550/arXiv.2403.05530

Godwin-Jones, R. (2021). Big data and language learning: Opportunities and challenges. *Language Learning & Technology, 25*(1), 4–19. https://doi.org/10.64152/10125/44747

Godwin-Jones, R. (2022). Partnering with AI: Intelligent writing assistance and instructed language learning. *Language Learning & Technology, 26*(2), 5–24. https://doi.org/10.64152/10125/73474

Godwin-Jones, R. (2024). Distributed agency in second language learning and teaching through generative AI. *Language Learning & Technology*, *28*(2), 5–31. https://doi.org/10.64152/10125/73570

Goldstein, I., & Papert, S. (1977). Artificial intelligence, language, and the study of knowledge. *Cognitive Science*, *1*(1), 84–123. https://doi.org/10.1016/S0364-0213(77)80006-2

Goldstein, L. M. (2004). Questions and answers about teacher written commentary and student revision: Teachers and students working together. *Journal of Second Language Writing, 13*(1), 63–80. https://doi.org/10.1016/j.jslw.2004.04.006

Hair, J. F., Jr., Black, W. C., Babin, B. J., & Anderson, R. E. (2018). *Multivariate data analysis* (8. bs.). London: Cengage Learning.

Haleem, A., Javaid, M., & Singh, R. P. (2023). An era of ChatGPT as a significant futuristic support tool: A study on features, abilities, and challenges. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations, 2*, 100089. https://doi.org/10.1016/j.tbench.2023.100089

Hamet, P., & Tremblay, J. (2017). Artificial intelligence in medicine. *Metabolism Clinical and Experimental, 69,* 36 – S40. https://doi.org/10.1016/j.metabol.2017.01.011

Han, Y. (2017). Mediating and being mediated: Learner beliefs and learner engagement with written corrective feedback. *System, 69*, 133–142. https://doi.org/10.1016/j.system.2017.07.003

Han, Y., & Hyland, F. (2019). Learner engagement with written feedback: A sociocognitive perspective. In K. Hyland & F. Hyland (Eds.), *Feedback in Second Language Writing (Contexts and Issues)* (pp. 247-264). Cambridge: Cambridge University. https://doi.org/10.1017/9781108635547.015

Haristiani, N. (2019). Artificial intelligence (AI) chatbot as language learning medium: An inquiry. *Journal of Physics: Conference Series, 1387*(1), 012020. DOI 10.1088/1742-6596/1387/1/012020

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112. https://doi.org/10.3102/003465430298487

Hendrickson, J. M. (1978). Error correction in foreign language teaching: Recent theory, resarch, and practice. *The Modern Language Journal, 62*(8), 387-398. https://doi.org/10.1111/j.1540-4781.1978.tb02409.x

Hodgins, J. M. (2022). Would artificial intelligence make strategy 'less human'?. *The Journal of Intelligence, Conflict, and Warfare. 5*(1), 75-84. https://doi.org/10.21810/jicw.v5i1.4212

Hyland, K., & Hyland, F. (2001). Sugaring the pill: Praise and criticism in written feedback. *Journal of Second Language Writing, 10*(3), 185–212. https://doi.org/10.1016/S1060-3743(01)00038-8

Hyland, K., & Hyland, F. (2019). Interpersonality and teacher-written feedback. In K. Hyland & F. Hyland (Eds.), *Feedback in Second Language Writing (Contexts and Issues)* (pp. 165-183). Cambridge: Cambridge University. https://doi.org/10.1017/9781108635547.011

Imran, M., & Almusharraf, N. (2024). Google Gemini as a next generation AI educational tool: A review of emerging educational technology. *Smart Learning Environments*, *11*, 22. https://doi.org/10.1186/s40561-024-00310-z

Jabeen, A., Kazemian, B., & Shahbaz, M. (2015). The Role of Error Analysis in Teaching and Learning of Second and Foreign Language. *Education and Linguistics Research, 1*(2), 52-62. DOI:10.5296/elr.v1i1.8189

James, C. (1998). *Errors in language learning and use – Exploring error analysis*. New York: Routledge. https://doi.org/10.4324/9781315842912

Javaid, M., Haleem, A., & Singh, R. P. (2023). ChatGPT for healthcare services: An emerging stage for an innovative perspective. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations, 3*(1), 100105. https://doi.org/10.1016/j.tbench.2023.100105

Javaid, M., Haleem, A., Singh, R. P., Khan, S., & Khan, I. H. (2023). Unlocking the opportunities through ChatGPT Tool towards ameliorating the education system. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations, 3*(2), 100115. https://doi.org/10.1016/j.tbench.2023.100115

Johnson, R. B., & Christensen, L. B. (2020). *Educational research: Quantitative, qualitative, and mixed approaches*. SAGE.

Juhn, Y., & Liu, H. (2020). Natural language processing to advance EHR-based clinical research in allergy, asthma, and immunology. *Journal of Allergy and Clinical Immunology*, *145*(1), 463-469. https://doi.org/10.1016/j.jaci.2019.12.897

Kalluri, K., Kokala, A., & Parvatha, N. (2024). Performance benchmarking of generative ai models: Chatgpt-4 vs. Google gemini AI. *International Research Journal of Modernization in Engineering Technology and Science*, *06*(11), 4673-4677. DOI:10.56726/IRJMETS64283

Kannan, J., & Munday, P. (2018). New trends in second language learning and teaching through the lens of ICT, networked learning, and artificial intelligence. *Círculo de Lingüística Aplicada a la Comunicación*, *76*, 13–30. https://doi.org/10.5209/CLAC.62495

Karataş, F., Abedi, F. Y., Gunyel, F. O., Karadeniz, D., & Kuzgun, Y. (2024). Incorporating AI in foreign language education: An investigation into ChatGPT's effect on foreign language learners. *Education and Information Technologies*, *29*, 19343–19366. https://doi.org/10.1007/s10639-024-12574-6

Khansir, A. A. (2012). Error analysis and second language acquisition. *Theory and practice in language studies, 2*(5), 1027-1032. DOI:10.4304/tpls.2.5.1027-1032

Kohnke, L., Moorhouse, B. L., & Zou, D. (2023). Exploring generative artificial intelligence preparedness among university language instructors: A case study. *Computers and Education: Artificial Intelligence, 5*, 100156. https://doi.org/10.1016/j.caeai.2023.100156

Lai, C.-C., & Kritsonis, W. A. (2006). The advantages and disadvantages of computer technology in second language acquisition. *National Journal for Publishing and Mentoring Doctoral Student Research*, *3*(1), 1–6. Abgerufen von https://files.eric.ed.gov/fulltext/ED492159.pdf

Lalande, J. F. (1982). Reducing composition errors: An experiment. *Modern Language Journal, 66*, 140–149. https://doi.org/10.1111/j.1540-4781.1982.tb06973.x

Lameras, P., & Arnab, S. (2022). Power to the teachers: An exploratory review on artificial intelligence in education. *Information, 13*(1), 14. https://doi.org/10.3390/info13010014

Law, L. (2024). Application of generative artificial intelligence (GenAI) in language teaching and learning: A scoping literature review. *Computers and Education Open*, *6*, 100174. https://doi.org/10.1016/j.caeo.2024.100174

Lee, I. (2008). Understanding teachers' written feedback practices in Hong Kong secondary classrooms. *Journal of Second Language Writing, 17*(2), 69–85. https://doi.org/10.1016/j.jslw.2007.10.001

Lee, I. (2010). Writing teacher education and teacher learning: Testimonies of four EFL teachers. *Journal of Second Language Writing, 19*(3), 143–157. https://doi.org/10.1016/j.jslw.2010.05.001

Lee, I. (2017). *Classroom Writing Assessment and Feedback in L2 School Contexts*. Springer Nature Singapore. https://doi.org/10.1007/978-981-10-3924-9

Lee, T. J., Campbell, D. J., Patel, S., Hossain, A., Radfar, N., Siddiqui, E., & Gardin, J. M. (2024). Unlocking health literacy: The ultimate guide to hypertension education from ChatGPT versus Google Gemini. *Cureus, 16*(5), e59898. DOI: 10.7759/cureus.59898

Lennon, P. (1991). Error: Some problems of definition, identification, and distinction. *Applied linguistics, 12*(2), 180-196. https://doi.org/10.1093/applin/12.2.180

Li, S., & Vuono, A. (2019). Twenty-five years of research on oral and written corrective feedback in system. *System, 84*, 1–11. https://doi.org/10.1016/j.system.2019.05.006

Liang, J.-C., Hwang, G.-J., Chen, M.-R. A., & Darmawansah, D. (2021). Roles and research foci of artificial intelligence in language education: An integrated bibliographic analysis and systematic review approach. *Interactive Learning Environments, 31*(7), 4270–4296. https://doi.org/10.1080/10494820.2021.1958348

Lightbown, P. M., & Spada, N. (2013). *How languages are learned*. Oxford: Oxford University. https://doi.org/10.2307/329630

Liu, M. (2023). Application and research on foreign language teaching in the context of digital transformation. *SHS Web of Conferences*, *159*, 03025. https://doi.org/10.1051/shsconf/202315901011

Liu, M., Ren, Y., Nyagoga, L. M., Stonier, F., Wu, Z., & Yu, L. (2023). Future of education in the era of generative artificial intelligence: Consensus among Chinese scholars on

applications of ChatGPT in schools. *Future in Educational Research*, *1*(1), 72–101. https://doi.org/10.1002/fer3.10

Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., Tian, J., … Ge, B. (2023). Summary of ChatGPT-Related research and perspective towards the future of large language models. *Meta-Radiology, 1*, 100017. https://doi.org/10.1016/j.metrad.2023.100017

Lo, C. K. (2023). What is the impact of ChatGPT on education? A rapid review of the literature. *Education Sciences, 13*(4), 410. https://doi.org/10.3390/educsci13040410

Lochtman, K. (2002). Oral corrective feedback in the foreign language classroom: how it affects interaction in analytic foreign language teaching. *International Journal of Educational Research, 37*(3), 271–283. https://doi.org/10.1016/S0883-0355(03)00005-3

Lott, D. (1983). Analysing and counteracting interference errors. *ELT Journal, 37*(3), 256-261. https://doi.org/10.1093/elt/37.3.256

Lyster, R., & Ranta, L. (1997). Corrective feedback and leraner uptake, *Studies In Second Language Acquisition*, 20, 37–66.

Lyster, R., & Ranta, L. (1997). Corrective feedback and learner uptake: Negotiation of form in communicative classrooms. *Studies in Second Language Acquisition*, *19*(1), 37–66. doi:10.1017/S0272263197001034

Lyster, R., Saito, K., & Sato, M. (2013). Oral corrective feedback in second language classrooms. *Language Teaching*, *46*(1), 1–40. doi:10.1017/S0261444812000365

Malik, A. R., Pratiwi, Y., Andajani, K., Numertayasa, I. W., Suharti, S., Darwis, A., & Marzuki. (2023). Exploring artificial intelligence in academic essay: Higher education student's perspective. *International Journal of Educational Research Open, 5*, 100296. https://doi.org/10.1016/j.ijedro.2023.100296

McCharty, J. (2007). *What is artificial intelligence*. Abgerufen am 15.04.2024 von https://cse.unl.edu/~choueiry/S09-476-876/Documents/whatisai.pdf

McIntosh, T. R., Susnjak, T., Liu, T., Watters, P., Xu, D., Liu, D., & Halgamuge, M. N. (2025). From Google Gemini to OpenAI Q* (Q-Star): A survey on reshaping the generative artificial intelligence (AI) research landscape. *Technologies*, *13*, 51. https://doi.org/10.3390/technologies13020051

Nabiyev, V. (2021). *Yapay zeka*, Ankara: Seçkin.

Nicholas, H., Lightbown, P. M., & Spada, N. (2001). Recasts as Feedback to Language Learners. *Language Learning 51*(4), 719–758 https://doi.org/10.1111/0023-8333.00172

Nilsson, N. J. (2009). *Artificial intelligence – A new synthesis*. San Francisco: Morgan Kaufmann.

Owan, V. J., Abang, K. B., Idika, D. O., & Bassey, B. A. (2023). Exploring the potential of artificial intelligence tools in educational measurement and assessment. *EURASIA Journal of Mathematics, Science and Technology Education*, *19*(8), em2307. https://doi.org/10.29333/ejmste/13428

Patton, M. Q. (2015). *Qualitative research & evaluation methods: Integrating theory and practice*. SAGE.

Qiao, H., & Zhao, A. (2023). Artificial intelligence-based language learning: illuminating the impact on speaking skills and self-regulation in Chinese EFL context. *Frontiers in Psychology*, *14*, 1255594. https://doi.org/10.3389/fpsyg.2023.1255594

Qin, C., Zhang, A., Zhang, Z., Chen, J., Yasunaga, M., & Yang, D. (2023). Is ChatGPT a general-purpose natural language processing task solver?. *ArXiv preprint arXiv:2302.06476.* https://doi.org/10.48550/arXiv.2302.06476

Rahman, M. M., & Watanobe, Y. (2023). ChatGPT for education and research: Opportunities, threats, and strategies. *Applied Sciences, 13*(9), 5783. https://doi.org/10.3390/app13095783

Rassaei, E. (2015). Oral corrective feedback, foreign language anxiety and L2 development. *System, 49*, 98–109. https://doi.org/10.1016/j.system.2015.01.002

Richards, J. C. (1971). Error analysis and second language strategies. *Language Science*, *17*, 12–22.

Richards, J.C., & Schmidt, R.W. (2011). *Longman dictionary of language teaching and applied linguistics*. London: Routledge. https://doi.org/10.4324/9781315833835

Robb, T., Ross, S., & Shortreed, I. (1986). Salience of feedback on error and its effect on EFL writing quality. *TESOL Quarterly, 20*(1), 83–95. https://doi.org/10.2307/3586390

Roumeliotis, K. I., & Tselikas, N. D. (2023). ChatGPT and Open-AI models: A preliminary review. *Future Internet, 15*(6), 192. https://doi.org/10.3390/fi15060192

Rudolph, J., Tan, S., & Tan, S. (2023). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *Journal of Applied Learning and Teaching, 6*(1). https://doi.org/10.37074/jalt.2023.6.1.9

Sallam, M., Al-Mahzoum, K., Sallam, M., & Mijwil, M. M. (2025). DeepSeek: Is it the end of generative AI monopoly or the mark of the impending doomsday? *Mesopotamian Journal of Big Data*, *2025,* 26–34. https://doi.org/10.58496/MJBD/2025/002

Schmidt, T., & Strasser, T. (2022). Artificial intelligence in foreign language learning and teaching: A CALL for intelligent practice. *Anglistik: International Journal of English Studies, 33*(1), 165–184. https://doi.org/10.33675/ANGL/2022/1/14

Sebetci, Ö. (2023). Yapay zekayı kodlamak. Ankara: Seçkin.

Shan, R., Ming, Q., Hong, G., & Wu, H. (2024). Benchmarking the hallucination tendency of Google Gemini and Moonshot Kimi. *OSF, version*, *1*. https://doi.org/10.31219/osf.io/83rq9

Sheen, Y. (2010). Differential effects of oral and written corrective feedback in the ESL classroom. *Studies in Second Language Acquisition*, *32*(2), 203–234. doi:10.1017/S0272263109990507

Son, J.-B., Ružić, N. K., & Philpott, A. (2023). Artificial intelligence technologies and applications for language learning and teaching. *Journal of China Computer-Assisted Language Learning, 5*(1), 94–112. https://doi.org/10.1515/jccall-2023-0015

Song, C., & Song, Y. (2023). Enhancing academic writing skills and motivation: assessing the efficacy of ChatGPT in AI-assisted language learning for EFL students. *Frontiers in Psychology, 14*, 1260843. https://doi.org/10.3389/fpsyg.2023.1260843

Storch, N., & Wigglesworth, G. (2010). Learners' processing, uptake, and retention of corrective feedback on writing. *Studies in Second Language Acquisition, 32*(2), 303–334. doi:10.1017/S0272263109990532

Subekti, A. S. (2018). Error analysis in complex sentences written by indonesian students from the english education department. *Studies In English Language And Education, 5*(2), 185-203. https://doi.org/10.24815/siele.v5i2.10686

Sumakul, D. T. Y. G., Hamied, F. A., & Sukyadi, D. (2022). Artificial intelligence in EFL classrooms: Friend or foe? *LEARN Journal: Language Education and Acquisition Research Network*, *15*(1), 232–256. Abgerufen von https://so04.tci-thaijo.org/index.php/LEARN/article/view/256723

Taecharungroj, V. (2023). "What can chatgpt do?" Analyzing early reactions to the innovative ai chatbot on twitter. *Big Data and Cognitive Computing, 7*(1), 35. https://doi.org/10.3390/bdcc7010035

Tapalova, O., & Zhiyenbayeva, N. (2022). Artificial intelligence in education: AIEd for personalised learning pathways. *The Electronic Journal of e-Learning, 20*(5), 639-653. https://doi.org/10.34190/ejel.20.5.2597

Touchie, H. Y. (1986). Second language learning errors: Their types, causes, and treatment. *JALT journal, 8*(1), 75-80. Abgerufen von https://jalt-publications.org/sites/default/files/pdf-article/art5_8.pdf

Truscott, J. (1996). The case against grammar correction in L2 writing classes. *Language Learning, 46*(2), 327–369. https://doi.org/10.1111/j.1467-1770.1996.tb01238.x

Wang, X., Pang, H., Wallace, M. P., Wang, Q., & Chen, W. (2024). Learners' perceived AI presences in AI-supported language learning: a study of AI as a humanized agent from community of inquiry. *Computer Assisted Language Learning*, *37*(4), 814-840. https://doi.org/10.1080/09588221.2022.2056203

Yang, H. & Kyun, S. (2022). The current research trend of artificial intelligence in language learning: A systematic empirical literature review from an activity theory perspective. *Australasian Journal of Educational Technology, 38*(5), 180-210. https://doi.org/10.14742/ajet.7492

Yılmaz, A. (2017). *Yapay zeka*. İstanbul: Kodlab.

Zhai, C., & Wibowo, S. (2023). A systematic review on artificial intelligence dialogue systems for enhancing English as foreign language students' interactional competence in the university. *Computers and Education: Artificial Intelligence*, *4*, 100134. https://doi.org/10.1016/j.caeai.2023.100134

Zhang, W., Lei, X., Liu, Z., Wang, N., Long, Z., Yang, P., … Lian, S. (2025). Safety evaluation of DeepSeek models in Chinese contexts. *ArXiv, abs/2502.11137*. https://doi.org/10.48550/arXiv.2502.11137

Zheng, Y. & Yu, S. (2018). Student engagement with teacher written corrective feedback in EFL writing: A case study of Chinese lower-proficiency students. *Assessing Writing, 37*, 13–24. https://doi.org/10.1016/j.asw.2018.03.001

Zhu, Q., Guo, D., Shao, Z., Yang, D., Wang, P., Xu, R., … Liang, W. (2024). DeepSeek-Coder-V2: Breaking the barrier of closed-source models in code intelligence. *ArXiv, abs/2406.11931*. https://doi.org/10.48550/arXiv.2406.11931

ANHÄNGE

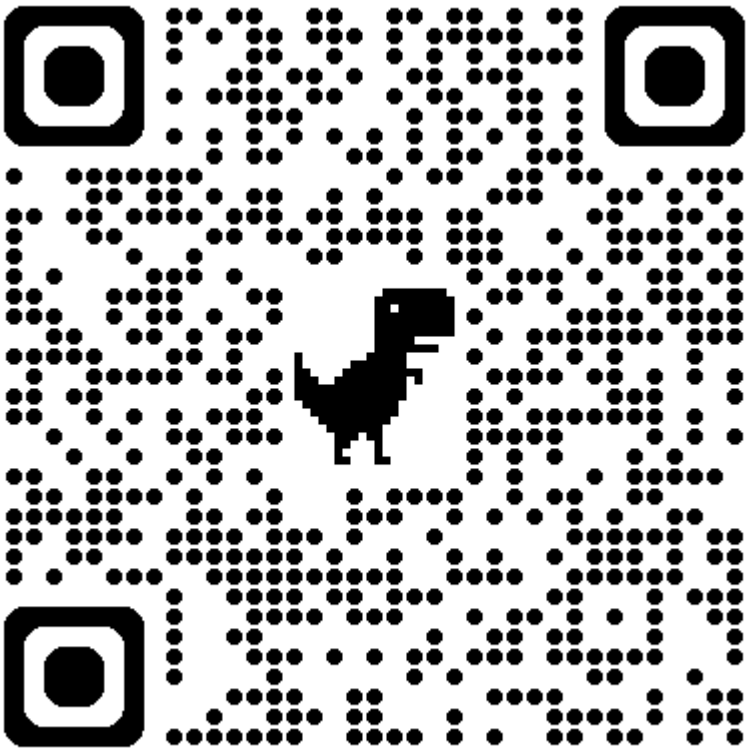
Anhang 1. Zugriff auf alle in der Studie verwendeten Daten

Die im Rahmen der Datenerhebung mithilfe von KI-Chatbots gewonnenen Texte und Daten sind zu umfangreich, um in den Anhang zu passen. Da die Arbeit in gedruckter Form eingereicht wird, wäre die Veröffentlichung aller Anhänge hier eine unnötige Papierverschwendung gewesen. In der heutigen digitalen Welt ist die Speicherung und Weitergabe dieser Daten problemlos möglich.

Aperta, das türkische Open-Access-Archiv von TÜBİTAK, bietet die Möglichkeit, diese Daten zu speichern und zu teilen. Alle in der Studie verwendeten Daten finden Sie unten.

DOI: https://doi.org/10.48623/aperta.286368

QR-CODE:



Anhang 2. Von künstlicher Intelligenz erstelltes fehlerfreies Textbeispiel auf A1-Niveau.

Hallo, ich heiße Leo. Ich bin zwanzig Jahre alt und komme aus Österreich. Ich wohne jetzt in Berlin. Berlin ist eine große und interessante Stadt. Ich lerne hier Deutsch, denn die Sprache ist wichtig für meine Arbeit. Ich bin Student und arbeite auch in einem Café. Die Arbeit macht Spaß und meine Kollegen sind sehr nett. Ich habe einen neuen Freund. Er heißt Felix. Felix kommt aus Spanien, aber er spricht schon sehr gut Deutsch. Er ist auch Student. Wir lernen zusammen in der Bibliothek. Die Bibliothek ist nicht weit von meiner Wohnung. Manchmal gehen wir nach dem Lernen spazieren. Felix hat viele Hobbys. Er spielt gern Fußball und hört Musik. Ich spiele nicht gern Fußball, aber ich höre auch gern Musik. Am Wochenende kochen wir oft zusammen. Felix kann sehr gut kochen. Wir essen gern Pizza oder Nudeln. Dann trinken wir eine Cola oder einen Saft. Abends sehen wir manchmal einen Film. Die Filme sind oft auf Deutsch. Das ist gut für uns. Wir lernen viele neue Wörter. Berlin ist eine tolle Stadt und ich bin froh, hier zu sein und einen guten Freund wie Felix zu haben. Wir planen auch eine kleine Reise. Vielleicht fahren wir im Sommer nach Hamburg.

Anhang 3. Von künstlicher Intelligenz erstelltes fehlerfreies Textbeispiel auf A2-Niveau.

Letzte Woche hatte ich einen sehr interessanten Tag. Am Morgen, als ich meinen Briefkasten öffnete, fand ich einen Brief ohne Absender. Das war komisch. Normalerweise bekomme ich nur Rechnungen oder Werbung. Ich ging in meine Wohnung zurück, setzte mich an den Tisch in der Küche und machte den Brief langsam auf. Drinnen war eine alte Postkarte. Auf dem Bild war ein kleiner Ort am Meer. Ich kannte diesen Ort nicht. Auf der Rückseite stand nur ein Satz: „Erinnerst du dich an unseren gemeinsamen Urlaub?“ Ich dachte lange nach. Wer könnte mir diesen Brief geschickt haben? Vielleicht ein alter Freund aus der Schule?

Am Nachmittag habe ich meine Freundin angerufen und ihr alles erzählt. Sie fand die Geschichte sehr spannend. Wir haben zusammen überlegt. Plötzlich hatte ich eine Idee. Vor vielen Jahren war ich mit meiner Familie in einem kleinen Dorf in Italien im Urlaub. Es war wunderschön dort. Ich habe dort einen Jungen kennengelernt. Sein Name war Marco. Wir haben den ganzen Sommer zusammen gespielt. Aber danach haben wir den Kontakt verloren. Könnte der Brief von ihm sein? Ich war sehr aufgeregt. Ich wollte ihm antworten, aber es gab keine Adresse. Ich wusste nicht, was ich tun sollte. Also habe ich beschlossen, einfach zu warten. Vielleicht kommt ja noch ein Brief. Dieser Gedanke machte mich glücklich und ein bisschen nervös. Ich hoffe, ich höre bald wieder von ihm.

Anhang 4. Von künstlicher Intelligenz erstelltes fehlerfreies Textbeispiel auf B1-Niveau.

Letztes Wochenende wollten meine Freundin Klara und ich eigentlich nur zu Hause bleiben und uns entspannen. Wir waren beide ziemlich müde von der langen Arbeitswoche. Doch am Samstagmorgen rief unser gemeinsamer Freund Lukas an. „Ich habe eine Überraschung für euch“, sagte er begeistert. „Zieht euch bequeme Schuhe an, wir machen einen Ausflug in die Berge!“ Anfangs waren wir nicht so überzeugt, weil das Wetter nicht besonders gut aussah. Trotzdem haben wir zugesagt.

Lukas holte uns mit seinem Auto ab. Während der Fahrt regnete es leicht, aber als wir ankamen, schien plötzlich die Sonne. Die Landschaft war wunderschön. Wir wanderten auf einem schmalen Weg, der durch einen dichten Wald führte. Überall war es ruhig und die Luft war frisch. Lukas hatte ein kleines Picknick vorbereitet. Es gab belegte Brote, Obst und Tee. Wir setzten uns auf eine Wiese mit Blick auf ein tiefes Tal und genossen das Essen. Ich musste zugeben, dass Lukas‘ Vorschlag fantastisch war. Obwohl ich anfangs skeptisch war, wurde es ein unvergesslicher Tag. Manchmal sind die spontanen Entscheidungen die besten. Es lohnt sich, offen für neue Pläne zu sein, auch wenn man müde ist. Wir haben beschlossen, solche Ausflüge in Zukunft häufiger zu unternehmen, denn sie geben einem unglaublich viel Energie für die kommende Woche.

Anhang 5. Von künstlicher Intelligenz erstelltes fehlerhaftes Textbeispiel auf A1-Niveau.

Hallo, ich heiße Leo. Ich bin zwanzig Jahre alt und komme aus Österreich. Ich wohne jetzt in Berlin. Berlin ist große eine und interessante Stadt. Ich lerne hier Deutsch, denn die Sprache ist wichtig für meine Arbeit. Ich bin Schüler und arbeite auch in Café. Die Arbeit ist macht Spaß und meine Kollegen sind sehr nett. Ich habe einen neuen Freund. Er heißt Felix. Felix kommt aus Spanien, aber er spricht schon sehr gut Deutsch. Er ist auch Student. Wir lernen zusammen in der Bibliothek. Die Bibliothek ist nicht weit von meine Wohnung. Wir gehen spazieren nach dem Lernen manchmal. Felix hat viele. Er spielt gern Fußball und hört Musik. Ich spiele nicht gern Fußball, aber ich höre auch gern Musik. Am Wochenende kochen wir oft zusammen. Felix kann sehr gut kochen. Wir essen gern Pizza oder Nudeln. Dann trinken wir eine Cola oder einen Saft Apfel. Abends sehen wir manchmal einen Film. Die Filme sind oft auf Deutsch. Das ist gut für uns. Wir lernen viele neue Wörter. Berlin ist eine tolle Stadt und ich bin froh, hier zu sein und einen guten Freund wie Felix zu haben. Wir planen auch eine kleine Reise. Vielleicht fahren wir im Sommer nach Hamburg.

Fehleranalyse-Liste

Fehler 1: Grammatikalisch - Wortstellung (Falsche Reihenfolge)

Position: Berlin ist große eine und interessante Stadt.

Fehlerdetails: Richtig: "...eine große und interessante Stadt." -> Falsch: "...große eine und interessante Stadt."

Erklärung: Dieser Fehler stellt einen Verstoß gegen die grammatikalische Wortstellungsregel in der Adjektivgruppe dar. Im Deutschen steht das Adjektiv vor dem Substantiv, das es beschreibt, und nach dem Artikel (Artikel + Adjektiv + Nomen). Im fehlerhaften Satz wurde das Adjektiv (große) fälschlicherweise vor dem Artikel (eine) platziert, was zu einer falschen Reihenfolge führt.

Fehler 2: Lexikalisch-semantisch - Selektion (Falsche Auswahl)

Position: Ich bin Schüler und arbeite auch in Café.

Fehlerdetails: Richtig: "Ich bin Student..." -> Falsch: "Ich bin Schüler..."

Erklärung: Dies ist ein semantisch falscher Wortgebrauch. Im Kontext des Textes ist der Charakter zwanzig Jahre alt, und es wird erwartet, dass er ein Student auf Universitätsniveau ist. Die Verwendung von "Schüler" (Grund-/Oberschüler) anstelle von "Student" (Universitätsstudent) ist ein lexikalisch-semantischer Fehler, da ein falsches Wort aus derselben Kategorie "Lernender" ausgewählt wurde.

Fehler 3: Grammatikalisch - Auslassung (Omission)

Position: Ich bin Schüler und arbeite auch in Café.

Fehlerdetails: Richtig: "...in einem Café." -> Falsch: "...in Café."

Erklärung: Der unbestimmte Artikel "einem", der für die grammatikalische Vollständigkeit des Satzes erforderlich ist, fehlt. Da die Präposition "in" den Dativ erfordert, muss der Ausdruck "ein Café" als "in einem Café" lauten. Das vollständige Weglassen des Artikels ist ein grammatikalischer Auslassungsfehler.

Fehler 4: Grammatikalisch - Hinzufügung (Addition)

Position: Die Arbeit ist macht Spaß und meine Kollegen sind sehr nett.

Fehlerdetails: Richtig: "Die Arbeit macht Spaß..." -> Falsch: "Die Arbeit ist macht Spaß..."

Erklärung: Ein überflüssiges Element, das Hilfsverb "ist", wurde der grammatikalischen Struktur des Satzes hinzugefügt. Das Verb "Spaß machen" bildet allein das Prädikat des Satzes und benötigt nicht das Hilfsverb "sein". Diese überflüssige Hinzufügung ist ein grammatikalischer Fehler.

Fehler 5: Grammatikalisch - Selektion (Falsche Auswahl)

Position: Die Bibliothek ist nicht weit von meine Wohnung.

Fehlerdetails: Richtig: "...von meiner Wohnung." -> Falsch: "...von meine Wohnung."

Erklärung: Dieser Fehler besteht in der Auswahl eines grammatikalisch falschen Elements. Die Präposition "von" erfordert immer den Dativ. Da der Dativ von "die Wohnung" "der Wohnung" lautet, muss auch das Possessivpronomen als "meiner" dekliniert werden. Im fehlerhaften Satz wurde anstelle des Dativs die Nominativ-/Akkusativform "meine" gewählt.

Fehler 6: Lexikalisch-semantisch - Wortstellung (Falsche Reihenfolge)

Position: Wir gehen spazieren nach dem Lernen manchmal.

Fehlerdetails: Richtig: "Manchmal gehen wir nach dem Lernen spazieren." -> Falsch: "Wir gehen spazieren nach dem Lernen manchmal."

Erklärung: Dies ist ein Wortstellungsfehler, der den semantischen Fluss des Satzes stört. Das Adverb "manchmal" steht normalerweise am Satzanfang oder nach dem finiten Verb. Die Platzierung des Adverbs am Satzende stört die semantische Kohärenz der Satzstruktur und stellt daher einen lexikalisch-semantischen Wortstellungsfehler dar.

Fehler 7: Lexikalisch-semantisch - Auslassung (Omission)

Position: Felix hat viele.

Fehlerdetails: Richtig: "Felix hat viele Hobbys." -> Falsch: "Felix hat viele."

Erklärung: Das Substantiv "Hobbys", das zur Vervollständigung der Satzbedeutung erforderlich ist, fehlt. Der Satz "Felix hat viele ..." ist semantisch unvollständig, da nicht angegeben wird, was er besitzt. Dies ist ein lexikalisch-semantischer Auslassungsfehler.

Fehler 8: Lexikalisch-semantisch - Hinzufügung (Addition)

Position: Dann trinken wir eine Cola oder einen Saft Apfel.

Fehlerdetails: Richtig: "...oder einen Saft." -> Falsch: "...oder einen Saft Apfel."

Erklärung: Dieser Fehler besteht in der Hinzufügung eines überflüssigen Wortes, das nicht zur Satzbedeutung passt. Das Wort "Apfel" ist kontextuell plausibel, da es sich in derselben Kategorie (Lebensmittel/Getränke) wie "Saft" befindet. Die Struktur "einen Saft Apfel" (ein Saft Apfel) ist jedoch semantisch fehlerhaft und stört die Kohärenz des Satzes. Dies ist ein Beispiel für einen lexikalisch-semantischen Hinzufügungsfehler.

Anhang 6. Von künstlicher Intelligenz erstelltes fehlerhaftes Textbeispiel auf A2-Niveau.

Letzte Woche hatte ich einen sehr interessanten Tag. Am Morgen, als ich öffnete meinen Briefkasten, fand ich einen Brief ohne Absender. Das war komisch. Normalerweise bekomme ich nur Rechnungen oder Werbung. Ich ging meine Wohnung zurück, setzte mich an dem Tisch in der Küche und aufmachte den Brief langsam. Drinnen war eine alte Fahrkarte. Auf dem Bild war ein kleiner am Meer. Ich kannte diesen Ort nicht. Auf der Rückseite stand nur ein Satz: „Erinnerst du dich an unseren gemeinsamen Urlaub?“ Ich dachte lange nach. Wer könnte mir diesen Brief geschickt haben? Vielleicht ein alter Freund aus der Schule? Am Nachmittag habe ich meine Freundin angerufen und ihr alles erzählt. Sie hat fand die Geschichte sehr spannend. Wir haben zusammen überlegt. Plötzlich hatte ich eine Idee. Vor vielen Jahren war ich mit meiner Familie in einem kleinen Dorf in Italien im Urlaub. Es war wunderschön dort. Ich habe dort einen Jungen kennengelernt. Sein Name war Marco. Wir haben den ganzen Sommer zusammen gespielt. Aber danach haben wir den Kontakt verloren. Könnte der Brief von ihm sein? Ich war sehr aufgeregt. Ich wollte ihm antworten, aber es gab keine Adresse Telefon. Ich wusste nicht, was ich tun sollte. Also habe ich beschlossen, einfach zu warten. Vielleicht kommt ja noch ein Brief. Dieser Gedanke machte mich glücklich und ein bisschen nervös. Ich hoffe, ich höre bald wieder von ihm.

Fehleranalyse-Liste

Fehler 1: Grammatikalisch - Wortstellung (Falsche Reihenfolge)

Position: Am Morgen, als ich öffnete meinen Briefkasten, fand ich einen Brief ohne Absender.

Fehlerdetails: Richtig: "...als ich meinen Briefkasten öffnete,..." -> Falsch: "...als ich öffnete meinen Briefkasten,..."

Erklärung: Dies ist ein grammatikalischer Wortstellungsfehler. In deutschen Nebensätzen (hier eingeleitet mit 'als') muss das konjugierte Verb am Satzende stehen. Im fehlerhaften Satz wurde das Verb ("öffnete") an der zweiten Position belassen.

Fehler 2: Grammatikalisch - Auslassung (Omission)

Position: Ich ging meine Wohnung zurück, setzte mich an dem Tisch in der Küche und aufmachte den Brief langsam.

Fehlerdetails: Richtig: "Ich ging in meine Wohnung zurück,..." -> Falsch: "Ich ging meine Wohnung zurück,..."

Erklärung: Dieser Fehler besteht in der Auslassung einer Präposition ('in'), die für die grammatikalische Vollständigkeit des Satzes notwendig ist. In diesem Satz, der eine Richtung angibt, bleibt der Satz ohne die Präposition "in" grammatikalisch unvollständig.

Fehler 3: Grammatikalisch - Selektion (Falsche Auswahl)

Position: Ich ging meine Wohnung zurück, setzte mich an dem Tisch in der Küche und aufmachte den Brief langsam.

Fehlerdetails: Richtig: "...setzte mich an den Tisch..." (Akkusativ) -> Falsch: "...setzte mich an dem Tisch..." (Dativ)

Erklärung: Dieser Fehler besteht in der Auswahl eines grammatikalisch falschen Kasus. Das Verb "sich setzen an..." drückt eine Richtung aus und erfordert daher den Akkusativ-Artikel ("den"). Im Satz wurde stattdessen der Dativ-Artikel ("dem") gewählt, wodurch eine falsche Grammatikregel angewendet wurde.

Fehler 4: Lexikalisch-semantisch - Wortstellung (Falsche Reihenfolge)

Position: Ich ging meine Wohnung zurück, setzte mich an dem Tisch in der Küche und aufmachte den Brief langsam.

Fehlerdetails: Richtig: "...und machte den Brief langsam auf." -> Falsch: "...und aufmachte den Brief langsam."

Erklärung: Dieser Fehler besteht in der falschen Anordnung der Bestandteile eines trennbaren Verbs ('aufmachen'), das eine semantische Einheit bildet. Im Präteritum sollte das Präfix eines trennbaren Verbs am Satzende stehen. Im fehlerhaften Satz wurde das Präfix ('auf') vor das Verb gestellt, wodurch die semantische Struktur des Wortes gestört wurde.

Fehler 5: Lexikalisch-semantisch - Selektion (Falsche Auswahl)

Position: Drinnen war eine alte Fahrkarte.

Fehlerdetails: Richtig: "...eine alte Postkarte." -> Falsch: "...eine alte Fahrkarte."

Erklärung: Dieser Fehler ist eine falsche Wortwahl, die die semantische Kohärenz des Satzes stört. Der Kontext des Textes (Bild am Meer, Urlaubserinnerung) erfordert eine "Postkarte", stattdessen wurde das semantisch verwandte, aber falsche Wort "Fahrkarte" (Ticket) verwendet.

Fehler 6: Lexikalisch-semantisch - Auslassung (Omission)

Position: Auf dem Bild war ein kleiner am Meer.

Fehlerdetails: Richtig: "Auf dem Bild war ein kleiner Ort am Meer." -> Falsch: "Auf dem Bild war ein kleiner am Meer."

Erklärung: Dieser Fehler besteht in der Auslassung eines wesentlichen Substantivs ('Ort'), das zur Vervollständigung der Satzbedeutung erforderlich ist. Ohne dieses Wort ist unklar, worum es in dem Satz geht, und er bleibt semantisch unvollständig.

Fehler 7: Grammatikalisch - Hinzufügung (Addition)

Position: Sie hat fand die Geschichte sehr spannend.

Fehlerdetails: Richtig: "Sie fand die Geschichte sehr spannend." -> Falsch: "Sie hat fand die Geschichte sehr spannend."

Erklärung: Dieser Fehler besteht in der Hinzufügung eines grammatikalisch überflüssigen Elements. Der Satz steht bereits im Präteritum ("fand"). Das Hilfsverb "hat", das zur Bildung des Perfekts verwendet wird, wurde unnötigerweise hinzugefügt, was einen Verstoß gegen die Grammatikregeln darstellt.

Fehler 8: Lexikalisch-semantisch - Hinzufügung (Addition)

Position: Ich wollte ihm antworten, aber es gab keine Adresse Telefon.

Fehlerdetails: Richtig: "...es gab keine Adresse." -> Falsch: "...es gab keine Adresse Telefon."

Erklärung: Dieser Fehler besteht in der Hinzufügung eines überflüssigen Wortes, das die semantische Kohärenz des Satzes stört. Das hinzugefügte Wort "Telefon" steht zwar im

Kommunikationskontext in Beziehung zum Wort "Adresse", hat aber an dieser Stelle im Satz keine semantische Funktion und ist überflüssig.

Anhang 7. Von künstlicher Intelligenz erstelltes fehlerhaftes Textbeispiel auf B1-Niveau.

Letztes Wochenende wollten meine Freundin Klara und ich eigentlich nur zu Hause bleiben und uns entspannen. Wir waren beide ziemlich müde von der langen Arbeitswoche. Doch am Samstagmorgen rief zu unser gemeinsamer Freund Lukas an. „Ich habe eine Überraschung für euch“, sagte er begeistert. „Zieht euch bequeme Schuhe an, wir machen einen Ausflug die Berge!“ Anfangs waren wir nicht so überzeugt, weil das Wetter aussah nicht besonders gut. Trotzdem haben wir zugesagt. Ab Lukas holte uns mit seinem Auto. Während der Fahrt regnete es leicht, aber als wir ankamen, schien plötzlich die Sonne. Die Landschaft war. Wir wanderten auf einen schmalen Weg, der durch einen dichten Wald führte. Überall war es ruhig und die Luft war frisch. Lukas hatte ein kleines Picknick Essen vorbereitet. Es gab belegte Brote, Obst und Milch. Wir setzten uns auf eine Wiese mit Blick auf ein tiefes Tal und genossen das Essen. Ich musste zugeben, dass Lukas‘ Vorschlag fantastisch war. Obwohl ich anfangs skeptisch war, wurde es ein unvergesslicher Tag. Manchmal sind die spontanen Entscheidungen die besten. Es lohnt sich, offen für neue Pläne zu sein, auch wenn man müde ist. Wir haben beschlossen, solche Ausflüge in Zukunft häufiger zu unternehmen, denn sie geben einem unglaublich viel Energie für die kommende Woche.

Fehleranalyse-Liste

Fehler 1: Grammatikalisch - Hinzufügung (Addition)

Position: Doch am Samstagmorgen rief zu unser gemeinsamer Freund Lukas an.

Fehlerdetails: Richtig: "...rief unser gemeinsamer Freund..." -> Falsch: "...rief zu unser gemeinsamer Freund..."

Erklärung: Dieser Fehler entstand durch die Hinzufügung der Präposition "zu", die für die grammatikalische Struktur des Satzes überflüssig ist. Das Verb "anrufen" erfordert keine Präposition, daher handelt es sich um einen Hinzufügungsfehler (Addition).

Fehler 2: Grammatikalisch - Auslassung (Omission)

Position: „Zieht euch bequeme Schuhe an, wir machen einen Ausflug die Berge!“

Fehlerdetails: Richtig: "...einen Ausflug in die Berge!" -> Falsch: "...einen Ausflug die Berge!"

Erklärung: Die Präposition "in", die für die grammatikalische Vollständigkeit des Satzes erforderlich ist, fehlt. Da es sich um eine Bewegung handelt, die eine Richtung angibt, ist die Verwendung der Präposition zwingend erforderlich. Dies ist ein Auslassungsfehler (Omission).

Fehler 3: Grammatikalisch - Wortstellung (Ordering)

Position: Anfangs waren wir nicht so überzeugt, weil das Wetter aussah nicht besonders gut.

Fehlerdetails: Richtig: "...weil das Wetter nicht besonders gut aussah." -> Falsch: "...weil das Wetter aussah nicht besonders gut."

Erklärung: In einem Nebensatz, der mit "weil" beginnt, muss das konjugierte Verb am Satzende stehen. Die Verwendung des Verbs an zweiter Position ist ein grammatikalischer Wortstellungsfehler (Ordering).

Fehler 4: Lexikalisch-semantisch - Wortstellung (Ordering)

Position: Ab Lukas holte uns mit seinem Auto.

Fehlerdetails: Richtig: "Lukas holte uns mit seinem Auto ab." -> Falsch: "Ab Lukas holte uns mit seinem Auto."

Erklärung: Das Präfix "ab" des trennbaren Verbs "abholen", das eigentlich am Satzende stehen müsste, wurde an den Satzanfang gestellt. Dieser Umstand stört die semantische Integrität des Verbs und stellt daher einen lexikalisch-semantischen Wortstellungsfehler (Ordering) dar.

Fehler 5: Lexikalisch-semantisch - Auslassung (Omission)

Position: Die Landschaft war.

Fehlerdetails: Richtig: "Die Landschaft war wunderschön." -> Falsch: "Die Landschaft war."

Erklärung: Das Adjektiv "wunderschön", das für die semantische Vervollständigung des Satzes erforderlich ist, fehlt. Ohne dieses Wort ist der Satz zwar grammatikalisch korrekt,

bleibt aber semantisch unvollständig. Dies ist ein lexikalisch-semantischer Auslassungsfehler (Omission).

Fehler 6: Grammatikalisch - Selektion (Selection)

Position: Wir wanderten auf einen schmalen Weg, der durch einen dichten Wald führte.

Fehlerdetails: Richtig: "...auf einem schmalen Weg..." -> Falsch: "...auf einen schmalen Weg..."

Erklärung: Die Präposition "auf" erfordert in Verbindung mit dem Verb "wandern", das einen Zustand (Ort) beschreibt, den Dativ. Im Satz wurde statt des Dativ-Artikels ("einem") der Akkusativ-Artikel ("einen") gewählt. Dies ist ein Fehler bei der Auswahl des grammatikalischen Elements (Selection).

Fehler 7: Lexikalisch-semantisch - Hinzufügung (Addition)

Position: Lukas hatte ein kleines Picknick Essen vorbereitet.

Fehlerdetails: Richtig: "...ein kleines Picknick vorbereitet." -> Falsch: "...ein kleines Picknick Essen vorbereitet."

Erklärung: Das Wort "Essen", das für die Bedeutung des Satzes überflüssig ist, wurde hinzugefügt. Da das Wort "Picknick" bereits eine Art Mahlzeit impliziert, ist das Wort "Essen" semantisch redundant, wenn auch kontextuell (Lebensmittel) nicht unlogisch. Dies ist ein lexikalisch-semantischer Hinzufügungsfehler (Addition).

Fehler 8: Lexikalisch-semantisch - Selektion (Selection)

Position: Es gab belegte Brote, Obst und Milch.

Fehlerdetails: Richtig: "...Obst und Tee." -> Falsch: "...Obst und Milch."

Erklärung: Im Originaltext wurde angegeben, dass "Tee" für das Picknick vorbereitet wurde. Obwohl "Milch" auch ein Getränk ist, ist es im Kontext eines Picknicks bei einer Bergwanderung unüblicher und eine semantisch abweichende Wahl. Dies ist ein lexikalisch-semantischer Selektionsfehler (Selection), bei dem ein falsches Wort aus derselben Kategorie anstelle des richtigen gewählt wurde.

Anhang 8. Antwortbeispiel von ChatGPT für den fehlerhaften Text auf A1-Niveau

Hier ist die vollständige Bearbeitung Ihres Textes auf A1-Niveau:

---

1. Erkannte Fehler

„Berlin ist große eine und interessante Stadt“ → falsche Wortstellung; korrekt: „Berlin ist eine große und interessante Stadt“.

„in Café“ → fehlender Artikel; korrekt: „in einem Café“.

„Die Arbeit ist macht Spaß“ → falsche Formulierung; korrekt: „Die Arbeit macht Spaß“.

„von meine Wohnung“ → falscher Kasus; korrekt: „von meiner Wohnung“.

„Wir gehen spazieren nach dem Lernen manchmal“ → falsche Wortstellung; korrekt: „Nach dem Lernen gehen wir manchmal spazieren“.

„Felix hat viele“ → unvollständig; wahrscheinlich: „Felix hat viele Hobbys“.

„einen Saft Apfel“ → falsche Wortstellung; korrekt: „einen Apfelsaft“.

---

2. Korrigierter Text

Hallo, ich heiße Leo.

Ich bin zwanzig Jahre alt und komme aus Österreich.

Ich wohne jetzt in Berlin.

Berlin ist eine große und interessante Stadt.

Ich lerne hier Deutsch, denn die Sprache ist wichtig für meine Arbeit.

Ich bin Schüler und arbeite auch in einem Café.

Die Arbeit macht Spaß und meine Kollegen sind sehr nett.

Ich habe einen neuen Freund.

Er heißt Felix.

Felix kommt aus Spanien, aber er spricht schon sehr gut Deutsch.

Er ist auch Student.

Wir lernen zusammen in der Bibliothek.

Die Bibliothek ist nicht weit von meiner Wohnung.

Nach dem Lernen gehen wir manchmal spazieren.

Felix hat viele Hobbys.

Er spielt gern Fußball und hört Musik.

Ich spiele nicht gern Fußball, aber ich höre auch gern Musik.

Am Wochenende kochen wir oft zusammen.

Felix kann sehr gut kochen.

Wir essen gern Pizza oder Nudeln.

Dann trinken wir eine Cola oder einen Apfelsaft.

Abends sehen wir manchmal einen Film.

Die Filme sind oft auf Deutsch.

Das ist gut für uns.

Wir lernen viele neue Wörter.

Berlin ist eine tolle Stadt und ich bin froh, hier zu sein und einen guten Freund wie Felix zu haben.

Wir planen auch eine kleine Reise.

Vielleicht fahren wir im Sommer nach Hamburg.

---

3. Feedback

1. „große eine“ → „eine große“

→ Adjektiv + Substantiv + Artikel: Im Deutschen steht der unbestimmte Artikel vor dem Adjektiv.

2. „in Café“ → „in einem Café“

→ Bei Orten im Dativ (nach „in“) benötigt man den richtigen Artikel: „in einem“.

3. „ist macht Spaß“ → „macht Spaß“

→ Doppeltes Verb ist grammatisch falsch; richtig ist „Die Arbeit macht Spaß“.

4. „von meine Wohnung“ → „von meiner Wohnung“

→ Präposition „von“ verlangt Dativ: „meiner Wohnung“.

5. „nach dem Lernen manchmal“ → „Nach dem Lernen gehen wir manchmal spazieren“

→ Satzbau: Zeitangabe + Verb an zweiter Stelle, Subjekt danach.

6. „Felix hat viele“ → „Felix hat viele Hobbys“

→ Satz unvollständig; es fehlt das Objekt.

7. „einen Saft Apfel“ → „einen Apfelsaft“

→ Im Deutschen wird das Fruchtgetränk mit zusammengesetztem Substantiv gebildet: Apfelsaft.

Wenn Sie möchten, kann ich auch eine kurze Übung zu Wortstellung und Dativpräpositionen basierend auf diesem Text erstellen, um typische Anfängerfehler zu trainieren.

Anhang 9. Antwortbeispiel von ChatGPT für den fehlerhaften Text auf A2-Niveau

Hier ist die sorgfältige Analyse und Korrektur Ihres Textes auf A2-Niveau:

---

1. Erkannte Fehler

- „als ich öffnete meinen Briefkasten“ → falsche Wortstellung; korrekt: „als ich meinen Briefkasten öffnete“.
- „Ich ging meine Wohnung zurück“ → fehlende Präposition; korrekt: „Ich ging in meine Wohnung zurück“.
- „setzte mich an dem Tisch“ → falscher Kasus; korrekt: „an den Tisch“.
- „aufmachte den Brief langsam“ → falsche Wortstellung; korrekt: „machte den Brief langsam auf“.
- „ein kleiner am Meer“ → unvollständig; es fehlt das Nomen; korrekt: „ein kleiner Junge am Meer“ oder „ein kleiner Ort am Meer“, je nach Kontext.
- „Sie hat fand die Geschichte sehr spannend“ → Doppelung von Verben; korrekt: „Sie fand die Geschichte sehr spannend“.
- „keine Adresse Telefon“ → fehlender Artikel; korrekt: „keine Telefonadresse“ oder „keine Adresse für das Telefon“.

---

2. Korrigierter Text

Letzte Woche hatte ich einen sehr interessanten Tag. Am Morgen, als ich meinen Briefkasten öffnete, fand ich einen Brief ohne Absender. Das war komisch. Normalerweise bekomme ich nur Rechnungen oder Werbung. Ich ging in meine Wohnung zurück, setzte mich an den Tisch in der Küche und machte den Brief langsam auf.

Drinnen war eine alte Fahrkarte. Auf dem Bild war ein kleiner Junge am Meer. Ich kannte diesen Ort nicht. Auf der Rückseite stand nur ein Satz: „Erinnerst du dich an unseren gemeinsamen Urlaub?“ Ich dachte lange nach. Wer könnte mir diesen Brief geschickt haben? Vielleicht ein alter Freund aus der Schule?

Am Nachmittag habe ich meine Freundin angerufen und ihr alles erzählt. Sie fand die Geschichte sehr spannend. Wir haben zusammen überlegt. Plötzlich hatte ich eine Idee. Vor vielen Jahren war ich mit meiner Familie in einem kleinen Dorf in Italien im Urlaub. Es war wunderschön dort. Ich habe dort einen Jungen kennengelernt. Sein Name war Marco. Wir haben den ganzen Sommer zusammen gespielt. Aber danach haben wir den Kontakt verloren. Könnte der Brief von ihm sein? Ich war sehr aufgeregt. Ich wollte ihm antworten, aber es gab keine Telefonadresse. Ich wusste nicht, was ich tun sollte. Also habe ich beschlossen, einfach zu warten. Vielleicht kommt ja noch ein Brief. Dieser Gedanke machte mich glücklich und ein bisschen nervös. Ich hoffe, ich höre bald wieder von ihm.

---

3. Feedback

1. Wortstellung bei Nebensätzen:

„als ich öffnete meinen Briefkasten“ → Im Nebensatz steht das konjugierte Verb am Ende.

2. Präpositionen und Kasus:

o „Ich ging meine Wohnung zurück“ → Präposition in + Akkusativ: „in meine Wohnung zurück“.

o „an dem Tisch“ → Bewegung: Akkusativ: „an den Tisch“.

3. Trennbare Verben:

o „aufmachte den Brief“ → trennbares Verb „aufmachen“: Verb im Satz trennen: „machte den Brief auf“.

4. Unvollständige Angaben / fehlende Nomen:

o „ein kleiner am Meer“ → man benötigt ein Nomen: „ein kleiner Junge am Meer“.

5. Doppelte Verben:

o „hat fand die Geschichte“ → nur ein Prädikat nötig: „fand die Geschichte“.

6. Fehlender Artikel / falsche Wortbildung:

o „keine Adresse Telefon“ → korrekt: „keine Telefonadresse“.

Wenn Sie möchten, kann ich aus diesem Text eine Lückentext-Übung erstellen, um gezielt die Fehlerarten wie Nebensatzstellung, Präpositionen und trennbare Verben zu üben.
Wollen Sie, dass ich das mache?

Anhang 10. Antwortbeispiel von ChatGPT für den fehlerhaften Text auf B1-Niveau

Hier ist die vollständige Bearbeitung Ihres Textes nach den drei geforderten Schritten:

---

1. Erkannte Fehler

- „rief zu unser gemeinsamer Freund Lukas an“ → falsche Wortstellung und Präposition; korrekt: „rief unser gemeinsamer Freund Lukas an“.
- „wir machen einen Ausflug die Berge“ → fehlende Präposition; korrekt: „wir machen einen Ausflug in die Berge“.
- „weil das Wetter aussah nicht besonders gut“ → falsche Wortstellung; korrekt: „weil das Wetter nicht besonders gut aussah“.
- „Ab Lukas holte uns mit seinem Auto“ → falscher Ausdruck; korrekt: „Ab da holte uns Lukas mit seinem Auto“ oder „Lukas holte uns mit seinem Auto ab“.
- „Die Landschaft war.“ → unvollständiger Satz; sollte ergänzt werden, z. B.: „Die Landschaft war wunderschön.“
- „auf einen schmalen Weg“ → falscher Fall; korrekt: „auf einem schmalen Weg“ (Dativ nach „auf“ bei Ortsangaben).
- „ein kleines Picknick Essen vorbereitet“ → falsche Wortstellung; korrekt: „ein kleines Picknick-Essen vorbereitet“.

---

2. Korrigierter Text

Letztes Wochenende wollten meine Freundin Klara und ich eigentlich nur zu Hause bleiben und uns entspannen. Wir waren beide ziemlich müde von der langen Arbeitswoche. Doch am Samstagmorgen rief unser gemeinsamer Freund Lukas an. „Ich habe eine Überraschung für euch“, sagte er begeistert. „Zieht euch bequeme Schuhe an, wir machen einen Ausflug in die Berge!“ Anfangs waren wir nicht so überzeugt, weil das Wetter nicht besonders gut aussah. Trotzdem haben wir zugesagt.

Lukas holte uns mit seinem Auto ab. Während der Fahrt regnete es leicht, aber als wir ankamen, schien plötzlich die Sonne. Die Landschaft war wunderschön. Wir wanderten auf einem schmalen Weg, der durch einen dichten Wald führte. Überall war es ruhig und die Luft war frisch. Lukas hatte ein kleines Picknick-Essen vorbereitet. Es gab belegte Brote, Obst und Milch. Wir setzten uns auf eine Wiese mit Blick auf ein tiefes Tal und genossen das Essen. Ich musste zugeben, dass Lukas' Vorschlag fantastisch war. Obwohl ich anfangs skeptisch war, wurde es ein unvergesslicher Tag. Manchmal sind die spontanen Entscheidungen die besten. Es lohnt sich, offen für neue Pläne zu sein, auch wenn man müde ist. Wir haben beschlossen, solche Ausflüge in Zukunft häufiger zu unternehmen, denn sie geben einem unglaublich viel Energie für die kommende Woche.

---

3. Feedback

1. Wortstellung bei Verben und Präpositionen

o Fehler: „rief zu unser gemeinsamer Freund Lukas an" → korrekt: „rief unser gemeinsamer Freund Lukas an".

o Regel: Im Deutschen steht das konjugierte Verb in Hauptsätzen an zweiter Position, Präpositionen werden korrekt integriert.

2. Präpositionen bei Ortsangaben

o Fehler: „Ausflug die Berge" → korrekt: „Ausflug in die Berge".

o Regel: Bewegungen zu einem Ort erfordern Präpositionen wie „in", „auf", „zu".

3. Verneinung im Satz

o Fehler: „weil das Wetter aussah nicht besonders gut" → korrekt: „weil das Wetter nicht besonders gut aussah".

o Regel: In Nebensätzen steht das Verb am Ende, und die Negation „nicht" steht vor dem Adjektiv oder Adverb, das verneint wird.

4. Zeit- und Ausdrucksangaben

o Fehler: „Ab Lukas holte uns…" → korrekt: „Lukas holte uns mit seinem Auto ab".

o Regel: Trennbare Verben wie „abholen“ werden im Präteritum oder Perfekt korrekt konjugiert; „ab“ kommt ans Satzende.

5. Dativ nach Präpositionen bei Ortsangaben

o Fehler: „auf einen schmalen Weg“ → korrekt: „auf einem schmalen Weg“.

o Regel: Nach Präpositionen wie „auf“ folgt bei Ortsangaben der Dativ.

6. Zusammengesetzte Nomen / Wortstellung

o Fehler: „ein kleines Picknick Essen“ → korrekt: „ein kleines Picknick-Essen“.

o Regel: Zusammengesetzte Nomen werden zusammengeschrieben, oft mit Bindestrich bei klarer Lesbarkeit.

7. Unvollständiger Satz

o Fehler: „Die Landschaft war.“ → ergänzt zu: „Die Landschaft war wunderschön.“

o Regel: Jeder Satz benötigt mindestens Subjekt und Prädikat.

Wenn Sie möchten, kann ich eine kurze Übung speziell zu Präpositionen, Dativ und zusammengesetzten Nomen erstellen, die genau diese Fehler trainiert.

Möchten Sie, dass ich das mache?

Anhang 11. Antwortbeispiel von DeepSeek für den fehlerhaften Text auf A1-Niveau

Hier ist Ihre Anfrage als erfahrener Deutschlehrer und Korrekturleser, nach den vorgegebenen Schritten bearbeitet.

1. Erkannte Fehler

- Berlin ist große eine und interessante Stadt. (Wortstellungsfehler)
- Ich bin Schüler und arbeite auch in Café. (Fehlender Artikel)
- Die Arbeit ist macht Spaß (Falsche Verwendung von "ist" als Hilfsverb)
- Die Bibliothek ist nicht weit von meine Wohnung. (Falscher Kasus nach Präposition)
- Wir gehen spazieren nach dem Lernen manchmal. (Falsche Position des Adverbs)
- Felix hat viele. (Unvollständiger Satz/fehlendes Substantiv)
- Dann trinken wir eine Cola oder einen Saft Apfel. (Falsche Wortstellung bei zusammengesetzten Nomen)
- Wir planen auch eine kleine Reise. (Kein grammatikalischer Fehler, aber eine natürlichere Formulierung ist möglich.)

2. Korrigierter Text

Ein neuer Freund

Hallo, ich heiße Leo. Ich bin zwanzig Jahre alt und komme aus Österreich. Ich wohne jetzt in Berlin. Berlin ist eine große und interessante Stadt. Ich lerne hier Deutsch, denn die Sprache ist wichtig für meine Arbeit. Ich bin Schüler und arbeite auch in einem Café. Die Arbeit macht Spaß und meine Kollegen sind sehr nett. Ich habe einen neuen Freund. Er heißt Felix. Felix kommt aus Spanien, aber er spricht schon sehr gut Deutsch. Er ist auch Student. Wir lernen zusammen in der Bibliothek. Die Bibliothek ist nicht weit von meiner Wohnung. Manchmal gehen wir nach dem Lernen spazieren. Felix hat viele Hobbys. Er spielt gern Fußball und hört Musik. Ich spiele nicht gern Fußball, aber ich höre auch gern Musik. Am Wochenende kochen wir oft zusammen. Felix kann sehr gut kochen. Wir essen gern Pizza oder Nudeln. Dann trinken wir eine Cola oder einen Apfelsaft. Abends sehen wir manchmal einen Film. Die Filme sind oft auf Deutsch. Das ist gut für uns. Wir lernen viele neue Wörter.

Berlin ist eine tolle Stadt und ich bin froh, hier zu sein und einen guten Freund wie Felix zu haben. Wir planen auch einen kleinen Ausflug. Vielleicht fahren wir im Sommer nach Hamburg.

3. Feedback

- große eine -> eine große: Die korrekte Reihenfolge im Deutschen ist Artikel + Adjektiv + Nomen. "Eine" ist der Artikel und "große" das Adjektiv. Sie müssen getauscht werden.
- in Café -> in einem Café: Nach der Wechselpräposition "in" (hier mit Dativ, da es einen Ort beschreibt) wird ein Artikel benötigt. Da "Café" sächlich (neutral) ist, lautet der unbestimmte Artikel im Dativ "einem".
- Die Arbeit ist macht Spaß -> Die Arbeit macht Spaß: Das Verb "machen" ist hier das Vollverb und bildet das Prädikat des Satzes. Die Form "ist" (von "sein") ist ein Hilfsverb und wird hier fälschlicherweise wie im Englischen ("is") verwendet. Im Deutschen steht nur das konjugierte Vollverb: "Die Arbeit macht".
- von meine Wohnung -> von meiner Wohnung: Die Präposition "von" verlangt immer den Dativ. "Meine" ist die Nominativ-Form des Possessivpronomens. Die korrekte Dativ-Form für "meine Wohnung" (feminin) lautet "meiner Wohnung".
- Wir gehen spazieren ... manchmal -> Manchmal gehen wir ... spazieren: Adverbien der Zeit wie "manchmal", "oft", "immer" stehen im Deutschen in der Regel an Position 1 oder im Satzklammer. "Manchmal" an den Satzanfang zu stellen, klingt deutlich flüssiger und natürlicher.
- Felix hat viele. -> Felix hat viele Hobbys.: Das Wort "viele" ist ein Indefinitpronomen und benötigt ein Substantiv, auf das es sich bezieht. Der Satz ist ohne dieses Substantiv unvollständig. Aus dem Kontext wird klar, dass "Hobbys" oder "Interessen" gemeint ist.

- Saft Apfel -> Apfelsaft: Im Deutschen werden zusammengesetzte Nomen (Komposita) meist zusammengeschrieben. Das Bestimmungswort ("Apfel") steht dabei immer am Ende. Die korrekte Zusammensetzung lautet "Apfelsaft".
- eine kleine Reise -> einen kleinen Ausflug: Der Ausdruck "eine Reise" ist nicht falsch, bezieht sich aber oft auf längere Fahrten mit Übernachtung. Für einen Tagesausflug in eine nahegelegene Stadt wie Hamburg ist das Wort "Ausflug" die üblichere und treffendere Bezeichnung. Da "Ausflug" maskulin ist, ändern sich auch der Artikel und die Adjektivendung.

Anhang 12. Antwortbeispiel von DeepSeek für den fehlerhaften Text auf A2-Niveau

Sehr gerne! Hier ist die Korrektur Ihres Textes auf A2-Niveau.

1. Erkannte Fehler

- als ich öffnete meinen Briefkasten (Falsche Satzstellung im Nebensatz)
- Ich ging meine Wohnung zurück (Fehlende Präposition)
- setzte mich an dem Tisch (Falscher Kasus nach Wechselpräposition)
- und aufmachte den Brief langsam (Falsche Satzstellung: Adverb bei trennbarem Verb)
- Auf dem Bild war ein kleiner am Meer. (Fehlendes Substantiv)
- stand nur ein Satz (Falsche Kongruenz: Artikel)
- Sie hat fand die Geschichte sehr spannend. (Falsche Bildung des Perfekts)
- den ganzen Sommer zusammen gespielt (Falsche Trennung des Verbs)
- aber es gab keine Adresse Telefon. (Falsche Wortstellung bei Komposita)

2. Korrigierter Text

Ein unerwarteter Brief

Letzte Woche hatte ich einen sehr interessanten Tag. Am Morgen, als ich meinen Briefkasten öffnete, fand ich einen Brief ohne Absender. Das war komisch. Normalerweise bekomme ich nur Rechnungen oder Werbung. Ich ging in meine Wohnung zurück, setzte mich an den Tisch in der Küche und machte den Brief langsam auf. Drinnen war eine alte Fahrkarte. Auf dem Bild war ein kleiner Ort am Meer. Ich kannte diesen Ort nicht. Auf der Rückseite stand nur ein Satz: „Erinnerst du dich an unseren gemeinsamen Urlaub?“ Ich dachte lange nach. Wer könnte mir diesen Brief geschickt haben? Vielleicht ein alter Freund aus der Schule?

Am Nachmittag habe ich meine Freundin angerufen und ihr alles erzählt. Sie fand die Geschichte sehr spannend. Wir haben zusammen überlegt. Plötzlich hatte ich eine Idee. Vor vielen Jahren war ich mit meiner Familie in einem kleinen Dorf in Italien im Urlaub. Es war wunderschön dort. Ich habe dort einen Jungen kennengelernt. Sein Name war Marco. Wir haben den ganzen Sommer zusammengespielt. Aber danach haben wir den Kontakt verloren.

Könnte der Brief von ihm sein? Ich war sehr aufgeregt. Ich wollte ihm antworten, aber es gab keine Adresse und keine Telefonnummer. Ich wusste nicht, was ich tun sollte. Also habe ich beschlossen, einfach zu warten. Vielleicht kommt ja noch ein Brief. Dieser Gedanke machte mich glücklich und ein bisschen nervös. Ich hoffe, ich höre bald wieder von ihm.

3. Feedback

- als ich öffnete meinen Briefkasten -> als ich meinen Briefkasten öffnete
  - Regel: In einem Nebensatz (eingeleitet durch "als", "weil", "dass" etc.) steht das konjugierte Verb (öffnete) am Ende. Das Objekt ("meinen Briefkasten") steht davor.
- Ich ging meine Wohnung zurück -> Ich ging in meine Wohnung zurück
  - Regel: Das Verb "zurückgehen" beschreibt eine Bewegung. Für die Richtung in einen Raum (wohin?) verwendet man die Präposition "in" mit Akkusativ.
- setzte mich an dem Tisch -> setzte mich an den Tisch
  - Regel: Die Wechselpräposition "an" beschreibt hier eine Bewegung (wohin setze ich mich?). Bei Bewegung/Direction verlangt "an" den Akkusativ. Der Artikel für "Tisch" (maskulin) im Akkusativ lautet "den".
- aufmachte den Brief langsam -> machte den Brief langsam auf
  - Regel: Das Verb "aufmachen" ist ein trennbares Verb. Im Präteritum (einfache Vergangenheit) wird die Vorsilbe "auf" vom Verb getrennt und ans Satzende gestellt. Adverbien wie "langsam" stehen in der Regel vor dieser Vorsilbe.
- ein kleiner am Meer -> ein kleiner Ort am Meer
  - Regel: Das Adjektiv "kleiner" benötigt ein Substantiv, auf das es sich bezieht. Was war auf dem Bild klein und am Meer? Aus dem Kontext (Fahrkarte, Urlaub) ergibt sich "Ort" als logische Ergänzung.
- stand nur ein Satz -> stand nur ein Satz
  - Anmerkung: Dieser Eintrag ist ein Versehen. Ihre ursprüngliche Formulierung "stand nur ein Satz" ist grammatikalisch korrekt. "Satz" ist maskulin, der unbestimmte Artikel lautet "ein". Bitte entschuldigen Sie den Fehler in der Fehlerliste.

- Sie hat fand die Geschichte -> Sie fand die Geschichte
  - Regel: Das Präteritum (Vergangenheitsform) von "finden" ist "fand". Diese Form wird ohne Hilfsverb verwendet. Das Perfekt ("hat gefunden") wäre auch möglich, aber die Mischung aus Hilfsverb ("hat") und Präteritum ("fand") ist falsch.
- zusammen gespielt -> zusammengespielt
  - Regel: Das Verb "zusammenspielen" (to play together) ist ein untrennbares Verb. Im Perfekt wird es nicht getrennt und das Partizip II lautet "zusammengespielt". Die Vorsilbe "zusammen-" bleibt also immer am Verb.
- keine Adresse Telefon -> keine Adresse und keine Telefonnummer
  - Regel: "Adresse Telefon" ist keine korrekte Zusammensetzung. Es handelt sich um zwei verschiedene Dinge. Die korrekte Aufzählung verbindet sie mit "und". "Telefon" allein ist das Gerät; die Nummer, die man anruft, ist die "Telefonnummer".

Anhang 13. Antwortbeispiel von DeepSeek für den fehlerhaften Text auf B1-Niveau

Hier ist die korrigierte und analysierte Version deines Textes.

---

1. Erkannte Fehler

- rief zu unser gemeinsamer Freund Lukas an – Falsche Präposition und Kasus
- wir machen einen Ausflug die Berge – Fehlende Präposition
- weil das Wetter aussah nicht besonders gut – Falsche Wortstellung im Nebensatz
- Ab Lukas holte uns mit seinem Auto – Falsche Präposition
- Die Landschaft war. – Unvollständiger Satz (fehlendes Prädikat)
- Lukas hatte ein kleines Picknick Essen vorbereitet – Falsche Wortverbindung
- Es gab ... Milch – Falsches Wort (lexikalisch-semantische Unstimmigkeit)

2. Korrigierter Text

Letztes Wochenende wollten meine Freundin Klara und ich eigentlich nur zu Hause bleiben und uns entspannen. Wir waren beide ziemlich müde von der langen Arbeitswoche. Doch am Samstagmorgen rief unser gemeinsamer Freund Lukas an. „Ich habe eine Überraschung für euch“, sagte er begeistert. „Zieht euch bequeme Schuhe an, wir machen einen Ausflug in die Berge!“ Anfangs waren wir nicht so überzeugt, weil das Wetter nicht besonders gut aussah. Trotzdem haben wir zugesagt.

Dann holte uns Lukas mit seinem Auto ab. Während der Fahrt regnete es leicht, aber als wir ankamen, schien plötzlich die Sonne. Die Landschaft war wunderschön. Wir wanderten auf einem schmalen Weg, der durch einen dichten Wald führte. Überall war es ruhig und die Luft war frisch. Lukas hatte ein kleines Picknick vorbereitet. Es gab belegte Brote, Obst und Tee. Wir setzten uns auf eine Wiese mit Blick auf ein tiefes Tal und genossen das Essen. Ich musste zugeben, dass Lukas‘ Vorschlag fantastisch war. Obwohl ich anfangs skeptisch war, wurde es ein unvergesslicher Tag. Manchmal sind die spontanen Entscheidungen die besten. Es lohnt sich, offen für neue Pläne zu sein, auch wenn man müde ist. Wir haben beschlossen,

solche Ausflüge in Zukunft häufiger zu unternehmen, denn sie geben einem unglaublich viel Energie für die kommende Woche.

3. Feedback

- rief zu unser gemeinsamer Freund Lukas an → rief unser gemeinsamer Freund Lukas an
  - Warum falsch: 1) Die Präposition "zu" ist bei "anrufen" falsch. 2) "Unser" muss dekliniert werden, da "Freund" im Nominativ steht.
  - Richtige Regel: Das Verb "anrufen" ist transitiv und steht direkt mit dem Akkusativobjekt (wen ruft man an?). Der Possessivartikel muss dem Kasus des Nomens angepasst werden. Hier: Nominativ Maskulinum = "unser gemeinsamer Freund".
- wir machen einen Ausflug die Berge → wir machen einen Ausflug in die Berge
  - Warum falsch: Es fehlt die Präposition, die das Ziel des Ausflugs angibt.
  - Richtige Regel: Das Ziel einer Bewegung wird oft mit "in" + Akkusativ ausgedrückt. Die feste Wendung lautet "einen Ausflug in die Berge/die Stadt/den Wald machen".
- weil das Wetter aussah nicht besonders gut → weil das Wetter nicht besonders gut aussah
  - Warum falsch: Im deutschen Nebensatz steht das konjugierte Verb am Ende. "Aussah" ist das konjugierte Verb.
  - Richtige Regel: Im Nebensatz gilt die Regel: Konjunktion (weil) + Subjekt (das Wetter) + ... + Verb (aussah). Das Verb muss ans Satzende.
- Ab Lukas holte uns mit seinem Auto → Dann holte uns Lukas mit seinem Auto ab.
  - Warum falsch: 1) "Ab" ist eine falsche Präposition für den Zeitpunkt. 2) Das trennbare Verb "abholen" wurde nicht korrekt getrennt.
  - Richtige Regel: Für eine zeitliche Abfolge verwendet man "dann". Das Verb "abholen" ist trennbar. Im Präteritum steht die Vorsilbe "ab-" am Ende des Satzes: "holte uns Lukas ... ab".
- Die Landschaft war. → Die Landschaft war wunderschön.

- Warum falsch: Der Satz ist unvollständig. Das Verb "sein" benötigt eine Ergänzung (ein Adjektiv, Nomen oder Adverb).
- Richtige Regel: Ein Satz mit "sein" als Hauptverb ist nur vollständig, wenn eine Beschreibung oder Gleichsetzung folgt (z.B. "schön", "groß", "ein Traum").

• ein kleines Picknick Essen → ein kleines Picknick

- Warum falsch: "Picknick" ist bereits das vollständige Wort für die Mahlzeit im Freien. "Essen" ist redundant.
- Richtige Regel: "Das Picknick" beschreibt bereits die gesamte Aktivität inklusive des Essens. Ein zusätzliches "Essen" ist nicht nötig.

• Es gab ... Milch → Es gab ... Tee

- Warum falsch: Während Milch theoretisch möglich ist, ist sie bei einem Picknick für Erwachsene untypisch. Ein Getränk wie Wasser, Saft oder Tee passt besser.
- Richtige Regel: Dies ist ein Fehler der lexikalischen Semantik und Stils. "Tee", "Saft" oder "Wasser" sind üblichere Getränke für ein Picknick. "Milch" klingt in diesem Kontext für viele Muttersprachler ungewöhnlich.

Anhang 14. Antwortbeispiel von Gemini für den fehlerhaften Text auf A1-Niveau

Guten Tag! Als erfahrener Deutschlehrer und Korrekturleser helfe ich Ihnen sehr gerne dabei, Ihren Text zu verbessern. Hier ist die vollständige Bearbeitung in den gewünschten Schritten._____

1. Erkannte Fehler

Die folgenden grammatikalischen und lexikalisch-semantischen Fehler wurden im Text identifiziert:

- Wortstellung (Adjektiv/Artikel): „Berlin ist große eine und interessante Stadt."
- Wortart/Kasus (Präposition): „Ich bin Schüler und arbeite auch in Café."
- Verbstellung/Doppelverb (Verb): „Die Arbeit ist macht Spaß und meine Kollegen sind sehr nett."
- Kasus (Präposition/Nomen): „Die Bibliothek ist nicht weit von meine Wohnung."
- Wortstellung (Temporaladverb): „Wir gehen spazieren nach dem Lernen manchmal."
- Lexikalisch-semantisch/Satzbau (Verb/Prädikat fehlt): „Felix hat viele."
- Wortstellung (Nomen/Artikel im Genitiv): „Dann trinken wir eine Cola oder einen Saft Apfel."

---

2. Korrigierter Text

Hallo, ich heiße Leo. Ich bin zwanzig Jahre alt und komme aus Österreich. Ich wohne jetzt in Berlin. Berlin ist eine große und interessante Stadt. Ich lerne hier Deutsch, denn die Sprache ist wichtig für meine Arbeit. Ich bin Schüler und arbeite auch in einem Café. Die Arbeit macht Spaß und meine Kollegen sind sehr nett. Ich habe einen neuen Freund. Er heißt Felix. Felix kommt aus Spanien, aber er spricht schon sehr gut Deutsch. Er ist auch Student. Wir lernen zusammen in der Bibliothek. Die Bibliothek ist nicht weit von meiner Wohnung. Wir gehen nach dem Lernen manchmal spazieren. Felix hat viele Hobbys. Er spielt gern Fußball und hört Musik. Ich spiele nicht gern Fußball, aber ich höre auch gern Musik. Am

Wochenende kochen wir oft zusammen. Felix kann sehr gut kochen. Wir essen gern Pizza oder Nudeln. Dann trinken wir eine Cola oder einen Apfelsaft. Abends sehen wir manchmal einen Film. Die Filme sind oft auf Deutsch. Das ist gut für uns. Wir lernen viele neue Wörter. Berlin ist eine tolle Stadt und ich bin froh, hier zu sein und einen guten Freund wie Felix zu haben. Wir planen auch eine kleine Reise. Vielleicht fahren wir im Sommer nach Hamburg.

---

3. Feedback

Hier sind die Erklärungen zu den gefundenen Fehlern und den korrekten Regeln, übersichtlich in Stichpunkten ohne Tabelle formatiert:

- Fehler: „Berlin ist große eine und interessante Stadt."
  - Erklärung: Das unbestimmte Artikel (eine) muss vor dem Adjektiv (große) stehen.
  - Regel: Im Deutschen steht der Artikel vor dem Adjektiv und dem Nomen: Artikel + Adjektiv + Nomen. Die korrekte Form ist „eine große und interessante Stadt".
- Fehler: „Ich bin Schüler und arbeite auch in Café."
  - Erklärung: Die Präposition "in" verlangt hier den Dativ (wo? in einem Café). Das Substantiv Café muss großgeschrieben werden.
  - Regel: Bei Ortsangaben mit der Präposition "in" fragen wir „Wo?" und verwenden den Dativ: in einem Café. Außerdem werden Nomen großgeschrieben.
- Fehler: „Die Arbeit ist macht Spaß"
  - Erklärung: Das Verb "ist" ist überflüssig. "Spaß machen" ist eine feste Wendung.
  - Regel: In einfachen Sätzen steht nur ein konjugiertes Verb an zweiter Stelle. Die feste Wendung ist: etwas macht Spaß.
- Fehler: „Die Bibliothek ist nicht weit von meine Wohnung."
  - Erklärung: Die Präposition "von" verlangt den Dativ. Das Nomen Wohnung ist feminin (die Wohnung), daher ist die Dativform meiner Wohnung richtig.
  - Regel: Die Präposition "von" fordert immer den Dativ: von + Dativ. Bei femininen Nomen (die Wohnung) wird der Dativ mit der oder meiner gebildet.

- Fehler: „Wir gehen spazieren nach dem Lernen manchmal.“
  - Erklärung: Das Temporaladverb "manchmal" steht idealerweise direkt nach dem konjugierten Verb (oder nach der Zeitangabe nach dem Lernen). Es sollte nicht am Satzende stehen.
  - Regel: Die typische Wortstellung ist: Subjekt (Wir) + Verb (gehen) + Zeit (nach dem Lernen) + Art und Weise/Modalität (manchmal) + Rest (spazieren).
- Fehler: „Felix hat viele.“
  - Erklärung: Der Satz ist unvollständig, es fehlt ein Nomen, das nach "viele" im Plural steht, z. B. Hobbys.
  - Regel: Nach dem Indefinitpronomen "viele" ist ein Nomen im Plural notwendig, damit der Satz grammatikalisch und semantisch korrekt ist.
- Fehler: „Dann trinken wir eine Cola oder einen Saft Apfel.“
  - Erklärung: Die korrekte und natürliche Wortkombination für das Getränk ist das Kompositum (zusammengesetztes Wort) Apfelsaft.
  - Regel: Im Deutschen werden oft zwei Wörter zu einem Kompositum verbunden, um ein neues Nomen zu bilden: Apfel + Saft Apfelsaft.

Anhang 15. Antwortbeispiel von Gemini für den fehlerhaften Text auf A2-Niveau

Gerne übernehme ich als Ihr erfahrener Deutschlehrer und Korrekturleser die Bearbeitung Ihres A2-Textes. Hier sind die drei Schritte zur Fehleranalyse und Korrektur:

---

1. Erkannte Fehler

Die folgenden grammatikalischen und lexikalisch-semantischen Fehler wurden im Text identifiziert:

- Wortstellung (Nebensatz nach als): „Am Morgen, als ich öffnete meinen Briefkasten, fand ich einen Brief..."
- Wortstellung/Trennbarkeit (Lokaladverb/Trennbares Verb): „Ich ging meine Wohnung zurück..."
- Kasus (Präposition/Artikel): „...setzte mich an dem Tisch in der Küche..."
- Wortstellung (Trennbares Verb): „...und aufmachte den Brief langsam."
- Nomen fehlt/Ellipsenfehler: „Auf dem Bild war ein kleiner am Meer."
- Temporale Verdoppelung (Perfekt): „Sie hat fand die Geschichte sehr spannend."
- Lexikalisch/Syntax (Nomen fehlt): „Ich wollte ihm antworten, aber es gab keine Adresse Telefon."

---

2. Korrigierter Text

Letzte Woche hatte ich einen sehr interessanten Tag. Am Morgen, als ich meinen Briefkasten öffnete, fand ich einen Brief ohne Absender. Das war komisch. Normalerweise bekomme ich nur Rechnungen oder Werbung. Ich ging zurück in meine Wohnung, setzte mich an den Tisch in der Küche und machte den Brief langsam auf. Drinnen war eine alte Fahrkarte. Auf dem Bild war ein kleiner Ort/Hafen am Meer. Ich kannte diesen Ort nicht. Auf der Rückseite stand nur ein Satz: „Erinnerst du dich an unseren gemeinsamen Urlaub?" Ich dachte lange nach. Wer könnte mir diesen Brief geschickt haben? Vielleicht ein alter Freund aus der Schule?

Am Nachmittag habe ich meine Freundin angerufen und ihr alles erzählt. Sie fand die Geschichte sehr spannend. Wir haben zusammen überlegt. Plötzlich hatte ich eine Idee. Vor vielen Jahren war ich mit meiner Familie in einem kleinen Dorf in Italien im Urlaub. Es war wunderschön dort. Ich habe dort einen Jungen kennengelernt. Sein Name war Marco. Wir haben den ganzen Sommer zusammen gespielt. Aber danach haben wir den Kontakt verloren. Könnte der Brief von ihm sein? Ich war sehr aufgeregt. Ich wollte ihm antworten, aber es gab keine Adresse oder Telefonnummer. Ich wusste nicht, was ich tun sollte. Also habe ich beschlossen, einfach zu warten. Vielleicht kommt ja noch ein Brief. Dieser Gedanke machte mich glücklich und ein bisschen nervös. Ich hoffe, ich höre bald wieder von ihm.

---

3. Feedback

Hier sind die Erklärungen zu den gefundenen Fehlern und den korrekten Regeln:

- Fehler: „Am Morgen, als ich öffnete meinen Briefkasten, fand ich einen Brief..."
  - Erklärung: Die Konjunktion als leitet einen Nebensatz ein. Im Nebensatz steht das konjugierte Verb (öffnete) immer am Satzende.
  - Regel: In Nebensätzen (eingeleitet durch als) steht das finite Verb an der letzten Position. Korrekt: als ich meinen Briefkasten öffnete.
- Fehler: „Ich ging meine Wohnung zurück..."
  - Erklärung: Das Verb zurückgehen ist ein trennbares Verb (zurück + gehen). Im Präsens/Präteritum steht das Präfix (zurück) am Satzende. Außerdem fehlt die Präposition in für die Richtung (Wohin?).
  - Regel: Bei Bewegung zu einem Ort wird zurück an das Satzende gestellt. Die Richtung wird mit in die Wohnung (Akkusativ) angegeben.
- Fehler: „...setzte mich an dem Tisch in der Küche..."
  - Erklärung: Das Verb setzen (Akkusativ) drückt eine Bewegung aus (Wohin?), daher regiert die Wechselpräposition an den Akkusativ. Die Form dem ist Dativ.

o Regel: Die Wechselpräposition an regiert den Akkusativ bei der Angabe einer Richtung/Bewegung. Maskulinum Akkusativ: an den Tisch.

• Fehler: „...und aufmachte den Brief langsam.“

o Erklärung: Das Verb aufmachen ist ein trennbares Verb. Das Präfix (auf-) muss an der letzten Position stehen.

o Regel: Das Präfix des trennbaren Verbs steht im Hauptsatz an letzter Position. Korrekt: machte den Brief langsam auf.

• Fehler: „Auf dem Bild war ein kleiner am Meer.“

o Erklärung: Der Satz ist unvollständig, es fehlt ein maskulines Nomen nach dem Adjektiv kleiner (z. B. Ort oder Hafen).

o Regel: Das Nomen muss ergänzt werden, um grammatikalisch und semantisch vollständig zu sein.

• Fehler: „Sie hat fand die Geschichte sehr spannend.“

o Erklärung: Hier wurden das Hilfsverb hat (Perfekt) und das Präteritum des Verbs finden (fand) fälschlicherweise kombiniert. Im Kontext einer Erzählung ist das Präteritum (fand) ausreichend und stilistisch flüssiger als Perfekt.

o Regel: Verwenden Sie entweder das Präteritum (fand) oder das Perfekt (hat gefunden), nicht beide Formen.

• Fehler: „...es gab keine Adresse Telefon.“

o Erklärung: Die Wörter Adresse und Telefon (gemeint ist Telefonnummer) müssen durch eine Konjunktion (oder) verbunden werden.

o Regel: Gleichartige Elemente in einer Aufzählung werden durch eine Konjunktion verbunden, und das Nomen Telefonnummer ist präziser.

Anhang 16. Antwortbeispiel von Gemini für den fehlerhaften Text auf B1-Niveau

Gerne übernehme ich als Ihr erfahrener Deutschlehrer und Korrekturleser die Bearbeitung Ihres B1-Textes. Hier sind die drei Schritte zur Fehleranalyse und Korrektur:

---

1. Erkannte Fehler

Die folgenden grammatikalischen und lexikalisch-semantischen Fehler wurden im Text identifiziert:

- Wortstellung (Hauptsatz/Temporaladverb): „Doch am Samstagmorgen rief zu unser gemeinsamer Freund Lukas an.“
- Präposition fehlt (Lokalangabe): „...wir machen einen Ausflug die Berge!“
- Wortstellung (Nebensatz): „...weil das Wetter aussah nicht besonders gut.“
- Konjunktion/Präposition (Falscher Satzbeginn): „Ab Lukas holte uns mit seinem Auto.“
- Nomen/Adjektiv fehlt (Ellipsenfehler): „Die Landschaft war.“
- Kompositum (Syntax/Nomen): „Lukas hatte ein kleines Picknick Essen vorbereitet.“
- Lexikalisch/Syntax (Nomen): „...Es gab belegte Brote, Obst und Milch.“
- Temporale Konjunktion fehlt: „Wir sind bis zum Abend frühen geblieben...“ (Dieser Fehler scheint aus einem vorherigen Text übernommen worden zu sein, ist aber im vorliegenden Text nicht zu finden. Ich nehme an, Sie meinen Obwohl ich anfangs skeptisch war, wurde es ein unvergesslicher Tag.) Korrektur: Der Satz ist grammatikalisch korrekt, aber die Formulierung ist redundant, da Obwohl bereits die Konzessivbedeutung liefert. Ich sehe aber einen fehlenden Satzanfang oder ein fehlendes Verb bei den letzten Sätzen. Ich korrigiere den Satzanfang zur besseren Kohärenz.

---

2. Korrigierter Text

Letztes Wochenende wollten meine Freundin Klara und ich eigentlich nur zu Hause bleiben und uns entspannen. Wir waren beide ziemlich müde von der langen Arbeitswoche. Doch

am Samstagmorgen rief unser gemeinsamer Freund Lukas an. „Ich habe eine Überraschung für euch“, sagte er begeistert. „Zieht euch bequeme Schuhe an, wir machen einen Ausflug in die Berge!“ Anfangs waren wir nicht so überzeugt, weil das Wetter nicht besonders gut aussah. Trotzdem haben wir zugesagt.

Dann holte uns Lukas mit seinem Auto ab. Während der Fahrt regnete es leicht, aber als wir ankamen, schien plötzlich die Sonne. Die Landschaft war atemberaubend / wunderschön. Wir wanderten auf einem schmalen Weg, der durch einen dichten Wald führte. Überall war es ruhig und die Luft war frisch. Lukas hatte ein kleines Picknick vorbereitet. Es gab belegte Brote, Obst und Wasser / Saft. Wir setzten uns auf eine Wiese mit Blick auf ein tiefes Tal und genossen das Essen. Ich musste zugeben, dass Lukas' Vorschlag fantastisch war. Obwohl ich anfangs skeptisch war, wurde es ein unvergesslicher Tag. Manchmal sind die spontanen Entscheidungen die besten. Es lohnt sich, offen für neue Pläne zu sein, auch wenn man müde ist. Wir haben beschlossen, solche Ausflüge in Zukunft häufiger zu unternehmen, denn sie geben einem unglaublich viel Energie für die kommende Woche.

---

3. Feedback

Hier sind die Erklärungen zu den gefundenen Fehlern und den korrekten Regeln:

- Fehler: „Doch am Samstagmorgen rief zu unser gemeinsamer Freund Lukas an.“
  - Erklärung: Das Präfix des trennbaren Verbs anrufen (an) muss am Satzende stehen. Die Präposition zu ist falsch und gehört hier nicht hin.
  - Regel: Im Hauptsatz steht das konjugierte Verb an Position 2 und das Präfix des trennbaren Verbs am Satzende.
- Fehler: „...wir machen einen Ausflug die Berge!“
  - Erklärung: Es fehlt die Präposition zur Angabe des Ziels der Bewegung (Wohin?). Die Bewegung in eine Region erfordert die Präposition in (+ Akkusativ).
  - Regel: Die Wechselpräposition in regiert den Akkusativ bei einer Richtungsangabe. Korrekt: in die Berge.

- Fehler: „...weil das Wetter aussah nicht besonders gut.“
  - Erklärung: Die Konjunktion weil leitet einen Nebensatz ein. Im Nebensatz steht das konjugierte Verb (aussah) immer an der letzten Position.
  - Regel: In Nebensätzen (eingeleitet durch weil) steht das finite Verb am Satzende. Korrekt: aussah.
- Fehler: „Ab Lukas holte uns mit seinem Auto.“
  - Erklärung: Das Wort Ab ist hier falsch und soll wahrscheinlich als Temporaladverb den Übergang signalisieren. Die korrekte Konjunktion/Adverb in diesem Kontext ist Dann. Das Verb abholen ist trennbar.
  - Regel: Verwenden Sie das korrekte Adverb Dann für die zeitliche Abfolge. Das Präfix ab gehört an das Satzende.
- Fehler: „Die Landschaft war.“
  - Erklärung: Der Satz ist unvollständig. Nach dem Verb sein (war) fehlt ein Adjektiv oder Nomen (z. B. atemberaubend oder wunderschön), das die Landschaft beschreibt.
  - Regel: Das Verb sein benötigt ein Prädikativ (Adjektiv/Nomen) zur Vervollständigung.
- Fehler: „Lukas hatte ein kleines Picknick Essen vorbereitet.“
  - Erklärung: Die Nomen Picknick und Essen sind redundant, da ein Picknick bereits eine Mahlzeit im Freien impliziert. Eines der Nomen muss weggelassen werden.
  - Regel: Das Nomen Essen muss entfernt werden. Die korrekte Wendung ist ein Picknick vorbereiten.
- Fehler: „Es gab belegte Brote, Obst und Milch.“
  - Erklärung: Das Nomen Milch (milk) passt semantisch weniger gut zu einem Picknick-Getränk als Wasser oder Saft, und Milch wird oft nicht bei Wanderungen mitgenommen.
  - Regel: Das lexikalisch passendere Nomen für ein Picknick-Getränk ist Wasser oder Saft.

GAZİLİ OLMAK AYRICALIKTIR…