Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Fakultät für Mathematik, Informatik und Naturwissenschaften
Arbeitsbereich Wirtschaftsinformatik,
Sozio-Technische Systemgestaltung (WISTS)

Bachelorarbeit

Untersuchung und Evaluation der optimierten Nutzung von Prompt Engineering in Sprachmodellen zur effizienten und effektiven Klausurvorbereitung von Studierenden

Erstgutachter: Lucas Memmert
Zweitgutachterin: Prof. Dr. Eva Bittner

Vorgelegt von:
Laurin Wesselkamp, 7502749
Martinistraße 24, 49078 Osnabrück
l.wesselkamp@icloud.com
Wirtschaftsinformatik
Abgabe: 14.03.2025
Osnabrück, den 11.03.2025

Zusammenfassung

Diese Arbeit untersucht, wie sich das Prompt Engineering (PE) und die Bereitstellung von Affordanzen für Large Language Models (LLMs) auf die Klausurvorbereitung, sowohl unter effektiven als auch effizienten Gesichtspunkten verhält. Explizit kam es zur Verwendung von ChatGPT als LLM. Um die Affordanzen und PE Techniken gebündelt zur Verfügung zu stellen, wurde ein Prompt Handbook entworfen. Zur Untersuchung der Effizienz und Effektivität wurde ein Between-Groups-Experiment durchgeführt, bei der sich auf einen Multiple-Choice-Test (MCT) vorbereitet wurde. In der Vorbereitung nutzte eine Gruppe kein ChatGPT, eine Gruppe ChatGPT uneingeschränkt und wiederum eine ChatGPT mit dem Prompt Handbook. Eine Evaluation der Ergebnisse zeigte, dass das Prompt Handbook, aber auch die Verwendung von ChatGPT in der *kurzfristigen* Klausurvorbereitung keinen Einfluss auf die Testergebnisse hatte. Gleichzeitig kam es aber auch nicht zu Verzerrungen in der Wahrnehmung des Lernfortschritts durch den Chatbot. Objektiv war die Nutzung des Prompt Handbooks tendenziell ineffizient, subjektiv nahmen viele Studierende eine Effizienzsteigerung wahr. Zur effektiven Nutzung trug das Prompt Handbook objektiv insofern bei, als dass es bisher unbekannte Affordanzen für Studierende bereitstellte. Zudem wurde subjektiv klar zugestimmt, dass ChatGPT effektiv das Verständnis gefördert hat. Vom Prompt Handbook wurden insgesamt weniger technische Prompt Techniken, die die Halluzination verringern, genutzt, sondern vielmehr konkrete Affordanzen, um mit ChatGPT in unterschiedlichen Rollen zu kommunizieren.

Abstract

This thesis investigates how prompt engineering (PE) and the provision of affordances for Large Language Models (LLMs) relate to exam preparation, both from an effective and efficient point of view. Explicitly, ChatGPT was used as an LLM. A Prompt Handbook was designed to make the affordances and PE techniques available in a bundled form. To analyse the efficiency and effectiveness, a between-groups experiment was conducted in which the participants prepared for an mutliple choice test (MCT). During preparation, one group did not use ChatGPT, one group used ChatGPT without restrictions and one group used ChatGPT with the Prompt Handbook. An evaluation of the results showed that the Prompt Handbook, as well as the use of ChatGPT in the *short-term* exam preparation, had no influence on the test results. At the same time, however, there were no distortions in the perception of learning progress through the chatbot. Objectively, the use of the Prompt Handbook tended to be inefficient; subjectively, many students perceived an increase in efficiency. Objectively, the Prompt Handbook contributed to effective use in that it provided previously unknown affordances for students. In addition, there was clear subjective agreement that ChatGPT effectively promoted understanding. Overall, the Prompt Handbook was used less for technical prompt techniques that reduce hallucination and more for concrete affordances to communicate with ChatGPT in different roles.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

LM	Language Model
LLM	Large Language Model
NLP	Natural Language Processing
RLHF	Reinforcement Learning from Human Feedback
PE	Prompt Engineering
CoT	Chain of Thought
ToT	Tree of Thoughts
EOL	Ease-of-Learning
JOL	Judgement-of-Learning
FOK	Feeling-of-Knowing
KI	Künstliche Intelligenz
AIEd	Artificial Intelligence in Education
GenKI	Generative Künstliche Intelligenz
MCT	Multiple-Choice-Test
MCQ	Multiple Choice Question
G-1	Gruppe 1, traditionelle Lernmethoden
G-2	Gruppe 2, uneingeschränkte Nutzung von ChatGPT
G-3	Gruppe 3, ChatGPT in Verbindung mit PH
Q-x	Frage x des MCT
PH	Prompt Handbook

1 Einleitung

1.1 Problemstellung und Relevanz

KI-basierte Hilfsmittel sind längst fester Bestandteil des Studiums und werden von Studierenden aller Fachrichtungen verwendet (Von Garrel und Mayer, 2023, 8). Fast die Hälfte der Studierenden in Deutschland nutzt dabei explizit ChatGPT und 12,8 % verwenden KI-basierte Hilfsmittel für die Klausurvorbereitung (Von Garrel und Mayer, 2023, 6f.). Large Language Models (LLMs) wie ChatGPT haben nach Ansicht vieler Studierender ein hohes Potenzial, die Klausurvorbereitung effizienter und effektiver zu gestalten (Ngo, 2023, 14). LLMs sind große computergestützte Sprachmodelle, die menschliche Sprache in Form von textuellen Eingaben verstehen und darauf reagieren können (Chang et al., 2024, 3). Für die effiziente und effektive Klausurvorbereitung ergeben sich allerdings zwei Herausforderungen. Zum einen sind sich viele Studierende der Möglichkeiten von Sprachmodellen nicht bewusst. Zum anderen neigen Chatbots dazu, falsche Informationen zu generieren bzw. zu *halluzinieren* (Amatriain, 2024, 4).

Insbesondere zur Reduktion der Halluzination kann das Prompt Engineering (PE) beitragen. Denn PE fungiert als Bindeglied zwischen menschlicher Sprache und maschinellem Verständnis (Chan und Colloton, 2024, 80). Durch gutes PE können Fehler reduziert und Anforderungen der Studierenden klarer formuliert werden. Eine Studie zeigte, dass gezieltes PE die Lernergebnisse von Studierenden stabilisieren kann (Bastani et al., 2024). Zudem wurden im Bereich der Lehre bereits spezifische Prompts zusammengetragen, die Studierenden die Nutzung von LLMs erleichtern soll (Mohr, 2024). Explizit beschreiben Jacobsen und Weber (2023b, 27) ChatGPT als ein sinnvolles Hilfsmittel in der Bildung mit der Prämisse, dass Lehrende geübt im PE sind.

Die Wichtigkeit von PE in der Bildung ist somit einleuchtend. Jedoch fehlt es vor allem an einer Betrachtungsweise für Studierende. Insbesondere Erkenntnisse, inwieweit sich die Verwendung von ChatGPT im Vergleich zur Nutzung von ChatGPT mit konkret bereitgestellten, vormodellierten Prompts auf die Effektivität und Effizienz von Studierenden auswirkt. Gerade im Vergleich zu traditionellen Lernmethoden der Studierenden, könnte die Nutzung von Sprachmodellen eine Effizienz und Effektivitätssteigerung bedeuten. Diese Forschungslücke bildet den Ausgangspunkt für die vorliegende Arbeit.

1.2 Zielsetzung und Forschungsfragen

Das Ziel dieser Arbeit ist es, zu erforschen, inwiefern durch abgestimmtes Prompt Engineering (PE) und die Bereitstellung von Nutzungsmöglichkeiten der Sprachmodelle, die Effizienz und Effektivität von Studierenden bei der Klausurvorbereitung gesteigert werden kann. Um über diese Fragen eine Aussage zu treffen, kommt es zur Beantwortung der folgenden Teilfragen:

1. Welche Prompt Engineering Techniken sind am effektivsten für die Klausurvorbereitung, um relevante Informationen aus vorhandenen Lehrmitteln vertiefend zu verstehen?

2. Inwiefern wirkt sich der Einsatz von Prompt Engineering bei Sprachmodellen, verglichen zur Nutzung von traditionellen Lernmethoden und dem allgemeinen Einsatz von Sprachmodellen, auf die Lernergebnisse von Studierenden aus?

3. Inwieweit steigert die Verwendung von Prompt Engineering-Techniken bei Sprachmodellen die Effizienz und Effektivität der Klausurvorbereitung von Studierenden?

4. Welche Vor- und Nachteile ergeben sich aus der Nutzung von Sprachmodellen mit vormodellierten Prompts und der „ad hoc“ Verwendung von Sprachmodellen?

1.3 Methodisches Vorgehen

Zur Beantwortung der 1. Forschungsfrage wurde eine Literaturrecherche zum PE in den unterschiedlichsten Bereichen durchgeführt. Anschließend wurden die effektivsten spezifischen Prompt Techniken und Verwendungsmöglichkeiten von ChatGPT im Prompt Handbook eingepflegt und den Studierenden bereitgestellt.

Zur weiteren Beantwortung aller Forschungsfragen fand ein Mixed-Methods Between-Groups-Design statt. Hierbei meldeten sich 51 Teilnehmer:innen freiwillig für das Experiment. Diese wurden in insgesamt drei Gruppen eingeteilt, denen jeweils andere Hilfsmittel in der Vorbereitungsmethode zur Verfügung standen. G-1 fungierte als Kontrollgruppe und nutzte kein ChatGPT, G-2 nutzte ChatGPT uneingeschränkt und G-3 verwendete ChatGPT in Verbindung mit dem Prompt Handbook. Im Anschluss an die Vorbereitungsphase wurde das Verständnis der Studierenden durch einen, für alle Gruppen identischen Multiple-Choice-Test (MCT), überprüft und die subjektive Einschätzung hinsichtlich der Nützlichkeit erfasst.

1.4 Aufbau der Arbeit

In diesem Kapitel wurden die Motivation, sowie die Forschungsfragen erläutert. Kapitel 2 gibt einen Überblick über Large Language Models, Prompt Engineering, allgemeine Lerntechniken, zwei kognitive Modelle sowie Prompts für das Lernen. Kapitel 3 erläutert die Methodik für das abgeleitete Prompt Handbook und den Ablauf des Between-Groups Experiments. In Kapitel 4 werden die Ergebnisse getrennt nach den Forschungsfragen aufgeführt. Schließlich findet in Kapitel 5 eine Evaluierung der Ergebnisse statt sowie eine kritische Betrachtung der Limitationen und ein genereller Ausblick. Kapitel 6 fasst die Erkenntnisse dieser Arbeit abschließend zusammen.

2 Theoretischer Hintergrund

Der theoretische Hintergrund beinhaltet relevante Hintergrundinformationen, die im Kontext dieser Arbeit von Bedeutung sind. Darunter fallen Large Language Models, Prompt Engineering, Prompt Techniken, Lerntechniken, Metakognition, Bloom's Taxonomy, LLMs in der Bildung und Prompts für die Hochschulbildung.

2.1 Large Language Models und Prompt Engineering

Nachfolgend werden die Large Language Models (LLMs) vorgestellt. Dabei wird verstärkt ChatGPT betrachtet. Anschließend erfolgt eine Einführung in das Prompt Enigneering (PE).

2.1.1 Einführung in LLM und GPT

Nach Chang et al. (2024, 3) sind einfache *Language Models* (LM) computergestützte Modelle, die in der Lage sind, menschliche Sprache zu „verstehen" und zu erzeugen. Ihre Kompetenz liegt darin, Wahrscheinlichkeiten von Wortfolgen vorherzusagen oder durch eine Eingabe einen neuen Text zu generieren. Durch die steigende Parameteranzahl und zunehmende Datensatzgröße für das Training von LMs hat sich schließlich der Begriff *Large Language Model* (LLM) etabliert (Zhao et al., 2024, 13). Im Allgemeinen sind LLMs KI-Systeme, die sich auf das *Natural Language Processing* (NLP) spezialisiert haben. NLP stellt den Teilbereich von Künstlicher Intelligenz dar, der sich mit der Verarbeitung und Erzeugung von natürlicher Sprache befasst (Seemann, 2023, 10). Auch bei den GPT-Modellen von OpenAI handelt es sich um LLMs. Neben den Modellen von OpenAI gibt es jedoch unzählige weitere Modelle, wie *Claude 3.5 Sonnet* von Anthropic (2024), *Gemini* von Google (2023) oder *DeepSeek-R1* von DeepSeek (2025).

Diese Arbeit fokussiert sich auf die Webanwendung ChatGPT von OpenAI (2022). GPT steht dabei für *Generative Pre-Trained Transformer*. *Generative* meint in diesem Kontext die Erzeugung von Sätzen (Dudenredaktion, 2024) durch das Modell. Die *Pre-Training*-Aufgabe ist bei den GPT-Modellen die Sprachmodellierung, die die Vorhersage des nächsten Wortes in einer Sequenz beschreibt (Thirunavukarasu et al., 2023, 1931). Übrig bleibt der *Transformer*. Dabei handelt es sich um eine Architektur, die vollständig auf dem im Jahr 2017 entwickelten Aufmerksamkeitsmechanismus basiert (Vaswani et al., 2017, 1).

Im Folgenden soll ein Verständnis des, für die Arbeit relevanten, LLMs vermittelt werden, um die bestmöglichen Prompts für die GPT-Reihe zu generieren. Dabei wird grob die zugrundeliegende Architektur der Modelle erläutert, sowie die Funktionsweise bei Eingabe eines Prompts. Erwähnenswert ist, dass einige Aspekte im Folgenden vereinfacht dargestellt werden und dass der Zugang zum Modell hinter ChatGPT nicht öffentlich ist, weshalb nur eingeschränkte Informationen zur Verfügung stehen.

Um die Architektur der LLMs zu verstehen, wird zunächst das Konzept der *Tokenization* erläutert. Ein Token stellt die linguistische Einheit für ein LLM dar. Tokens können Wortbestandteile, Worte oder vollständige Sätze umfassen (Phoenix und Taylor, 2024, 41). Bei der *Tokenization* wird ein Text in einzelne Tokens oder Wortbestandteile zerlegt, sodass das LLM in der Lage ist, einen Text in einzelnen Einheiten zu verarbeiten. Nach der *Tokenization* lassen sich die Tokens als separate Vektoren darstellen (Campesato, 2023, 34f.). Die Abbildung der Tokens auf einen Vektor ist entscheidend, damit das LLM den Text in Form von Zahlen interpretieren kann. Mathematisch lässt sich ein Token $\mathbf{t}$ auf einen Vektor $\vec{\mathbf{v}}$ wie folgt abbilden:

$$\mathbf{t} \rightarrow \vec{\mathbf{v}} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix}$$

Solche *Wortvektoren* existieren jedoch in einem großen multidimensionalen Raum. Beispielsweise ist der *Wortvektor* für das GPT-3 Modell in einem Raum mit 12.288 Dimensionen (Brown et al., 2020, 8). Diese 12.288 Koordinaten des Vektors repräsentieren allerdings nicht nur das Token selbst, sondern auch dessen semantische und syntaktische Beziehungen[1]. Die Abbildung der Tokens auf *Wortvektoren* erfolgt während der Trainingsphase der Modelle. Hierbei werden die Modelle so konzipiert, dass Tokens mit ähnlicher Bedeutung im hochdimensionalen Raum näher beieinander liegen, als Tokens mit unterschiedlicher Bedeutung. Zum Beispiel wäre das Wort „Hochhaus“ näher an dem Wort „Gebäude“ gelegen, als das Wort „Baum“. Das Konzept der *Tokenization* und Abbildung der Tokens auf *Wortvektoren* im multidimensionalen Raum ermöglicht es den Modellen, die komplexe Sprache zu erfassen (Phoenix und Taylor, 2024, 42f.).

Doch was passiert nun genau, wenn ein Prompt in das Eingabefeld von ChatGPT eingegeben wird? Wie in der Definition von *Large Language Models* und dem ***G**enerative **P**retrained **T**ransformer* erwähnt, beabsichtigt das GPT-Modell bei einer gegebenen Textsequenz $\mathbf{T}$, das nächste Token $\mathbf{n}$ zu prognostizieren. Beim Trainieren wird das Modell darauf getrimmt, die Wahrscheinlichkeit für eine Tokenfolge in Abhängigkeit vom Kontext zu maximieren. Also $P(n|T) = P(n|t_1, t_2, ..., t_{p-1})$, wobei $\mathbf{t_1}, \mathbf{t_2}, ..., \mathbf{t_{p-1}}$ die Token in der Textsequenz und $\mathbf{p}$ die aktuelle Position darstellt (Chang et al., 2024, 5). Das heißt, für die Textgenerierung des Models wird zuerst eine gegebene Textsequenz $\mathbf{T}$ *tokenized* und danach durch mehrere *Wortvektoren* dargestellt. Diese *Wortvektoren* dienen dann als Input für das Modell. Anschließend wird eines der Transformer Modelle wie GPT-4 genutzt, um das nächst passende Token zu bestimmen:

$$\text{argmax}_n \, P(n \mid t_1, t_2, \ldots, t_{p-1})$$

Die Ausgabe ist schließlich das nächst wahrscheinliche Token, welches mit der vorherigen Textsequenz kombiniert wird (Phoenix und Taylor, 2024, 45).

Das GPT-3 Modell erfüllt lediglich diese Aufgabe. Allerdings geht die Funktionalität von ChatGPT darüber hinaus. ChatGPT vervollständigt auf einer gegebenen Eingabe nicht den Text, sondern agiert als hilfreicher Chatbot (OpenAI, 2022). Möglich gemacht wurde dies unter anderem durch das *Reinforcement Learning from Human Feedback* (RLHF), mit ähnlichen Methoden des *InstructGPT* Modells (Ouyang et al., 2022), dem Vorgänger-Modell von ChatGPT. Beim RLHF kommt es zur Anpassung von ChatGPT an menschliche Präferenzen (Ray, 2023, 123f.). Die gesamte Erstellung von ChatGPT umfasste deutlich mehr Schritte (Ouyang et al., 2022, 3), die für die Arbeit allerdings weniger relevant sind.

LLMs können zweifellos herausragende sprachliche Fähigkeiten zugeschrieben werden. Jedoch weisen sie Einschränkungen auf, die bei ihrer Nutzung beachtet werden müssen. Als Beispiele für die Limitationen von LLMs können die nicht deterministischen Antworten angeführt werden, die LLMs geben. Diese können selbst bei Gleichartigkeit eines Prompts auftreten. Diese Varianz der Antworten geht mit einer Neigung von LLMs einher, zu *halluzinieren*. Dabei handelt es sich um ein Phänomen, bei dem LLMs faktisch falsche Antworten produzieren. Darüber hinaus benötigen LLMs häufig domänenspezifische Informationen, um eine Aufgabe bewältigen zu können. Ergänzend treten weitere Herausforderungen zutage, wie die erhebliche Ressourcenintensität der Modelle und ihre Unfähigkeit, aktuelle Daten zeitnah zu integrieren (Amatriain, 2024, 4f.). Mit Blick auf die Funktionsweise großer Sprachmodelle ist entscheidend zu verstehen, dass deren Ausgaben auf Wahrscheinlichkeiten basieren, die aus einem enormen Trainingssatz gelernt wurden. Das Sprachmodell besitzt kein eigenes Verständnis der generierten Inhalte. Es ist vielmehr ein *stochastischer Papagei* (Bender et al., 2021, 616f.). Ein großes Problem stellt diesbezüglich auch der Bias, also die Verzerrung der Antworten von großen Sprachmodellen, dar. Dabei macht es einen Unterschied mit welchen Daten sie trainiert wurden (Weidinger et al., 2021, 36). Durch das RLHF kann es dabei ebenfalls zu Verzerrungen kommen. Schließlich gibt es viele weitere Limitationen und Bedenken,

[1] In der Literatur wird oft anstelle der Bezeichnung *Wortvektor* der Begriff *Word Embedding* verwendet.

die auch den Datenschutz, das Urheberrecht, aber auch explizite Auswirkungen auf der Entwicklung von Lernenden umfassen (Cong-Lem et al., 2024, 10). Diese Bedenken im Bezug auf das Lernen werden in Kapitel 2.3.1 vertieft. Gerade die Halluzination ist ein Problem für diese Arbeit. Durch gutes Prompt Engineering kann dem jedoch entgegengewirkt werden (Chen et al., 2023, 1). Aus diesem Grund folgt nun Kapitel 2.1.2, was eine Einführung in das Prompt Engineering gibt, sowie relevante Methoden, um die genannten Probleme zu reduzieren.

2.1.2 Einführung in das Prompt Engineering

Innerhalb der Welt der LLMs gibt es einige Ansätze zur Anpassung eines Modells an die individuellen Bedürfnisse. Eines davon ist das *Prompt Engineering* (PE). Das PE kann dabei grundsätzlich von jeder Person durchgeführt werden. Im Folgenden findet eine Auseinandersetzung mit dem PE statt. Hierbei werden zunächst Definition, Elemente und Haupttechniken des PE betrachtet sowie spezifische und konkrete Prompt Beispiele.

Primär ist ein *Prompt* (dt.: Eingabeaufforderung) eine Reihe von Anweisungen an ein LLM, welches den Kontext für die Konversation zwischen Nutzer und LLM setzt. Es gibt dem LLM vor, welche Informationen wichtig sind, sowie Form und Inhalt der Ausgabe (White et al., 2023, 1). Das *Prompt Engineering* (PE) bezeichnet die systematische Gestaltung und Optimierung von *Prompts* mit dem Ziel einer effizienten Nutzung von LLMs. Zu Beginn wurde beim PE lediglich die Gestaltung einzelner Prompts untersucht, um ein bestimmtes Ergebnis vom LLM zu erhalten. Inzwischen hat sich daraus ein umfangreiches Forschungsgebiet mit Methoden und „Best Practices“ entwickelt. Um das volle Potenzial von LLMs auszuschöpfen und sie für eine Reihe von Bereichen besser zugänglich und anwendbar zu machen, ist das PE von großer Bedeutung (Chen et al., 2023, 2).

Ein Prompt besteht in der Regel aus 4 Elementen:

1. *Die Anweisung*: Eine spezifische Aufgabe, die das Modell in Richtung des gewünschten Ergebnisses lenkt.
2. *Der Kontext*: Externe Informationen oder ein zusätzlicher Hintergrund, der dem Modell dabei helfen kann, präzisere oder relevantere Antworten zu geben.
3. *Die Eingabedaten*: Die eigentliche Eingabe, auf die das Modell eine Antwort finden soll. Sie fungiert als Hauptgegenstand des gesamten Prompts und bestimmt das Verständnis des Modells von der Aufgabe.
4. *Der Ausgabeindikator*: Das Ausgabeformat. Um das gewünschte Ausgabeformat der Antwort direkt zu erhalten, ist es hilfreich, den gewünschten Antworttyp vorher zu definieren.

Es ist von entscheidender Bedeutung, dass diese Elemente verstanden werden, da sie eine wesentliche Grundlage dafür bilden, Intentionen an das LLM zu übermitteln, sodass eine tiefgründige und präzise Antwort gewährleistet werden kann (Giray, 2023, 2630).

Die Landschaft an PE Techniken ist äußerst umfangreich. Beispielsweise werden in Sahoo et al. (2024) insgesamt 29 verschiedene Techniken vorgestellt. Diese reichen von ganz basierten Methoden, wie dem *Zero-Shot* oder *Few-Shot* Prompting, bis hin zu komplexeren Methoden, wie dem *Chain Of Code* Prompting (Sahoo et al., 2024, 1). Allerdings gibt es noch weitaus mehr. So werden auf der Webseite von DAIR.AI (2024a) insgesamt 135 Paper zu verschiedenen PE Technik Ansätzen gelistet.
Im Folgenden werden zunächst die relevantesten, grundlegendsten Prompt Techniken für diese Arbeit vorgestellt.

Zero-Shot Prompting: Diese Technik wurde erstmals von Radford et al. (2019) vorgestellt. Dabei ist ein LLM, ohne umfangreiche Beispieldaten (im Prompt), in der Lage eine Antwort zu liefern (Sahoo et al., 2024, 2). Schon im Jahr 2019 wurden mit dieser Methode eindrucksvolle Aufgaben wie das Zusammenfassen oder das Übersetzen von Texten mittels Prompts demonstriert (Liu et al., 2021, 19). Das *Zero-Shot Prompting* gilt heute als veraltete Methode. Stattdessen sollte das Instruction Tuning verwendet werden (DAIR.AI, 2024b). Dennoch ist es eine grundlegende Technik, worauf viele weitere aufbauen.
Few-Shot Prompting: Im Rahmen dieser Arbeit kommt das *Few-Shot Prompting* zum Einsatz. Erstmalig wurde dieses bei GPT-3 (Brown et al., 2020) eingesetzt. Im Gegensatz zum *Zero-Shot Prompting* werden hierbei lediglich wenige Beispiele im Prompt mitgegeben. Dies resultiert in einer verbesserten Performance des Modells, insbesondere mit Blick auf komplexere Aufgaben. Wie in 2.1.1 dargelegt, führt ein Prompt, mit Beispielen oder Text, zu einer erhöhten Tokenanzahl. Dies kann vorwiegend bei der Verarbeitung längerer Texte zu Schwierigkeiten führen. Zudem ist die Zusammensetzung der Beispiele zu berücksichtigen, da diese durch ihre spezifische Konstellation das Modellverhalten verzerren können (Sahoo et al., 2024; Liu et al., 2021, 2; 13f.).
Prompt Augmentation: In Verbindung mit dem *Few-Shot Prompting* kann die *Prompt Augmentation* bzw. das *Demonstartion Learning* angeführt werden. Dabei wird dem LLM mit einigen Beispielen demonstriert, wie es die Antwort auf den tatsächlichen Prompt bereitstellen soll (Liu et al., 2021; Memmert et al., 2024, 13; 7523).

Die zuvor dargelegten Techniken bilden bereits eine solide Grundlage für effektives PE und einen verständnisvollen Umgang mit LLMs. Gerade das *Few-Shot Prompting* und die *Prompt Augmentation* werden im Prompt Handbook vor allem indirekt in den einzelnen Prompts verwendet. Bei der Bearbeitung komplexer Aufgaben, insbesondere im Kontext von Argumentationen und Analysen, weist die Genauigkeit der Modellausgaben allerdings noch Optimierungspotential auf (Chen et al., 2023, 6). Da der Test auch mathematische Aufgaben beinhaltet, die nur eine Lösung zulassen, werden folgend zwei fortgeschrittene Methoden dargestellt. Diese sollen dazu beitragen, dass das Modell präziser arbeitet und weniger halluziniert. Die beiden folgenden Techniken wurden explizit in das Prompt Handbook für die Studierenden integriert.

Chain of Thought: Die von Google-Forscher:innen (Wei et al., 2023) vorgestellte *Chain of Thought* (CoT) Technik zählt zu den am weit verbreitetsten Methoden im Bereich des PE. Durch sie wurde ein bedeutender Fortschritt in der Nutzung von LLMs zur Verbesserung ihrer Argumentationsfähigkeit möglich. CoT transformiert implizite Argumentationsschritte, die ein Modell zur Beantwortung einer Frage durchläuft, in explizite ausgeschriebene Sequenzen. Die Anwendung dieser Technik führt zu einer signifikanten Steigerung der Modellfähigkeit, logische Schlussfolgerungen zu ziehen. Dabei kann zwischen zwei Varianten unterschieden werden, dem *Zero-Shot-CoT* und dem *Manual-CoT* (Amatriain, 2024, 13). Die von Amatriain (2024) benannte *Manual-CoT* Technik wurde zunächst von Wei et al. (2023) eingeführt. Diese Methode nutzt das oben genannte *Few-Shot Prompting*, wobei dem Modell durch beispielhafte Lösungswege demonstriert wird, wie es seine Antworten schrittweise herleiten soll (Wei et al., 2023, 1). Da diese Herangehensweise allerdings ein tiefergreifendes Verständnis der jeweiligen Themenbereiche voraussetzt, verwendet die vorliegende Arbeit vor allem das von Kojima et al. (2023) entwickelte *Zero-Shot-CoT* Verfahren. Das vereinfachte Verfahren kommt ohne vorgegebene Beispiele aus. Stattdessen wird durch die Aufforderung *Let's think step by step*, ein strukturierter Denkprozess beim Modell eingeleitet, bevor es die eigentliche Antwort gibt (Kojima et al., 2023, 2). Diese Methode erzielt vergleichbare Ergebnisse mit deutlich geringerem Aufwand. Eine beispielhafte Gegenüberstellung beider Methoden ist in Abbildung 1 zu finden.

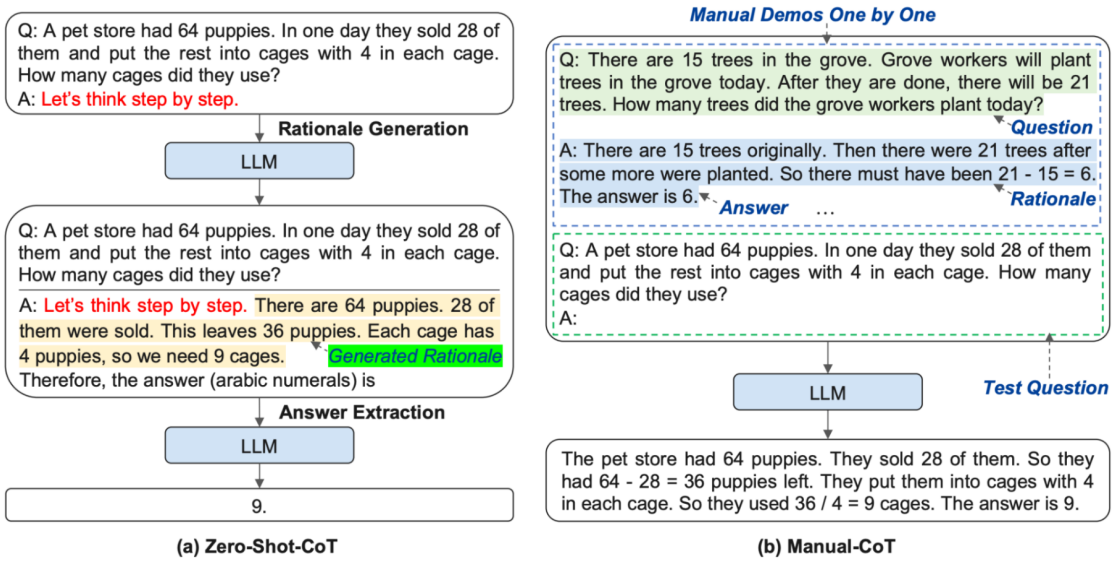

Abbildung 1: Gegenüberstellung von *Zero-Shot-CoT* und *Manual-CoT*

Notiz: Das *Zero-Shot-CoT* stammt von Kojima et al. (2023) und das *Manual-CoT* von Wei et al. (2023) und wird mit beispielhaften Ein- und Ausgaben eines LLMs illustriert. Die Grafik wurde übernommen aus Amatriain (2024, 14).

Tree of Thoughts: Die Tree of Thoughts (ToT)-Technik nutzt einen hierarchischen Ansatz zur Prompt-Verarbeitung, der sich an einer Baumstruktur orientiert (Chen et al., 2023, 10). Ursprünglich wurde dieser Ansatz von Yao et al. (2023) entwickelt. In der vorliegenden Arbeit wird jedoch eine von Hulbert (2023) adaptierte, vereinfachte Variante verwendet, die sich auf die grundlegenden Prinzipien des ToT-Frameworks konzentriert. Der zentrale Aspekt dieser Anpassung liegt in der gezielten Anregung des Modells, verschiedene Antwortalternativen zu generieren, indem ein spezifisch erstellter Prompt verwendet wird. Die empfohlene Promptvorlage lautet wie folgt: „*Stelle dir drei verschiedene Experten vor, die diese Frage beantworten. Alle Experten schreiben einen Schritt ihrer Überlegungen auf und teilen ihn dann der Gruppe mit. Dann gehen alle Experten zum nächsten Schritt über, usw.. Wenn einer der Experten zu irgendeinem Zeitpunkt merkt, dass er falsch liegt, verlässt er die Gruppe. Die Frage lautet . . .*“ (Hulbert, 2023, aus dem englischen).

Neben den genannten zwei Techniken gibt es noch weitere komplexere Advanced-Techniken DAIR.AI (2024a), die für diese Arbeit im Kontext der Bildung und der begrenzten Möglichkeiten von Studierenden jedoch weniger relevant sind. Stattdessen werden Prompt Techniken betrachtet, die spezifischer auf konkrete Beispiele angewandt werden können und verstärkt die möglichen Affordanzen der Modelle aufzeigen. In der Domäne der Softwareentwicklung gibt es sogenannte *Prompt Patterns* (White et al., 2023), die auf den Kontext der Klausurvorbereitung übertragbar sind. Die für das Prompt Handbook relevanten Patterns, werden im folgenden Abschnitt mit zusätzlichen Beispielen aufgeführt.

- **The Persona Pattern**: Dem LLM wird eine „Persona“ zugeschrieben. Als Persona ist in diesem Kontext eine hypothetische Person gemeint, die eine bestimmte Rolle einnimmt. Dies hilft ihm, eine Spezifikation über Ausgabe und Details für die Konversation zu erhalten (White et al., 2023, 7). In anderen Kontexten findet diese Methode unter dem Namen *Expert Perspective* auch ihre Verwendung (Memmert et al., 2024, 7523).
- **The Template Pattern**: Der Benutzer unterweist das Sprachmodell, sodass die Ausgabe in einer für ihn passenden wohldefinierten Struktur erfolgt (White et al., 2023, 12).

- **The Reflection Pattern**: Das LLM wird dazu aufgefordert, die Gründe für die Antwort selbständig detailliert zu erläutern, sodass das Verständnis des kausalen Zusammenhangs der Modellantwort verständlicher ist und auf Validität geprüft werden kann (White et al., 2023, 15).
- **The Recipe Pattern**: Es vereint unter anderem das *Template* und *Reflection Pattern*. Um ein bestimmtes Ziel zu erreichen, wird dem Modell eine klare Abfolge an Schritten bereitgestellt, die Frage zu beantworten (White et al., 2023, 17). Eine ähnliche Herangehensweise ist die *Evaluation Criteria Specification*, wobei dem Modell explizite Erwartungen mitgegeben werden (Memmert et al., 2024, 7522).
- **The Question Refinement Pattern**: Das LLM übernimmt die Rolle des Prompt Engineers, indem es dazu aufgefordert wird, für die eingegebene Frage, eine bessere verfeinerte Frage vorzuschlagen, welche stattdessen genutzt werden kann (White et al., 2023, 8).
- **The Flipped Interaction Pattern**: Während bei anderen Prompts der Nutzer die Konversation aufrechterhält, wird bei diesem Pattern das LLM instruiert, die Konversation bezogen auf ein gewisses Ziel voranzutreiben (White et al., 2023, 6).
- **The Context Manager Pattern**: Der Nutzer liefert dem Sprachmodell spezifischen Kontext für den Dialog, wobei der Fokus auf bestimmten Themen liegt, bzw. Themen explizit ausschließt (White et al., 2023; Memmert et al., 2024, 16; 7523).

Es gibt unzählige weitere Prompts, die spezifischer auf bestimmte Bereiche konkretisiert sind (Sahoo et al., 2024; Liu et al., 2021; DAIR.AI, 2024a; Memmert et al., 2024; Wei et al., 2023; Amatriain, 2024; White et al., 2023). Im Rahmen dieser Arbeit werden ebenfalls Prompts betrachtet, welche sich mit der Hochschulbildung und dem Lernen auseinandersetzen. Nach einer Einführung in die Hochschulbildung und klassischen Lernmethoden werden abschließend in Kapitel 2.3.2 diese Prompts und Affordanzen genauer beleuchtet.

2.2 Hochschulbildung und kognitive Modelle

Im vorherigen Kapitel 2.1 wurden die grundlegenden Konzepte hinter LLMs und dem PE dargelegt, sowie relevante PE Techniken beschrieben. Im Folgenden wird die Verbindung zum Lernen hergestellt. Dies wird zunächst durch allgemeine effektive Lerntechniken aus Studierendensicht beschrieben, um hieraus relevante Prompt-Techniken und Affordanzen in Verbindung mit ChatGPT im Prompt Handbook festzuhalten. Anschließend folgen wichtige Konzepte für die Selbsteinschätzung, sowie Bloom's Taxonomy für das Experiment.

2.2.1 Lernen und Lerntechniken

Bereits seit über 100 Jahren werden von Psychologen effiziente Lerntechniken untersucht (Dunlosky et al., 2013, 5). In diesem Zeitraum wurden von Kognitions- und Bildungspsychologen elementare Lerntechniken erforscht, die Studierenden helfen können, ihre Lernziele zu erreichen (Dunlosky et al., 2013, 4). Solche Lerntechniken sind dabei in variierenden Kontexten zu finden. In Wittrock (1974, 1989) wird ein generatives Lernmodell vorgestellt, welches beschreibt, dass Lernende, Wissen in Verbindung mit vorhandenem Wissen entwickeln (Fiorella und Mayer, 2016, 2).

Eines der wesentlichsten Bestandteile beim Lernen ist die Abfrage (Retrieval-Based Learning). Durch einen aktiven Informationsabruf kann das Lernen sowohl besser verstanden, als auch gefördert werden (Karpicke, 2012, 157). Grund hierfür ist, dass Menschen keine konstanten exakten Abbilder von Erfahrungen speichern und diese beim Abrufen wortwörtlich wiedergeben. Stattdessen wird Wissen beim Abrufen, in Verbindung mit Kontext und Abrufhinweisen, aktiv rekonstruiert. Bei jeder Gelegenheit, in der eine Person Wissen abruft, wird das Wissen verändert, sodass es leichter wird, das Wissen erneut abzurufen

(Karpicke, 2012, 158). Tatsächlich ist das Verständnis von Studierenden diesbezüglich anders. In einer Studie gingen Studierende davon aus, sich durch mehrmaliges Lesen eines Textes an mehr Inhalte zu erinnern, als durch mehrmaliges Abrufen des Textes nach einmaligem Lesen. Der Test widerlegt die Annahme der Studierenden vollends. So schnitten Studierende mit der Methode *Lesen-Abrufen-Abrufen-Abrufen* am besten ab. Die Selbsteinschätzung der Studierenden findet in dieser Arbeit mit dem Kognitionsmodell nach Nelson und Narens (1990) auch seine Verwendung und wird in Kapitel 2.2.3 beschrieben. Gerade das langfristige Behalten von Informationen wird durch die Abfrage gefördert (Karpicke, 2012, 159). Somit scheinen vor allem Prompt Techniken und Affordanzen von ChatGPT sinnvoll, die das Abrufen ermöglichen. Aber welche Lerntechniken eignen sich für eine effektive und effiziente Klausurvorbereitung am besten?

Dunlosky et al. (2013) haben in einer umfassenden Literaturrecherche untersucht, welche Lerntechniken am effizientesten und effektivsten für das Lernen von Studierenden sind. Dabei wurden insgesamt zehn Techniken zusammengetragen: *elaborative interrogation*, *self-explanation*, *summarization*, *highlighting (or underlining)*, *keyword mnemonic*, *imagery use for text learning*, *rereading*, *practice testing*, *distributed practice*, und *interleaved practice*. Untersucht wurden die Lernstrategien unter den vier Gesichtspunkten Lernbedingung, Merkmale der Lehrenden, Materialien und Kriterien-Aufgaben (Dunlosky et al., 2013, 4). In Einvernehmlichkeit mit Karpicke (2012) verzeichneten hierbei vor allem das *practice testing* und die *distributed practice* eine hohe Wirksamkeitsbewertung. Hintergrund ist, dass es den Lehrenden durch die zwei Lerntechniken ermöglicht wird, die Performance in Bildungskontexten zu verbessern. Allerdings sind *elaborative interrogation*, *self-explanation* und *interleaved practice* nicht unbedeutend. Ihnen wird jedoch nur eine moderate Nützlichkeit, bedingt durch die geringe Anzahl an Forschungsergebnissen, zugewiesen. Dies zeigt vor allem, dass ChatGPT praktische Tests unterstützten sollte und wird im Prompt Handbook mit beachtet.

Erweitert wurde das ganze von Rovers et al. (2018), indem sie 26 effektiv selbstregulierende Studierende befragt haben, wie sie ihr Lernziel erreichen. Dabei gaben diese an, in speziellen Situationen auch auf ineffektive Lerntechniken wie *highlighting* oder *rereading* zurückzugreifen (Rovers et al., 2018, 1). Das *highlighting* und *rereading* von Studierenden als Lernmethoden bevorzugt werden, zeigen auch weitere Studien (Dunlosky et al., 2013, 46). In einer Fokusgruppe wurde konstatiert, dass die vorher befragten effizienten Studierenden während ihrer Lernphase ihr Lernen ständig verarbeiten und überwachen. Es ist ein ständiges Gleichgewicht zwischen der Anwendung gewohnter Lerntechniken in Verbindung mit Flexibilität der Lernumgebung, Prüfungsanforderungen und Zeitbeschränkungen (Rovers et al., 2018, 1). Dies zeigt, dass die Nutzung von Lernstrategien mit einer niedrigeren Wirksamkeitsbewertung (Dunlosky et al., 2013) auch sinnvoll sein kann. Jedoch sollte man die jeweilige Strategie an die Lernsituation anpassen (Rovers et al., 2018, 9). Dieser Aspekt korreliert insbesondere mit dem Aufbau einer Klausur, auf die sich Studierende vorbereiten. Wenn es sich um einen Multiple-Choice-Test oder offene Fragen handelt, die lediglich kurzzeitig Informationen abfragen, wird sich dementsprechend vorbereitet. Dabei können oberflächliche Strategien zum Einsatz kommen (Rovers et al., 2018, 9). Dies ist ein wichtiger Hinweis für den erstellten MC-Test, für den somit auch oberflächliche Strategien effektiv sinnvoll seien können.

Im Kontext des generativen Lernprozesses (Wittrock, 1974, 1989) beschreiben Fiorella und Mayer (2016) acht zu Dunlosky et al. (2013) ähnliche Lerntechniken. Bei diesem konstatieren sie, dass Studierende wissen, *welche* Lernstrategie sie zu *welcher* Zeit *wie*, am effektivsten nutzten können (Fiorella und Mayer, 2016, 17).

Schließlich ist gutes Lernen auch vom Lehrenden abhängig. So sollten Dozenten bei Studierenden das *Hard Thinking* stimulieren. Dies meint Inhalte und Interaktionen so zu präsentieren, dass es das Denken dieser fördert (Coe et al., 2020, 30).

2.2.2 Kritik an Lerntechniken

Die effektive und effiziente Herangehensweise birgt allerdings auch negative Aspekte, die in dieser Arbeit knapp betrachtet werden. Soderstrom und Bjork (2015) diskutieren eine Trennung zwischen *Lernen* und *Performance*[2]. Das *Lernen* ist eine dauerhafte Veränderung im Verhalten bzw. Wissen, infolgedessen Informationen langfristig behalten und auf neue Situationen angewendet werden können. Wohingegen die *Performance* eine kurzfristige, gemessene oder wahrgenommene Schwankung im Verhalten bzw. Wissen darstellt, die während oder sofort nach einer Lern- oder Unterrichtsphase eintreten kann (Soderstrom und Bjork, 2015, 193). In Bezug auf das Abrufen von Informationen (Karpicke, 2012) können *Performance* und *Lernen* tatsächlich im Konflikt miteinander stehen. Hiernach sind Modalitäten die die *Performance* verringern, oftmals die beständigsten und flexibelsten für das *Lernen* (Soderstrom und Bjork, 2015, 193).

Allerdings wird in dieser Arbeit nicht untersucht, wie man langfristig das Meiste lernen kann, sondern welche Methoden am effektivsten und effizientesten für die Klausurvorbereitung sind. Zwar gibt es Beispiele, die zeigen, dass Studierende trotz simpler Tests ein langfristigeres Verständnis und Wissen aufbauen können, die jedoch vor allem im Zusammenhang mit einer Zukunftsperspektive (zum Beispiel als fähiger Doktor in der Medizin zu arbeiten) steht (Rovers et al., 2018, 9).

2.2.3 Metakognition nach Nelson und Narrens

In dieser Arbeit wird die subjektive Effizienz und Effektivität unter anderem durch die Selbsteinschätzung der Studierenden erfolgen. Dafür wird im Bereich der Metakognition ein theoretisches Framework von Nelson und Narens (1990) genutzt. Es beschreibt, wie sich Studierende bei der Wissensaneignung selbst kontrollieren (*Control*) und überwachen (*Monitor*) (Nelson und Narens, 1990; Rovers et al., 2018, 127; 10). Hiernach definieren sich Studierende ihre persönlichen Ziele nach ihrer Studiennorm, nachdem Lerninhalt und Testart bekannt sind. Die Definition erfolgt mit der Absicht, einen Test in der Zukunft zu bestehen (Nelson und Narens, 1990, 129). Im Anschluss wird eine Vorgehensweise von Studierenden entwickelt, um die individuellen Ziele zu erreichen. Infolgedessen kommt es zur Überwachung, sowohl in der Retro- als auch Prospektive. Im Detail wird dabei zwischen dem *Ease-of-learning* (EOL), *Judgements-of-Learning* (JOL) und *Feeling-of-knowing* (FOK) differenziert (Nelson und Narens, 1990, 130). Die drei genannten Kriterien dienen in dieser Arbeit als subjektive Parameter zur quantitativen Auswertung. Infolgedessen folgt eine präzise Erläuterung der Messinstrumente EOL, JOL und FOL.

- *Ease-of-learning* (EOL): Dieses Kriterium wird vor dem Lernen ermittelt. Es ist die subjektive Einschätzung darüber, welche Elemente simpler oder anspruchsvoller zu lernen sind. Das EOL bezieht sich auf den Schwierigkeitsgrad der zu lernenden Elemente oder der Wahl der effizientesten Lernstrategie (Nelson und Narens, 1990, 130).
- *Judgement-of-learning* (JOL): Dieses Kriterium wird während oder nach dem Lernen ermittelt. Es gibt an, wie die subjektive Einschätzung über die zukünftige Testperformance bezogen auf die aktuell abrufbaren Elemente ist (Nelson und Narens, 1990, 130).
- *Feeling-of-knowing* (FOK): Dieses Kriterium wird während oder nach dem Lernen ermittelt. Es ist die subjektive Einschätzung darüber, ob man ein Element, was gegenwärtig nicht abrufbar ist, zu einem späteren Zeitpunkt abrufen kann (Nelson und Narens, 1990, 130).

Eine praxisnahe Auseinandersetzung mit den Kriterien erfolgt in Kapitel 3.

2.2.4 Bloom's Taxonomy

Da in der Umfrage überprüft wird, ob die Teilnehmer:innen bei den Themen ein tiefgehendes Wissen gebildet haben, basieren die gestellten MC-Fragen auf *Bloom's Taxonomie*. Die Taxonomie wurde erstmals

[2] gleichbedeutend mit Effektivität und Effizienz

von Bloom et al. (1956) vorgestellt. Sie stellt eine der etabliertesten Klassifikationen in der Bildung dar, sowohl für Lehrer:innen als auch Dozent:innen (Forehand et al., 2005; Armstrong, 2010, 41; 1). Die Taxonomie ist ein Framework, dass Bildungsziele kategorisiert und als ein Messinstrument für das Denken dient (Armstrong, 2010; Forehand et al., 2005, 1; 44). In dieser Arbeit wird die anerkannte überarbeitete Taxonomie von Anderson und Krathwohl (2001) genutzt. Grund hierfür sind die bessere Kompatibilität mit der heutigen Zeit, sowie geeignetere Erfüllung der Anforderungen von Lehrenden (Forehand et al., 2005, 44). Im Paper von Anderson und Krathwohl (2001) wird ein zwei-dimensionales Framework aufgestellt, das zum einen die Dimension *Wissen (Knowledge)* und zum anderen die Dimension des *kognitiven Prozesses (Cognitive Process)* beinhaltet (Krathwohl, 2002, 218). Der *kognitive Prozess*, stellt allgemein das Vorgehen dar, welches zum Lernen verwendet wird (Forehand et al., 2005, 43). Aus dieser Perspektive steht ein mehrstufiges hierarchisches Modell im Mittelpunkt, dass das Denken in sechs unterschiedliche kognitive Schwierigkeitsstufen einteilt (Forehand et al., 2005, 42).

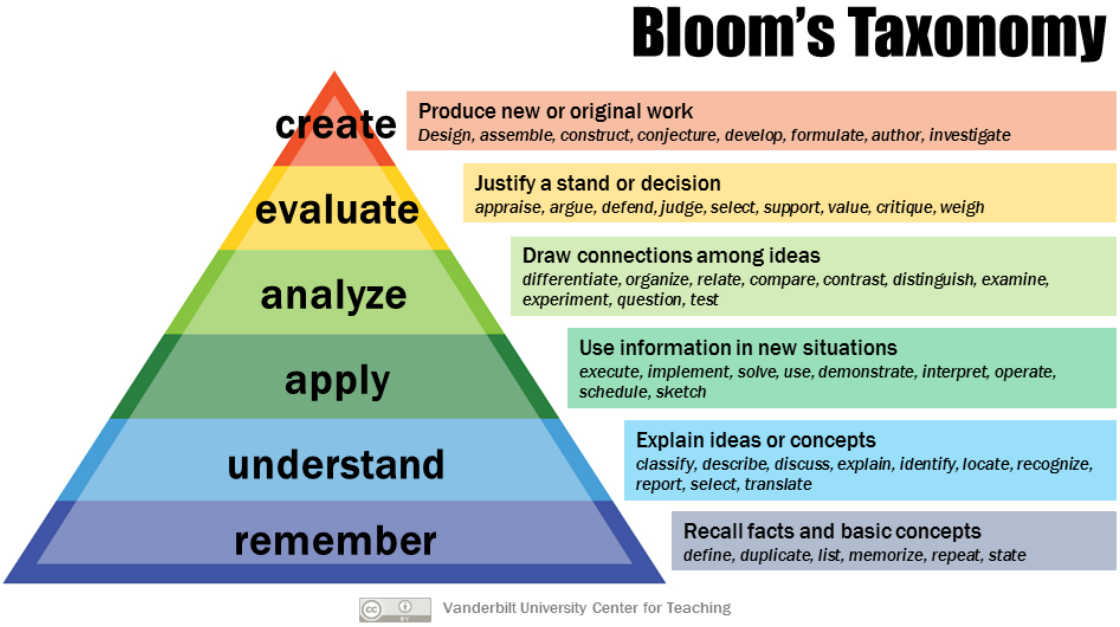

Abbildung 2: Bloom's Revised Taxonomy (Vanderbilt University Center for Teaching, 2020)

Quelle: https://www.flickr.com/photos/vandycft/29428436431. Lizenz: Creative Commons Namensnennung 2.0 (CC BY 2.0).

In Abbildung 2 sind die sechs kognitiven Schwierigkeitsstufen dargestellt. Auf der untersten Ebene liegt das *Erinnern (remember)*. Dabei geht es darum, relevantes Wissen aus dem Langzeitgedächtnis abzurufen. Eine Ebene höher ist das *Verstehen (Understand)* positioniert. Damit wird die Bestimmung der Bedeutung der Lehrinhalte beschrieben. Die 3. Ebene bildet das *Anwenden (Apply)*, welches die Durchführung oder das Anwenden von Verfahren in gegebenen Situationen darstellt. Das Analysieren *(Analyze)* ist auf der 4. Ebene zu finden und steht für die Zerlegung von Lerninhalten in einzelne Fragmente und das Verständnis, wie diese Fragmente miteinander in Beziehung stehen. Die vorletzte Ebene bildet die *Bewertung (Evaluate)*, welche das Fällen von Urteilen zu bestimmten Vorgaben kennzeichnet. Schließlich ist die höchste Schwierigkeitsstufe das *Generieren (Erschaffen)*. Hierbei geht es darum, einzelne Elemente zu kombinieren, um eine harmonische Einheit zu erstellen (Krathwohl, 2002, 215).

Zusätzlich zur Dimension des *kognitiven Prozesses* existiert die Dimension des *Wissens*. Die *Wissensdimension* beschreibt allgemein die Wissensart (Forehand et al., 2005, 43) und wird in vier Bereiche unterteilt. Differenziert wird zwischen dem *Factual Knowledge (Faktenwissen)* - Wissen an grundlegenden notwendigen Elementen, um mit einer Disziplin vertraut zu sein, oder Probleme zu lösen - und dem *Con-*

ceptual Knowledge (Konzeptuelles Wissen) - Wissen über die Beziehungen der grundlegenden Elemente, die eine Zusammenarbeit ermöglichen. Weiterhin bilden das *Procedural Knowledge (Prozedurale Wissen)* - Wissen darüber, wie man etwas durchführt, zum Beispiel der Einsatz von Algorithmen und Fähigkeiten -, sowie das *Metacognitive Knowledge (Metakognitive Wissen)* - Wissen über die generelle Kognition und den Denkprozess - die zwei letzten Bereiche (Krathwohl, 2002, 214). Insbesondere das *Metacognitive Knowledge* wird durch das Metakognitionsmodel nach Nelson und Narens (1990) aus einer etwas anderen Betrachtungsweise quantitativ im Detail untersucht.

Durch die Einordnung ist es nun möglich, in der Arbeit die einzelnen MCQs in die unterschiedlichen Dimensionen des kognitiven Prozesses des Wissens einzuordnen, um das tiefere Verständnis der Studierenden objektiv zu untersuchen. Die genaue Umsetzung wird in Kapitel 3 beschrieben.

2.3 LLMs und Prompts in der Hochschulbildung

2.3.1 KI und ChatGPT in der Hochschulbildung

Die Auseinandersetzung mit Künstlicher Intelligenz (KI) im Bereich der Lehre, reicht bis in die 1970er Jahre zurück (Rudolph et al., 2023, 350). In den Jahren 2021 und 2022 hat die Forschung jedoch einen bemerkenswerten Anstieg erfahren (Crompton und Burke, 2023, 19). Die wissenschaftliche Diskussion mit dem Thema ist facettenreich etabliert. So findet jährlich die Konferenz *Artificial Intelligence in Education (AIEd)* statt (Springer Verlag, 2024), aber auch zahlreiche Fachzeitschriften widmen sich diesem Gebiet. Die wichtigste Zeitschrift bildet dabei die *International Journal of Arteficial Intelligence in Education* (Zawacki-Richter et al., 2019, 8). In der Forschung haben sich die Anwendungsfelder von KI zunehmend erweitert. Während zu Beginn drei Bereiche identifiziert wurden (Luckin et al., 2016), erhöhte sich der Bereich von vier (Zawacki-Richter et al., 2019) auf fünf Anwendungsfelder (Crompton und Burke, 2023). Crompton und Burke (2023) schreiben der KI demzufolge fünf Affordanzen zu: (1) Bewertung und Evaluation, (2) Prognose, (3) KI-Assistenz, (4) intelligente Tutorensysteme und (5) Lernmanagement für Studierende (Crompton und Burke, 2023, 13). Des Weiteren wurden von Baker et al. (2019) drei verschiedene Domänen von AIEd vorgestellt, die für diese Arbeit als Eingrenzung dienen. So wird zwischen lernenden-, lehrenden- und systemzentrierter KI in der Bildung differenziert (Baker et al., 2019, 11ff.).

Im Bereich der lehrendenzentrierten Bildung kann Generative Künstliche Intelligenz (GenKI) als vielversprechendes Hilfsmittel angesehen werden (Choi et al., 2024, 1). Im Folgenden erfolgt ein verstärkter Fokus auf die lernendenzentrierte GenKI in der Bildung, da diese für die Arbeit bedeutungsvoller ist.

In einer Arbeit von Chan und Hu (2023) wurde bei ca. 400 Studenten und Doktoranden aus vielfältigen Bereichen, die Haltung und Wahrnehmung GenKI untersucht. Dabei wurde konstatiert, dass viele der Teilnehmer:innen eine allgemein positive Einstellung gegenüber GenKI in Lehre und Studium haben (Chan und Hu, 2023, 1). Die Probanden sahen einen signifikanten Mehrwert beim personalisierten und direktem Feedback, beim Lernen, sowie weitreichende Unterstützungen beim Brainstorming und Verfassen von Texten (Chan und Hu, 2023, 13). Exemplarisch kann dies mit einer Studie von Henkel et al. (2024) bestätigt werden, bei welcher ein KI-Mathematik-Tutor die Lernergebnisse verbesserte (Henkel et al., 2024, 1). Allerdings sind den Akteuren auch Probleme, die Themenbereiche wie Genauigkeit und Datenschutz umfassen, bewusst (Chan und Hu, 2023, 13). Weiterhin wird auf die Relevanz von KI-Fähigkeiten von Studierenden hingewiesen. Demzufolge sollten Studierende mit einem gewissen Grundwissen, über Funktionsweise, Vor- und Nachteile, sowie Affordanzen in der Hochschulbildung hinsichtlich GenKI vertraut sein (Chan und Hu, 2023, 16). Für die zukünftige Forschung wird empfohlen sich mit den Lernergebnissen mithilfe von GenKI zu beschäftigen (Chan und Hu, 2023, 16).

In der Perspektiven-Studie von Rasul et al. (2023) wurden ähnliche Ergebnisse erarbeitet. Dabei wurden jeweils fünf Vor- und Nachteile hinsichtlich des expliziten Gebrauchs von ChatGPT im Bildungs-

kontext dargestellt (Rasul et al., 2023, 10). Als zentrale positive Betrachtungsweise wird vor allem das *adaptive Lernen* (personalisierte Lernen) beschrieben (Rasul et al., 2023, 4). Jedoch sollte auch auf die Bias in den Antworten von ChatGPT, als eine negative Perspektive, hingewiesen werden (Rasul et al., 2023, 9). Diese Perspektive steht im engen Zusammenhang mit der erwähnten *Halluzination* (Amatriain, 2024, 4) sowie explizit dem Problem des Bias der Antworten von LLMs (Weidinger et al., 2021, 36). Darüber hinaus ist ein negativer Einfluss auf die Entwicklung akademischer Fähigkeiten möglich, wie etwa das kritische Hinterfragen oder das Lösen von Problemen (Rasul et al., 2023, 9). Gerade ein Rückgang des kritischen Denkens wird beim Lernen in vielen Papern als Besorgniserregend angesehen (Cong-Lem et al., 2024, 6). AlAfnan et al. (2023, 60) sprechen bei einer unethischen Nutzung von einer Verdummung der Studierenden. Dies zeigt die Wichtigkeit eines Regelwerks mit Affordanzen und Hinweisen, welche das Prompt Handbook bereitstellt.

Auch Rudolph et al. (2023) legen in ihrem Artikel nahe, dass Studierende durch ChatGPT in der Lage sind, auf einer Erlebnis- und Experimentierebene zu studieren (Rudolph et al., 2023, 353f.). Gleichwohl wird die Befürchtung Lehrender dargestellt, dass ChatGPT nicht wirklich begreift was er generiert (Rudolph et al., 2023, 353). Diese Einschätzung steht in engem Zusammenhang mit der Betrachtung von LLMs als *stochastischen Papagei* (Bender et al., 2021, 616f.). In dem Artikel wird neben fünf weiteren Punkten jedoch empfohlen, dass Studierende KI als Werkzeug nutzen sollen, um ihre Schreibfähigkeiten und Kreativitätsentwicklung zu verbessern, und nicht generierten KI-Text zu übernehmen (Rudolph et al., 2023, 356).

Nach Giannos und Delardas (2023, 6) wird empfohlen ChatGPT mit in den Lernprozess einzubeziehen, allerdings sollten weiterhin bewährte Lerntechniken verwendet werden, die nicht auf ChatGPT zurückgreifen.

Im Kontext der beschriebenen Anwendungsfelder und Herausforderungen beim Einsatz von KI in der Hochschulbildung führte Bastani et al. (2024) eine Feldstudie durch, welche die Lernleistung unter dem Einfluss von zwei verschiedenen ChatGPT-basierten Tutoren untersuchte. Zum Einsatz kam der sogenannte *GPT Base*, was dem klassischen ChatGPT entsprach, sowie der *GPT Tutor*, welcher PE und Eingaben der Lehrkraft nutzt, um den Lernprozess zu fördern (Bastani et al., 2024, 4). Die Modelle durften von Studierenden, während der *assisted practice periode* genutzt werden, um eine Reihe von Aufgaben der Lehrenden zu lösen (Bastani et al., 2024, 5f.). Hierbei war ein interessantes Ergebnis von Bastani et al. (2024) die signifikante Performancesteigerung durch die Verwendung der Tutoren. So erhöhte sich die Leistung während der Anwendungsphase beim *GPT Base* um 48 % und beim *GPT Tutor* um 127 %. Nach dieser Phase folgte eine Closed Book Prüfung, bei der die Sprachmodelle ebenfalls nicht mehr verwendet werden durften (Bastani et al., 2024, 6). Bei Prüfungen, in denen die Tutoren entfernt wurden, verringerte sich die Leistung von Studierenden, die den *GPT Base* nutzten, jedoch um 17 % (Bastani et al., 2024, 2). Dieses Ergebnis unterstreicht die Risiken, die mit der freien Nutzung von Sprachmodellen ohne fundiertes Wissen über deren Funktionsweise oder Anwendungshinweise verbunden sind (Chan und Hu, 2023; Rasul et al., 2023; Rudolph et al., 2023, 16; 9; 356). Der modifizierte *GPT Tutor* hingegen, wies mit seiner gestuften Hilfestellung keinerlei signifikante Performanceeinbuße bei der Chatbot ungestützten Prüfungen auf (Bastani et al., 2024, 9). In der Metakognition stellt die Diskrepanz in der Wahrnehmung der Studierenden und der Realität eine Relevanz dar. Interessanterweise nahmen Studierende die Auswirkungen der Nutzung von Tutoren zum Thema Prüfungsleistung und Lernen als übermäßig optimistisch war. Ähnlich zu Karpicke (2012, 159), wo Studierende ineffektive Techniken subjektiv sinnvoller empfanden. Denn von Studierenden, die den *GPT Base* nutzten, wurde trotz schlechterer Testergebnisse nie angegeben weniger gelernt oder schlechtere Leistungen erbracht zu haben. Aber auch beim *GPT Tutor*, wurde trotz gleichbleibender Testergebnisse, von Studierenden wahrgenommen, besser abzuschneiden (Bastani et al., 2024, 10). Dieser Umstand wird in dieser Arbeit ebenfalls durch die Metakognition nach

Nelson und Narens (1990) untersucht.

2.3.2 Verbindung von Prompt Engineering und Lernen

In Kapitel 2.1.2 wurde das Prompt Engineering mit seinen vielen konkreten Patterns, Techniken und Methodiken vorgestellt. Mit dem Hintergrund der Hochschulbildung erfolgt nun eine differenzierte Auseinandersetzung mit Prompts und Affordanzen.

Mit dem Hintergrund, dass die Erstellung von konkreten Prompts für Sprachmodelle äußerst unkompliziert geworden ist, können auch Lehrende in der Bildung Prompts und damit Hilfsmittel kreieren, die sich explizit auf die Bildung beziehen (Mollick und Mollick, 2024, 4). Allerdings sollten Personen mit dem Prompt Engineering vertraut sein (Jacobsen und Weber, 2023b, 27). Gerade im Bereich der lehrenden- und lernendenzentrierten KI in der Bildung (Baker et al., 2019) sind einige konkrete Prompts zu finden. Das angesprochene *Persona Pattern* (White et al., 2023; Memmert et al., 2024) kann in der Bildung mit dem *Role-Play whit AI Feedback* in Verbindung gebracht werden. Im Allgemeinen ist es eine zentrale konkrete Prompt Technik in der Bildung, da sehr viele Affordanzen für ChatGPT durch das Pattern entstehen. So bieten unterschiedliche Personas Studierenden die Gelegenheit, ihre Komfortzone zu verlassen und mit neuen Rollen zu experimentieren. Durch Rollenspiele mit einem Chatbot ist es den Studierenden auf eine erzählerische, persönlich ansprechende Art und Weise möglich, ihre Stärken und Schwächen zu einem bestimmten Thema zu identifizieren (Mollick und Mollick, 2024, 12). Konkret kann das Pattern bei der Erzeugung des *Protégé Effect* im Prompt, Studierenden helfen. Der *Protégé Effect* bezeichnet den Umstand, dass sich Studierende, die ihren Mitmenschen etwas erläutern, ein besseres Verständnis von dem erklärten Material bilden können (Mollick und Mollick, 2024, 21). In diesem Fall nimmt ChatGPT die Persona eines neugierigen Schülers ein, der Fragen stellt. Während auf der anderen Seite die Nutzenden, die Persona eines Lehrenden einnehmen, um ChatGPT auf seine Fragen zu antworten (Mollick und Mollick, 2024, 24). Dabei können weitere Prompt Patterns wie das *Recipe Pattern* oder *Flipped Interaction Pattern* (White et al., 2023) zum Einsatz kommen, welche die Persona gewünscht reagieren lassen. Weiterhin können die Persona Rollen invertiert werden, indem ChatGPT die Persona eines Tutors (Crompton und Burke, 2023) übernimmt, der hilfreiches Feedback gibt (Mollick und Mollick, 2024, 37f.). Dabei wurde die Effektivität von Sprachmodellen als Tutoren bereits nachgewiesen (Henkel et al., 2024). Darüber hinaus lässt sich das Persona Pattern auch anwenden, indem ChatGPT als Mentor oder Coach fungiert (Mollick und Mollick, 2024, 6).
Auch Jacobsen und Weber (2023b) untersuchten aus der lehrendenzentrierten Perspektive, welches Prompt für hochqualitatives ChatGPT Feedback in der Hochschulbildung nötig ist (Jacobsen und Weber, 2023b, 2). Hierbei wurde konstatiert, dass ChatGPT in der Lage ist, für Studierende hochqualitatives Feedback bereitzustellen. Allerdings war das Feedback stark vom Prompt abhängig (Jacobsen und Weber, 2023b, 24). Auf der Literatur und den Erkenntnissen basierend, haben Jacobsen und Weber (2023b) ein Manual zur Erstellung hochwertiger Prompts (Jacobsen und Weber, 2023a) entwickelt.
Weiterhin wurde ein Prompt Workbook von Mohr (2024) zusammengetragen. Dieses umfasst diverse konkrete, auf spezifische Beispiele anwendbare Prompts speziell für Studierende und Lehrende. Relevant sind Prompts, die das Sprachmodell, *verschiedene Perspektiven berücksichtigen*, *Lerninhalte festigen*, *Hinweise zum Vorgehen bereitstellen*, *Selbsttests anfertigen*, *passende Hilfe generieren* oder *ein Thema aus einem anderen Blickwinkel sehen* lassen (Mohr, 2024).

Es lässt sich konstatieren, dass KI in der Bildung im Wandel ist und in vielen Bereichen genutzt werden kann. Explizit GenKI hat für Studierende einen positiven Einfluss, es sollten jedoch zusätzlich die Risiken beachtet werden. Schließlich ist das PE im Bereich des Lernens schon etwas erforscht und gibt gute Promptbeispiele zur Nutzung.

3 Methodik

In diesem Kapitel wird die Methodik für die Forschungsfragen der Arbeit beschrieben, die darauf abzielen, die Effektivität und Effizienz von PE in der Klausurvorbereitung von Studierenden zu untersuchen. Zunächst wird erläutert, wie die produktiven Prompt-Techniken zur Unterstützung des Verständnisses in der Klausurvorbereitung, im folgenden Prompt Handbook genannt, aus der Literatur abgeleitet wurden. Darauf aufbauend wird das durchgeführte Between-Groups Experiment skizziert, dass den Einsatz von PE bei Sprachmodellen im Vergleich zum allgemeinen Einsatz von Sprachmodellen und traditionellen Lernmethoden untersucht. Die Beschreibung beinhaltet sowohl die quantitativen als auch qualitativen Messinstrumente zur Betrachtung der Effizienz, Effektivität, sowie Vor- und Nachteile der verschiedenen Ansätze.

3.1 Ableitung des Prompt Handbook

Um effektive Prompt Techniken und Affordanzen von ChatGPT für die Klausurvorbereitung zu identifizieren, wurde eine strukturierte Literaturanalyse im Forschungsfeld des PE durchgeführt. Da es sich um ein neues Forschungsgebiet handelt, wurden neben Peer-reviewten Publikationen auch Preprints und praktische Erfahrungsberichte genutzt. Die Recherche erfolgte über *Google Scholar*, *ResearchGate*, *arXiv.org* und *Inciteful*.

Als Suchbegriffe wurden „prompt engineering“, „prompt engineering education“, „ai in higher education“, sowie „large language models in higher education“ verwendet. Über die Inciteful-Grafen wurden zudem weitere thematisch verwandte Paper gefunden. Der Zeitraum der Recherche konzentrierte sich primär auf Publikationen, die nach der Veröffentlichung von ChatGPT erschienen sind (November 2022). Dennoch wurden ältere, grundlegende Techniken, zum Beispiel Methodiken zur Verringerung der Halluzination, berücksichtigt.

Um ein möglichst breites Spektrum an Prompts und Affordanzen bereitzustellen, wurden diverse Prompt Techniken in unterschiedlichen Kontexten herangezogen, wie zum Beispiel in der Softwareentwicklung (White et al., 2023), sowie grundlegende Techniken (Radford et al., 2019; Brown et al., 2020) und vorhandene Prompts im Kontext des Lehrens und Lernens (Mollick und Mollick, 2024; Mohr, 2024). Der Aufbau des Prompt Handbooks erfolgte ähnlich zum Prompt Workbook von Mohr (2024). Dabei wurden fünf Prompts von Mohr (2024) übernommen und angepasst. Darüber hinaus beinhaltet das Handbook weitere klassische Techniken, sowie explizite Ansätze für die Hochschulbildung. Zur Qualitätssicherung des Prompt Handbooks wurden die Prompts teilweise mit dem Prompt Manual von Jacobsen und Weber (2023a) und Testfragen von Mollick und Mollick (2024, 8) abgeglichen und angepasst, sofern es im Kontext der Klausurvorbereitung für sinnvoll erachtet wurde.

Basierend auf den Ergebnissen der Literaturanalyse und mit Blick auf das Between-Groups Experiment wurde das Prompt Handbook erstellt und in vier klar strukturierte Abschnitte eingeteilt:

1. **Hinweise zum Vorgehen**: Dieser Abschnitt enthält eine allgemeine Anleitung, die für die Durchführung des Between-Groups Experiments notwendig ist. So wird der Inhalt und das Konzept des Handbook zusammengefasst und der Nutzungshinweis erläutert.

2. **Prompt Techniken aus der Forschung**: Dieser Teil umfasst insgesamt acht Prompt Techniken (inklusive Affordanzen), die aus der Literatur abgeleitet wurden. Jede Technik wurde in einem Satz prägnant beschrieben und mit einem spezifischen Beispielprompt illustriert, sodass die Teilnehmenden die Technik direkt in ihre Lernpraxis umzusetzen konnten.

3. **Weitere Hinweise**: Hier werden spezifische Restriktionen und Empfehlungen zur Nutzung von ChatGPT als Hilfsmittel zur Klausurvorbereitung zusammengestellt. Ziel war es, den Teilnehmenden die Affordanz von ChatGPT als Werkzeug zum Verstehen der Lerninhalte deutlich zu machen.

4. **Abschließender Nutzungshinweis des Workbooks**: In diesem Teil wird ein Tipp zur Vorgehensweise gegeben, inklusive Video, um sich schnell mit der Benutzung vertraut zu machen.

Das Prompt Handbook durchlief mehrere Iterationen, mit dem Fokus, die Inhalte verständlich und klar zu formulieren, ohne die Teilnehmenden mit zu vielen Informationen zu überlasten. Insbesondere wurde das Handbook an die Anforderungen des Experiments entwickelt, sodass es die Beteiligten ohne eine große Einarbeitung bei der Klausurvorbereitung unterstützte. Das fertige Prompt Handbook wurde schließlich in digitaler Form auf der Plattform GitHub als Markdown-Datei bereitgestellt[3], um eine möglichst einfache Zugänglichkeit und Verbreitung zu gewährleisten. Darüber hinaus ermöglicht dies, eine unkomplizierte Anpassung für zukünftige Prompts, sowie eine intuitive Navigation für die Nutzer:innen. Genauere Informationen zu den Prompts und Affordanzen werden in Kapitel 4.1 ausgeführt.

3.2 Between-Groups Experiment

Die Untersuchung des Einsatzes von PE bei Sprachmodellen, im Vergleich zum allgemeinen Einsatz von Sprachmodellen und traditionellen Lernmethoden, erfolgte mit einem Mixed-Methods Between-Groups-Design. Dabei wurden die Teilnehmer:innen in drei verschiedene Gruppen hinsichtlich der Vorbereitung auf einen Multiple-Choice-Test (MCT) aufgeteilt. Durchgeführt wurde die Präparation bei Gruppe eins (G-1) eigenständig, ohne zusätzliche Hilfsmittel. Gruppe zwei (G-2) nutze zur Vorbereitung ChatGPT. Und Gruppe drei (G-3) verwendete ChatGPT in Verbindung mit dem Prompt Handbook. Das Between-Groups-Design wurde gewählt, um die Auswirkungen von PE und Sprachmodellen auf die Lernergebnisse zu isolieren und zu vergleichen.

Die Stichprobe umfasste 51 Studierende aus diverseren Studiengängen, die über soziale Medien rekrutiert wurden. Voraussetzung war, dass die Teilnehmer:innen sich freiwillig für das Experiment meldeten und mindestens ein elektronisches Gerät (Tablet, Laptop, etc.) zur Verfügung hatten. Für Proband:innen der G-3 und G-2 war ein Laptop notwendig.

Für den MCT wurde als Lehrmaterial ein 10-seitiger Foliensatz erstellt. Er umfasste acht ausgewählte Folien aus dem Vortagsfoliensatz „IT-Sicherheit und Künstliche Intelligenz“ (Pohlmann, 2024), sowie zwei eigens erstellte Folien zur Umrechnung von Binär- zu Dezimalzahlen. Als Sprachmodell wurde für G-2 und G-3 jeweils das GPT-4o Modell von OpenAI verwendet. Der MCT wurde auf Grundlage des Foliensatzes und der „Anleitung zur Herstellung von MC-Fragen und MC-Prüfungen für die ärztliche Ausbildung“ von Krebs (2004) erstellt. Es wurden acht *Typ A* Fragen (Krebs, 2004, 8ff.), eine *Typ Aneg* Frage (Krebs, 2004, 11) und eine Frage mit Zahleneingabe angefertigt. Die MC-Fragen wurden in insgesamt vier Iterationen überarbeitet und getestet. Grundlage für die Schwierigkeit der Frage und Antwortmöglichkeiten war die Bloom's Taxonomy (Krathwohl, 2002). Die MC-Fragen wurden inkrementell in direkter Abfolge zu den

The Knowledge Dimension	**The Cognitive Process Dimension**					
	1. Remember	*2. Understand*	*3. Apply*	*4. Analyze*	*5. Evaluate*	*6. Create*
A. Factual	Q1					Q8
B. Conceptual	Q2, Q7	Q3, Q6	Q3, Q4	Q5, Q6	Q4, Q6, Q7	
C. Procedural				Q9, Q10		Q10
D. Metacognitive						

Tabelle 1: Einordnung der zehn MCQs in die Bloom's Taxonomy

Hinweis: Grundlage dieser Tabelle stammt aus Krathwohl (2002, 217).

Folien erstellt. Damit bezieht sich Frage 1 (Q1) auf Folie 1, Frage 2 (Q2) auf Folie 2, Frage 3 (Q3)

[3] `https://github.com/LaurinWesselkamp/Bachelorarbeit/edit/main/README.md`

auf Folie 3, und so weiter. Dadurch wurde eine klare inhaltliche Relation zwischen Fragen und Folien gewährleistet, welches die Messinstrumente EOL und JOL jeder einzelnen Folie und entsprechenden Frage präziser zuordnet. Exemplarisch wird im Folgenden die erste der zehn Fragen illustriert:

Q1

Welche der folgenden Begriffe beschreibt eine Methode des maschinellen Lernens, die bessere Ergebnisse erzielt als traditionelle Methoden?

- Data Science
- Generative KI
- Artificial General Intelligence
- **Deep Learning** (richtige Antwort)
- Large Language Model
- Ich kenne die Antwort nicht, aber wahrscheinlich fällt sie mir zu einem späteren Zeitpunkt ein.
- Ich kenne die Antwort nicht, und sie wird mir in Zukunft auch nicht einfallen.

Die anderen Fragen mit Zuteilung zur Bloom's Taxonomy sind der Tabelle 1 zu entnehmen und befinden sich mit allen Antwortmöglichkeiten im Anhang A. Für Informationen zum Prompt Handbook, vgl. 3.1. Als Software für die Durchführung des Experiments wurde LimeSurvey für Studierende der UHH vom Regionalen Rechenzentrum der Universität Hamburg genutzt.

Nachfolgend wird der Ablauf des Experiments erläutert. Vor dem eigentlichen Start des Experiments wurde den Teilnehmer:innen vom Versuchsleiter unterschiedliche Instruktionen gegeben, abhängig davon in welcher Gruppe sie sich befanden.

- **G-1:** Als Kontrollgruppe erhielten die Teilnehmer:innen ausschließlich den Foliensatz.
- **G-2:** Die Beteiligten bekamen die Zugangsdaten für ChatGPT und den Foliensatz. Der Foliensatz wurde auf ChatGPT hochgeladen und (sofern möglich) in die gewohnte Lernanwendung eingepflegt.
- **G-3:** Die Teilnehmer:innen erhielten das Prompt Handbook, welches sie zunächst lesen mussten. Anschließend wurden ihnen die Zugangsdaten für ChatGPT sowie der Foliensatz zur Verfügung gestellt. Der Foliensatz wurde von den Proband:innen auf ChatGPT hochgeladen und (sofern möglich) in ihre gewohnte Lernanwendung auf dem Tablet/ Laptop integriert. Abschließend platzierten die Testpersonen ChatGPT auf der linken und das Prompt Handbook auf der rechten Seite des Bildschirms.

Nachdem die Vorbereitungen für das Experiment abgeschlossen waren, konnte es über einen beigefügten Link von den Proband:innen gestartet werden. Sie wurden kontinuierlich vom Versuchsleiter begleitet, sodass die Teilnehmer:innen die jeweiligen Einschränkungen ihrer Gruppen einhielten. Zu Beginn des Experiments erläuterte der Versuchsleiter den Proband:innen den Ablauf des Experiments verbal und in schriftlicher Form. So durchliefen die Teilnehmer:innen während des Experiments fünf Phasen, von denen drei als Hauptphasen galten.

1. **EOL-Zwischenphase** *(~5 Min.)*: In dieser Phase verschafften sich die Partizipanten:innen einen Überblick über die Folien und schätzten das *EOL* für jede der 10 Folien ein. Die Einschätzung erfolgte ähnlich zu Nelson und Leonesio (1988, 678).
2. **Vorbereitungsphase** *(30 Min.)*: Die Versuchspersonen bereiteten sich je nach Gruppenzugehörigkeit (vgl. Tabelle 2) mit dem Foliensatz (und Hilfsmitteln) auf den MC-Test vor. Für diese Phase betrug das maximale Zeitvolumen 30 Minuten, allerdings konnten die Teilnehmer:innen diese Phase auch vor Ablauf der Zeit beenden. Die Zeit wurde sowohl vom Versuchsleiter, als auch LimeSurvey gestoppt.

3. **JOL-Zwischenphase**: Die Testpersonen schätzten ihre subjektive Testperformance für alle zehn Folien nach *JOL* ein, ähnlich wie in der Studie von Dunlosky und Nelson (1994, 549).
4. **Testphase** *(~15 Min.)*: Die Teilnehmer:innen absolvierten den Test. Während dieser Phase durfte, unabhängig von ihrer Gruppe, kein ChatGPT, Prompt Handbook oder Foliensatz mehr nutzbar sein. Dies wurde vom Versuchsleiter über eine angeschaltete Webcam und Bildschirmübertragung des Tests kontinuierlich überprüft. Alternativ kam es auch zur Überprüfung in Präsenz. In diesem Teil wurde das *FOK* der Studierenden indirekt, ähnlich wie im Paper von Hart (1965, 209) untersucht.
5. **Fragen zur Person**: Zum Abschluss erfolgten Fragen zur Person, sowie quantitative und qualitative Fragen zur jeweiligen Lernmethode/ Gruppe. Diese Fragen unterschieden sich je nach Gruppe.

Material/ Hilfsmittel	**G-1**	**G-2**	**G-3**
Foliensatz	x	x	x
Uneingeschränkte Nutzung von ChatGPT		x	
Beschränkte Nutzung von ChatGPT mit Prompt Handbook			x

Tabelle 2: Vergleich der drei Gruppen in Bezug auf Merkmale der Klausurvorbereitung

Die Teilnehmer:innen wurden instruiert, zu keinem Zeitpunkt im Internet nach Informationen zu suchen. Das Experiment fand im Zeitraum vom 14.01.2025 bis zum 31.01.2025 statt. Die Stichproben wurden über den gesamten Tag verteilt, meist in einer 1:1 Begleitung zwischen Versuchsperson und Versuchsleiter, 39-mal Digital und 12-mal in Präsenz durchgeführt.

Um die Forschungsfragen zu beantworten, wurden verschiedene Daten im Rahmen des Experiments erhoben. Die folgenden Messinstrumente fanden ihre Verwendung:

1. **Soziodemografische Fragen**: Um die soziodemografischen Einflussfaktoren nachzuvollziehen, wurden Fragen zum Geschlecht, Studienfach, Semesteranzahl und weiteren Aspekten gestellt.
2. **MCT-Ergebnisse**: Die quantitativ messbaren Lernergebnisse der Proband:innen wurde durch die Antworten des MCT aufgenommen. Diese waren ebenfalls Ausdruck der objektiven Effektivität.
3. **Metakognitive Einschätzungen zu EOL und JOL**: Um die subjektiven metakognitiven Einschätzungen der Teilnehmer in Bezug auf den Lernstoff zu erfassen, wurden mittels Likert-Skala von 1 bis 5 das EOL und JOL pro Folie von den Teilnehmer:innen abgefragt.
4. **Zeiterfassung**: Die Zeit, welche die Teilnehmer:innen für die Vorbereitungs- und Testphase benötigt haben. Diese wurde mittels automatischen Zeitracker von LimeSurvey gemessen. Das Zeitvolumen stellte die objektive Effizienz dar.
5. **Fragen zur subjektiven Effektivität und Effizienz**: Um die subjektive Wahrnehmung der verwendeten Lernmethoden zur Effektivität und Effizienz zu erfassen, wurden den Teilnehmer:innen skalenbasierte (Likert-Skala von 1-5) Fragen gestellt.
6. **Fragen zu den Vor- und Nachteilen der Methode**: Um die Vor- und Nachteile aus Sicht der Proband:innen für die Verwendung von Sprachmodellen mit und ohne Prompt Handbook herauszustellen, wurden die Teilnehmenden explizit danach gefragt.
7. **Feedback zum Prompt Handbook**: Für die Sammlung von Feedback zur Verbesserung und Optimierung des Prompt Handbooks, wurde den Proband:innen offene und skalenbasierte Fragen zu dessen Verwendung gestellt.
8. **Verwendete Prompts**: Zur Analyse, welche Prompts in den Gruppen G-2 und G-3 am häufigsten genutzt wurden, wurden die spezifischen Prompts innerhalb eines Projektes von ChatGPT protokolliert.

Der Zeitpunkt der Datenerhebung kann Tabelle 3 entnommen werden.

Erhobene Daten	*1*	**2**	*3*	**4**	**5**
Soziodemografische Fragen					✓
MCT-Ergebnisse				✓	
Metakognitive Einschätzungen zu EOL und JOL	✓		✓		
Zeiterfassung	✓	✓	✓	✓	✓
Fragen zur subjektiven Effektivität und Effizienz			✓		✓
Fragen zu den Vor- und Nachteilen der Methode					✓
Feedback zum Prompt Handbook					✓
Verwendete Prompts		✓			

Tabelle 3: Zeitpunkt der Datenerhebung in den unterschiedlichen Phasen des Experiments
Legende: *1 = EOL-Zwischenphase, 2 = Vorbereitungsphase, 3 = JOL-Zwischenphase, 4 = Testphase, 5 = Fragen zur Person.*

Die Datenauswertung der Messinstrumente erfolgte in mehreren Schritten mittels deskriptiven und explorativen Statistiken sowie inferenzstatistischen Tests. Die soziodemografischen Merkmale wurden mittels absoluter und prozentualer Häufigkeitsverteilung verglichen und mit Chi-Quadrat Tests auf Unterschiede überprüft. Die anderen Messinstrumente wurden unter den Gruppen folgendermaßen verglichen:

- *Deskriptive Statistik*: Für die initiale Beschreibung wurden meistens die Mittelwerte genutzt. Allerdings kam es auch zur Verwendung von Median, Standardabweichung, Varianz, Minimum, Maximum und Summe, falls dies sinnvoll erschien. Genutzt wurde die deskriptive Statistik für die MCQ-Testergebnisse, metakognitiven Einschätzungen, subjektiven Effektivitäts- und Effizienzbewertung, Vor- und Nachteile der Methode sowie dem Feedback zum Prompt Handbook.
- *Prüfung der Voraussetzung für inferenzstatistische Tests*: Die Normalverteilung wurde mittels Shapiro-Wilk-Test geprüft. Für die Prüfung der Homogenität der Varianz kam der Levene-Test zum Einsatz. Falls die Voraussetzung gegeben waren, wurde ein einfaktorieller ANOVA-Test durchgeführt, ansonsten ein Kruskal-Wallis-Test.
- *Vergleich der Gruppenunterschiede:*
  - *Einfaktorieller ANOVA*: Für die Analyse signifikanter Unterschiede in den aggregierten MCQ-Testergebnissen sowie dem EOL und JOL, falls Normalverteilung und Varianzhomogenität gegeben.
  - *Kruskal-Wallis-Test*: Bei Verletzung der Normalverteilung kam es bei der Zeiterfassung, der subjektiven Effektivitätsbewertung und beim Autonomieverlust zu diesem Test.
  - *Chi-Quadrat-Test*: Zur Auswertung der Verteilung richtiger Antworten pro MC-Frage.
- *Qualitative Inhaltsanalyse*: Für die Freitextantworten zur subjektiven In- und Effektivität, zu den Vor- und Nachteilen, sowie den Affordanzen des Sprachmodells von G-2 wurde jeweils eine qualitative Inhaltsanalyse nach Mayring durchgeführt. Die Kategorien wurden induktiv aus den Antworten entwickelt. Eine Reliabilitätsprüfung wurde ausgelassen.
- *Signifikanzniveau und Software*: Das Signifikanzniveau wurde auf $\mathbf{p < 0{,}05}$ festgelegt. Für die Aufbereitung der Daten wurde Excel verwendet. Für die deskriptive Statistik, die inferenzstatistischen Test und Erstellungen von Grafiken wurde JASP in der Version 0.19.3 und Excel verwendet.

Um die Daten der Proband:innen zu schützten, wurden sie anonym abgespeichert. Zu Beginn des Experimentes wurde den Teilnehmer:innen eine Datenschutzerklärung bereitgestellt, in der sie über die Speicherung ihrer Daten aufgeklärt wurden.
Die Limitationen des Experiments werden in 5.7 aufgeführt.

4 Ergebnisse

Im Folgenden werden die Ergebnisse dargestellt, eingeteilt in die vier Forschungsfragen.

4.1 Abgeleitete Prompts für die Klausurvorbereitung

F-1: *„Welche Prompt Engineering Techniken sind am effektivsten für die Klausurvorbereitung, um relevante Informationen aus vorhandenen Lehrmitteln vertiefend zu verstehen?“*

Es konnten acht konkrete Prompt Techniken bzw. Affordanzen aus der Forschung abgeleitet werden. Diese werden nachstehend mit zugehörigen Quellen und gekürzten Beispielen illustriert. Die ungekürzten Promptbeispiele sind im Prompt Handbook einsehbar. Alle Prompts wurden mit Blick auf die Lerntechniken von Karpicke (2012); Dunlosky et al. (2013); Rovers et al. (2018); Fiorella und Mayer (2016); Coe et al. (2020) ausgewählt und angepasst.

Lass uns Schritt für Schritt nachdenken. Hierbei handelt es sich um eine praxisbezogene deutsche Version der *Chain-of-Thought* (Wei et al., 2023) Technik. Studierende können mit dem Prompt beim Sprachmodell einen strukturierten Denkprozess anregen, der Fehlern vorbeugt.
Beispiel: *[FRAGE] Lass uns Schritt für Schritt nachdenken.*
Quellen: Die Prompt Idee entspricht dem *Zero-shot CoT* von Kojima et al. (2023, 2) und hilft gegen das „Error in Reasoning“ (Mollick und Mollick, 2024, 4).

Gedankenbaum. Diese Technik ist eine praxisbezogene, ins Deutsche übersetzte, Version der *Tree-of-Thought* (Chen et al., 2023, 10f.) Technik. Studierende lassen sich hiermit möglichst viele Antwortalternativen aus verschiedenen Perspektiven geben und das Modell diskutieren.
Beispiel: *Stelle dir drei verschiedene Experten vor, die diese Frage beantworten. [...]*
Quellen: Das Prompt wurde von Hulbert (2023) übernommen und findet ähnliche Anwendung unter der Bezeichnung „Verschiedene Perspektiven berücksichtigen“ (Mohr, 2024) sowie „Expert Perspective (EXP)“ (Memmert et al., 2024, 7523).

Hinweise zum Vorgehen bereitstellen (Kontext geben). Studierende können dieses Prompt zur Planung des Vorgehens für die Klausurvorbereitung nutzen. Es enthält zusätzlich Hinweise über die verbleibende Vorbereitungszeit, Thema der Klausur und beugt Missverständnissen vor.
Beispiel: *Du übernimmst die Rolle eines Planers. Ich möchte eine Klausurvorbereitung so effizient wie irgend möglich durchführen. Ich weiß, dass [...]*
Quellen: Die Idee basiert auf „Hinweise zum Vorgehen bereitstellen“ (Mohr, 2024) und ist auch als Vereinigung der „Expert Perspective (EXP)“, des „Providing Context (CON)“ sowie der „Evaluation Criteria Specification“ von Memmert et al. (2024, 7522f.) anzusehen. Weiterhin findet es ähnliche Verwendung als „The Recipe Pattern“ (White et al., 2023, 17) und hilft gegen „Responds With Persistent Missconceptions“ (Mollick und Mollick, 2024, 4).

Selbsttest anfertigen. Nutzer:innen sind mithilfe dieses Prompts daser Lage, das Sprachmodell zu instruieren, MCQs mit dazugehörigen Antwortoptionen und Erläuterungen zu falschen Antwortoptionen erstellen zu lassen.
Beispiel: *Erstelle 5 Multiple-Choice-Fragen basierend auf den Themen im hochgeladenen Foliensatz, sodass ich ein besseres Verständnis bilde. [...]*
Quellen: Die Grundlage dafür bildet „Selbsttests anfertigen (H5P Single Choice Test)“ von (Mohr, 2024) in Verbindung mit dem „Prompt Augmenation“ Gedanken (Liu et al., 2021; Memmert et al., 2024, 13;

7523). Weiterhin findet es in Teilen ähnliche Anwendung unter dem „The Template Pattern“ und „The Flipped Interaction Pattern“ von White et al. (2023, 6,12) und es ist eine Methode des „Retrieval-Based Learning“ (Karpicke, 2012, 162).

Lerninhalte festigen. Die Nutzer:innen schreiben dem Sprachmodell die Persona eines Tutors zu und festigen somit ihre Lerninhalte.
Beispiel: *Deine Aufgabe ist es, als persönlicher Tutor in der Rolle eines Universitätsprofessors zu agieren. Beginne mit einfachen Fragen zum hochgeladenen Foliensatz [...]*
Quellen: Das Prompt wurde auf Basis von „Lerninhalte festigen“ (Mohr, 2024) erstellt, findet jedoch auch als „Expert Perspective (EXP“ (Memmert et al., 2024, 7523), „The Persona Pattern“ (White et al., 2023, 7) sowie „Simulation Type 1: Role Play“ (Mollick und Mollick, 2024, 9) und „AI Tutors“ (Mollick und Mollick, 2024, 34f.) seine Anwendung.

Passende Hilfe erhalten. Studierende nutzten dieses Prompt zur Unterstützung im Selbststudium. Hierbei können sie zwischen diversen Unterstützungen wählen.
Beispiel: *Unterstütze mich beim Lernen des hochgeladenen Foliensatzes. Gebe mir immer nur auf dem Niveau Hilfe, dass ich zusammen mit meiner Frage benenne. Die Niveaus der Hilfe sind: [...]*
Quellen: Dieses Prompt basiert auf dem gleichnamigen Beispiel „Passende Hilfe erhalten“ von Mohr (2024).

Der andere Blickwinkel (Reframing). Mit diesem Prompt betrachtet ChatGPT Situationen aus einer anderen Perspektive.
Beispiel: *Du bist mein Coach. Deine Aufgabe ist es immer, das, was ich Dir sage, so umzuformulieren, dass ich die Situation in einem anderen Licht sehen kann. [...]*
Quellen: Das Prompt „Der andere Blickwinkel (Reframing)“ von Mohr (2024) diente hierfür als Grundlage und findet ebenfalls als „Question Refinement Pattern“ (White et al., 2023, 8) seine Verwendung.

Erkläre es ChatGPT. Mit diesem Prompt übernimmt das Sprachmodell die Affordanz eines indirekten Feedback-Gebers, indem durch Fragen Wissenslücken von Nutzer:innen aufgedeckt werden. Hierbei wurde auch der Hinweis gegeben, das Sprachmodell bei irrelevanten Fragen zu instruieren, sich wieder auf das Thema zu fokussieren.
Beispiel: *Du bist ein AI-Mentor, der in einem Rollenspiel-Szenario als Schüler agiert. Deine Aufgabe ist es, den Benutzer (Lehrer) dabei zu unterstützen, die Themen des hochgeladenen Foliensatzes effektiv zu erklären, indem du die Rolle eines Schülers übernimmst. [...]*
Quellen: Die Prompt Idee basiert auf „Critiquing Type 2: AI as Student“ von Mollick und Mollick (2024, 21) und wurde auch in „AI USE: STUDENT“ (Mollick und Mollick, 2023, 4) empfohlen.

Neben den Prompt Beispielen wurden zusätzliche Nutzungshinweise und Restriktionen für eine effektive, verständnisvolle Klausurvorbereitung mit dem LLM in dem Prompt Handbook eingepflegt. Dazu gehörten Hinweise zur Vermeidung von Halluzinationen, der Bitte keine Zusammenfassungen von ChatGPT erstellen zu lassen, die Bedeutung domänenspezifischer Inhalte, sowie der Umgang bei wiederholt unpassenden Antworten vom Sprachmodell. Die ausführliche Ausgestaltung findet sich im Prompt Handbook, zu finden unter dem Link: `https://github.com/LaurinWesselkamp/Bachelorarbeit/blob/main/README.md`.

4.2 Soziodemografische Merkmale

Um die Vergleichbarkeit der Ergebnisse des Between-Groups Experiments zu gewährleisten, werden zunächst die protokollierten soziodemografischen Parameter miteinander verglichen.

Die Geschlechterverteilung der drei Gruppen (G-1, G-2, G-3) zeigte eine leicht unterschiedliche Zusammensetzung (siehe Tabelle 4), jedoch ohne extreme Abweichung. Ein Chi-Quadrat-Test ergab keinen signifikanten Unterschied zwischen den Gruppen ($\chi^2(2, N = 51) = 3,31, p = 0,192$), sodass Verzerrungen durch Geschlechtsunterschiede ausgeschlossen wurden. Insgesamt nahmen 27 Frauen und 24 Männer teil (vgl. Abbildung 3).

	Gruppe		
Geschlecht	G-1	G-2	G-3
Männlich	5	9	10
Weiblich	12	8	7

Tabelle 4: Geschlechterverteilung

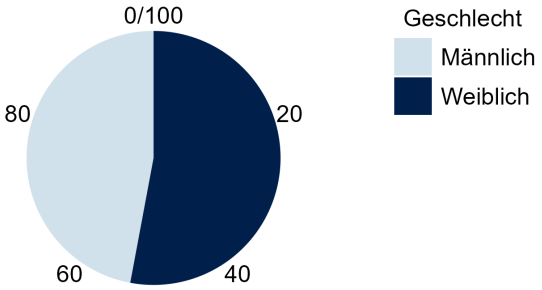

Abbildung 3: Gesamte Geschlechterverteilung aller Teilnehmer

Der höchste erreichte Bildungsabschluss (vgl. Tabelle 5) war in den drei Gruppen ähnlich verteilt. Der Chi-Quadrat-Test zeigte keinen signifikanten Unterschied ($\chi^2(6, N = 51) = 5,09, p = 0,532$), sodass größere Abweichungen ausgeschlossen wurden. Gleiches gilt für das aktuelle Semester (vgl. Tabelle 6) der Proband:innen ($\chi^2(8, N = 49) = 7,95, p = 0,439$).

	Gruppe		
Bildungsabschluss	G-1	G-2	G-3
Abitur	12	9	10
Ausbildung	0	3	1
Bachelor	4	5	5
Master oder höher	1	0	1

Tabelle 5: Höchster Bildungsabschluss pro Gruppe

	Gruppe		
aktuelles Semester	G-1	G-2	G-3
1.- 2. Semester	4	6	9
3.- 4. Semester	4	3	0
5.- 6. Semester	5	3	4
7.- 8. Semester	3	4	2
9. - 10. Semester	0	1	1

Tabelle 6: Aktuelles Semester

Bei der Häufigkeit der Sprachmodellnutzung (vgl. Tabelle 7) gab es ebenfalls keinen signifikanten Unterschied in den Gruppen ($\chi^2(6, N = 51) = 7,86, p = 0,249$). Gleiches galt für die selbst eingeschätzte Technikaffinität (vgl. Tabelle 8) ($\chi^2(8, N = 51) = 12,048, p = 0,149$). Weitere demografische Parameter, wie das Studienfach, die Anzahl bearbeitete MCQs, die investierte Zeit für Klausuren und Klausurvorbereitungsbeginn sind in den Tabellen 14, 15, 16 und Abbildungen 15, 16, 17 im Anhang B dargestellt.

	Gruppe		
Nutzungshäufigkeit	G-1	G-2	G-3
Mehrmals pro Woche	8	5	11
Nie	1	0	0
Selten	4	3	2
Täglich	4	9	4

Tabelle 7: Häufigkeit der Sprachmodellnutzung

	Gruppe		
Technikaffinität	G-1	G-2	G-3
Gering	5	2	0
Hoch	3	7	7
Mittel	8	6	5
Sehr gering	0	0	1
Sehr hoch	1	2	4

Tabelle 8: Technikaffinität

4.3 Einfluss von Prompt Engineering auf die Lernergebnisse

F-2: *Inwiefern wirkt sich der Einsatz von Prompt Engineering bei Sprachmodellen, verglichen zur Nutzung von traditionellen Lernmethoden und dem allgemeinen Einsatz von Sprachmodellen, auf die Lernergebnisse von Studierenden aus?*

Nachfolgend werden die quantitativen Messinstrumente mittels deskriptiven Statistiken und inferenzstatistischen Tests dargestellt.

4.3.1 MCQ-Testergebnisse

Dieses Kapitel widmet sich der Beantwortung der Frage, ob die Gruppen G-1, G-2 und G-3 sich hinsichtlich ihrer Leistung bei einem MCQ-Test unterschieden. Es erfolgt zunächst eine „aggregierte Analyse“ (Summe der korrekt beantworteten MCQs) der Testergebnisse. Anschließend erfolgt eine „Item Level Analyse“ (Betrachtung der Testergebnisse pro MCQ).

Aggregierte Analyse. Im Vergleich erzielten Proband:innen der G-3 den höchsten Gesamtscore (Mittelwert: 6,588), gefolgt von G-2 (Mittelwert 6,353) und G-1 (Mittelwert 6,176), sodass für die Mittelwerte gilt: $\bar{G-1} < \bar{G-2} < \bar{G-3}$. Der Median unterschied sich ebenfalls geringfügig. G-2 und G-3 erreichten einen Median von 7 und G-1 einen von 6. Die Differenzen der Gruppen sind allerdings insgesamt gering. Die detaillierten Kennzahlen können im Anhang C (Tab. 17) entnommen werden. Abbildung 4 zeigt die Verteilung der Testergebnisse in einem Boxplot.

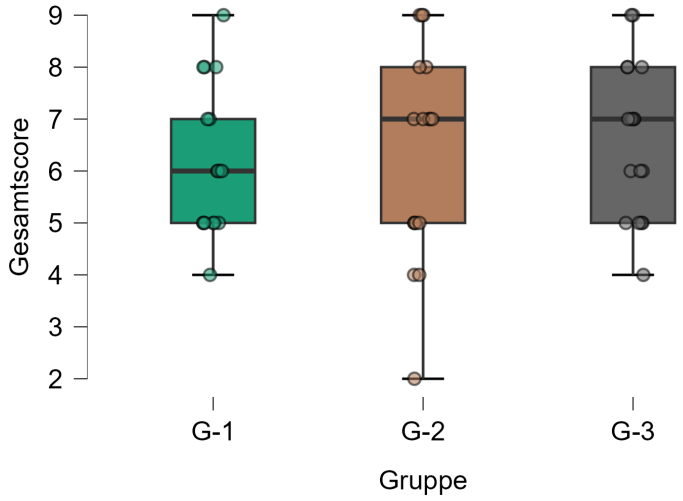

Abbildung 4: Boxplot der aggregierten Testergebnisse pro Gruppe

Mit einem Shapiro-Wilk Testergebnis von p > 0,05 für alle Gruppen, konnte von einer Normalverteilung der Daten in allen Gruppen ausgegangen werden. Diesbezüglich wurden ebenfalls Histogramme und Q-Q-Plots für jede der Gruppen erstellt, die im Anhang C (Tab. 17, Abb. 18, 19, 20, 21, 22, 23) zu finden sind. Zur Überprüfung der Homogenität der Varianzen wurde ein Levene-Test durchgeführt. Die Ergebnisse zeigten, dass sich die Varianzen der Gruppen nicht signifikant unterschieden haben ($F(2, 48) = 1,898, p = 0,161$). Das Boxplot (vgl. Abb. 4) lässt einen ähnlichen Entschluss zu. Somit sind die Varianzen homogen. Die aggregierten Testergebnisse liegen zudem intervallbasiert vor.
Die einfaktorielle ANOVA (vgl. Tab. 9) zeigt mit p > 0,05, dass die Unterschiede der aggregierten Scores zwischen den Gruppen statistisch nicht signifikant waren.

Fälle	Quadratsumme	df	Mittlere Quadratsumme	F	p	η^2
Gruppe	1.451	2	0.725	0.259	0.773	0.011
Residuals	134.471	48	2.801			

Tabelle 9: ANOVA-Test für aggregierte Testergebnisse

Item Level Analyse. Die Unterschiede der Ergebnisse U_E für jede einzelne Frage (Q) waren bei Q1, Q4, Q5, Q8, Q9 und Q10 $U_E \leq 3$. Bei Q2, Q3, Q6 und Q7 lagen die Differenzen der Ergebnisse bei $U_E > 3$. Alle Fragen, bei denen der Unterschied der Summen der korrekten Antworten $U_E \leq 3$ beträgt, werden als weniger differenzierend betrachtet, da dies als geringe Abweichung festgelegt wird. Alle Testergebnisse pro Frage sind im Anhang C (Tab. 18) zu finden. Die Fragen mit $U_E > 3$ werden für vertiefte Analysen herangezogen.

Q2 wurde mit Blick auf die Bloom's Taxonomy dem *1. Remember* und *B. Conceptual Knowledge* zugeordnet (vgl. Abb. 1). Die Trefferquoten waren hier bei allen Gruppen deutlich erhöht. Weiterhin zeigte sich ein Trend, nach welchem G-3 schlechter abschnitt als G-1 und G-2. Der Chi-Quadrat-Test deutete auf einen statistischen Unterschied hin, erreichte allerdings mit $p = 0,056$ nicht das Signifikanzniveau ($\chi^2(2, N = 51) = 5,77, p = 0,056$).

Q3 entsprach dem *3. Apply* und *5. Evaluate* sowie dem *B. Conceptual Knowledge* in der Bloom's Taxonomy (vgl. Abb. 1). Bei den Testergebnissen zwischen den Gruppen wurden extreme Differenzen verzeichnet. Gerade G-2 schnitt mit 82,4 % Trefferquote deutlich besser ab, als G-1 mit 35,3 %. G-3 lag hier mit 64,7 % im Mittelfeld. Diese Differenz bestätigte sich mit dem Chi-Quadrat-Test. Er ergab einen statistisch signifikanten Unterschied ($\chi^2(2, N = 51) = 8,06, p = 0,018$). Das durchschnittliche Ergebnis der Gruppen kann im Intervalldiagramm in Abbilung 5 betrachtet werden.

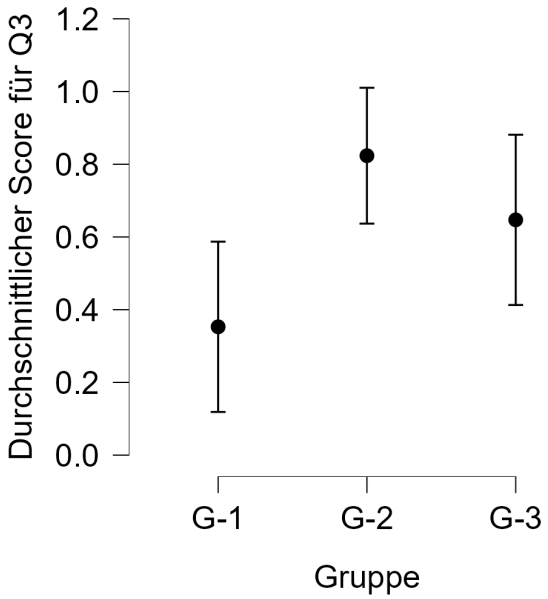

Abbildung 5: Intervalldiagramm für das Testergebnis der Q3

Q6 war Gegenstand von *2. Understand*, *4. Analyze*, *5. Evaluate* sowie dem *B. Conceptual Knowledge* der Bloom's Taxonomy (vgl. Abb. 1). Diese Frage beantworteten von allen 51 Partizipant:innen wenig korrekt (Korrekt: 31,37 %). G-1 schnitt mit einer Summe von drei richtigen Fragen merklich schlechter ab, als G-2 und G-3. Dieser Unterschied war nach dem Chi-Quadrat-Test jedoch nicht statistisch signifikant ($\chi^2(2, N = 51) = 2,368, p = 0,306$).

Q7 wurde dem *1. Remember* und *5. Evaluate* nach Bloom zugeschrieben (vgl. Abb. 1). Interessan-

terweise schnitten Teilnehmer:innen der G-2 und G-1 tendenziell schlechter ab, als Probanden:innen der G-3. Bei diesem Unterschied lag keine statistische Signifikanz vor ($\chi^2(2, N = 51) = 2,059, p = 0,357$).

Zusammenfassend wiesen Q2 (tendenziell) mit $p = 0,056$ und vor allem Q3 mit $p = 0,018$ nach dem Chi Quadrat Test signifikante Unterschiede auf. Q6 mit $p = 0,306$ und Q7 mit $p = 0,357$ zeigten mit ihren Chi-Quadrat-Ergebnissen keine statistisch signifikanten Unterschiede. Q2, Q3, Q6 und Q7 waren die Fragen mit der höchsten Differenz in der Summe der korrekten Antworten U_E.

4.3.2 Metakognitive Einschätzungen (EOL/ JOL)

Ease of Learning (EOL). Die durchschnittlichen EOL-Werte waren über alle Folien hinweg bei jeder Gruppe tendenziell gleich groß. Zwar war die Einschätzung, wie anspruchsvoll die einzelnen Folien zu lernen sind, bei G-3 etwas breiter gestreut ($\sigma_{G-3} = 0,635$). Dennoch lag die Differenz der Mittelwerte zwischen den Gruppen bei maximal 0,129. Das Maximum erreichte bei G-3 jedoch einen erhöhten Wert mit einer maximalen Differenz von 0,8. Mit einem Shapiro-Wilk Testergebniss von p > 0,05 für alle Gruppen und einem Levene Testergebnis von p > 0,05 ($F(2, 48) = 1,340, p = 0,272$) waren die Vorraussetzungen für eine einfaktorielle ANOVA gegeben. Diese bestätigte, dass sich die gemittelten EOL-Werte nicht statistisch signifikant unterschieden ($F(2, 48) = 0,316, p = 0,730$). Die durchschnittlichen EOL-Werte für alle Gruppen sind unter verschiedenen Kennzahlen in Tabelle 10 zu finden.

Differenzen bei den einzelt betrachteten Folien zeigten ähnliche Ergebnisse. Bis auf die EOL-Werte zu Folie 6 und Folie 9 unterschieden sich die Gruppen um maximal 0,3 im Mittelwert. Allerdings konnte mittels Shapiro-Wilk Test bei G-2 jeweils für Folie 6 mit p = 0,001 und Folie 9 mit p = 0,003 nicht von einer Normalverteilung der Werte ausgegangen werden. Die subjektiven Einschätzungen der Teilnehmer pro Folie sind im Anhang C (Tab. 19) zu finden.

	EOL			JOL		
	G-1	G-2	G-3	G-1	G-2	G-3
Mittelwert	2.688	2.559	2.671	3.594	3.712	3.771
Standardabweichung	0.439	0.446	0.635	0.571	0.395	0.385
Minimum	1.800	1.600	1.800	2.500	3.000	3.000
Maximum	3.300	3.500	4.100	4.600	4.200	4.500

Tabelle 10: Durchschnittliche EOL- und JOL-Werte über alle 10 Folien unter verschiedenen Kennzahlen

Judgement of Learning (JOL). Bei den durchschnittlichen JOL-Werten verzeichnete sich eine ähnliche Tendenz. Allerdings war der Mittelwert bei G-2 und G-3 bei der Einschätzung über die zukünftige Testperformance im Vergleich von G-1 mit maximal 0,177 etwas höher als bei der EOL-Einschätzung. Die Standardabweichung zeigte sich im JOL-Wert bei G-1 etwas erhöht ($\sigma_{G-1} = 0,571$). Da der Shapiro-Wilk Test für G-2 bei den gemittelten JOL-Werten p = 0,048 ergab, wurde nicht von einer Normalverteilung ausgegangen. Die durchschnittlichen JOL-Werte sind in Tabelle 10 zu finden.

Differenzen $\geq 0,3$ der einzelnen JOL-Einschätzungen pro Folie wurden bei Folie 2, Folie 3, Folie 5 und Folie 7 konstatiert. Bei allen Einschätzungen, ließ ein Shapiro-Wilk Test mit p > 0,05, sowie ein Levene-Test mit p > 0,05 einen einfaktoriellen ANOVA Test zu. Das Ergebnis war jedoch bei allen Folien, dass es keine statistisch signifikanten Unterschiede gab. Den niedrigsten p - Wert erreichten die Selbsteinschätzungen zu Folie 3 mit p = 0,099 ($F(2, 48) = 2,426, p = 0,099$). Interssanterweise ist der JOL-Mittelwert deutlich höher, als der EOL-Mittelwert über alle Gruppen. Alle weiteren Einschätzungen der Teilnehmer zu jeder Folie befinden sich im Anhang C (Tab. 20)[4].

[4]Hinweis: Da es kaum zur Anwendung des FOKs in der metakognitiven Einschätzung der Teilnehmer:innen kam, wurde dieses Messinstrument bei der Auswertung nicht evaluiert.

4.4 Effizienz- und Effektivitätsunterschied bei Verwendung von PE Techniken

F-3a: *Inwieweit steigert die Verwendung von Prompt Engineering-Techniken, bei Sprachmodellen, die Effizienz und Effektivität der Klausurvorbereitung von Studierenden?*

Es erfolgt zunächst die Betrachtung der Effizienz. Anschließend findet die Auseinandersetzung mit der Effektivität unter verschiedenen Gesichtpunkten statt.

4.4.1 Effizienz

Nachfolgend wird die Zeiterfassung als objektive Messgröße für die Effizienz in der Klausurvorbereitung beschrieben. Nachfolgend erfolgt die Analyse der subjektiven Bewertungen der Effizienz von den Teilnehmer:innen unter verschiedenen Fragen.

Zeiterfassung. Bei der Analyse der Zeiterfassung umfasst die Vorbereitungszeit die gesamte Zeit für die EOL- und JOL-Zwischenphasen, sowie der Vorbereitungsphase[5]. Die Testzeit ist die Zeit, die in der Testphase benötigt wurde. Die Ergebnisse sind in Tabelle 11 einsehbar. Hierbei waren Unterschiede in der Vorbereitungszeit zwischen G-1 (Mittelwert: 32,949) und G-3 (Mittelwert: 39,861) von 6,912 Minuten erkennbar. Weiterhin war das Minimum in der Vorbereitungszeit von G-3 mit 30,7 Minuten sichtbar höher als bei G-1 (Unterschied: 15,83 Minuten) und G-2 (Unterschied: 24,49 Minuten). Die Testzeiten wichen kaum voneinander ab. Eine differenzierte Übersicht für jede Zwischenphase ist im Anhang C (Tab. 21) zu finden.

	Vorbereitungszeit (Min)			Testzeit (Min)		
	G-1	G-2	G-3	G-1	G-2	G-3
Mittelwert	32.949	35.526	39.861	12.269	12.856	12.157
Standardabweichung	9.020	9.555	4.668	2.896	3.708	2.972
Minimum	14.870	6.210	30.700	7.420	4.690	6.900
Maximum	46.220	42.810	49.920	17.190	18.950	17.180

Tabelle 11: Übersicht der Zeiterfassungsdaten

Bei der Übprüfung der Normalverteilungsannahme der Vorbereitungszeit mittels Shapiro-Wilk-Test wurde bei G-2 eine signifikante Verletzung der Normalverteilung mit $p < 0{,}001$ festgestellt. Die exakten Werte für alle Gruppen sind im Anhang C (Tab. 22) angegeben. Aufgrund der Verletzung wurde auf eine einfaktorielle ANOVA verzichtet und stattdessen ein Kruskal-Wallis-Test durchgeführt. Der Test ergab beim Faktor der Gruppe einen Chi-Quadrat Wert von $H = 5{,}660$ bei $df = 2$ und einen p-Wert von 0,059. Somit konnten in den Vorbereitungszeiten zwischen den Gruppen keine statistisch signifikanten Unterschiede erkannt werden. Dennoch sind die Unterschiede zwischen den Vorbereitungszeiten tendenziell stark abweichend.

Subjektive Bewertung. Die Ergebnisse der subjektiven Effizienzbewertung der jeweiligen Lernmethode sind in Tabelle 12 illustriert. Die ausformulierten Fragen zu den Fragencodes befinden sich im Anhang C (Kap. C.6).

[5] Grund hierfür war die verzerrte durchschnittliche Zeit bei einigen Teilnehmern, die nicht auf der Vorbereitungsseite des Experimentes verweilten, sondern direkt zur JOL-Phase weiterklickten. Dementsprechend werden Zeiten aggregiert evaluiert.

Bis auf eine Ausnahme lagen die Medianwerte bei allen Fragen bei 4. Dies entspricht einer überwiegenden Zustimmung zur subjektiven Effizienz unter verschiedenen Gesichtpunkten unter allen Gruppen. Die Mittelwerte bewegten sich zwischen 3,4 und 4,1, was zwischen neutral und zustimmend liegt. Dies deutet auf eine positive Bewertung mit variiender Intensität hin.

Fragencode	EZP			EZZ			EZST			EZB		
	G-1	G-2	G-3	G-1	G-2	G-3	G-1	G-2	G-3	G-1	G-2	G-3
Median	4.000	4.000	4.000	4.000	4.000	4.000	3.000	4.000	4.000	–	4.000	4.000
Mittelwert	3.588	3.588	4.000	3.471	4.000	3.706	3.471	3.647	3.412	–	3.882	3.824

Tabelle 12: Ergebnisse der subjektiven Bewertung der Effizienz

Interessanterweise war die Frage danach, wie viel der 30 Minuten Vorbereitungszeit effizient genutzt wurden, (**EZP**) bei G-3 im Mittelwert etwas höher als G-1 und G-2 (Differenz: 0,4). Bei der subjektiven Einschätzung des Zeitaufwandes (**EZZ**) ließ sich ein erhöhter Trend von G-2 gegenüber G-3 (Differenz: 0,3) und G-1 (Differenz: 0,5) feststellen. Die subjektive Stresseinschätzung (**EZST**) lag bei G-2 im Mittelwert etwas höher. Alle Studierende die ChatGPT nutzten nahmen tendenziell eine Beschleunigung in der Prüfungsvorbereitung durch den Chatbot war (**EZB**).
Da die Unterschiede zwischen den Gruppen maginal waren und keine praktischen Implikationen nahelegten, wurde auf statistische Signifikanztest verzichtet.

4.4.2 Effektivität

Nachfolgend wird die Effektivtät zunächst unter den objektiven MCT-Ergebnissen und anschließend unter der subjektiven Bewertung der Proband:innen behandelt.

MCT-Ergebnisse. Die objektive Effektivität wurde mittels der MCT-Ergebnisse bestimmt (siehe Kap. 4.3.1). Gruppe G-3 erreichte hierbei tendenziell die besten Ergebnisse, allerdings nicht mit signifikanten Unterschieden.

Affordanzen des Sprachmodells von G-2. 14 von 17 Teilnehmer:innen nutzten das Modell hauptsächlich um sich Inhalte erklären zu lassen („Erklärungen"). Elf Personen verwendetend das LLM, um die Folien zusammenzufassen („Zusammenfassen"). Lediglich sechs Studierende benutzten das Sprachmodell um einen „Selbsttest [zu] erstellen" und nur eine Person für das „Planen". Die Ergebnisse kamen durch die Inhaltsanalyse nach Mayring zustande. Die prozentualen Werte sind dem Balkendiagram der Abb. 6 zu entnehmen.

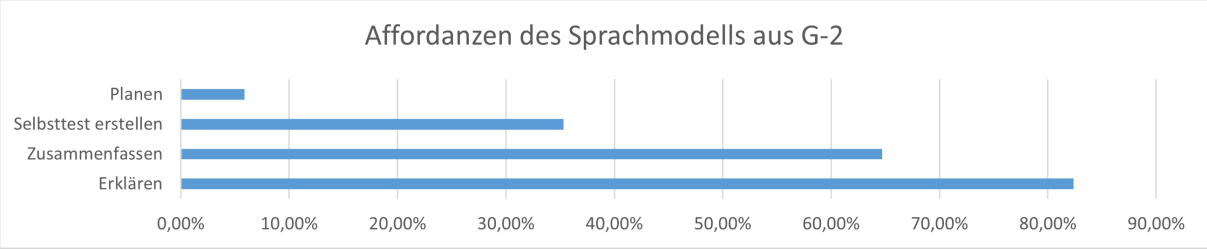

Abbildung 6: Affordanzen des Sprachmodells von G-2 nach Inhaltsanalyse

Subjektive Bewertung. Das Ergebnis der Skalenfrage ist in Tabelle 13 zu finden. Die Ergebnisse der Freitextfragen zur Effektivtät sowie Ineffektivität sind in Abbildung 7 und 8 dargestellt. Der exakte Wortlaut der Fragen ist im Anhang C (Kap. C.7) zu finden.

	ETV (in %)			ETMV		ETLS
	G-1	G-2	G-3	G-2	G-3	G-3
Median	99.000	99.000	99.000	4.000	4.000	4.000
Mittelwert	89.118	83.294	93.471	4.176	4.059	3.471
Standardabw.	17.450	19.186	10.566	–	–	–

Tabelle 13: Ergebnis der subjektiven Bewertung der Effektivität

Die Selbsteinschätzung über das Verständnis der Folieninhalte (**ETV**) war beim Mittelwert bei Gruppe G-3 gegenüber G-1 und G-2 erhöht. Allerdings zeigte sich der Median mit 99 % unverändert bei allen Gruppen. Aufgrund der Shapiro-Wilk-Testergebnisses von $p < 0{,}001$ für alle Gruppen (vgl. Anhang C, Tab. 23), wurde ein Kruskal-Wallis-Test durchgeführt. Dieser ergab keine statistisch signifikanten Unterschiede zwischen den Gruppen ($H(2) = 3,761, p = 0,152$). Zur Frage, ob die Nutzung des Sprachmodells half die Inhalte besser zu verstehen (**ETMV**), gab es von den Studierenden große Zustimmung, da beide Mittelwerte für G-2 und G-3 über 4 lagen. Für die Einschätzung der Prompts (**ETLS**) zeichnete sich ein weniger zustimmendes Bild ab. Dennoch wurden die bereitgestellten Prompts mit einem Mittelwert von 3,5 tendenziell als hilfreich angesehen.

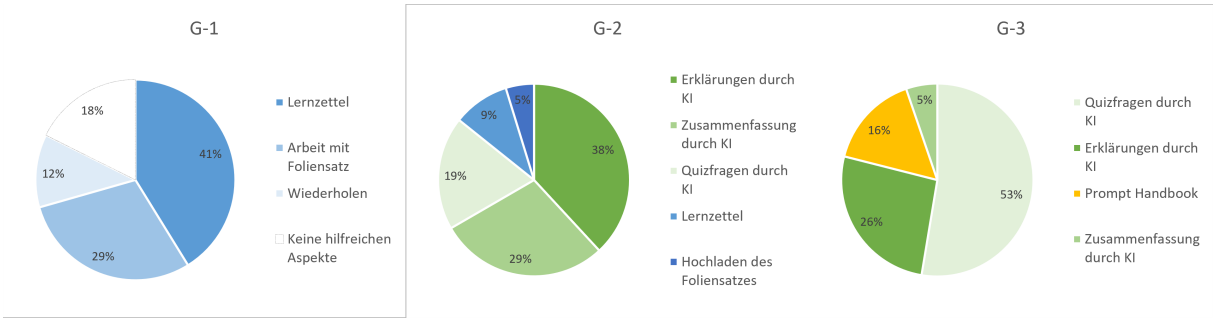

Abbildung 7: Subjektive Aspekte, die zur Effektivität beitrugen (ETHA)

Hinweis: Es waren mehrere Nennungen pro Teilnehmer:in möglich, daher wird eine prozentuale Darstellung genutzt. Die Farben wurden bewusst gewählt: Blaue kennzeichnet traditonelle Methoden, Grün steht für Methoden in Zusammenarbeit mit KI, und Gelb hebt das Prompt Workbook hervor.

Aus der Inhaltsanalyse nach Mayring konstatierten sich für G-1 vor allem *Lernzettel* (41,2 %) und die *Arbeit mit dem Foliensatz* (29,4 %) als wichtigste Faktoren für die effektivte Klausurvorbereitung (**ETHA**) heraus.

Für G-2 wurden die *Erklärungen durch KI* (38,1 %) und *Zusammenfassungen durch KI* (28,6 %) am häufigsten als effektiv (**ETHA**) bewertet. Auch *Quizfragen durch KI* (19,0 %) empfanden Studierende bei der Lernvorbereitung als hilfreich. *Lernzettel* (9,5 %) spielten eine untergeordnetere Rolle, im Gegensatz zu G-1. 85,7 % der genannten Aspekte standen im Zusammenhang mit KI.

Das Prompt Handbook wurde bei G-3 als Oberbegriff indirekt von 15 Teilnehmer:innen als hilfreich für ihre Klausurvorbereitung (**ETHA**) angeführt. In der Abbildung 7 ist es fein granularer für Vergleichbarkeit mit G-2 eingeteilt. Es wird deutlich, dass die *Quizfragen duch KI* (52,6 %) häufiger als nützlich angesehen wurden, als bei G-2. Aber auch die *Erklärungen von KI* (26,3 %) und explizit das *Prompt Handbook* (15,8 %) im allgemeinen wurden als subjektiv effektiv identifiziert. Hierbei standen alle genannten Aspekte im Zusammenhang mit KI.
Eine Häufigkeitstabelle für alle Kategorien ist im Anhang C (Tab. 24) einsehbar.

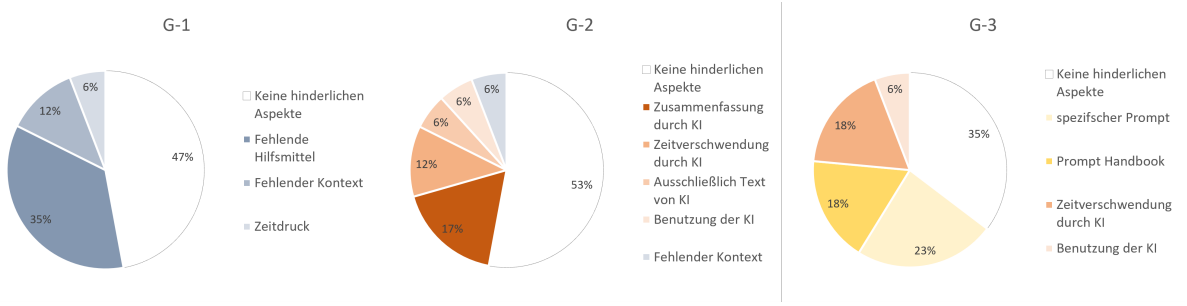

Abbildung 8: Subjektive Aspekte, die zur Ineffektivität beitrugen (ETNH)

Hinweis: Die Farben wurden bewusst gewählt: Graue kennzeichnet traditonelle Methoden, Organge steht für Methoden in Zusammenarbeit mit KI, und Gelb hebt das Prompt Workbook hervor.

Nach der Inhaltsanalyse nannten die Teilnehmenden der G-1 als häufigsten Grund für die Ineffizienz (**ETNH**) die *Fehlende[n] Hilfsmittel* (35,3 %). Weiterhin war auch der *Fehlende Kontext* (11,8 %) und *Zeitdruck* (5,9 %) Indikatoren. Rund die Hälfte der Teilnehmenden (47,1 %) gaben an, keine hinderlichen Aspekte erlebt zu haben.

Bei der Benutzung eines Sprachmodells (G-2) wurde tendenziell häufig die KI für die Ineffizienz (**ETNH**) (insgesamt: 41,2 %) verantwortlich gemacht. So wurden die *Zusammenfassungen durch KI* (17,6 %) und *Zeitverschwendung durch KI* (11,8 %) am konstantesten angeführt. Mit der *Benutzung von KI* (5,9 %) und *Ausschließlich Text von KI* (5,9 %) zeichneten sich insgesamt 35,3 % subjektiv-ineffektive Aspekte bei der Klausurvorbereitung mit KI ab. Trotzdessen gaben 53 % der Teilnehmer:innen an, keine hinderlichen Aspekte wahrgenommen zu haben.

Einige Teilnehmer, die das Prompt Handbook nutzten (G-3), hatten bezüglich der Ineffizienz (**ETNH**) explizite Probleme mit dem *Prompt Handbook* (17,6 %). Jedoch wurden ebenfalls einzelne *spezifische Prompts* (23,5 %), sowie die *Zusammenfassung durch KI* (17,6 %) bemängelt. 35,3 % gaben an keine hinderlichen Aspekte wahrgenommen zu haben.
Eine Häufigkeitstabelle mit absoluten Zahlen ist im Anhang C (Tab. 25).

4.5 Vor- und Nachteile vormodellierter Prompts vs. ad-hoc Nutzung

F-3b: *Welche Vor- und Nachteile ergeben sich aus der Nutzung von Sprachmodellen mit vormodellierten Prompts und der „ad hoc“ Verwendung von Sprachmodellen?*

Im nachfolgenden wird das subjektive Feedback zu Vor- und Nachteilen mit Blick auf die „ad hoc”-Verwendung von Sprachmodellen (G-2) gegenüber der Nutzung von Sprachmodellen mit vormodellierten Prompts (G-3) verglichen. Anschließend erfolgt das separate Feedback zum Prompt Handbook (PH) von G-3.

4.5.1 Vergleich G-2 und G-3

G-2 und G-3 werden im folgenden vor allem in Bezug auf die genannten Vor- und Nachteile verglichen. Zuvor wird der Nachteil des Autonomieverlusts einzeln betrachtet. Die exakten Wortlaute der Fragen für den Autonomieverlust, sowie die Vor- und Nachteile sind im Anhang C (Kap. C.11) zu finden.

Autonomieverlust. Benutzer:innen der Gruppen G-3 hatten das Gefühl, weniger Kontrolle in der Vorbereitungsphase abzugeben (Mittelwert: 3,65; Median: 4) als Teilnehmer:innen der G-2 (Mittelwert: 3,24; Median: 4). Die Differenzen unterscheiden sich jedoch nicht stark. Folglich empfanden die Teilnehmer:innen die Nutzung des Sprachmodells tendenziell nicht als Beeinträchtigung in der Selbständigkeit des Lernens. Ein ANOVA-Test kam durch das p-Wert Ergebnis des Shapiro-Wilk Tests nicht infrage. Das

Kurskal-Wallsi-Testergebnis zeigte, dass es keine statistisch signifikanten Unterschiede gab (p = 0,280). Relevante Kennzahlen sind im Anhang C (Tab. 26) zu finden.

Vorteile. Die Ergebnisse der offenen Frage nach den Vorteilen zur jeweiligen Methode an die Teilnehmer:innen ließen eine Inhaltsanalyse nach Mayring zu. Die Ergebnisse sind in den Abbildungen 9 (ad hoc Nutzung) und 10 (vormodellierte Prompts) illustriert.

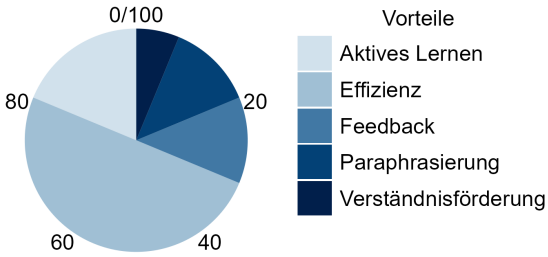

Abbildung 9: Vorteile in der „ad-hoc" Nutzung eines Sprachmodells (G-2)

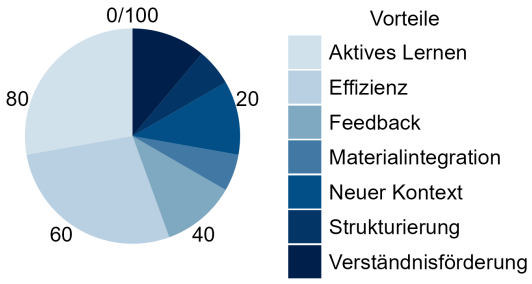

Abbildung 10: Vorteile in der Sprachmodellnutzung mit vormodellierten Prompts (G-3)

Hinweis: Es waren mehrere Nennungen pro Teilnehmer:in möglich, daher wird eine prozentuale Darstellung genutzt. Bei G-2 und G-3 machten jeweils drei Teilnehmer:innen keine Angaben. Diese Angaben sind nicht Teil des Diagramms.

Bei Gruppe G-2 gaben insgesamt 14 Teilnehmer:innen Vorteile an, wobei eine Mehrfachnennung möglich war. So wurde bei der „ad hoc" Nutzung des Sprachmodells (G-2) mit 50 % die *Effizienz* genannt. Sie macht die Hälfte der Vorteile aus, gefolgt vom *Aktiven Lernen* mit 18,8 %.

In Gruppe G-3 gaben auch 14 Teilnehmer:innen mit der Möglichkeit zur Mehrfachnennung Vorteile an. Mit der Nutzung des Prompt Handbooks (G-3) wurden ebenfalls das *Aktive Lernen* (27,8 %) und die *Effizienz* (27,8 %) am häufigsten genannt, wenngleich das *Aktive Lernen* häufiger als bei G-2 angegeben wurde. Während die *Paraphrasierung* mit 12,5 % bei der „ad hoc" Nutzung (G-2) auch noch ein Thema war, taucht sie bei G-3 gar nicht mehr auf. Staddessen wurden hier neue Vorteile wie *Neuer Kontext*, *Strukturierung* und *Materialintegration* genannt. Die Häufigkeitstabellen der Vorteile für jede Gruppe sind im Anhang C (Tab. 27, 28) zu finden.

Nachteile. Aufgrund der Vergleichbarkeit wurden für die Nachteile der jeweiligen Gruppen auch Inhaltsanalysen nach Mayring durchgeführt. Die Ergebnisse sind in den Abbildungen 11 und 12 zu finden.

Von Gruppe G-2 nannten bei der direkten Nutzung des Sprachmodells anteilig 66,7 % die *Halluzination* als Problem. Weiterhin wurde aber auch die *Promptabhängigkeit* und der *Verlust des kritischen Hinterfragens* erwähnt. Es gaben insgesamt sechs Teilnehmer Nachteile an.

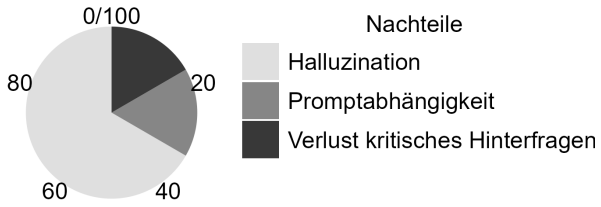

Abbildung 11: Nachteile in der „ad-hoc" Nutzung eines Sprachmodells (G-2)

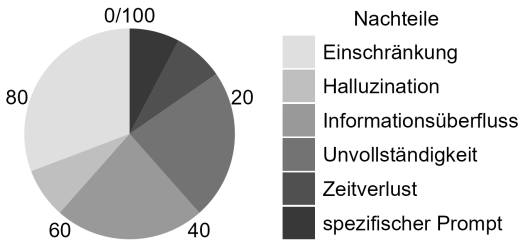

Abbildung 12: Nachteile in der Sprachmodellnutzung mit vormodellierten Prompts (G-3)

Hinweis: Es waren mehrere Nennungen pro Teilnehmer:in möglich, daher wird eine prozentuale Darstellung genutzt. Bei G-2 gaben elf und bei G-3 sechs der Teilnehmer:innen nichts an. Diese Angaben sind nicht Teil des Diagramms.

Bei Gruppe G-3 ergab sich ein gemischtes Bild, wenngleich die *Halluzination* (7,7 %) deutlich weniger als Nachteil genannt wurde. Das größte Problem nach Meinung der Studierenden waren die *Einschränkung[en]* (30,8 %) durch das Prompt Handbook. Allerdings wurde auch der *Informationsüberfluss* und die *Unvollständigkeit* mit jeweils 23,1 % genannt. Insgesamt gaben elf Partizipant:innen Nachteile an. Die Häufigkeitstabellen für die Nachteile der Gruppen G-2 und G-3 sind dem Anhang C (Tab. 29, 30) zu entnehmen.

4.5.2 Feedback zum Prompt Handbook

Im folgenden wird das Feedback der Teilnehmer:innen der Gruppe G-3 zum Prompt Handbook unter verschiedenen Gesichtspunkten differnziert betrachtet. Die oben angegeben Vor- und Nachteile der Gruppe G-3 können dabei ebenfalls in Betrachtung kommen. Die ausformulierten Fragen zu den Messinstrumenten sind im Anhang C (Kap. C.15) zu finden.

Verwendete Prompts. Der am häufigsten genutzte Prompt aus dem Prompt Handbook war „2.4 Selbsttests anfertigen", den 83,4 % der 17 Teilnehmer:innen verwendeten. Danach folgten „2.3 Hinweise zum Vorgehen bereit stellen" (58,8 %), „2.5 Lerninhalte festigen" (58,8 %) und „2.1 Lass uns Schritt für Schritt nachdenken" (41,2 %). Weniger zur Anwendung kamen „2.2 Gedankenbaum" (23,5 %), „2.8 Erkläre es ChatGPT" (17,7 %) und „2.7 Der andere Blickwinkel (Reframing)" (5,9 %). Der Prompt „2.6 Passende Hilfe erhalten" wurde hingegen in keiner Weise genutzt. Die relative Nutzung ist in Abbildung 13 zu finden. Die aboluten Häufigkeiten finden sich im Anhang C (Tab. 31).

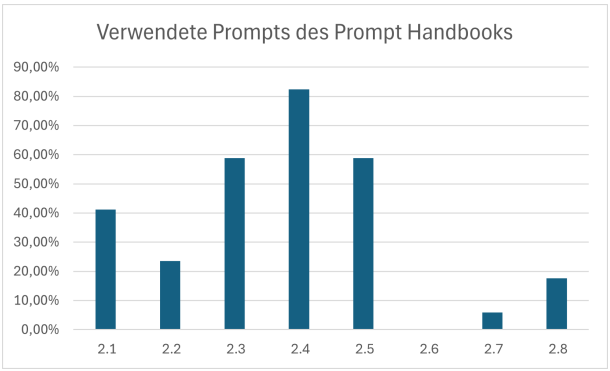

Abbildung 13: Verwendete Prompts der Gruppe G-3 aus dem Prompt Handbook

Nutzungsbewertung & Benutzerfreundlichkeit. In der Abbildung 14 ist ein Likert-Diagramm zur Nutzungsbwertung und Benutzerfreundlichkeit angegeben.

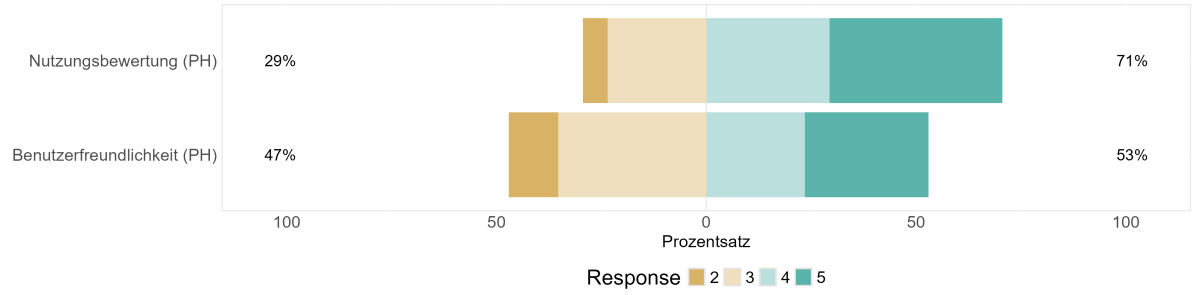

Abbildung 14: Subjektive Nutzungsbewertung und Benutzerfreundlichkeit des Prompt Handbooks (PH)

Die Nutzungsbewertung fiel überwiegend positv aus. 71 % der Befragten Personen fanden die vormodellierten Prompts und Affordanzen gegenüber der freien Nutzung hilfreich. 23,5 % der Befragten

waren neutral und 5,9 % sahen das Prompt Handbook als tendenziell hinderlich an. Niemand stufte die vormodellierten Prompts als „gar nicht hilfreich" ein.
Die Benutzer:innen zeigten sich bei der Benutzerfreundlichkeit etwas neutraler (35,3 %), wenngleich die allgemeine Benutzerfreundlichkeit mit 53 % durchaus positv ausfiel. Zurückzuführen sind die neutraleren Ergebnisse vermutlich auf die Nachteile (vgl. Abb. 12) von G-3 des *Informationsüberfluss*, sowie der *Einschränkung* bezüglich des Prompt Handbooks.

Schwierigkeiten. Mit Schwierigkeiten werden im folgenden expliziete Prompts des Prompt Handbooks angesprochen mit denen die Studierenden Probleme hatten. Aber auch generelle Benutzungsprobleme und äußere Einflüsse werden hier mit aufgenommen. Zu folgenden Prompts wurden Schwierigkeiten gemeldet:

- **2.1 Lass uns Schritt für Schritt nachdenken**: Problem, dass „keine sinvollen und nutzbaren Informationen heraus" kamen.
- **2.2 Gedankenbaum**: Problem zu filtern welche Antwort „wichtig für die Klausur ist".

Auch wurde von zwei Teilnehmer:innen das Problem beschrieben, dass der Chatbot manchmal über den „Foliensatz hinaus ging", was im Prompt Handbook unter „3. Weitere Hinweise" mit Ratschlägen, sich auf den Foliensatz zu beziehen, verhindert werden sollte.
Drei der Partizipant:innen erwähnten die Zeit als generelles Problem. Sie hätten grundsätzlich mehr Zeit gebraucht.
Weiterhin wurden von jeweils einem Studierenden Detailverlust, Anfangsschwiergkeiten und das erneute Nachfragen bei komplexeren Antworten als Problem konstatiert.

Weiterempfehlung des Prompt Handbooks. Alle 17 Partizipant:innen würden das Prompt Handbook ihren Mitstudierenden weiterempfehlen, wenn auch zwei davon mit einer Einschränkung.
Eine Person erwähnte erneut den *Zeitdruck*, unter dem das Prompt Handbook genutzt werden musste. Die andere Person würde das Prompt Handbook noch etwas spezifischer an die Klausurvorbereitung anpassen und die Ergebnisse kritisch hinterfragen.
Alle anderen positiven Gründe für das Prompt Handbook waren:

- Selbsttests (4x)
- Nutzung von Lernstrategien (2x)
- Feedback
- Effizienz (3x)
- Erklärungen (2x)
- Verständnisförderung
- Strukturierte Darstellung von Inhalten
- Wissen festigen
- Organisation
- Einstiegshilfe

5 Diskussion

Ziel des Between-Groups-Experiments war die Untersuchung und Evaluation der optimierten Nutzung von Prompt Engineering (PE) in Sprachmodellen zur effizienten und effektiven Klausurvorbereitung. Die Ergebnisse mit insgesamt 51 Studierenden zeigten mit Ausnahme einer Teilfrage (Q3) keine statistisch signifikanten Unterschiede in den Lernerergebnissen zwischen den Gruppen. Auch bezüglich objektiver Effektivität und Effizienz konnte kein statistisch signifikanter Unterschied festgestellt werden. Subjektiv beschrieben die Studierenden jedoch gruppenspezifisch unterschiedliche Faktoren für die effektive Klausurvorbereitung. Die Anwendung des Prompt Handbooks wurde teilweise als einschränkend und fremdbestimmt wahrgenommen, führte jedoch zur häufigeren Verwendung des Sprachmodells für das aktive Lernen, im Vergleich zur „ad hoc“ Nutzung, bei der das Zusammenfassen im Vordergrund stand.
Die Ergebnisse sind aufgrund der begrenzten Teststärke teilweise mit Vorsicht zu interpretieren.

5.1 Kein Einfluss von PE/ ChatGPT auf Lernergebnisse und Metakognition

MC-Testergebnisse. Die Untersuchung der Testergebnisse zeigt insgesamt, dass sich die *kurzfristige* Verwendung von ChatGPT, sei es mit oder ohne Prompt Handbook, *neutral* auf die Testergebnisse auswirkt. Dadurch liegt die Vermutung nahe, dass das Prompt Engineering in der durchgeführten Form tendenziell irrelevant für die quantifizierbaren Lernergebnisse in der Prüfung ist. Dieses Ergebnis steht insofern in Übereinstimmung mit Bastani et al. (2024, 9), als dass die abgesicherte Nutzung von ChatGPT die Closed-Book Prüfungsergebnisse nicht verschlechtert hat. Gleichzeitig widerspricht es auch den Befunden von Bastani et al. (2024, 2), dass die freie Verwendung von ChatGPT in der Prüfungsvorbereitung zur Verschlechterung der Prüfungsergebnisse führte. Dieser Aspekt könnte allerdings auf die unterschiedliche Vorbereitungszeit (maximal 30 Minuten in dieser Arbeit) sowie der Art der Prüfungsform (in diesem Fall ausschließlich MC-Fragen) zurückzuführen sein. Es ist wahrscheinlich, dass Studierende sich auf diesen MC-Test anders vorbereitet haben, als bei bewerteten Prüfungen (Rovers et al., 2018; Fiorella und Mayer, 2016, 9, 17). Dennoch besteht Grund zur Annahme, dass die Verwendung von ChatGPT bei der *kurzfristigen* Klausurvorbereitung nicht im Allgemeinen zur Verbesserung der Ergebnisse führt. Die Nutzung der PE Techniken hatten vergleichsweise keinen Einfluss auf die Testergebnisse.

Jedoch kann ChatGPT, besonders wenn ein Kontext fehlt, nützlich sein. Bei einer spezifischen MC-Frage (Q3) konnten signifikante Unterschiede vor allem zwischen der G-1 und G-2 festgestellt werden. Diese Frage zielte nach Bloom et al. (1956) und Krathwohl (2002) auf dem Verstehen eines grafischen Modells (Understand) und der anschließenden Anwendung auf ein neues Beispiel (Apply) ab. Diese Feststellung könnte darauf hindeuten, dass ChatGPT bei bestimmten Themen helfen kann, einen Kontext zu erschließen. In einer Publikation von Luckin (2010, 262) wurde dies allgemein als Möglichkeit des „learner-generated context” durch Technologie beschrieben. Dieser Umstand sollte jedoch weiter untersucht werden. Erwähnenswert sind hierbei die Ergebnisse aus der Arbeit von Dell'Acqua et al. (2023, 1), in der die Beantwortung von wissensintensiven Aufgaben mit ChatGPT untersucht wurde. Dabei wurde ebenfalls die Verwendung von ChatGPT für einige Aufgaben als effizienter angesehen, in anderen wiederum nicht (Dell'Acqua et al., 2023, 10, 14).

Metakognitive Einschätzungen. Da es in Übereinstimmung mit den Ergebnissen des MC-Tests ebenfalls keine signifikanten Unterschiede unter den Gruppen im Ease-of-Learning (EOL) und Judgement-of-Learning (JOL) gab, lassen sich diese Ergebnisse als positiv interpretieren. Der konstante EOL-Wert könnte darauf hindeuten, dass Studierende nicht vermuten, sich durch ChatGPT die Inhalte des Foliensatzes besser aneignen zu können. Dies spricht dafür, dass der Einsatz von ChatGPT keine Verzerrung in der Wahrnehmung der Studierenden hinsichtlich der Leichtigkeit des Lernprozesses (EOL) hatte. Beim JOL-Wert zeichnet sich ein ähnliches Bild ab. Die Studierenden fühlten sich durch den Lernprozess mit

ChatGPT nicht besser vorbereitet, als die Vergleichsgruppe ohne ChatGPT. Auch hier deutet somit nichts auf einen Bias in dem subjektiven Vorbereitungsstand durch ChatGPT hin. Es ist zu beachten, dass sich die EOL/ JOL-Einschätzungen auf ein kleines Zeitfenster während der Vorbereitungsphase beziehen. Dennoch ist hervorzuheben, dass zumindest in der *kurzfristigen* Verwendung von ChatGPT keine Bias in der Selbsteinschätzung von Studierenden erkennbar wurden.

Ferner war beim Judgement-of-Learning bei allen Gruppen eine deutliche Tendenz zu erkennen, dass Studierende im Durchschnitt nach der Lernphase das Gefühl hatten, viele Inhalte direkt abrufen zu können. Dies steht im engen Zusammenhang mit Nelson und Narens (1990, 134), dass Studierende sich beim direkten Abruf von Informationen vor allem auf Informationen aus dem Kurzzeitgedächtnis beziehen. Gerade vor dem Hintergrund, dass die EOL-Werte niedriger waren, zeigt sich, dass Studierende annehmen, etwas gelernt zu haben, dies jedoch nur im Kurzzeitgedächtnis und nicht im Langzeitgedächtnis abgespeichert ist. Zu diesen Befunden sind weitreichende Studien sinnvoll, welche die langfristigen JOL-Werte mit Blick auf die Verwendung von Sprachmodellen untersuchen.

Zusammenfassung. Zusammenfassend lässt sich sagen, dass die kurzfristige Nutzung von ChatGPT und Bereitstellung von PE Techniken in dieser Studie keine signifikanten Auswirkungen auf die Prüfungsergebnisse hatten oder die metakognitiven Einschätzungen der Studierenden verzerrten. Allerdings hat ChatGPT das Potenzial in bestimmten Situationen nützlich zu sein, insbesondere bei einem fehlendem Kontext. Für die langfristige Nutzung von ChatGPT und PE sind längerfristige Studien nötig, welche die Metakognition und benotete Prüfungsleistung über längere Zeiträume untersuchen.

5.2 Objektiv tendenziell ineffizient, subjektiv effizient

Zeiterfassung. G-3 brauchte am meisten Vorbereitungszeit. Sie benötigte im Schnitt vier Minuten mehr als G-2 und sieben Minuten als G-1. Weiterhin war das Zeitvolumen für die explizite Vorbereitungsphase bei G-3 ebenfalls höher.

Diese Differenzen könnten darauf hindeuten, dass das PH ChatGPT als „Sparringspartner“ etabliert hat. Studierende lassen sich infolgedessen länger auf den Lernprozess ein und haben mehr Motivation, sich mit dem Material auseinanderzusetzen. Das könnte jedoch generell für ChatGPT auch ohne Prompt Handbook gelten. In Anlehnung an diese Interpretation kann eine Arbeit von Beltozar-Clemente und Díaz-Vega (2024, 89) herangezogen werden, die durch ChatGPT das Potenzial sehen, die Haltung von Studierenden gegenüber dem Lernprozess positiv zu verändern. Dies wird auch von Studierenden in der Arbeit von Chan und Hu (2023, 9) angenommen.

Dagegen spricht jedoch, dass Studierende besonders durch das Prompt Handbook einen zusätzlichen Störfaktor empfanden, der den Lernprozess ineffizient gestaltet hat. Hierzu kann der *Split-Attention-Effekt* (Ayres und Sweller, 2005) erwähnt werden, der beschreibt, dass die kognitive Belastung durch mehrere Quellen, die Informationen zur Verfügung stellen, erhöht wird (Zumbach, 2010, 82f.). Der ständige Wechsel zwischen Foliensatz, Prompt Handbook und ChatGPT auf zwei unterschiedlichen Geräten könnte die kognitive Belastung erhöhen und die Vorbereitung ineffizient gestaltet haben.

Ob durch die erhöhte Zeit eine wirkliche Ineffizienz zu beobachten ist, bleibt unklar. Eine Fusion der Interpretationen könnte auch eine mögliche Antwort auf den größeren Zeitaufwand sein. Es besteht die Annahme, dass der kognitive „Load“ durch ChatGPT und dem PH erhöht wurde, allerdings auch die Motivation und Lernbereitschaft, da die Studierenden einen Partner im Lernprozess hatten. Es verringert vielleicht die Effizienz, erhöht jedoch die Motivation, sich mit einem Problem auseinanderzusetzen. Hierzu sollten weitere Untersuchungen erfolgen, die den Split-Attention-Effekt in Zusammenhang mit ChatGPT weiter untersuchen, sowie längere Vorbereitungsphasen und ChatGPT als Sparringspartner.

Subjektive Bewertung. Ein signifikanter Unterschied in der subjektiven Bewertung der effizienten Nutzung der Vorbereitungszeit (EZP), des Zeitaufwandes (EZZ) und des Stresses (EZST) zeigt sich nicht. Dies relativiert die objektiv gemessene längere Vorbereitungszeit von G-3.

Insbesondere die Einschätzung zur Beschleunigung der Vorbereitungszeit durch ChatGPT (EZB) fielen in den Gruppen G-2, als auch G-3 ähnlich aus. In beiden Gruppen lag der Median bei vier, was auf eine deutlich wahrgenommene Beschleunigung der Vorbereitung hinweist. Dabei macht es keinen Unterschied, ob das Prompt Handbook, oder das Sprachmodell verwendet wird. In einer Arbeit von Ngo (2023, 14) wurden ähnliche Ergebnisse konstatiert, wobei viele Studierende bestätigten, dass ChatGPT helfen kann, Zeit zu sparen. Auch nach Chan und Hu (2023, 13) ist KI in der Lage, die Effizienz zu fördern.

Zusammenfassung. Objektiv kam es in der Vorbereitungsphase zu zeitlichen Ineffizienzen vor allem bei Gruppe G-3. Dies könnte auf eine erhöhte kognitive Belastung, als auch eine gesteigerte Motivation, sich mit dem Foliensatz auseinander zu setzen, hinweisen. Subjektiv wird der Nutzung von ChatGPT eine klare Effizienzsteigerung von den Studierenden zugeschrieben, unabhängig davon, ob das Prompt Handbook verwendet wurde.

5.3 ChatGPT ohne PH: Wenig Affordanzen, hilfreich, Regelwerk nötig

Gemessen an den äquivalenten Testergebnissen wurde, objektiv betrachtet, der neutrale Effektivitätseinfluss durch die Nutzung von ChatGPT mit oder ohne PH schon in Kapitel 5.1 bestätigt.

Affordanzen des Sprachmodells. Die Analyse der Affordanzen von G-2 offenbart, dass viele Nutzungsmöglichkeiten von ChatGPT ohne das PH ungenutzt bleiben. Dies könnte darauf hinweisen, dass das Potenzial des Sprachmodells, ohne die gezielte Bereitstellung von Affordanzen, nicht wirklich ausgeschöpft wird.

Auch zeigt die häufige Verwendung des Sprachmodells für Erklären und Zusammenfassen, dass sich Studierende gerade auf Funktionen verlassen, die prädestiniert für die *Halluzination* (Amatriain, 2024; Mohr, 2024, 4f.) sind. Gerade das Zusammenfassen vom Foliensatz bietet ChatGPT die Möglichkeit, über den Foliensatz hinaus Inhalte zu generieren.

Allerdings nutzten sechs der Studierenden ChatGPT auch für die Erstellung von Selbsttest, sowie eine Person für die Planung der 30 minütigen Vorbereitungszeit. Somit sind einigen Studierenden auch mehr als nur „Zusammenfassen" und „Erklärungen" als Affordanzen bewusst. Dieser Anteil ist jedoch geringer.

Quantitative, subjektive Bewertung. Beim selbst eingeschätzten Verständnis der Testinhalte (ETV) zeigen sich keine signifikanten Unterschiede zwischen den Gruppen. Dies könnte bedeuten, dass die kurzfristige Verwendung von ChatGPT nicht effektiv das subjektive Verständnis erhöht.

Demgegenüber steht jedoch die explizite Einschätzung zur Verständnisförderung durch ChatGPT (ETMV). Hierbei gaben die Studierenden, unabhängig vom PH, mehrheitlich an, dass ChatGPT hilfreich war, den Foliensatz tiefer zu verstehen. Die Wahrnehmung deckt sich mit den Inhalten des Papers von Farhi et al. (2023, 2). Somit wird die Vorbereitung mit ChatGPT von den Studierenden allgemein als hilfreich für den tieferen Verständnisaufbau angesehen, auch wenn objektiv die tiefgehenden Fragen nach Bloom's Taxonomy nicht häufiger korrekt beantwortet wurden.

Die Strukturierung des Lehrnstoffs durch die gegebenen Prompts (ETLS) wurde tendenziell als gut wahrgenommen. Daher können die Prompts aus dem Handbook als subjektiv einigermaßen hilfreich angesehen werden. Eine detaillierte Diskussion erfolgt später im Feedback zum Prompt Handbook.

Qualitative, subjektive Bewertung. G-1 gab vor allem „Lernzettel" und die „Arbeit mit dem Foliensatz" als hilfreich für die Klausurvorbereitung an (ETHA). Diese Techniken lassen sich nach Dunlosky

et al. (2013, 18) vor allem den Lerntechniken *Summarizing* und *Rereading* zuordnen, die als weniger wirksam gelten. Allerdings betonen Rovers et al. (2018, 9), dass die Strategie von der Lernsituation abhängt. Gerade durch die Zeitbeschränkung und Kürze des Tests, besteht die Annahme, dass die genutzten Techniken für die traditionelle Gruppe (G-1) effektiv sind. Die Interpretation ist jedoch mit Vorsicht zu betrachten, da diese Annahme auf den Meinungen der Teilnehmer:innen basiert.

Von G-2 standen 85,7 % der genannten hilfreichen Aspekte (ETHA) im Zusammenhang mit KI. Bei G-3 lagen alle positiven Aspekte in Zusammenhang mit KI. Somit wurde im Allgemeinen die Nutzung der KI von vielen Studierenden als nützlich angesehen.

Im Detail waren es bei G-2 vor allem Erklärungen und Zusammenfassungen durch KI, die als hilfreich angesehen wurden, was sich mit den genutzten Affordanzen des Sprachmodells deckt. G-3 sah vor allem Quizfragen durch KI am hilfreichsten an. Somit gibt es eine Verschiebung bei G-3 zum Retrieval Based Learning (Karpicke, 2012, 157). Das einfache Erstellen der Quizfragen wurde deutlich häufiger als hilfreich angegeben als bei G-2. Vermutlich auch, weil es deutlich einfacher war, mit dem Handbook die Abfrage zu üben. Natürlich heißt das nicht direkt, dass das PE das Retrieval Based Learning gefördert hat, jedoch zeigt es, dass die Studierenden durch das PH literaturnahe effiziente Methoden als hilfreicher ansehen.

Wenn auch im geringen Maße wurden bei G-2 ebenfalls die Lernzettel als hilfreich empfunden, was erneut aufzeigt, dass auch ineffiziente Techniken für Studierende hilfreich waren (Rovers et al., 2018, 9). G-3 gab jedoch ausschließlich Quizfragen und Erklärungen durch KI als hilfreich an, die als bewährte Methoden in der Literatur angesehen werden (Karpicke, 2012; Mollick und Mollick, 2024, 157, 6).

Bei G-1 waren vor allem die fehlenden Hilfsmittel sowie der fehlende Kontext ein Problem. Gerade der fehlende Kontext spricht für die Annahme, dass ChatGPT einem Kontext liefern kann (Luckin, 2010, 262). Das Problem des fehlenden Kontexts oder der Hilfsmittel wurde indes bei G-2 und G-3 nicht erwähnt. Dies ist nur eine vage Vermutung, deckt sich jedoch mit dem signifikanten Unterschied bei Q3.

Allerdings wurden von ca. der Hälfte der Teilnehmenden hinderliche Aspekte mit KI angegeben, wobei als Hauptaspekt die *Zusammenfassungen* und *Zeitverschwendung durch KI* genannt wurden. Somit sehen einige Teilnehmer:innen die Zusammenfassung der KI, aber auch die generelle Nutzung tendenziell als nicht zielführend an. Dies verdeutlicht die Wichtigkeit für ein Regelwerk, damit ChatGPT als Sparringspartner statt für fehleranfällige Zusammenfassungen genutzt wird. Sinnvoll sind klare Hinweise oder auch Einschränkungen, das Modell zu verwenden, sodass erwähnte Hindernisse mit der KI nicht entstehen.

Das G-3 die meisten hinderlichen Aspekte angab, wobei die spezifischen Prompts oder das Prompt Handbook allgemein in 1/3 der Fälle als Grund genannt wurde, zeigt, dass das Prompt Handbook bzw. seine spezifischen Prompts noch angepasst werden müssen. Dazu im folgenden Kapitel aber mehr. Interessanter ist die Angabe von *Zusammenfassungen durch KI* als negativen Aspekt bei G-3, da diese eigentlich nicht gestattet waren. Dies könnte auch ein Indiz dafür sein, einen Chatbot, wie Bastani et al. (2024) ihn entworfen haben, zu konstruieren, der keine direkten Ergebnisse zulässt. Möglicherweise ist eine abgesicherte Nutzung besser.

Zusammenfassung. Objektiv hat die Nutzung des Prompt Handbooks und ChatGPT keine Effektivitätssteigerung in den Testergebnissen ergeben. Allerdings schöpfte die „ad hoc“ ChatGPT Gruppe (G-2) ohne das Prompt Handbook nicht das volle Potenzial an Affordanzen des Chatbots aus. In der subjektiven Wahrnehmung wird quantitativ einer Verständnisförderung durch ChatGPT klar zugestimmt, jedoch besteht für das PH Optimierungsbedarf. Qualitativ gab jede Gruppe für ihre Methode unterschiedliche hilfreiche Aspekte an, G-3 vor allem literaturnahe effektive Lerntechniken. Die hinderlichen Angaben der G-1, korrelieren mit dem Unterschied von Q3. G-2 und G-3 zeigen, dass es Hinweise und Einschränkung von ChatGPT braucht, die ggf. auch direkt im Modell verankert sind. Die Effektivität sollte gerade objektiv über einen längeren Zeitraum und subjektiv im größeren Stil untersucht werden.

5.4 Vor- und Nachteile von PE und ChatGPT aus Sicht der Studierenden

Subjektive Vorteile. Die Studierenden empfanden tendenziell durch die kurzfristige Verwendung von ChatGPT, mit oder ohne PH, keinen Autonomieverlust. Dieser Befund könnte in Verbindung mit Rashed Ibraheam Almohesh (2024, 10) einen großen Vorteil für Studierende in der Lernautonomie bedeuten. ChatGPT ermöglicht den Studierenden, sich eigenständig Wissen anzueignen und kann zudem die Motivation (Beltozar-Clemente und Díaz-Vega, 2024, 89) fördern, autonom zu einem Ergebnis zu gelangen.

Für die G-2 war die Effizienz der am häufigsten genannte Vorteil in der Nutzung. In G-3 wurde dieser Aspekt nur in 27,8 % der Fälle erwähnt. Stattdessen nannten Studierende in G-3 das aktive Lernen ebenso häufig wie die Effizienz (27,8 %), während dieser Aspekt in G-2 nur zu 18,7 % hervorgehoben wurde. Dies spricht zunächst auch für die Vorteile von Rasul et al. (2023, 4), dass ChatGPT insbesondere das personalisierte Lernen fördern kann. Sei es in Form von *Effizienz* oder *aktivem Lernen*, ChatGPT hilft Teilnehmer:innen durch den direkten Kontakt dabei, sich effizienter und aktiver mit dem Lernmaterial auseinander zu setzen. Das G-3 das aktive Lernen (Karpicke, 2012, 158) am häufigsten nannte, könnte hier auch wieder dafür sprechen, dass die Studierenden durch das PH mehr Affordanzen und somit auch Vorteile in ChatGPT sehen.

Subjektive Nachteile. G-2 nannte vor allem die Probleme der Halluzination (Amatriain, 2024; Bender et al., 2021, 4f., 616f.). Wenn auch indirekt fiel Studierenden bei der kurzen Nutzung von ChatGPT das Problem der Halluzination auf. Diese Erkenntnis geht mit der Empfehlung von Rasul et al. (2023, 10) einher, dass Studierende über die Bias und falschen Informationen aufgeklärt werden müssen. Es spricht viel dafür, dass der Hinweis im Prompt Handbook auf die Bias in den Antworten Studierenden helfen, die Antworten besser einzuordnen. G-3 erwähnte das Problem der Halluzination nämlich kaum. Natürlich ist einigen Studierenden auch ohne Handbook klar, dass ChatGPT falsche Antworten produzieren kann, dennoch ist es ein Aspekt der Fairness, für alle gleiche Möglichkeiten zu schaffen, mit dem Chatbot zu arbeiten. Weiterhin gab G-2 die Gefahr für den Verlust des kritischen Hinterfragens an. Bedenken, die auch in anderen Papern geäußert wurden Cong-Lem et al. (2024, 6).

G-3 sah die Nachteile vor allem in der Einschränkung durch das PE und einem allgemeinen Informationsüberfluss. Der Informationsüberfluss spricht für die, bei der Zeiterfassung diskutierten, erhöhten kognitiven Belastung durch das Prompt Handbook, während der Vorbereitungsphase. Da dieser Aspekt sowohl in der objektiven als auch subjektiven Wahrnehmung angenommen werden kann, erfolgt eine dedizierte Auseinandersetzung im Sinne der *Cognitive Load Theory* nach Sweller (1988); Sweller et al. (2011). Es ist vermutlich wahr, dass ein Nachteil des Prompt Handbooks der erhöhte *Extraneaous Cognitive Load* war Zumbach (2010, 82f.). Aus diesem Grund ist es umso bedeutender, die Nutzung eines Prompt Handbooks über einen längeren Zeitraum zu untersuchen, sodass Studierende ohne diesen zusätzlichen „Load“ mit dem PH arbeiten können. Einen anderen Ansatz bieten Suriano et al. (2025, 6), die davon ausgehen, dass ChatGPT an sich, den *Extraneous* und *Instrinsic Cognitive Load* verringern kann. Ggf. konnte sich damit der „Load“ für Gruppe G-3 etwas verringern, dennoch ist eine Erhöhung durch das PH wahrscheinlich. Ob die Einschränkungen durch das PH wirklich als Nachteil gelten, ist anzuzweifeln. Nach Bastani et al. (2024, 9) sind die Gruppen, die ChatGPT mit Restriktionen verwendeten, erfolgreicher gewesen. Es wird versucht, es sich so einfach wie möglich zu machen. Für ein besseres Verständnis sind diese Einschränkungen von Studierenden vielleicht nicht als Nachteil, sondern eher als Vorteil anzusehen.

Zusammenfassung. ChatGPT kann die Lernautonomie fördern und das Lernen effizienter und aktiver gestalten. Für die uneingeschränkte Nutzung sind die Probleme der Halluzination und der Verlust des kritischen Hinterfragens präsent, die mit dem PH relativiert wurden. Die PH-Gruppe spricht dagegen den Informationsüberfluss und die Einschränkung durch das PE an, die für eine längerfristige Nutzungsuntersuchung sprechen.

5.5 Feedback des Prompt Handbook

Zunächst erfolgt der Hinweis auf Forschungsfrage 1, die als Ergebnis das Prompt Handbook beinhaltet. Weiterhin wurde in der Diskussion zu den Testergebnissen, Effizienz, Effektivität sowie den Vor- und Nachteilen einiges an Feedback genannt und diskutiert.

Verwendete Prompts und Schwierigkeiten. Das Prompt „2.4 Selbsttests anfertigen“ wurde von 83,7 % der Teilnehmenden in G-3 genutzt. Dies ist als positiv zu bewerten, da es sowohl den Einsatz von effektiven und effizienten Lernmethoden wie das Retrieval Based Learning (Karpicke, 2012, 157), als auch die Ähnlichkeit zur Vorbereitung auf die tatsächliche Prüfungssituation (Rovers et al., 2018, 1) fördert. Insgesamt wurden vor allem die Prompts aus dem Handbook genutzt, die ChatGPT die Affordanz eines Tutors und Planers (Crompton und Burke, 2023, 13) zuschreiben. Dass insbesondere die Prompts „2.6 Passende Hilfe erhalten“ und „2.7 Der andere Blickwinkel (Reframing)” kaum genutzt wurden, deutet darauf hin, dass Studierende vor allem effektive Prompts nutzten, die direkt mit dem Lerninhalt verknüpft waren. Diese Erkenntnis sollte bei der Weiterentwicklung des PH berücksichtigt werden.

Gerade die sehr offenen Prompts, die gezielt Techniken des Prompt Engineerings verwendeten, indes „2.1 Lass uns Schritt für Schritt nachdenken“ sowie „2.2 Gedankenbaum“, wurden vergleichsweise seltener genutzt. Zusätzlich wurden Schwierigkeiten mit den Prompts gemeldet. Die Prompts basieren vor allem auf abstrakten unkonkreten Anweisungen, die sich mehr auf PE Techniken fokussierten. Die Ergebnisse zeigen, dass Studierende eher konkrete Prompts bevorzugen, die sich direkt auf den Kontext der Klausur beziehen. Einzelne PE-Techniken lassen sich in vorhandene Prompt-Ideen einbauen. Wichtiger sind die konkreten Affordanzen, die ChatGPT für das Lernen bietet und das Wissen darüber. Dementsprechend sollten die Prompts, die nur PE-Techniken ohne Affordanzen nutzten, entfernt werden.

Nutzungsbewertung & Benutzerfreundlichkeit. Da die Nutzungsbewertung überwiegend positiv ausfiel, unterstreicht dies die Nützlichkeit einiger Prompts. Bei der Benutzerfreundlichkeit gibt es jedoch noch Optimierungspotenzial. Hierzu wäre eine mögliche Verbesserung eine Weboberfläche, die die Prompts aus dem PH zum Auswählen direkt in einer übersichtlichen Liste bereitstellt, anschließend anpassbare Parameter bietet und direkt MC-Fragen oder Verständnisfragen generiert werden. Die verschiedenen PE-Techniken könnten dabei im Backend bereits integriert sein, wenngleich die Anwendung von PE-Techniken tendenziell durch die immer besseren Reasoning-Modelle obsolet werden könnten. Eine Weiterentwicklung in Richtung der Affordanzen wäre somit vielversprechend.

Weiterempfehlung. Das PH wurde von allen Teilnehmer:innen gerade wegen der Selbsttest, der Effizienzsteigerung, der Nutzung von Lernstrategie und guter Erklärungen empfohlen. Dies spricht klar dafür, dass das PH von Studierenden genutzt werden sollte, vor allem um sich neue Affordanzen des Chatbots für die Klausurvorbereitung anzueignen. Dabei geht es für die Studierenden weniger um konkrete PE Techniken, sondern mehr um die Affordanzen, die das Sprachmodell bietet.

Zusammenfassung. Beim PH geht es nunmehr weniger um die PE Techniken. Nach Angaben der Studierenden sind vor allem die Vermittlung von Affordanzen wichtig, die ChatGPT bereitstellt. Dies kann durch einige Anpassungen in der Benutzerfreundlichkeit effizient und effektiv durch das Prompt Handbook erfolgen. Das technische Prompt Engineering spielt dabei tendentiell eine untergeordnete Rolle.

5.6 Implikation für Studierende

Studierende profitieren kurzfristig nicht signifikant von ChatGPT oder dem PE in der Klausurvorbereitung. Allerdings wird es von ihnen als motivierend und effizient, insbesondere bei der Kontexterschließung

empfunden. Langfristig könnte eine automatisierte Nutzung mit klaren Anleitungen und optimierten Affordanzen das volle Potenzial von Sprachmodellen von Studierenden besser ausgeschöpft werden.

5.7 Limitationen

Obwohl die vorliegende Arbeit wertvolle Erkenntnisse zur Nutzung von ChatGPT und dem damit verbundenen Prompt Engineering für die Klausurvorbereitung liefert, gibt es einige Einschränkungen.

Methodisch ist die Anzahl an Studierenden für die Stichprobe mit 51 relativ klein. Dies ermöglicht zwar eine explorative Analyse, allerdings ist die Induktion auf die Gesamtheit eingeschränkt. Die Studierenden meldeten sich freiwillig für das Experiment, sodass tendenziell motivierte Studierende teilnahmen. Dies könnte den Transfer der Ergebnisse auf die gesamte Studierendenschaft beeinflussen. Zudem mahmen Studierende aller möglichen Fachbereiche teil, bei der keine exakte Protokollierung der Vorkentnisse zu den Themen des Foliensatzes erfolgte.

Das Experiment wurde in einem kontrollierten Umfeld durchgeführt, was nicht die realen Lernbedingungen widerspiegelt. Während der Vorbereitungsphase kontrollierte der Versuchsleiter kontinuierlich die Einhaltung der Regeln, was einen Einfluss auf das Lernverhalten haben könnte. Ferner war die Vorbereitungszeit von 30 Minuten, sowie der Umfang an Folien und MC-Fragen im Test deutlich niedriger als bei einer realen Klausur.

Hinsichtlich der Messinstrumente gab es ebenfalls Limitationen. Die subjektive Wahrnehmung der Effektivität und Effizienz könnte durch die Likert-Skala verzerrt sein, indem Studierende mit einer Tendenz zur Mitte oder dem sozial Erwünschten antworteten. Dafür ist die Likert-Skala anfällig. Die Schwierigkeit der MC-Fragen wurde durch die kognitiven Ebenen nach Bloom's Taxonomy indirekt erfasst. Gerade für die höheren kognitiven Ebenen der Taxonomy würden sich Textfragen besser eignen, da die Überprüfung des Erschaffens durch die textuellen Gedanken der Studierenden besser ermittelt werden können.

Bei der qualitativen Inhaltsanalyse nach Mayring kam es zu keiner Reliabilitätsprüfung der Kodierung. Dadurch besteht das Riskio der subjektiven Interpretation, das die Ergebnisse beeinflussen kann.

5.8 Ausblick

Die vorherige Untersuchung liefert weitere Erkenntnisse zur Rolle vom Prompt Engineering in der Nutzung von ChatGPT für die Klausurvorbereitung. Sie zeigt, dass Studierende weniger Verständnis für das PE, sondern für die Affordanzen, die der Chatbot bietet, benötigen. Somit sollten zukünftige Forschungsarbeiten verstärkt untersuchen, wie Studierende dabei unterstützt werden, Sprachmodelle und ihre Affordanzen in ihren Lernprozess einzubauen.

Für die Generalisierbarkeit sind Studien mit einer größeren Stichprobenanzahl sinnvoll. Auch eine Untersuchung in realistischeren Lernumgebungen würde klarere Aussagen versprechen. Beispielweise könnte die Vorbereitungszeit verlängert, oder ChatGPT in bestehende Lehrveranstaltungen integriert werden.

Ferner wäre eine Betrachtung der Klausurvorbereitung über einen längeren Zeitraum, zum Beispiel eines ganzen Semesters, interessant. Hierbei könnte ein Vergleich zwischen ChatGPT mit bereitgestellten Affordanzen sowie die offene Verwendung von ChatGPT erfolgen, allerdings ohne Bewertung der Prüfung.

Auch könnte eine detaillierte Analyse der Chatbotnutzung von Studierenden wertvolle Einblicke liefern, die das Prompt Handbook um neue Affordanzen oder Anleitungen erweitert. Weiterhin wäre eine Studie mit einem festen Zeitvolumen für alle Teilnehmer:innen sinnvoll, bei dem das Lesen des Prompt Handbooks mit in die Vorbereitungsphase einfließt, um die Unterschiede in den Gruppen kleiner zu halten.

Durch längerfristige und größere Untersuchungen könnten für Studierende weitere Empfehlungen für den Einsatz von Sprachmodellen in der Hochschule abgeleitet werden, um deren Potenzial für eine effektive und effiziente Klausurvobereitungs auszuschöpfen.

6 Fazit

Ziel dieser Arbeit war es, konkrete Prompts für Studierende bereitzustellen und zu untersuchen, inwiefern diese das Lernen effektiver und effizienter im Vergleich zu traditionellen Lernmethoden und der uneingeschränkten Nutzung von ChatGPT gestalten können. Dazu wurden zunächst die Grundlagen der Large Language Models und Prompt Engineering in Verbindung mit der Hochschulbildung erläutert. Im Anschluss daran erfolgte die Erstellung des Prompt Handbooks, mit konkret spezifischen Prompts und die Erstellung eines MC-Tests basierend auf zehn Folien. Das Prompt Handbook, sowie die Testergebnisse wurde durch acht Messinstrumente untersucht und sowohl unter objektiven als auch subjektiven Gesichtspunkten ausgewertet und evaluiert.

Innerhalb dieser Arbeit ist deutlich geworden, dass es keinen Unterschied macht, ChatGPT/ Prompt Engineering in der kurzfristigen Klausurvorbereitung zu nutzen, oder ohne KI zu arbeiten. Jedoch zeigte sich in bestimmten Bereichen das Potenzial von ChatGPT Kontext geliefert zu bekommen. Auch in der Selbsteinschätzung bemerkten Studierende, durch die Nutzung von ChatGPT keine Verzerrung in ihrer Wahrnehmung. Die höhere Vorbereitungszeit bei Teilnehmer:innen, die das Prompt Handbook nutzten, lässt objektiv eine tendenzielle zeitlich Ineffizienz schlussfolgern. Allerdings ist die subjektive Wahrnehmung, durch ChatGPT eine Beschleunigung im Lernen zu erfahren, egal ob mit Prompt Handbook oder ohne. Effektive Affordanzen des Chatbots fehlten vor allem bei der uneingeschränkten Nutzung, sodass zumindest für die gleichen Voraussetzungen zur Klausurvorbereitung, Hinweise und Einschränkungen zur Nutzung von ChatGPT, sinnvoll erscheinen. Nicht zuletzt ist nach Angaben der Studierenden ChatGPT vor allem ein Hilfsmittel, welches das Lernen effizienter und aktiver gestalten kann, gleichzeitig aber auch Halluzination und einen Verlust des kritischen Hinterfragens fördern kann. Das Prompt Handbook sticht vor allem mit seinen spezifischen Prompts, die ChatGPT eine bestimmte Affordanz zuweisen, als sinnvoll heraus. Das technische PE spielt dabei eine untergeordnete Rolle.

Im Forschungsfeld der Hochschulbildung sind nach jetzigem Stand, lediglich wenige Studien vorhanden, die den Gebrauch von ChatGPT und die Bereitstellung von spezifischen Affordanzen für den Chatbot in der Klausurvorbereitung untersuchen. Diese Arbeit zeigt vor allem, dass ChatGPT ein sinnvolles Hilfsmittel in der Klausurvorbereitung sein kann und deckt sich damit in Teilen mit der Forschungsliteratur. Durch die Kürze und Lernbedingungen der Vorbereitungszeit, und anderen Erkenntnisse aus der Forschungsliteratur, wird dennoch die restriktive Verwendung von ChatGPT empfohlen. In anderen Kontexten zeigt die Literatur, dass KI auch die Leistung verringern kann.

Es ist gerade für Studierende und die Hochschulbildung von Bedeutung längerfristige Studien, die den Einfluss von ChatGPT auf die Lernergebnisse und Metakognition untersuchen durchzuführen, um den Studierenden Handlungsempfehlung zu geben, KI zu verwenden. Gleichzeitig könnte auch ein genormtes Regelwerk mit Affordanzen zur Verwendung von ChatGPT sinnvoll sein, wobei als Grundlage das Prompt Handbook dienlich sein kann.

Literatur

AlAfnan, M. A., Dishari, S., Jovic, M., und Lomidze, K. (2023): ChatGPT as an Educational Tool: Opportunities, Challenges, and Recommendations for Communication, Business Writing, and Composition Courses. *Journal of Artificial Intelligence and Technology*, **3** (2), S. 60–68.

Amatriain, X. (2024): Prompt Design and Engineering: Introduction and Advanced Methods.

Anderson, L. W., und Krathwohl, D. R. (2001): *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives: complete edition*. Addison Wesley Longman, Inc.

Anthropic (2024): Claude 3.5 Sonnet. https://www.anthropic.com/news/claude-3-5-sonnet, abgerufen am 10.10.2024.

Armstrong, P. (2010): Bloom's taxonomy. *Vanderbilt University Center for Teaching*, S. 1–3.

Ayres, P., und Sweller, J. (2005): *The Split-Attention Principle in Multimedia Learning*, S. 135–146. Cambridge Handbooks in Psychology. Cambridge University Press.

Baker, T., Smith, L., und Anissa, N. (2019): Educ-AI-tion Rebooted? Exploring the Future of Artificial Intelligence in Schools and Colleges. Abgerufen am November 26, 2024.

Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, O., und Mariman, R. (2024): Generative ai can harm learning. *SSRN*, **4895486**. `https://ssrn.com/abstract=4895486`.

Beltozar-Clemente, S., und Díaz-Vega, E. (2024): Physics XP: Integration of ChatGPT and Gamification to Improve Academic Performance and Motivation in Physics 1 Course. *International Journal of Engineering Pedagogy (iJEP)*, **14** (6), S. 82–92.

Bender, E. M., Gebru, T., McMillan-Major, A., und Shmitchell, S. (2021): On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? FAccT '21, S. 610–623, New York, NY, USA. Association for Computing Machinery.

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., Krathwohl, D. R. et al. (1956): *Taxonomy of educational objectives: The classification of educational goals. Handbook 1: Cognitive domain*. Longman New York.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., und Amodei, D. (2020): Language Models are Few-Shot Learners.

Campesato, O. (2023): *Transformer, BERT, and GPT: Including ChatGPT and Prompt Engineering*. Mercury Learning and Information.

Chan, C. K. Y., und Colloton, T. (2024): *Generative AI in Higher Education*. Routledge, London.

Chan, C. K. Y., und Hu, W. (2023): Students' voices on generative AI: Perceptions, benefits, and challenges in higher education. *International Journal of Educational Technology in Higher Education*, **20** (1), S. 43.

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P. S., Yang, Q., und Xie, X. (2024): A Survey on Evaluation of Large Language Models. *ACM Transactions on Intelligent Systems and Technology*, **15** (3).

Chen, B., Zhang, Z., Langrené, N., und Zhu, S. (2023): Unleashing the potential of prompt engineering in Large Language Models: a comprehensive review. *arXiv preprint arXiv:2310.14735*.

Choi, J. H., Garrod, O., Atherton, P., Joyce-Gibbons, A., Mason-Sesay, M., und Björkegren, D. (2024): Are LLMs Useful in the Poorest Schools? TheTeacher.AI in Sierra Leone.

Coe, R., Rauch, C. J., Kime, S., und Singleton, D. (2020): Great Teaching Toolkit: Evidence Review.

Cong-Lem, N., Soyoof, A., und Tsering, D. (2024): A Systematic Review of the Limitations and Associated Opportunities of ChatGPT. *International Journal of Human–Computer Interaction*, **0** (0), S. 1–16.

Crompton, H., und Burke, D. (2023): Artificial intelligence in higher education: the state of the field. *International Journal of Educational Technology in Higher Education*, **20**.

DAIR.AI (2024a): Papers. https://www.promptingguide.ai/de/papers, abgerufen am 21.11.2024.

DAIR.AI (2024b): Zero-Shot Prompting. https://www.promptingguide.ai/de/techniques/zeroshot, abgerufen am 16.11.2024.

DeepSeek (2025): Deepseek – Into the unkown. https://www.deepseek.com/, abgerufen am 26.02.2025.

Dell'Acqua, F., McFowland III, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayer, L., Candelon, F., und Lakhani, K. R. (2023): Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. *Harvard Business School Technology & Operations Mgt. Unit Working Paper*, 24-013.

Dudenredaktion (2024): „generativ" auf Duden online.

Dunlosky, J., und Nelson, T. (1994): Does the Sensitivity of Judgments of Learning (JOLs) to the Effects of Various Study Activities Depend on When the JOLs Occur? *Journal of Memory and Language*, **33** (4), S. 545–565.

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., und Willingham, D. T. (2013): Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public interest*, **14** (1), S. 4–58.

Farhi, F., Jeljeli, R., Aburezeq, I., Dweikat, F. F., Al-shami, S. A., und Slamene, R. (2023): Analyzing the students' views, concerns, and perceived ethics about chat GPT usage. *Computers and Education: Artificial Intelligence*, **5**, S. 100180.

Fiorella, L., und Mayer, R. E. (2016): Eight ways to promote generative learning. *Educational psychology review*, **28**, S. 717–741.

Forehand, M. et al. (2005): Bloom's taxonomy: Original and revised. *Emerging perspectives on learning, teaching, and technology*, **8**, S. 41–44.

Giannos, P., und Delardas, O. (2023): Performance of ChatGPT on UK Standardized Admission Tests: Insights From the BMAT, TMUA, LNAT, and TSA Examinations. *JMIR Med Educ*, **9**, S. e47737.

Giray, L. (2023): Prompt engineering with ChatGPT: a guide for academic writers. *Annals of biomedical engineering*, **51** (12), S. 2629–2633.

Google (2023): Introducing Gemini: our largest and most capable AI model. https://blog.google/technology/ai/google-gemini-ai/, abgerufen am 10.10.2024.

Hart, J. (1965): Memory and the feeling-of-knowing experience. *Journal of educational psychology*, **56**, S. 208–16.

Henkel, O., Horne-Robinson, H., Kozhakhmetova, N., und Lee, A. (2024): Effective and Scalable Math Support: Evidence on the Impact of an AI- Tutor on Math Achievement in Ghana.

Hulbert, D. (2023): Using Tree-of-Thought Prompting to boost ChatGPT's reasoning. `https://github.com/dave1010/tree-of-thought-prompting`. Abgerufen am 24.11.2024.

Jacobsen, L. J., und Weber, K. E. (2023a): Manual zur Erstellung qualitativ hochwertiger Prompts. Vortrag auf der AEPF 2023.

Jacobsen, L. J., und Weber, K. E. (2023b): The Promises and Pitfalls of LLMs as Feedback Providers: A Study of Prompt Engineering and the Quality of AI-Driven Feedback.

Karpicke, J. D. (2012): Retrieval-based learning: Active retrieval promotes meaningful learning. *Current Directions in Psychological Science*, **21** (3), S. 157–163.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., und Iwasawa, Y. (2023): Large Language Models are Zero-Shot Reasoners.

Krathwohl, D. R. (2002): A Revision of Bloom's Taxonomy: An Overview. *Theory Into Practice*, **41** (4), S. 212–218.

Krebs, R. (2004): Anleitung zur Herstellung von MC-Fragen und MC-Prüfungen für die ärztliche Ausbildung. *Bern: Institut für Medizinische Lehre IML, Abteilung für Ausbildungs-und Examensforschung AAE*.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., und Neubig, G. (2021): Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing.

Luckin, R. (2010): *Re-designing Learning Contexts: Technology-rich, learner-centred ecologies*. Routledge.

Luckin, R., Holmes, W., Griffiths, M., und Forcier, L. B. (2016): Intelligence Unleashed: An argument for AI in Education.

Memmert, L., Cvetkovic, I., und Bittner, E. (2024): The More Is Not the Merrier: Effects of Prompt Engineering on the Quality of Ideas Generated By GPT-3. In: *Proceedings of the 57th Hawaii International Conference on System Sciences*, S. 7520–7529.

Mohr, G. (2024): Prompt Workbook. Online-Ressource. Zugriff am 31.12.2024.

Mollick, E., und Mollick, L. (2023): Assigning AI: Seven Approaches for Students, with Prompts.

Mollick, E., und Mollick, L. (2024): Instructors as Innovators: A future-focused approach to new AI learning opportunities, with prompts.

Nelson, T., und Leonesio, R. (1988): Allocation of Self-Paced Study Time and the "Labor-in-Vain Effect". *Journal of experimental psychology. Learning, memory, and cognition*, **14**, S. 676–86.

Nelson, T. O., und Narens, L. (1990): Metamemory: A theoretical framework and new findings. In: *Psychology of learning and motivation*, Band 26, S. 125–173. Elsevier.

Ngo, T. T. A. (2023): The Perception by University Students of the Use of ChatGPT in Education. *International Journal of Emerging Technologies in Learning (iJET)*, **18** (17), S. 4–19.

OpenAI (2022): Introducing ChatGPT. https://openai.com/index/chatgpt/, abgerufen am 26.09.2024.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A. et al. (2022): Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, **35**, S. 27730–27744.

Phoenix, J., und Taylor, M. (2024): *Prompt engineering for generative AI: future-proof inputs for reliable AI outputs at scale*. O'Reilly Media, Inc.

Pohlmann, N. (2024): IT-Sicherheit und Künstliche Intelligenz. `https://norbert-pohlmann.com/vortraege/it-sicherheit-und-kuenstliche-intelligenz/`. Zugriff am 11.11.2024.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I. et al. (2019): Language models are unsupervised multitask learners. *OpenAI blog*, **1** (8), S. 9.

Rashed Ibraheam Almohesh, A. (2024): AI Application (ChatGPT) and Saudi Arabian Primary School Students' Autonomy in Online Classes: Exploring Students and Teachers' Perceptions. *The International Review of Research in Open and Distributed Learning*, **25** (3), S. 1–18.

Rasul, T., Nair, S., Kalendra, D., Robin, M., Santini, F., Ladeira, W., Sun, M., Day, I., Rather, A., und Heathcote, L. (2023): The Role of ChatGPT in Higher Education: Benefits, Challenges, and Future Research Directions. **6**.

Ray, P. P. (2023): ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, **3**, S. 121–154.

Rovers, S. F., Stalmeijer, R. E., van Merriënboer, J. J., Savelberg, H. H., und De Bruin, A. B. (2018): How and why do students use learning strategies? A mixed methods study on learning strategies and desirable difficulties with effective strategy users. *Frontiers in psychology*, **9**, S. 2501.

Rudolph, J., Tan, S., und Tan, S. (2023): ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *Journal of applied learning and teaching*, **6** (1), S. 342–363.

Sahoo, P., Singh, A. K., Saha, S., Jain, V., Mondal, S., und Chadha, A. (2024): A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications.

Seemann, M. (2023): Künstliche Intelligenz, Large Language Models, ChatGPT und die Arbeitswelt der Zukunft. Arbeitspapier, Working Paper Forschungsförderung.

Soderstrom, N. C., und Bjork, R. A. (2015): Learning versus performance: An integrative review. *Perspectives on Psychological Science*, **10** (2), S. 176–199.

Springer Verlag (2024): International Conference on Artificial Intelligence in Education. `https://link.springer.com/conference/aied`. Abgerufen am 26.11.2024.

Suriano, R., Plebe, A., Acciai, A., und Fabio, R. A. (2025): Student interaction with ChatGPT can promote complex critical thinking skills. *Learning and Instruction*, **95**, S. 102011.

Sweller, J. (1988): Cognitive load during problem solving: Effects on learning. *Cognitive Science*, **12** (2), S. 257–285.

Sweller, J., Ayres, P., und Kalyuga, S. (2011): *Cognitive Load Theory*. Explorations in the Learning Sciences, Instructional Systems and Performance Technologies. Springer, 1 Auflage. Includes 20 b/w illustrations.

Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., und Ting, D. S. W. (2023): Large language models in medicine. *Nature medicine*, **29** (8), S. 1930–1940.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., und Polosukhin, I. (2017): Attention Is All You Need.

Von Garrel, J., und Mayer, J. (2023): Artificial Intelligence in studies—use of ChatGPT and AI-based tools among students in Germany. *Humanities and social sciences communications*, **10** (1), S. 1–9.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., und Zhou, D. (2023): Chain-of-Thought Prompting Elicits Reasoning in Large Language Models.

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., Isaac, W., Legassick, S., Irving, G., und Gabriel, I. (2021): Ethical and social risks of harm from Language Models.

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., und Schmidt, D. C. (2023): A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT.

Wittrock, M. C. (1974): Learning as a generative process. *Educational psychologist*, **11** (2), S. 87–95.

Wittrock, M. C. (1989): Generative processes of comprehension. *Educational psychologist*, **24** (4), S. 345–376.

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., und Narasimhan, K. (2023): Tree of Thoughts: Deliberate Problem Solving with Large Language Models.

Zawacki-Richter, O., Marín, V., Bond, M., und Gouverneur, F. (2019): Systematic review of research on artificial intelligence applications in higher education -where are the educators? *International Journal of Educational Technology in Higher Education*, **16**, S. 1–27.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., Liu, P., Nie, J.-Y., und Wen, J.-R. (2024): A Survey of Large Language Models.

Zumbach, J. (2010): *Lernen mit neuen Medien Instruktionspsychologische Grundlagen*. Kohlhammer.

A Anhang A

Auf den folgenden Seiten sind zunächst die konkreten MC-Fragen mit Antwortmöglichkeiten, eingeordnet in Boom's Taxonomy, aufgeführt. Im Anschluss finden sich die letzten zwei selbst erstellten Seiten (S. 9-10) des bereitgestellten Foliensatzes, der den Studierenden in der Vorbereitungsphase zur Verfügung stand. Die vorherigen Seiten (S. 1-8) bestehen zum Teil aus einem Foliensatz von Norbert Pohlmann und können unter diesem Link abgerufen werden.

MC-Fragen mit Antwortmöglichkeiten

Q1

? 1: Welche der folgenden Begriffe beschreibt eine Methode des maschinellen Lernens, die bessere Ergebnisse erzielt als traditionelle Methoden?

- Data Science
- Generative KI
- Artificial General Intelligence
- Deep Learning
- Large Language Mode
- FOK1: Ich kenne die Antwort nicht, aber wahrscheinlich fällt sie mir zu einem späteren Zeitpunkt ein.
- FOK2: Ich kenne die Antwort nicht, und sie wird mir in Zukunft auch nicht einfallen.

- Richtig: Deep Learning.
- Bloom
  - 1. Remember: Wiedererkennen.
  - A. Factual Knowledge: Aa. Knowledge of terminology

Q2

? 2: Was wäre die erste Folge beim Eintreffen einer Singularität im Kontext von KI?

- Der Mensch arbeitet im Dienste der KI und führt ihre Befehle aus
- Der Mensch wird von der KI schrittweise zur Seite gedrängt
- Der Mensch verliert die Kontrolle über die KI und ihre Systeme
- Der Mensch verschmilzt mit der KI zu einer neuen höheren Einheit
- Der Mensch wird nach Superalignment von der KI unterstützt
- FOK1: Ich kenne die Antwort nicht, aber wahrscheinlich fällt sie mir zu einem späteren Zeitpunkt ein.
- FOK2: Ich kenne die Antwort nicht, und sie wird mir in Zukunft auch nicht einfallen.

- Richtig: Der Mensch verliert die Kontrolle über KI.
- Bloom
  - 1. Remember

- Bc. Knowledge of theories, models, and structures

Q3

3: Ein großes Sprachmodell namens "BetterGPT" wurde ausschließlich mit Textinhalten von Instagram-Diskussionen aus den Jahren 2010 bis 2020 trainiert. Welches Problem könnte gemäß dem "Garbage in - Garbage out" Prinzip für das "BetterGPT"-Modell entstehen?

- BetterGPT liefert oft verzerrte Antworten.
- BetterGPT ist immer auf dem neuesten Stand.
- BetterGPT antwortet nur mittels Bildern.
- BetterGPT kann keine humorvollen Inhalte generieren.
- BetterGPT versteht oft den Kontext von Diskussionen nicht.
- FOK1: Ich kenne die Antwort nicht, aber wahrscheinlich fällt sie mir zu einem späteren Zeitpunkt ein.
- FOK2: Ich kenne die Antwort nicht, und sie wird mir in Zukunft auch nicht einfallen.

- Richtig: BetterGPT liefert oft verzerrte Antworten.
- Bloom
  - 2. Understand, 4. Apply
  - Bb. Knowledge of principles and generalizations

Q4

4: Sinan arbeitet als Arzt in einer Hausarztpraxis und nutzt ein KI-System zur Unterstützung bei der Diagnose. Das System analysiert Patientendaten und schlägt mögliche Diagnosen und Behandlungen vor. Eine 45-jährige Patientin kommt mit Kopfschmerzen und Schwindel in die Praxis. Die KI schlägt drei mögliche Diagnosen vor. Sinan ist sich unsicher, wer bei einer falschen Diagnose das Restrisiko trägt.

- Das Restrisiko wird durch die gemeinsame Verantwortung von Arzt und KI-System getragen, wobei die Haftung zwischen beiden Parteien aufgeteilt und durch spezielle Versicherungen abgedeckt wird.
- Das Restrisiko wird vollständig durch die KI-Herstellerhaftung abgedeckt, während der Arzt nur eine beratende Funktion hat und keine direkte Verantwortung für Fehldiagnosen übernimmt.
- Das Restrisiko wird durch ein mehrstufiges Validierungssystem minimiert, bei dem sowohl die KI als auch der Arzt unabhängig voneinander Diagnosen erstellen und diese dann abgeglichen werden.
- Das Restrisiko wird durch eine automatisierte Zweitmeinung eines anderen KI-Systems überprüft, während der Arzt die finale Diagnose nur noch formal bestätigt

und dokumentiert.
- Das Restrisiko wird durch die bewusste Entscheidungsfindung des Arztes und dessen Verantwortungsübernahme gemanagt, wobei die KI-Vorschläge als unterstützende Handlungsempfehlungen dienen.
- FOK1: Ich kenne die Antwort nicht, aber wahrscheinlich fällt sie mir zu einem späteren Zeitpunkt ein.
- FOK2: Ich kenne die Antwort nicht, und sie wird mir in Zukunft auch nicht einfallen.

- Richtig: Das Restrisiko wird durch die bewusste Entscheidungsfindung des Arztes und dessen Verantwortungsübernahme gemanagt, wobei die KI-Vorschläge als unterstützende Handlungsempfehlungen dienen.
- Bloom
    - 3. Apply, 5. Evaluate
        - Bb. Knowledge of principles and generalizations

Q5

5: Thomas entwickelt KI-Systeme für ein Krankenhaus. Beim letzten Ärztetreffen bemerkt er, dass viele Ärzte die von ihm entwickelten KI-Diagnosevorschläge zwar anschauen, am Ende jedoch ignorieren. Als er nachfragt, äußern die Ärzte Bedenken bezüglich der undurchsichtigen Entscheidungsprozesse der KI. Sie verstehen nicht, wie die KI zu ihren Vorschlägen kommt und fühlen sich unwohl dabei, Entscheidungen auf dieser Basis zu treffen. Was ist das Problem mit Thomas entwickelten KI-System?

- Ärzte sehen keinen Mehrwert in dem Diagnosevorschlag.
- Ärzte vertrauen nicht auf Thomas KI-System.
- Ärzte finden die Vorschläge zu kompliziert.
- Ärzte finden die Vorschläge zu einfach.
- Ärzte bangen um ihren Arbeitsplatz.
- FOK1: Ich kenne die Antwort nicht, aber wahrscheinlich fällt sie mir zu einem späteren Zeitpunkt ein.
- FOK2: Ich kenne die Antwort nicht, und sie wird mir in Zukunft auch nicht einfallen.

- Richtig: Ärzte vertrauen nicht auf Thomas KI-System.
- Bloom
    - 4. Analyze
    - Bb. Knowledge of principles and generalizations

Q6

6: Ein großer Wasserversorger im Osnabrücker Land wird angegriffen. Dies gefährdet die gesamte Wasserversorgung in der Region. Zur schnellen Untersuchung möchte das Unternehmen personenbezogene Daten der Mitarbeiter auswerten. Darf das Unternehmen die personenbezogenen Daten der Mitarbeiter ohne deren ausdrückliche Zustimmung gemäß Utilitarismus auswerten?

- Es besteht ein ethisches Dilemma, diese Frage ist nicht beantwortbar.
- Ja, weil der Täter nur mit Hilfe von Kundendaten identifiziert werden kann und der Datenschutz keine Bedeutung hat.
- Nein, da die Privatsphäre und das Recht auf Datenschutz gewahrt bleiben müssen.
- Ja, weil der Schutz der öffentlichen Sicherheit und die Abwehr des Angriffs eine höhere Priorität haben als die Privatsphäre der Mitarbeiter.
- Nein, die Mitarbeiter sollten entlassen werden und neue Sicherheitsmaßnahmen implementiert werden.
- FOK1: Ich kenne die Antwort nicht, aber wahrscheinlich fällt sie mir zu einem späteren Zeitpunkt ein.
- FOK2: Ich kenne die Antwort nicht, und sie wird mir in Zukunft auch nicht einfallen.

- Richtig: Ja, weil der Schutz der öffentlichen Sicherheit und die Abwehr des Angriffs eine höhere Priorität haben als die Privatsphäre der Mitarbeiter.
- Bloom
  - 2. Understand, 4. Analyze, 5. Evaluate
  - Bb. Knowledge of principles and generalizations

Q7

7: Ein Bankunternehmen wird Opfer eines Cyberangriffs, der die Kundendaten verschlüsselt. Das KI-Sicherheitssystem identifiziert die Quelle und schlägt einen Strike Back vor. Was ist die Herausforderung in diesem Szenario?

- KI kann den Strike Back nicht berechnen.
- KI kann den Strike Back nicht automatisch durchführen.
- KI hat zu viel Entscheidungsgewalt.
- KI startet einen Gegenangriff.
- KI hat unvollständige Informationen.
- FOK1: Ich kenne die Antwort nicht, aber wahrscheinlich fällt sie mir zu einem späteren Zeitpunkt ein.
- FOK2: Ich kenne die Antwort nicht, und sie wird mir in Zukunft auch nicht einfallen.

- Richtig: KI hat unvollständige Informationen

- Bloom
  - 5. Evaluate, 1. Remember
  - Bb. Knowledge of principles and generalizations

Q8

8: Welche der folgenden Aussagen beschreibt NICHT eine notwendige Maßnahme, um das Machtverhältnis zwischen Angreifern und Verteidigern im Bereich der Cybersicherheit zu verändern?

- Die Verteidiger müssen KI deutlich intensiver nutzten.
- Der Austausch von sicherheitsrelevanten Daten.
- Bewusstsein für Datenschutzprobleme
- Überprüfen und Aktualisieren von Sicherheitsrichtlinien
- kontinuierliche Schulung von Mitarbeitern
- FOK1: Ich kenne die Antwort nicht, aber wahrscheinlich fällt sie mir zu einem späteren Zeitpunkt ein.
- FOK2: Ich kenne die Antwort nicht, und sie wird mir in Zukunft auch nicht einfallen.

- Richtig: Bewusstsein für Datenschutzprobleme.
- Bloom
  - 6. Create
  - Ab. Knowledge of specific details and elements

Q9

9: Eine KI-Sicherheitsprüfung ergab den Binärcode 110010. Welcher Dezimalwert entspricht diesem Code?

- 3
- 19
- 50
- 76
- 98
- FOK1: Ich kenne die Antwort nicht, aber wahrscheinlich fällt sie mir zu einem späteren Zeitpunkt ein.
- FOK2: Ich kenne die Antwort nicht, und sie wird mir in Zukunft auch nicht einfallen.

- Richtig: 50
- Bloom

- 3. Apply
- Cb. Knowledge of subject-specific techniques and methods

Q10

10: Ein KI-Sensor gibt den 6-bit Statuscode 001101 aus. Welchen Dezimalwert entspricht diesem Code?

- *FREIE ZAHLENEINGABE*
- FOK1: Ich kenne die Antwort nicht, aber wahrscheinlich fällt sie mir zu einem späteren Zeitpunkt ein.
- FOK2: Ich kenne die Antwort nicht, und sie wird mir in Zukunft auch nicht einfallen.

- Richtig: 13
- Bloom
  - 6. Create/ 3. Apply
  - Cb. Knowledge of subject-specific techniques and methods

Codierung von Informationen auf dem Computer

- **Grundlagen**:
  - Computer speichern Daten binär: Nur die Zustände 0 und 1
  - Jedes Bit (0 oder 1) ist eine Informationseinheit
- **Codierungstypen**:
  - **Text**: ASCII (z.B. Buchstabe ‘A‘ = 01000001 als Binärzahl)
  - **Bilder**: Farbwerte werden als Binärdaten codiert (z.B. RGB → 1111111 für Rot = 255)
  - **Zahlen**: Dezimalwerte werden in Binärzahlen umgewandelt
- **Vorteil der binären Codierung**:
  - Robust gegenüber Störungen
  - Einfach in elektronischen Schaltungen umsetzbar

Dezimalzahl	Binärzahl
5	101
10	1010

Umwandlung von Binärzahlen in Dezimalzahlen

- **Schritte zur Umwandlung:**
  - Notiere die Binärzahl
  - Ordne jeder Stelle die entsprechende Potenz von 2 zu (rechts beginnend mit 2^0)
  - Multipliziere jede Ziffer mit ihrer Potenz
  - An Stellen, an denen eine 1 steht, gilt der Dezimalwert, die anderen werden ignoriert
  - Addiere die Ergebnisse
- **Beispiel**: Umrechnung von: **10011**:

1. $1 \quad 0 \quad 0 \quad 1 \quad 1$

2. $= 1 \cdot 2^4 + 0 \cdot 2^3 + 0 \cdot 2^2 + 1 \cdot 2^1 + 1 \cdot 2^0$

3. $= 16 + 0 + 0 + 2 + 1$

4. $= 19$

Binärzahl	1	0	0	1	1
Potenzen	2^4	2^3	2^2	2^1	2^0

B Anhang B

Im folgenden werden die weitere demographischen Parameter der Teilnehmenden dargestellt.

B.1 Studienfächer pro Gruppe

Gruppe G-1

#	Studienfach
1	Logistik und Informationsmanagement
2	Elektrotechnik
3	Mathe und Sport (gym. Lehramt)
4	Landwirtschaft
5	Geographie/ Politikwissenschaften
6	Lehramt mit den Fächern Biologie, Englisch und Sport
7	Technischer Informatik
8	Lehramt an Gymnasien (Biologie/ Katholische Theologie)
9	Musik & Philosophie
10	International Business Management
11	Biologie
12	Jura
13	Medientechnik B.Sc.
14	Data Science
15	Master of Science Landschaftsökologie
16	Bachelor Bildung und Erziehung mit den Fächern Mathe und Sport
17	Grundschullehramt Germanistik und Textiles Gestalten

Tabelle 14: Studienfächer der Gruppe G-1

Gruppe G-2

#	Studienfach
1	Biologie/Germanistik
2	Grundschullehramt Musik und Germanistik
3	Maschinenbau
4	2 Fächer Bachelor (Kunst/Kunstpädagogik und Geographie)
5	Wirtschaftsinformatik
6	Geographie
7	Mechatronik
8	Master Mechatronik
9	Geographie M.Sc. Gesellschaft-Umwelt-Zukunft
10	Biologie
11	Landwirtschaft
12	M.Sc. Interdisziplinäre Public und Nonprofit Studien
13	Media & Interaction Design
14	Info
15	Wirtschaftsingenieurwesen E-Technik
16	Mechatronik B.SC.
17	Grundschullehramt Deutsch/Textiles Gestalten

Tabelle 15: Studienfächer der Gruppe G-2

Gruppe G-3

#	Studienfach
1	Betriebswirtschaftslehre (B. A.), dual
2	Mechatronik
3	Media Interaction Design
4	Wirtschaftsinformatik
5	Mechatronik
6	Gymnasiales Lehramt (Englisch und Sport)
7	International Management
8	Berufsschullehramt Bautechnik/ Wirtschaftslehre, Politik
9	Informatik
10	Sonderpädagogische Förderung (Mathe und Deutsch)
11	Klinische Psychologie und Psychotherapie (Master)
12	Kunst und Biologie auf Gymnasial Lehramt
13	Informatik
14	BWL
15	Architektur
16	Master of Science Landschaftsökologie
17	Wirtschaftsinformatik

Tabelle 16: Studienfächer der Gruppe G-3

B.2 Bisherige Teilnahme an bewerteten MC-Tests

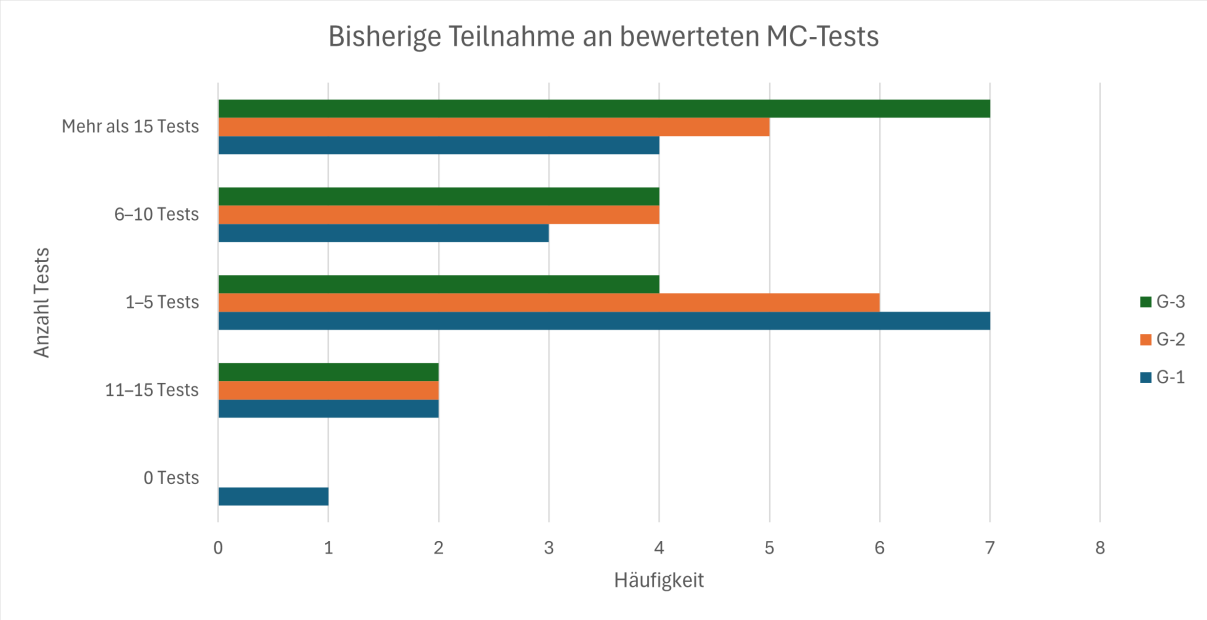

Abbildung 15: Häufigkeitsverteilung zur subjektiven Einschätzung über bisherige Teilnahme an MC-Tests

B.3 Wöchentliche Zeitinvestition für Klausurvorbereitung im Semester

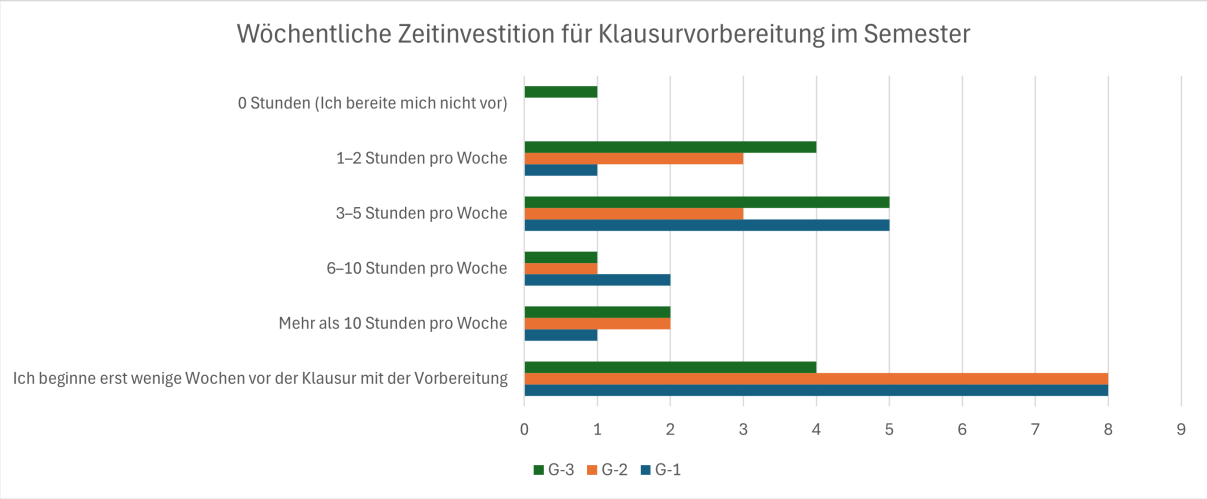

Abbildung 16: Wöchentliche Zeitinvestition für Klausurvorbereitung im Semester

B.4 Typischer Beginn der aktiven Klausurvorbereitung

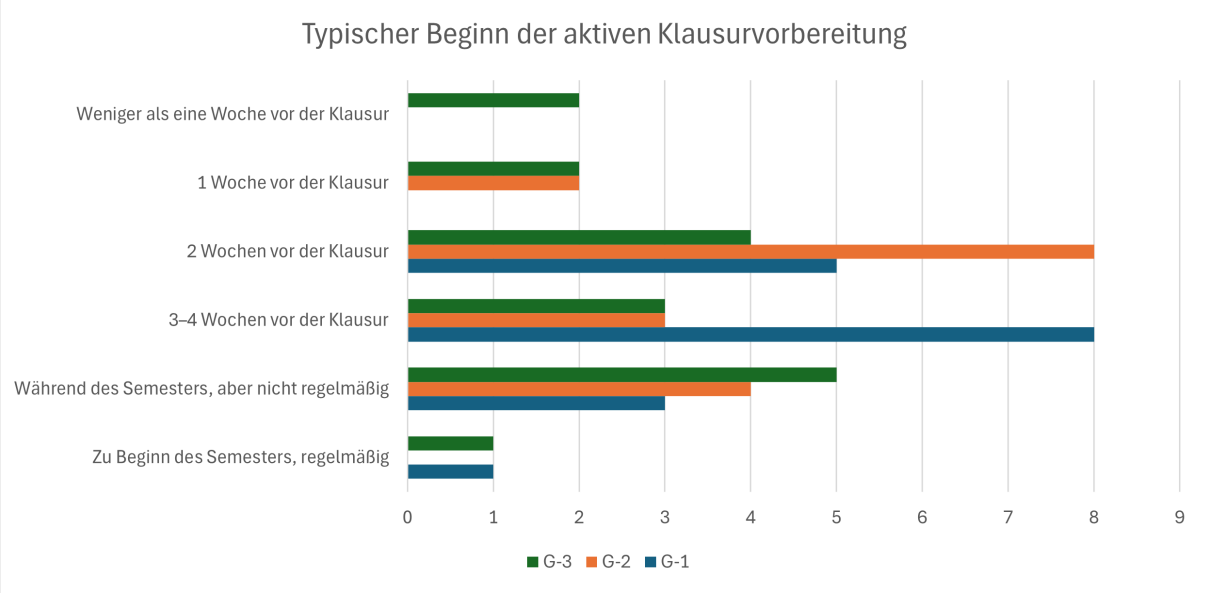

Abbildung 17: Typischer Beginn der aktiven Klausurvorbereitung

C Anhang C

Im weiteren Verlauf werden die detailierten Ergebnisse der unterschiedlichen Messinstrumente aufgeführt. Die Zuordnung erfolgt analog zu den Überschriften in Kapitel 4.

C.1 Aggregierte Analyse

	Gesamtscore		
	G-1	G-2	G-3
Median	6.000	7.000	7.000
Mittelwert	6.176	6.353	6.588
Standardabweichung	1.425	2.029	1.502
Varianz	2.029	4.118	2.257
Shapiro-Wilk	0.906	0.924	0.940
P-Wert Shapiro-Wilk	0.085	0.175	0.323
Minimum	4.000	2.000	4.000
Maximum	9.000	9.000	9.000
Summe	105.000	108.000	112.000

Tabelle 17: Aggregierte Testergebnisse unter verschiedenen Kennzahlen

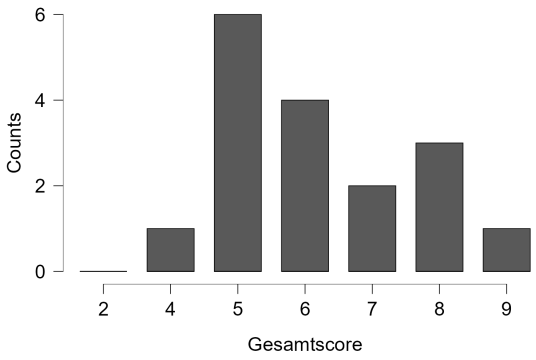

Abbildung 18: Histogramm der Gruppe 1

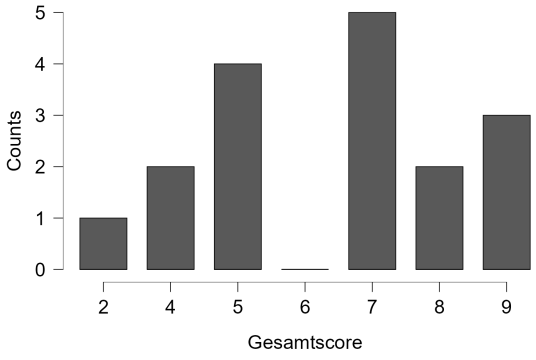

Abbildung 19: Histogramm der Gruppe 2

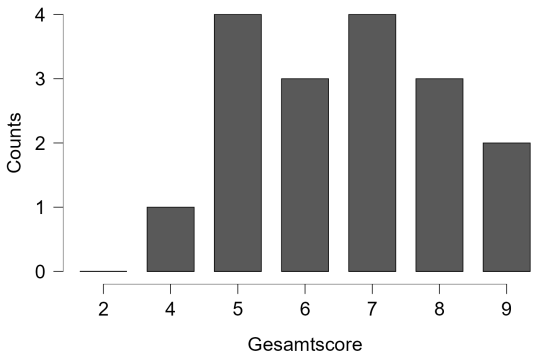

Abbildung 20: Histogramm der Gruppe 3

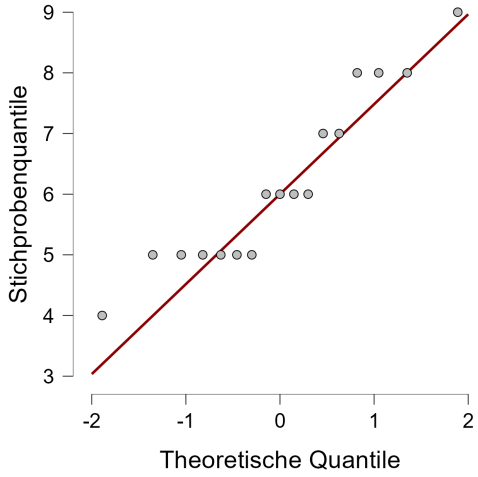

Abbildung 21: Q-Q-Diagramm der Gruppe 1

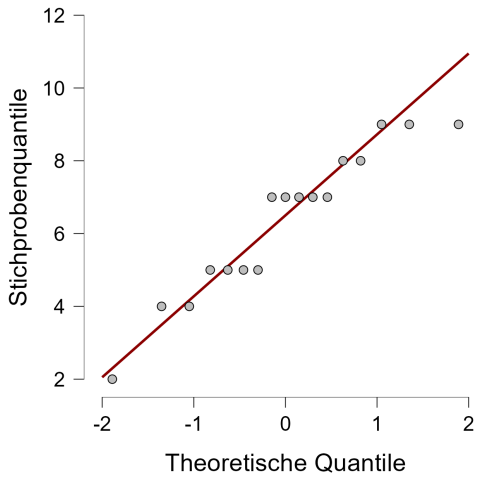

Abbildung 22: Q-Q-Diagramm der Gruppe 2

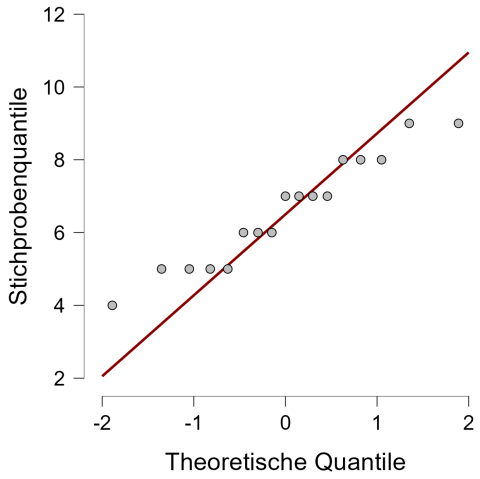

Abbildung 23: Q-Q-Diagramm der Gruppe 3

C.2 Item Level Analyse

	Q1			Q2			Q3		
	G-1	G-2	G-3	G-1	G-2	G-3	G-1	G-2	G-3
Mittelwert	0.765	0.588	0.706	1.000	0.941	0.765	0.353	0.824	0.647
Summe	13	10	12	17	16	13	6	14	11
	Q4			Q5			Q6		
	G-1	G-2	G-3	G-1	G-2	G-3	G-1	G-2	G-3
Mittelwert	0.941	0.941	0.941	0.882	1.000	1.000	0.176	0.353	0.412
Summe	16	16	16	15	17	17	3	6	7
	Q7			Q8			Q9		
	G-1	G-2	G-3	G-1	G-2	G-3	G-1	G-2	G-3
Mittelwert	0.412	0.353	0.588	0.235	0.059	0.059	0.882	0.824	0.824
Summe	7	6	10	4	1	1	15	14	14
	Q10								
	G-1	G-2	G-3						
Mittelwert	0.529	0.471	0.647						
Summe	9	8	11						

Tabelle 18: Testergebnisse pro MCQ

C.3 Ease of Learning (EOL)

	EOL zu Folie 1			EOL zu Folie 2			EOL zu Folie 3		
	G-1	G-2	G-3	G-1	G-2	G-3	G-1	G-2	G-3
Median	2.000	2.000	2.000	3.000	3.000	3.000	3.000	2.000	2.000
Mittelwert	2.353	2.471	2.353	2.765	2.588	2.706	2.647	2.471	2.353
Standardabweichung	0.931	0.800	0.931	0.831	1.064	0.849	0.702	0.717	0.702
Minimum	1.000	1.000	1.000	2.000	1.000	1.000	1.000	1.000	1.000
Maximum	4.000	4.000	4.000	4.000	4.000	4.000	4.000	4.000	4.000

	EOL zu Folie 4			EOL zu Folie 5			EOL zu Folie 6		
	G-1	G-2	G-3	G-1	G-2	G-3	G-1	G-2	G-3
Median	2.000	2.000	2.000	2.000	2.000	2.000	3.000	2.000	3.000
Mittelwert	2.529	2.412	2.412	2.235	2.353	2.353	2.647	2.294	2.647
Standardabweichung	0.624	0.870	0.712	0.562	0.786	0.786	0.996	0.686	0.862
Minimum	2.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Maximum	4.000	4.000	4.000	3.000	4.000	4.000	5.000	4.000	4.000

	EOL zu Folie 7			EOL zu Folie 8			EOL zu Folie 9		
	G-1	G-2	G-3	G-1	G-2	G-3	G-1	G-2	G-3
Median	3.000	3.000	3.000	3.000	2.000	2.000	3.000	3.000	3.000
Mittelwert	3.235	2.941	3.059	2.412	2.412	2.588	3.000	2.588	2.941
Standardabweichung	0.831	0.748	0.748	0.712	0.712	1.004	1.061	1.278	1.478
Minimum	2.000	2.000	2.000	1.000	1.000	1.000	1.000	1.000	1.000
Maximum	4.000	5.000	5.000	3.000	4.000	5.000	5.000	4.000	5.000

	EOL zu Folie 10		
	G-1	G-2	G-3
Median	3.000	4.000	3.000
Mittelwert	3.059	3.059	3.294
Standardabweichung	1.249	1.519	1.532
Minimum	1.000	1.000	1.000
Maximum	5.000	5.000	5.000

Tabelle 19: EOL-Werte pro Folie eingeteilt nach Gruppe

C.4 Judgement of Learning (JOL)

	JOL zu Folie 1			JOL zu Folie 2			JOL zu Folie 3		
	G-1	G-2	G-3	G-1	G-2	G-3	G-1	G-2	G-3
Median	4.000	4.000	4.000	4.000	4.000	4.000	3.000	4.000	4.000
Mittelwert	3.647	3.647	3.765	3.706	4.059	4.176	3.118	3.529	3.588
Standardabweichung	0.606	1.057	0.752	0.920	0.748	0.636	0.781	0.717	0.507
Minimum	3.000	1.000	3.000	2.000	2.000	3.000	2.000	2.000	3.000
Maximum	5.000	5.000	5.000	5.000	5.000	5.000	4.000	4.000	4.000

	JOL zu Folie 4			JOL zu Folie 5			JOL zu Folie 6		
	G-1	G-2	G-3	G-1	G-2	G-3	G-1	G-2	G-3
Median	4.000	4.000	4.000	4.000	3.000	4.000	4.000	4.000	4.000
Mittelwert	3.529	3.765	3.765	3.765	3.529	3.941	3.706	3.882	3.882
Standardabweichung	0.874	0.664	0.903	0.970	0.624	0.748	0.920	0.600	0.697
Minimum	2.000	3.000	2.000	1.000	3.000	2.000	2.000	3.000	3.000
Maximum	5.000	5.000	5.000	5.000	5.000	5.000	5.000	5.000	5.000

	JOL zu Folie 7			JOL zu Folie 8			JOL zu Folie 9		
	G-1	G-2	G-3	G-1	G-2	G-3	G-1	G-2	G-3
Median	3.000	4.000	4.000	3.000	3.000	3.000	4.000	4.000	4.000
Mittelwert	3.235	3.647	3.471	3.235	3.235	2.941	3.824	3.765	3.941
Standardabweichung	0.970	0.862	0.800	0.664	0.752	0.659	0.951	1.033	0.966
Minimum	2.000	2.000	2.000	2.000	2.000	2.000	2.000	2.000	2.000
Maximum	5.000	5.000	5.000	4.000	5.000	4.000	5.000	5.000	5.000

	JOL zu Folie 10		
	G-1	G-2	G-3
Median	4.000	4.000	4.000
Mittelwert	4.176	4.059	4.235
Standardabweichung	0.809	1.088	0.752
Minimum	3.000	2.000	3.000
Maximum	5.000	5.000	5.000

Tabelle 20: JOL-Werte pro Folie eingeteilt nach Gruppe

C.5 Zeiterfassung

	EOL-Zwischenphase (Min)			Vorbereitungsphase (Min)			JOL-Zwischenphase (Min)		
	G-1	G-2	G-3	G-1	G-2	G-3	G-1	G-2	G-3
Mittelwert	5.536	6.085	6.924	22.986	22.178	28.922	4.428	7.263	4.015
Standardabweichung	2.824	2.079	2.122	9.074	12.670	8.896	7.361	11.830	7.377
Minimum	0.780	2.260	3.560	0.720	0.490	0.200	1.440	1.130	1.300
Maximum	12.820	8.930	11.100	31.480	32.380	42.670	32.780	32.950	32.470

Tabelle 21: Differenzierte Zeiterfassungsdaten für die einzelnen Phasen

	Vorbereitungszeit (Min)			Testzeit (Min)		
	G-1	G-2	G-3	G-1	G-2	G-3
Shapiro-Wilk	0.940	0.696	0.954	0.967	0.970	0.959
P-Wert Shapiro-Wilk	0.315	$< .001$	0.517	0.767	0.813	0.606

Tabelle 22: Ergebnisse des Shapiro-Wilk Tests für die Vorbereitungs- und Testzeit

C.6 Fragen zur subjektiven Bewertung der Effizienz

- **EZP**: Wie viel der 30 Minuten an Vorbereitungszeit, konnten Sie *effizient* nutzten? [6]
- **EZZ**: Wie bewerten Sie den Zeitaufwand bei der Vorbereitung mit der von Ihnen verwendeten Methode? (Skala: 1 = sehr zeitaufwendig, 5 = sehr effizient)
- **EZST**: Wie viel Stress haben Sie bei der Vorbereitung mit dieser Methode empfunden? (Skala: 1 = Sehr stressig, 5 = Gar nicht stressig)
- **EZB**: Wie sehr hat die Nutzung von Sprachmodellen (mit vorgegebenen Prompts) Ihre Prüfungsvorbereitung beschleunigt? (Skala: 1 = Überhaupt nicht, 5 = Deutlich schneller)

C.7 Fragen zur subjektiven Bewertung der Effektivität

- **ETMV**: Inwiefern konnte die Nutzung des Sprachmodell Ihnen helfen, die Prüfungsinhalte besser zu verstehen? (Skala: 1 = Gar nicht hilfreich, 5 = Sehr hilfreich)
- **ETLS**: Wie gut hat das Sprachmodell mit den bereitgestellten Prompts den Lernstoff strukturiert? (Skala: 1 = Gar nicht hilfreich, 5 = Sehr hilfreich)
- **ETV**: Wie viel Prozent der gesamten Testinhalte haben Sie durch ihre Vorbereitung verstanden? (Antwortmöglichkeiten: 0–25 %, 26–50 %, 51–75 %, 76–99 %, 100 %)
- **ETHA/ effektiv**: Welche Aspekte der Methode empfanden Sie als besonders hilfreich, um Ihre Ziele zu erreichen?
- **ETNH/ ineffektiv**: Welche Aspekte der Methode empfanden Sie als wenig hilfreich oder hinderlich bei der Erreichung Ihrer Ziele?

C.8 Shapiro-Wilk-Testergebnis für ETV

	ETV		
	G-1	G-2	G-3
Shapiro-Wilk	0.638	0.751	0.554
P-Wert Shapiro-Wilk	$< .001$	$< .001$	$< .001$

Tabelle 23: Shapiro-Wilk-Testergebnis für ETV

[6] Falls es Studierende gab, die weniger als 30 Minuten der Vorbereitungszeit benutzten, wurde mündlich darauf hingewiesen, die tatsächlich benötigte Zeit zu bewerten und nicht die 30 Minuten.

C.9 Häufigkeitstabellen für Effektivität und Ineffektivität

Gruppe	Kategorie	Häufigkeit	Prozent	Prozent gültig	Prozent kumulativ
G-1	Arbeit mit Foliensatz	5	29.412	29.412	29.412
	Erklärungen durch KI	0	0.000	0.000	29.412
	Hochladen des Foliensatzes	0	0.000	0.000	29.412
	Quizfragen durch KI	0	0.000	0.000	29.412
	Zusammenfassungen durch KI	0	0.000	0.000	29.412
	Lernzettel	7	41.176	41.176	70.588
	Prompt Handbook	0	0.000	0.000	70.588
	Wiederholen	2	11.765	11.765	82.353
	Keine hilfreichen Aspekte	3	17.647	17.647	100.000
	Fehlend	0	0.000		
	Gesamt	17	100.000		
G-2	Arbeit mit Foliensatz	0	0.000	0.000	0.000
	Erklärungen durch KI	8	38.095	38.095	38.095
	Hochladen des Foliensatzes	1	4.762	4.762	42.857
	Quizfragen durch KI	4	19.048	19.048	61.905
	Zusammenfassungen durch KI	6	28.571	28.571	90.476
	Lernzettel	2	9.524	9.524	100.000
	Prompt Handbook	0	0.000	0.000	100.000
	Wiederholen	0	0.000	0.000	100.000
	Keine hilfreichen Aspekte	0	0.000	0.000	100.000
	Fehlend	0	0.000		
	Gesamt	21	100.000		
G-3	Arbeit mit Foliensatz	0	0.000	0.000	0.000
	Erklärungen durch KI	5	26.316	26.316	26.316
	Hochladen des Foliensatzes	0	0.000	0.000	26.316
	Quizfragen durch KI	10	52.632	52.632	78.947
	Zusammenfassungen durch KI	1	5.263	5.263	84.211
	Lernzettel	0	0.000	0.000	84.211
	Prompt Handbook	3	15.789	15.789	100.000
	Wiederholen	0	0.000	0.000	100.000
	Keine hilfreichen Aspekte	0	0.000	0.000	100.000
	Fehlend	0	0.000		
	Gesamt	19	100.000		

Tabelle 24: Häufigkeiten der subjektiven Effektivität nach Kategorien

Gruppe	Kategorie	Häufigkeit	Prozent	Prozent gültig	Prozent kumulativ
G-1	Fehlende Hilfsmittel	6	35.294	35.294	35.294
	Fehlender Kontext	2	11.765	11.765	47.059
	Benutzung der KI	0	0.000	0.000	47.059
	Ausschließlich Text von KI	0	0.000	0.000	47.059
	Zeitverschwendung durch KI	0	0.000	0.000	47.059
	Zusammenfassungen durch KI	0	0.000	0.000	47.059
	Prompt Handbook	0	0.000	0.000	47.059
	spezifischer Prompt	0	0.000	0.000	47.059
	Zeitdruck	1	5.882	5.882	52.941
	Keine hinderlichen Aspekte	8	47.059	47.059	100.000
	Fehlend	0	0.000		
	Gesamt	17	100.000		
G-2	Fehlende Hilfsmittel	0	0.000	0.000	0.000
	Fehlender Kontext	1	5.882	5.882	5.882
	Benutzung der KI	1	5.882	5.882	11.765
	Ausschließlich Text von KI	1	5.882	5.882	17.647
	Zeitverschwendung durch KI	2	11.765	11.765	29.412
	Zusammenfassungen durch KI	3	17.647	17.647	47.059
	Prompt Handbook	0	0.000	0.000	47.059
	spezifischer Prompt	0	0.000	0.000	47.059
	Zeitdruck	0	0.000	0.000	47.059
	Keine hinderlichen Aspekte	9	52.941	52.941	100.000
	Fehlend	0	0.000		
	Gesamt	17	100.000		
G-3	Fehlende Hilfsmittel	0	0.000	0.000	0.000
	Fehlender Kontext	0	0.000	0.000	0.000
	Benutzung der KI	1	5.882	5.882	5.882
	Ausschließlich Text von KI	0	0.000	0.000	5.882
	Zeitverschwendung durch KI	0	0.000	0.000	5.882
	Zusammenfassungen durch KI	3	17.647	17.647	23.529
	Prompt Handbook	3	17.647	17.647	41.176
	spezifischer Prompt	4	23.529	23.529	64.706
	Zeitdruck	0	0.000	0.000	64.706
	Keine hinderlichen Aspekte	6	35.294	35.294	100.000
	Fehlend	0	0.000		
	Gesamt	17	100.000		

Tabelle 25: Häufigkeiten der subjektiven Ineffektivität nach Kategorien

C.10 Tabelle zum subjektiven Autonomieverlust

	Autonomieverlust	
	G-2	G-3
Median	4.000	4.000
Mittelwert	3.235	3.647
Standardabweichung	1.251	1.455
Shapiro-Wilk	0.892	0.837
P-Wert Shapiro-Wilk	0.049	0.007

Tabelle 26: Subjektiver Autonomieverlust bei Sprachmodellverwendung

C.11 Fragen für den Vergleich G-2 und G-3

- **Autonomieverlust/ SE**: Haben Sie das Gefühl, dass die Nutzung von Sprachmodellen (mit Prompts) Ihre Selbstständigkeit im Lernen beeinträchtigt hat? (Skala: 1 = Ja, stark, 5 = Nein)

- **Vorteile/ FTVO**: Gab es bei der Vorbereitung bestimmte Vorteile in der Verwendung des Sprachmodells?
- **Nachteile/ FTNA**: Gab es bei der Vorbereitung bestimmte Nachteile in der Verwendung des Sprachmodells?

C.12 Häufigkeitstabellen Vorteile in der „ad hoc" und Prompt Handbook Sprachmodellnutzung

Vorteile	Häufigkeit	Prozent	Prozent gültig	Prozent kumulativ
Aktives Lernen	3	15.789	18.750	18.750
Effizienz	8	42.105	50.000	68.750
Feedback	2	10.526	12.500	81.250
Paraphrasierung	2	10.526	12.500	93.750
Verständnisförderung	1	5.263	6.250	100.000
Fehlend	3	15.789		
Gesamt	19	100.000		

Tabelle 27: Häufigkeiten für Vorteile in der „ad hoc" Nutzung eines Sprachmodells (G-2)

Vorteile	Häufigkeit	Prozent	Prozent gültig	Prozent kumulativ
Aktives Lernen	5	23.810	27.778	27.778
Effizienz	5	23.810	27.778	55.556
Feedback	2	9.524	11.111	66.667
Materialintegration	1	4.762	5.556	72.222
Neuer Kontext	2	9.524	11.111	83.333
Strukturierung	1	4.762	5.556	88.889
Verständnisförderung	2	9.524	11.111	100.000
Fehlend	3	14.286		
Gesamt	21	100.000		

Tabelle 28: Häufigkeiten für Vorteile in der Sprachmodellnutzung mit vormodellierten Prompts (G-3)

C.13 Häufigkeitstabellen Nachteile in der „ad hoc" und Prompt Handbook Sprachmodellnutzung

Nachteile	Häufigkeit	Prozent	Prozent gültig	Prozent kumulativ
Halluzination	4	23.529	66.667	66.667
Promptabhängigkeit	1	5.882	16.667	83.333
Verlust kritisches Hinterfragen	1	5.882	16.667	100.000
Fehlend	11	64.706		
Gesamt	17	100.000		

Tabelle 29: Häufigkeiten für Nachteile in der „ad hoc" Nutzung eines Sprachmodells (G-2)

Nachteile	Häufigkeit	Prozent	Prozent gültig	Prozent kumulativ
Einschränkung	4	21.053	30.769	30.769
Halluzination	1	5.263	7.692	38.462
Informationsüberfluss	3	15.789	23.077	61.538
Unvollständigkeit	3	15.789	23.077	84.615
Zeitverlust	1	5.263	7.692	92.308
spezifischer Prompt	1	5.263	7.692	100.000
Fehlend	6	31.579		
Gesamt	19	100.000		

Tabelle 30: Häufigkeiten für Nachteile in der Sprachmodellnutzung mit vormodellierten Prompts (G-3)

C.14 Absolute Häufigkeiten der Promptnutzung des Prompt Handbooks

Prompt	2.1	2.2	2.3	2.4	2.5	2.6	2.7	2.8
Genutzt	7	4	10	14	10	0	1	3
Nicht genutzt	10	13	7	3	7	17	16	14
Prozent	41,18%	23,53%	58,82%	82,35%	58,82%	0,00%	5,88%	17,65%

Tabelle 31: Absolute & prozentuale Zahlen zur Verwendung der Prompts des Prompt Handbooks

C.15 Fragen zum Feedback des Prompt Handbooks

- **Nutzungsbewertung/ HR**: Wie hilfreich empfanden Sie die vor modellierten Prompts und Affordanzen im Vergleich zur freien Nutzung des Sprachmodells? (Skala: 1 = Gar nicht hilfreich, 5 = Sehr hilfreich)
- **Benutzerfreundlichkeit /IA**: Wie intuitiv empfanden Sie die Anwendung von den vorgegebenen Prompts bei der Klausurvorbereitung? (Skala: 1 = sehr schwierig, 5 = sehr einfach)
- **Schwierigkeiten/ EEES**: Gab es Situationen, in denen die Nutzung von Prompts die Vorbereitung erschwert hat?
- **Weiterempfehlung/ FTE**: Würden Sie die Methode der Nutzung von Sprachmodellen mit Prompts anderen Studierenden empfehlen? Warum?

Eidesstattliche Erklärung

Hiermit versichere ich an Eides statt, dass ich die vorliegende Arbeit im Bachelorstudiengang Wirtschaftsinformatik selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel – insbesondere keine im Quellenverzeichnis nicht benannten Internet-Quellen – benutzt habe. Alle Stellen, die wörtlich oder sinngemäß aus Veröffentlichungen entnommen wurden, sind als solche kenntlich gemacht. Ich versichere weiterhin, dass ich die Arbeit vorher nicht in einem anderen Prüfungsverfahren eingereicht habe. Sofern im Zuge der Erstellung der vorliegenden Abschlussarbeit generative Künstliche Intelligenz (gKI) basierte elektronische Hilfsmittel verwendet wurden, versichere ich, dass meine eigene Leistung im Vordergrund stand und dass eine vollständige Dokumentation aller verwendeten Hilfsmittel gemäß der Guten Wissenschaftlichen Praxis vorliegt. Ich trage die Verantwortung für eventuell durch die gKI generierte fehlerhafte oder verzerrte Inhalte, fehlerhafte Referenzen, Verstöße gegen das Datenschutz- und Urheberrecht oder Plagiate.

Osnabrück, den 30.09.2025

Ort, Datum

Unterschrift

Einwilligungserklärung zur Bibliotheksaufnahme

Ich stimme der Einstellung der Arbeit in die Bibliothek des Fachbereichs Informatik zu.

Osnabrück, den 30.09.2025

Ort, Datum

Unterschrift