Das Concept Image in Studierendenantworten mit Large Language Models klassifizieren

AENNE KNIERIM, HILDESHEIM & THERESA KRUSE, HILDESHEIM

***Zusammenfassung**: Gerade bei fächerübergreifenden Konzepten ist es für Lehrende wichtig, das Concept Image der Lernenden zu einem Begriff zu erfassen. Wir untersuchen, wie Large Language Models (LLMs) dabei unterstützen können. Dazu arbeiten wir mit Studierendentexten zum Thema Graphentheorie und untersuchen, ob die Studierenden Graphen aus einer graphentheoretischen Perspektive betrachten oder sich auf andere Begriffsbilder stützen. Die Klassifizierung wurde mit vier LLMs durchgeführt, um das Concept Image automatisiert zu erfassen. In unserem Vergleich schneidet GPT 4.0 mit einem F1-Score von 0,56 zwar am besten ab, insgesamt erkennen die LLMs das Concept Image jedoch noch nicht zufriedenstellend.*

***Abstract:** Especially in the case of cross-curricular concepts, it is important for teachers to capture the learners' concept image of a term. We are investigating how Large Language Models (LLMs) can support this. To this end, we work with student texts on graph theory and investigate whether they view graphs from a graph-theoretical perspective or rely on other concept images. The classification was carried out with four LLMs to automatically capture the concept image. In our comparison, GPT 4.0 performs best with an F1 score of 0.56, but overall, the LLMs do not yet recognize the concept image satisfactorily.*

1. Einleitung

Innerhalb der Mathematik begegnen Lernenden ähnliche Konzepte in verschiedenen Teilgebieten als cross-curricular concepts (dt. fächerübergreifende Konzepte). Abhängig von den individuellen Lernvoraussetzungen stellt dies eine Herausforderung für Lehrende und Lernende da (Kontorovich, 2018). Gerade im Hochschulkontext mit sehr großen Lerngruppen in Vorlesungen ist es jedoch kaum möglich, einen Eindruck der Vorstellungen der Lernenden zu bekommen (Renkl et al., 2020). Ziel des vorliegenden Beitrages ist es, zu ermitteln, wie der Einsatz von Large Language Models (LLMs) hier den Prozess der Lernstandermittlung unterstützen kann, um somit gerade die Hochschullehre in Vorlesungen weiter zu verbessern. Wir untersuchen dies am Beispiel der Graphentheorie.

Der Zugriff auf das rein kognitive Concept Image kann nur über die Verbalisierung der Concept Definition erfolgen (Tall & Vinner, 1981). Deshalb arbeiten wir mit kurzen Texten von Studierenden, ausgehend von einem graphischen Stimulus. Für diesen Datensatz vergleichen wir eine manuelle Annotation mit der durch verschiedene LLMs. Wir untersuchen also wie LLMs ohne weiteres Finetuning den Prozess unterstützen können, zu ermitteln, welche Vorstellungen die Studierenden bei neu eingeführten Konzepten haben, die sie bereits aus anderen Teildisziplinen in ähnlicher Form kennen. Beim Finetuning werden die Parameter eines Modells optimiert, um die Leistung des Modells auf spezifischen Aufgaben zu verbessern (X.-K. Wu et al., 2025).

Das Concept Image der Studierenden zu kennen, könnte den Workload für Lehrende reduzieren (Pepin et al., 2025), und somit die Lehrplanung verbessern und im nächsten Schritt auch eine Rolle bei der automatisierten und personalisierten Generierung von Feedback spielen. Wir arbeiten mit bereits annotierten Studierendendaten, bei denen die Teilnehmenden die Aufgabe hatten, Eigenschaften verschiedener Graphen anzugeben (N=182). Die Annotation gibt an, ob die Studierenden den Graphen aus einer graphentheoretischen Perspektive betrachten oder ob sie sich auf andere, z. B. geometrisch oder stochastisch geprägte, Begriffsbilder stützen (Kruse, 2025). Es soll nun überprüft werden, ob ein LLM bei der Annotation zu den gleichen Ergebnissen kommt wie Expert:innen bei der manuellen Annotation.

Konkret soll folgende Forschungsfrage beantwortet werden: Wie unterscheiden sich verschiedene LLMs in der Bestimmung des Concept Images von Lernenden ausgehend von der Concept Definition?

Dieses Paper trägt auf dreifache Weise zur Forschung bei: Erstens handelt es sich, nach unserem besten Wissen, um die erste Studie zur automatisierten Erkennung des Concept Images mit Large Language Models und ergänzt damit Studien, die die Anwendung und Adaption von LLMs auf domänenspezifische Fragestellungen, hier im Bereich der Mathematikdidaktik, erforschen. Zweitens widmen sich mathematikdidaktische Studien zu Potentialen von LLMs oft ausschließlich Modellen aus der GPT-Familie, nicht aber anderen Modellen. Drittens können

die Ergebnisse dieser Studie dabei helfen, einen weiteren Schritt in die Richtung automatisierter Feedbackgenerierung zu machen, wie es beispielsweise bereits im Bereich Programmieren erprobt wurde (Kiesler et al., 2023).

Automatisiert generierte Erkenntnisse über den Wissensstand der Studierenden können auch hilfreiche Einblicke für weitergehende Unterrichts- und Lehrplanung bieten. Aufgrund der Vielzahl an Vorteilen und Herausforderungen, die LLMs für die Mathematikdidaktik bieten, ist es auch wichtig, die Limitationen dieser Technologie zu erforschen (Pepin et al., 2025).

In Abschnitt 2 definieren wir zunächst das Concept Image nach Tall & Vinner (1981) und betrachten die Anwendung von Large Language Models in der Bildungsforschung. Anschließend betrachten wir Graphentheorie als Lerngegenstand sowie die hier untersuchten fächerübergreifenden Konzepte. In Abschnitt 3 erfolgt die Erläuterung des methodischen Vorgehens. Erst wird die manuelle Annotation beschrieben, dann folgt die Beschreibung des Versuchs der Automatisierung mithilfe der Modelle Meta Llama 3.1 8B Construct, Meta Llama 3.1 7B Construct (Grattafiori et al., 2024; Touvron et al., 2023), EM German Mistral AI (Harries & Wetter, o. J.) und GPT 4.0 (OpenAI et al., 2024). Dafür formulieren die Autorinnen das Problem als Klassifikationsaufgabe. In den weiteren Sektionen folgen die Dokumentation unserer Resultate (Abschnitt 4), die Limitationen (Abschnitt 5) und ein Ausblick (Abschnitt 6).

2. Theoretischer Hintergrund & Forschungsstand

In diesem Abschnitt erfolgt ein kurzer Überblick über das Concept Image und zur Verwendung von Large Language Models in der Didaktik. Danach geben wir eine didaktische Einordnung der Graphentheorie und stellen die verwendeten fächerübergreifenden Konzepte vor.

2.1 Konzeptvorstellungen von Lernenden

Unter dem Concept Image werden die individuellen nonverbalen Assoziationen von Lernenden mit mathematischen Begriffen verstanden (Tall & Vinner, 1981; Vinner, 1991). Dabei kann es sich beispielsweise um visuelle Repräsentationen oder persönliche Erfahrungen handeln. Entsprechend wandelt sich das Concept Image auch im Lernprozess und es ist möglich, dass für einen Begriff konkurrierende Concept Images bei einer Person vorliegen, die dann situationsabhängig evoziert werden.

Das Concept Image kann als latente Variable nur implizit erfasst werden, nämlich anhand der sogenannten Concept Definition. Dabei werden zwei Arten von Concept Definition unterschieden: Die formale und die persönliche. Während die formale diejenige ist, die in Lehrbüchern steht und in der Regel von der mathematischen Community insgesamt anerkannt wird, gilt das für die persönliche Concept Definition eines Lernenden nicht unbedingt. Sie kann zudem situationsabhängig und über die Zeit variieren.

Abhängig vom Konzept selbst sowie von den Voraussetzungen der Lernenden, können verschiedene Herausforderungen gerade im Hinblick auf die Entstehung eines Concept Images bestehen. Ein Aspekt, der das erschweren kann, sind cross-curricular concepts, also Konzepte, die in verschiedenen mathematischen Teilgebieten auf ähnliche Weise verwendet werden (Kontorovich, 2018). Zusätzlich liegen für diese fächerübergreifenden Konzepte mitunter auch gleiche Bezeichnungen vor. Wir fokussieren hier auf die visuelle Ebene als Ausgangspunkt.

Dass die Verbalisierung einen guten Zugang zum Ermitteln der Concept Images von Lernenden liefert, wird in der Mathematikdidaktik und auch in der Psychologie schon sehr lange diskutiert (Austin & Howson, 1979; Vygotsky, 1962). Obwohl es aus der Computerlinguistik bereits lange vor LLMs Werkzeuge gab, um große Textmengen zu analysieren, erfolgte die Untersuchung von Concept Images in der Mathematikdidaktik vor allem qualitativ für kleinere Kohorten, wie beispielsweise zu Differentialgleichungen (McCarty & Sealey, 2025), Ableitungen (Bingolbali & Monaghan, 2008) und Funktionen (Panaoura et al., 2017) oder in der Linearen Algebra (Kontorovich, 2018).

2.2 Large Language Models in der didaktischen Forschung

Eine Schlüsselentwicklung in der Sprachtechnologie war die Nutzung von Transformer-Modellen und dem zugrundeliegenden Self-Attention-Mechanismus (Devlin et al., 2019; Floridi & Chiriatti, 2020; Kasneci et al., 2023; Vaswani et al., 2017). Mithilfe von Finetuning werden und wurden solche vortrainierten, transformer-basierten Encoder-Modelle auf verschiedenen Klassifikations- und Generierungsaufgaben angewendet (Kasneci et al., 2023). Ein Nachteil dieser Vorgehensweisen ist, dass für das Training und die Evaluation dieser Modelle große Goldstandarddatensätze benötigt werden und die Modelle domänenspezifisch und meist nicht für vielfältige Aufgaben einsetzbar sind. Diese Datensätze zu erstellen ist teuer und aufwendig. Zudem sind

Programmierkenntnisse erforderlich, was die Anwendung für Nichtexpert:innen erschwert.

Dem Einsatz von Large Language Models und anderen Anwendungen des maschinellen Lernens wird für die Didaktik allgemein und die Mathematikdidaktik im Speziellen ein großes Potential zugeschrieben. Bei der Verwendung der sogenannten Decoder-Modelle sind große Datenmengen nicht mehr notwendig, stattdessen kann man natürlichsprachlich, ohne große Datensätze und ohne Programmierkenntnisse, mit ihnen interagieren. Die Modelle können sowohl Lernende als auch Lehrende unterstützen; für die Lehrenden kann es beispielsweise bei der Evaluation und dem Feedback für Lernende unterstützen (Kasneci et al., 2023). Konkret für den Bereich der Mathematikdidaktik sind LLMs bei der Bewertung vor allem für sofortiges Feedback und damit verbundenes passendes Aufgabendesign eingesetzt worden (Pepin et al., 2025). Im Folgenden geben wir einen Überblick über konkrete vorhandene Forschungsarbeiten zur Verwendung von LLMs in der Didaktik. Dabei betrachten wir sowohl Arbeiten aus der Mathematikdidaktik als auch aus anderen Bereichen, die Anknüpfungspunkte für unseren Untersuchungsgegenstand liefern.

Wir führen hier zunächst nur eine Annotation der Concept Images von Lernenden durch. Wenn jedoch mehrere konkurrierende Concept Images evoziert werden, können zwischen diesen Logikfehler entstehen, was besonders dann passieren kann, wenn das Concept Image der Lernenden aus einer Mischung verschiedener fächerübergreifender Konzepte besteht. Dass solche Logikfehler in Texten von Lernenden durch LLMs erkannt werden können, wurde bereits für Experimentierprotokolle im Naturwissenschaftsunterricht gezeigt (Bewersdorff et al., 2023).

Weiterhin können durch LLMs produzierte Feedbacktexte zur Motivationssteigerung dienen, wie es für Englischlernende gezeigt wurde (Meyer et al., 2024; Schiller et al., 2024). In Meyer et al. (2024) wird GPT-3.5-turbo benutzt, um Feedback zu für die Verfassung eines Aufsatzes auf Englisch zu geben. Die Kontrollgruppe erhält kein Feedback. Genau wie in Schiller et al. (2024) zeigt sich das KI-generierte Feedback als nützlich für die Schüler:innen und kann ihre Motivation steigern. Das Concept Image von Lernenden zu erkennen, birgt Potential für die zielgerichtete Bereitstellung von Feedback.

Des Weiteren könnten basierend auf LLM-Annotationen passende Aufgaben und Lehrpläne generiert werden, wie es bereits untersucht wurde (Buchholtz & Huget, 2024; Dilling, 2024; Schorcht et al., 2025): Schorcht et al. (2025) führen eine Studie zu verschiedenen LLMs durch, die eingesetzt werden, um mathematische Aufgaben zu evaluieren und zu modifizieren. Die Ergebnisse werden mithilfe einer Chat-Kette von LLMs erzeugt und können in mehreren Iterationen modifiziert werden. Die generierten Aufgaben werden einerseits mithilfe von Likert-Skalen und andererseits mithilfe menschlicher Lehrkräfte in Ausbildung vorgenommen. Die LLMs liefern dort insgesamt brauchbare, aber nicht immer optimale Ergebnisse.

Nach Dilling (2024) liegen die Potentiale von LLMs in der Anwendung für automatisiertes Feedback zu mathematischen Aufgaben in der Adaptivität in Bezug auf die Aufgabe und die Lernenden (Dilling, 2024). Darüber hinaus stellt Dilling (2024) fest, dass die Anwendung von Large Language Models hilfreich sein kann, um Bereiche zu identifizieren, in denen Studierende Hilfe brauchen. Auch diese Ergebnisse können zukünftig mit eingesetzt werden, um an das Concept Image angepasste Aufgaben stellen zu kennen, die auch direkt auf mögliche Fehlvorstellungen gerade bei fächerübergreifenden Konzepten eingehen. Buchholtz und Huget (2024) testen verschiedene Prompting-Strategien für die Erstellung von Lehrplänen. Während strukturiertes Prompting ein nützliches Reflexionstool für die Unterrichtsplanung sein kann, waren Modifizierungen des Outputs von Menschen für die Qualität der Lehrpläne entscheidend. Der Output der LLMs variiert in diesem Setting insgesamt stark. In der Zukunft könnten die hier erzeugten Ergebnisse auch genutzt werden, um nicht nur Aufgaben, sondern auch Lehrpläne an das Concept Image anzupassen.

Arbeiten, die das Concept Image selbst mittels automatischer Annotation bestimmen, liegen bislang nicht vor. Eine explorative qualitative Studie zu Herausforderungen in der Anwendung von LLMs für die Bewertung des konzeptuellen Verständnisses mathematischer Inhalte von Lernenden wurde jedoch bereits durchgeführt (Hankeln, 2024). Sie zeigt, dass die Unterschiede in der Herausbildung von Verständnis zwischen LLMs und Menschen die Bewertung menschlichen Verständnisses von LLMs negativ beeinflussen. LLMs nutzen Wortkookkurenzen in natürlicher Sprache, um kontextsensitives linguistisches Wissen zu erlernen und generieren dann probabilistische Vorhersagen wahrscheinlicher nächster Wörter (Suresh et al., 2023). Das „Verständnis“ dieser Maschinen beruhe also auf Wortrelationen und Wahrscheinlichkeiten, während Menschen Konzeptelemente in größere Netzwerke einbetten und übertragen (Hankeln, 2024; Hiebert & Carpenter,

1992). Am Beispiel von ChatGPT (OpenAI et al., 2024) zeigt Hankeln (2024), dass der Output nicht solide und kohärent fundiert ist, sondern auf Mustern aus den Trainingsdaten basiert. Sie zeigt zwei grundlegende Probleme auf: Statistiken über für das Pre-Training genutzte Texte sind meist nicht bekannt: daher fehlen Informationen darüber, wie viele mathematikspezifische Texte von Studierenden in den Trainingsdaten vorhanden sind. Eine zweite von Hankeln (2024) genannte Herausforderung ist die der Kontextsensitivität, die widerspiegelt, dass das Modell Probleme haben kann, Aufgaben zu lösen, die ein konzeptuelles Verständnis erfordern. Diese Tendenz kann zu unlogischen, nicht kohärenten Antworten oder Halluzinationen führen (Hankeln, 2024; T. Wu et al., 2023). In der vorliegenden Studie wird die automatische Klassifikation des Concept Images von Studierenden mithilfe von LLMs untersucht und sie ergänzt damit Hankelns (2024) explorativen Ansatz um eine quantitative Studie.

2.3 Graphentheorie

Im Folgenden betrachten wir zunächst Graphentheorie als Lerngegenstand. Anschließend führen wir die hier relevanten graphentheoretischen und die damit verbundenen fächerübergreifenden Konzepte ein.

2.3.1 Graphentheorie als Lerngegenstand

In der mathematikdidaktischen Literatur wurden immer wieder Vorzüge von Graphentheorie als Lerngegenstand hervorgehoben: Die Arbeit mit Graphentheorie hat beispielsweise positive motivationale Effekte für Schüler:innen (Windler, 2018). Außerdem hat Graphentheorie ein großes Anwendungspotential, beispielsweise für die Ingenieursausbildung (Ferrarello et al., 2022; Sabra & Tabchi, 2024), und bietet gute Visualisierungsmöglichkeiten (Sandefur et al., 2022). Trotz dieser Vorzüge ist diskrete Mathematik insgesamt und damit insbesondere auch Graphentheorie nur selten Teil der schulischen Curricula (Sandefur et al., 2022). Das Thema ist somit gut für Forschungsarbeiten geeignet, da die Studienteilnehmenden einerseits gleichermaßen wenig Vorwissen haben, im Rahmen einer Erhebung aber andererseits gut mit den Konzepten arbeiten können. Gerade aufgrund der Anschaulichkeit ist die Graphentheorie ein Gegenstand, der ohne viele Voraussetzungen erlernt werden kann (Greefrath et al., 2022).

Für den hier gewählten Untersuchungsbereich der Verbalisierung des Concept Images für fächerübergreifende Konzepte ist Graphentheorie also gut geeignet. Neben dem einheitlich nicht vorhandenen Vorwissen der Studienteilnehmenden werden graphentheoretische Konzepte sowie die dazugehörigen Bezeichnungen oft auch in anderen mathematischen Teilgebieten verwendet. Dies betrifft beispielsweise die Begriffe Baum, Isomorphismus und Kreis, die wir für die vorliegende Untersuchung ausgewählt haben. Im folgenden Abschnitt werden diese Konzepte aus fachmathematischer Perspektive eingeführt und didaktisch analysiert.

2.3.2 Konzepte der Graphentheorie in verwandten Disziplinen

Graphentheoretische Einführungen gibt es beispielsweise mit einem eher anschaulichen Zugang (Nitzsche, 2009) oder aus einer strenger mathematischen Perspektive (Diestel, 2017). Daran orientieren sich auch die folgenden Definitionen.

Anders als bei Graphen als Darstellungsform von Funktionen, wie sie aus der Schule bekannt sind, handelt es sich bei graphentheoretischen Graphen um ein geordnetes Paar zweier Mengen: Ein Graph *G=(V,E)* besteht aus einer Eckenmenge *V* (engl. *vertices*) und einer Kantenmenge *E* (engl. *edges*). Dabei ist *E* eine Teilmenge der zweielementigen Teilmengen von *V*. Um die beiden Bedeutungen des Wortes *Graph* voneinander abzugrenzen, wird in englischer Literatur von *vertex-edge graphs* gesprochen (Kontorovich, 2018). Somit könnte man im Deutschen zwischen Funktionsgraphen und Ecken-Kanten-Graphen unterscheiden, sofern sich die intendierte Bedeutung nicht direkt aus dem Kontext erschließt.

In der Ebene werden die Ecken eines Graphen in der Regel als Punkte dargestellt und die Kanten als Verbindungslinien zwischen ihnen gezeichnet (vgl. Abb. 1 bis 4). Dabei spielen weder die Länge der Linien noch die Winkel zwischen ihnen eine Rolle. Entsprechend ist die Darstellung eines Graphen in der Ebene nicht eindeutig (vgl. Abb. 4). Zudem heißen zwei Graphen *G* und *G'* isomorph, wenn es eine bijektive Abbildung $\varphi\colon \mathrm{V} \to V'$ gibt, sodass für alle Paare $v_i, v_j \in V$ gilt, dass $\{v_i, v_j\} \in E$ genau dann, wenn $\{\varphi(v_i), \varphi(v_j)\} \in E'$. Abbildung 4 zeigt zwei zueinander isomorphe Graphen.

Die zweidimensionalen Repräsentationen von Graphen erinnern an geometrische Objekte, bei denen Aspekte wie Länge und Winkel in der Darstellung durchaus relevant sind. Somit können entsprechende Concept Images bei den Lernenden evoziert werden. Auf der sprachlichen Ebene können diese unpassenden evozierten Concept Images noch unterstützt werden, da in der Graphentheorie auch bereits aus der Geometrie bekannte Bezeichnungen

mit anderen Definitionen verwendet werden. Ein Kreis beispielsweise ist ein Graph mit $|V| = |E|$ und $E = \{\{v_0, v_1\}, \{v_1, v_2\}, \ldots, \{v_{n-1}, v_n\}, \{v_n, v_0\}\}$. In Abbildung 2 ist der Kreis mit sechs Ecken und sechs Kanten dargestellt. Diese Repräsentation ähnelt zwar dem geometrischen Konzept eines Kreises, gleicht in der Darstellung aber vielmehr einem Hexagon. Hier liegt neben der terminologischen Polysemie das fächerübergreifende Konzept eines Hexagons vor.

Ähnliches gilt für den sogenannten Petersengraph, der in Abbildung 3 gezeigt wird: In ihm scheinen die geometrischen Konzepte des Pentagons und des Pentagramms kombiniert, auch wenn die Darstellung aus graphentheoretischer Perspektive ganz anders aussehen könnte und ohne die beiden geometrischen Objekte auskommen würde.

Abschließend betrachten wir Graphen, die als Bäume bezeichnet werden: Dabei handelt es sich um Graphen, in denen es für je zwei paarweise verschiedene Knoten jeweils genau einen Weg gibt, der diese beiden Knoten verbindet (Bondy & Murty, 2008)[1]. Ein Beispiel für einen solchen Baum zeigt Abbildung 1. Neben der Polysemie des aus der Gemeinsprache bekannten Wortes *Baum*, ist im Hinblick auf fächerübergreifende Konzepte vor allem die Nutzung von Bäumen als Visualisierungsstütze in der Stochastik in Form von Baumdiagrammen relevant: Den Studierenden sind Bäume als Baumdiagramme zur Darstellung mehrstufiger Zufallsexperimente bekannt. Somit evoziert ein Baum vermutlich eher Concept Images von Zufallsexperimenten als solche aus der Graphentheorie.

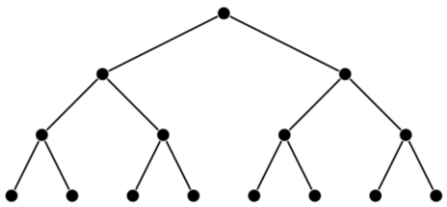

Abb. 1: Baum

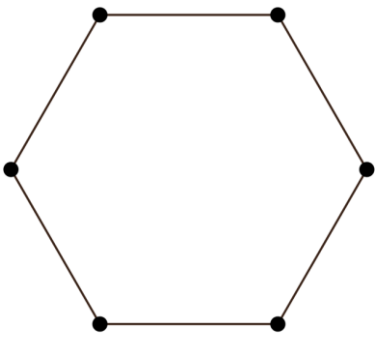

Abb. 2: Kreis

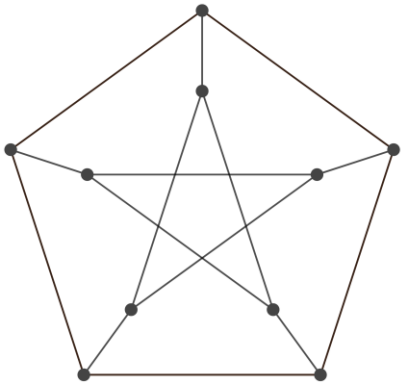

Abb. 3: Petersen-Graph

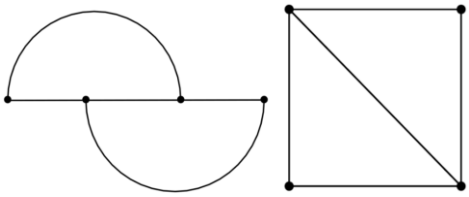

Abb. 4: Zwei isomorphe Graphen

3. Methodisches Vorgehen

In dieser Sektion erläutern die Autorinnen das methodische Vorgehen. Zuerst wird der Prozess der Datenerhebung und manuellen Annotation beschrieben. Danach werden die hier verwendeten Promptingstrategien dargestellt.

3.1 Datenerhebung und manuelle Annotation

Die verwendeten Daten sind Teil einer größeren Studie zum Erlernen der Fachsprache mit lexikografischen Hilfsmitteln (Kruse, 2025). Im vorliegenden Beitrag wird geprüft, ob ein Teil der dort durchgeführten manuellen Annotation mithilfe von LLMs ersetzt werden kann.

Die Erhebung wurde im Sommer 2020 mit Mathematik-Lehramtsstudierenden an der Universität Hildesheim durchgeführt (N=182). Diese wurden über eine Lehrveranstaltung zu *Algebra und Zahlentheorie* rekrutiert und haben als Belohnung für die Teilnahme Bonuspunkte in der dazugehörigen Klausur erhalten. Gemäß Studienordnung sind die Teilnehmenden zu diesem Zeitpunkt im vierten Bachelorsemester und haben noch keinerlei graphentheoretische Veranstaltungen besucht.

In den hier betrachteten Teilaufgaben wurden den Studierenden vier Graphen im graphentheoretischen Sinne gezeigt, insbesondere ein Baum (vgl. Abb. 1), ein Kreisgraph (vgl. Abbildung 2), der Petersengraph (vgl. Abb. 3) sowie zwei zueinander isomorphe Graphen (vgl. Abb. 4). Zu den Graphen aus

Abbildung 1 bis 3 wurde jeweils die Frage gestellt: *Welche Eigenschaften hat der abgebildete Graph?* Zu Abbildung 4 lautet die Frage *Welcher Zusammenhang besteht zwischen den beiden Graphen? - abgesehen von der Tatsache, dass beide jeweils fünf Kanten und vier Ecken haben.* Als Hilfsmittel bei der Bearbeitung durfte Wikipedia genutzt werden.

Da 182 Studierende an der Erhebung teilgenommen haben, sind insgesamt 728 Instanzen aus den vier Aufgaben vorhanden. Die Antworten der Studierenden auf die Fragen sind zwischen 6 und 1045 mit durchschnittlich 118 Zeichen lang. Teilweise haben sie in ganzen Sätzen geantwortet, teilweise in Stichpunkten. Die Studie wurde mit einem Online-Fragebogen durchgeführt und eine sprachliche Bereinigung der Antworten ist nicht erfolgt, um die Daten möglichst authentisch zu lassen.

Die 728 Antworten wurden semi-manuell von einer der Autorinnen des vorliegenden Beitrags einer der folgenden vier Kategorien zugeordnet:

1) Graphentheoretisches Beschreibung (z. B. „Beide Graphen sind ungerichtet und nicht gewichtet. Es sind beides zusammenhängende Graphen.“ zu Abb. 4)
2) Nicht-graphentheoretische Beschreibung (z. B. „Ein Sechseck oder Hexagon ist ein Polygon (Vieleck), bestehend aus sechs Ecken und sechs Seiten. Sind alle sechs Seiten gleich lang, spricht man von einem gleichseitigen Sechseck. Sind darüber hinaus alle Winkel an den sechs Ecken gleich groß, dann wird das Sechseck regulär oder regelmäßig genannt.“ zu Abb. 2)
3) Mischung aus graphentheoretischer und nicht-graphentheoretischer Beschreibung (z. B. „Baumdiagramm Zusammenhängender Graf“ zu Abb. 1)
4) Antwort gibt keine inhaltliche Beschreibung (z. B. „kann ich nicht beantworten, habe nichts gefunden“ zu Abb. 4)

Insgesamt gehören 54,0 Prozent der Aussagen zu Kategorie (1), 35,6 Prozent zu Kategorie (2), 9,6 Prozent zu Kategorie (3) und 0,9 Prozent zu Kategorie (4). DIe Verteilung ist also nicht ausgewogen.

Die Zuordnung erfolgte dabei auf Basis der Terminologie, die die Studierenden zur Beschreibung der Graphen verwendet haben. Konkret wurde geschaut, welche Begriffe zur Beschreibung genutzt werden.

Hinweise auf eine graphentheoretische Perspektive und das damit verbundene Concept Image geben Wörter, die im gegebenen Kontext nur eine graphentheoretische Bedeutung, aber keine geometrische oder stochastische haben. Dazu zählen: bipartit, gerichtet, isomorph, Kante, Knoten, Mehrfachkante, Multigraph, planar, Schleife, vollständig, Weg, zusammenhängend, zyklisch sowie die Namen von Personen, deren Resultate hier angewendet werden können: Euler, Hamilton und Petersen.

Für die nicht-graphentheoretische Perspektive greifen hier vor allem Begriffe aus der Geometrie oder Stochastik, die von den Teilnehmenden genutzt werden. Dazu gehören für die Stochastik Baumdiagramm, Ereignis, Experiment, Wahrscheinlichkeit, Zufall und erscheinen in der Regel zu Abbildung 1. Alle diese Wörter deuten auf eine Interpretation des Graphen als Baumdiagramm zur Darstellung eines Zufallsexperiments hin und haben keine graphentheoretische Bedeutung.

Bei den anderen Beispielen kommen vor allem Begriffe aus der Geometrie vor: Abstand, diagonal, Durchmesser, gleichseitig, Länge, n-Eck, parallel, spiegeln, Symmetrie, Winkel. All diese Eigenschaften sind aus einer graphentheoretischen Perspektive für die Darstellung nicht relevant, sondern Teil der Geometrie und deuten somit auf ein entsprechenden Concept Image hin.

Wenn sowohl Begriffe aus der Graphentheorie als auch aus Stochastik bzw. Geometrie gleichermaßen vorkommen, erfolgt die Zuordnung zu Kategorie 3), da kein eindeutiger Rückschluss auf das Conept Image möglich ist. Vermutlich werden in diesem Fall bei den Teilnehmenden konkurrierende Concept Images evoziert. Erscheinen hingegen keinerlei Fachbegriffe, wird der Kommentar der Kategorie 4) zugeordnet.

3.2 Automatische Annotation: Prompting und Modelle

Prompting. Einem explorativem Ansatz folgend nutzen die Autorinnen Few-Shot Learning und Zero-Shot Learning, um die Modelle Meta Llama 3.1 8B Construct, Meta Llama 3.1 7B Construct (Grattafiori et al., 2024; Touvron et al., 2023), EM German Mistral AI (Harries & Wetter, o. J.) und GPT 4.0[2] (OpenAI et al., 2024) für die Klassifikationsaufgabe anzuwenden.

Analog zur manuellen Annotation sollen die vier Klassen „GRAPHENTHEORETISCH“ für die graphentheoretische Perspektive, „ANDERE“ für die nicht-

graphentheoretische Perspektive sowie „UNSICHER“ für die Mischung und „KEINE“ detektiert werden. Wichtig ist, dass dieses Papier nicht auf Modelloptimierung abzielt, sondern die Möglichkeiten des Einsatzes von LLMs für die automatisierte Klassifizierung des Concept Images explorativ erforscht.

Insgesamt liegen 728 zu klassifizierende Antworten von den 182 Teilnehmenden aus vier Teilaufgaben vor. Als Goldstandard werden die Annotationen benutzt, die in Abschnitt 3.1 beschrieben sind. Zunächst haben die Autorinnen einen Zero-Shot Ansatz verwendet, bei dem derselbe Prompt für alle Modelle verwendet wurde:

> In einer Erhebung sollten Studierende zu einem abgebildeten Graphen folgende Fragen beantworten: ‚Welche Eigenschaften hat der abgebildete Graph?‘. Wir möchten nun wissen, ob aus der Antwort der Studierenden hervorgeht, ob der Antwort der Studierenden eine graphentheoretische oder eine andere Perspektive zugrunde liegt. Bitte label die folgenden Aussagen (1) bis (N) mit ‚GRAPHENTHEORETISCH‘, ‚ANDERE‘, ‚UNSICHER‘ oder ‚KEINE‘.

Zero-Shot Learning ist eine Aufgabe aus dem Bereich *transfer learning*. Die Idee ist, Wissen aus bekannten Trainingsinstanzen auf ungesehene Daten in der Zieldomäne zu transferieren (W. Wang et al., 2019). Statt anhand von Trainingsdaten soll das Modell ergänzte Informationen (engl. *auxiliary information*) zur Hilfe nehmen (W. Wang et al., 2019). In unserem Experiment konnten die Modelle GPT 4.0 und Llama 3.17B Construct mit der Methode Zero-Shot Learning keine evaluierbaren Klassifikationsergebnisse erzeugen. Deshalb wurde für diese Modelle mit zwei verschiedenen Prompts aus dem Bereich Few-Shot Learning gearbeitet.

Few-Shot Learning wird als die Fähigkeit eines Modells des Maschinellen Lernens definiert, aus wenigen Trainingsbeispielen zu lernen (Parnami & Lee, 2022). Few-Shot Learning Ansätze sind gegenüber klassischen Deep Learning Modellen eine effiziente Alternative, da sie keine Trainingsdatensätze benötigen, die aufwändig erstellt werden müssen. Daher wird Large Language Models, die gute Ergebnisse mit One-Shot- oder Few-Shot learning produzieren, großes Potential zugeschrieben. Die hier gewählte Vorgehensweise könnte als Baseline-Ansatz verstanden werden, auf dem weitere Modellanpassungen in späteren Arbeiten vorgenommen werden können. Da Few-Shot-Learning hier der vielversprechendere Ansatz ist, könnte in zukünftigen Studien Chain-of-Thought-Prompting durchgeführt und für diese Aufgabe getestet werden. Chain-of-Thought-Prompting fordert das Modell auf, eine Begründung mit den generierten Antworten zu liefern und bietet so einen interpretierbaren Einblick in das Verhalten des Modells (Hou et al., 2023; Wei et al., 2022).

Die Länge des Prompts hat eine entscheidende Rolle dabei gespielt, evaluierbare Ergebnisse zu erzeugen. Bei anderer Promptlänge kam es zu Halluzinationen. Halluzinationen sind ein übliches Problem bei der Arbeit mit LLMs (Huang et al., 2025; Li et al., 2024) und sind definiert als generierte Inhalte, die „entweder unsinnig sind oder mit dem vorgegebenen Quellinhalt nicht übereinstimmen“ und als „unsachlich, unsinnig oder inkonsistent“ (Huang et al., 2025; Waldo & Boussard, 2024). Die Autorinnen können nicht feststellen, ob das „Vergessen“ von Teilen des Prompts bei Nutzung der längeren Promptversion auf Probleme mit der Verarbeitung der *long-range dependency* zurückzuführen ist. Während EM-German Mistral AI und das jüngere Modell Llama 3.1 8B Construct nur mit dem kürzeren Prompt messbare Ergebnisse erzeugen, ist beim älteren Llama 3.1 7B Construct und bei GPT 4.0 der längere Prompt, also Few-Shot Learning notwendig, Die verschiedenen Prompts sind in Tabelle 1 dargestellt und wurden zur Vergleichbarkeit dokumentiert.

Large Language Models. Frühere Sprachmodelle haben der Modellierung und Generierung von Sprache gedient, während mit der Innovation neuer Sprachmodelle wie GPT-4 ein Schritt in Richtung von allgemeiner Problemlösung gemacht wurde (Ling et al., 2024; OpenAI et al., 2024).

Die Llama 2 Modelle (Touvron et al., 2023) sind auf zwei Billionen Token trainiert, während die Llama 3 Modelle (Grattafiori et al., 2024) auf einem Datensatz trainiert wurde, der sieben Mal so groß ist (Naveed et al., 2024). Die Llama 3 Modelle haben ihre Vorgänger übertroffen und haben in vielfältigen Aufgaben eine vergleichbare Performance im Vergleich zu den führenden GPT-4-Modellen (Grattafiori et al., 2024). EM-German basiert auf Mistral AI 7B (https://huggingface.co/jphme/em_german_leo_mistral). Das Finetuning wurde auf deutschsprachigen Daten durchgeführt. Mistral 7B wurde auf sieben Billionen Parametern trainiert und übertrifft Llama 2 Modelle.

Modelle	Prompt
Llama 3.1 7B Construct	In einer Erhebung sollten Studierende zu einem abgebildeten Graphen folgende Fragen beantworten: "Welche Eigenschaften hat der abgebildete Graph?" Wir möchten nun wissen, ob aus der Antwort der Studierenden hervorgeht, ob der Antwort der Studierenden eine graphentheoretische oder eine andere Perspektive zugrunde liegt. Bitte label die folgenden Aussagen (1) bis (N) mit "GRAPHENTHEORETISCH", "KEINE", "NICHT" oder "MISCHFORM". Bitte gib das Label chronologisch aus. Hier sind Beispiele: "KEINE": kann ich nicht beantworten, habe nichts gefunden. "GRAPHENTHEORETISCH": - 6 Knoten - 6 Kanten - Pro Knoten zwei Kantenverbindungen an je ein anderen Knoten "UNSICHER": Baumdiagramm - Entscheidungsbaum besteht immer aus einem Wurzelknoten und beliebig vielen inneren Knoten sowie mindestens zwei Blättern - geordnete, gerichtete Bäume - Darstellung von Entscheidungsregeln - eine Methode zur automatischen Klassifikation von Datenobjekten - übersichtliche Darstellung von formalen Regeln "ANDERE": Vermutung: 3*Radius = Diagonale
Llama 3.1 8B Construct	Studierende beantworten zu einem abgebildeten Graphen folgende Fragen: ‚Welche Eigenschaften hat der abgebildete Graph?'. Bitte label die folgende Aussage mit ‚GRAPHENTHEORETISCH', ‚ANDERE', ‚UNSICHER' oder ‚KEINE'. Gib nur das Label an.
EM-German Mistral AI	Studierende beantworten zu einem abgebildeten Graphen folgende Fragen: ‚Welche Eigenschaften hat der abgebildete Graph?'. Bitte label die folgende Aussage mit ‚GRAPHENTHEORETISCH', ‚ANDERE', ‚UNSICHER' oder ‚KEINE'. Gib nur das Label an. Aussage {idx}: '{response}'
GPT 4.0	In einer Erhebung sollten Studierende zu einem abgebildeten Graphen folgende Fragen beantworten: "Welche Eigenschaften hat der abgebildete Graph?" Wir möchten nun wissen, ob aus der Antwort der Studierenden hervorgeht, ob der Antwort der Studierenden eine graphentheoretische oder eine andere Perspektive zugrunde liegt. Bitte label die folgenden Aussagen (1) bis (N) mit "GRAPHENTHEORETISCH", "KEINE", "NICHT" oder "MISCHFORM". Bitte gib das Label chronologisch aus. Hier sind Beispiele: "KEINE": kann ich nicht beantworten, habe nichts gefunden. "GRAPHENTHEORETISCH": - 6 Knoten - 6 Kanten - Pro Knoten zwei Kantenverbindungen an je ein anderen Knoten "UNSICHER": Baumdiagramm - Entscheidungsbaum besteht immer aus einem Wurzelknoten und beliebig vielen inneren Knoten sowie mindestens zwei Blättern - geordnete, gerichtete Bäume - Darstellung von Entscheidungsregeln - eine Methode zur automatischen Klassifikation von Datenobjekten - übersichtliche Darstellung von formalen Regeln "ANDERE": Vermutung: 3*Radius = Diagonale

Tab. 1: Übersicht über die Prompts für die verschiedenen Modelle.

4. Ergebnisse der automatischen Annotation

Tabelle 2 zeigt die Klassifikationsergebnisse der von uns trainierten Modelle. Obwohl verschiedene Promptingansätze getestet wurden, konnten die Llama Modelle 3.1 7B Construct und Llama 3.1 8B mit der vorgestellten Methodik keine differenzierte Klassifikation erzielen und labeln alle Antworten als „graphentheoretisch“ (s. Tab. 2). Auch in Hankelns (2024) Ergebnissen, zeigt sich, dass LLMs Schwierigkeiten haben, Aufgaben zu lösen, die konzeptuelles Verständnis benötigen. Es lässt sich auch keine Verbesserung zwischen dem Modell Llama 3.1 7B Construct und Llama 3.1 8B Construct feststellen. Der einzige Unterschied besteht darin, dass das spätere Modell Llama 3.1 8B Construct mithilfe eines Zero-Shot Learning Ansatzes auswertbare Ergebnisse erzeugt, während das frühere Modell Llama 3.1 7B wie GPT 4.0 mit Beispielen gepromptet wird (vgl. Tab. 1).

Das Modell EM-German Mistral AI erfasst mit dem Recall-Wert von 0,29 einen großen Anteil der positiven Fälle (der graphentheoretischen Antworten) nicht. Dafür labelt es viele Aussagen, die keine graphentheoretische Perspektive haben, als positiv. Nur 52% der als positiv gelabelten Antworten sind tatsächlich positiv. Das Modell trifft sowohl falsch positive als auch falsch negative Aussagen und hat einen F1-Score von 0,29. GPT 4.0 übertrifft alle Modelle mit einem F1-Score von 0,56. Trotzdem erfasst das Modell 39% der tatsächlich positiven Fälle nicht.

LLMs kommen typischerweise mit einer prädefinierten Kontextlänge (context window): zum Beispiel der Input für Llama-Modelle muss kürzer als 2048 Tokens lang sein (Touvron et al. 2023, Chen et al., 2023). Das hier verwendete Mistral-Modell und das GPT-4.0 haben längere Kontextlängen als die Llama-Modelle. Dies kann eine Erklärung für die Unterschiede in der Performance der Modelle auf der betrachteten Aufgabe sein.

Modell	Precision	Recall	F1-Score
EM-German Mistral AI	0.52	0.29	0.29
GPT 4.0	0.63	0.61	**0.56**
Llama 3.1 7B Construct	1	0	0
Llama 3.1 8B Construct	1	0	0

Tab. 2: Klassifikationsergebnisse der verschiedenen Modelle.

5. Limitationen

Die vorliegenden Ergebnisse sind durch eine Reihe von Limitationen beschränkt. Zunächst ist die Verallgemeinerbarkeit der LLM-Nutzung limitiert durch den gewählten Gegenstandsbereich und durch die gestellten Aufgaben, sodass in einem anderen Kontext womöglich andere Ergebnisse erzielt werden würden. Andere LLMs sowie andere Versionen und andere Zugangswege als die hier verwendeten, können zudem ebenfalls andere Ergebnisse liefern. Durch die rasche Weiterentwicklung der Modelle können unsere Ergebnisse also nur einen punktuellen Einblick geben. Außerdem variieren die LLM-Ergebnisse mit jeder Durchführung, sodass mehrmalige Durchläufe und ein damit verbundenes Interrater-Agreement hilfreich wären, um die automatischen Annotationsergebnisse besser einschätzen zu können. Auch nur minimal abgewandelte andere Prompts oder auch eine andere Reihenfolge der Input-Daten können zu immer wieder anderen Ergebnissen führen. Aufgrund der Vielzahl an möglichen Formulierungen, ist es jedoch unmöglich, alle Varianten zu erproben und zu vergleichen.

Des Weiteren stammen die genutzten Daten aus einer künstlichen Erhebungssituation und sind nicht im Zuge eines authentischen Lernprozesses entstanden. Eine Rücksprache mit den Lernenden zu den ermittelten Concept Images ist ebenfalls nicht erfolgt, könnte aber gerade in den Zweifelsfällen genauere Ergebnisse liefern. Hilfreich wäre es dafür vermutlich auch, die Lernenden längere Texte als es hier der Fall war, produzieren zu lassen.

Zudem wurde die manuelle Annotation des Concept Images lediglich von einer Person kriteriengeleitet durchgeführt. Eine Annotation durch mehrere Personen in Kombination mit einer Berechnung des Interrater-Agreements sowie eine genaue Definition für den Umgang mit Zweifelsfällen würde hier validere Ergebnisse liefern. Dann könnten die Ergebnisse der verschiedenen Modelle mit den Rating-Ergebnissen von verschiedenen Personen verglichen werden.

6. Diskussion und Ausblick

In diesem Papier untersuchen wir das Potential von Large Language Models, um das Concept Image von Studierenden zu klassifizieren. Dies geschieht am Beispiel einer Aufgabe aus dem Bereich der Graphentheorie, in dem Studierende Eigenschaften von Graphen beschreiben sollten. Wir können somit sowohl aus didaktischer als auch aus technischer Perspektive Schlussfolgerungen ziehen.

Zunächst betrachten wir die technische Perspektive: Explorative qualitative Forschung hat bereits Herausforderungen identifiziert, die mit der Kontextsensitivität und den Trainingsdaten von Modellen einhergehen (Hankeln, 2024). Unsere Ergebnisse bestätigen diese Erkenntnisse und sind weiterhin im Einklang mit Studien, die Schwierigkeiten in der Adaption von Large Language Models auf domänenspezifische Fragestellungen aufzeigen, welche durch die Komplexität von Domänendaten und vielfältige andere Beschränkungen auftauchen (Ling et al., 2024; C. Wang et al., 2023).

Aus didaktischer Perspektive zeigt sich, dass LLMs langfristig gesehen ein gutes Hilfsmittel sein können, um in großen Lerngruppen wie Vorlesungen einen Eindruck der Concept Images zu erhalten. Hierfür werden textbasierte Daten benötigt. Für diesen Zweck stellen wir den hier verwendeten Datensatz zur Verfügung (Kruse-Kurbach & Knierim, 2025). Eine weitere Idee wäre, schriftliche Reflexionen über den Gegenstand in die Lehre einzubauen, um Textdaten für die LLM-basierte Forschung zu kreieren. Möglich wäre dann der Einsatz von LLMs bei der Generierung von Feedback zu Beweisaufgaben, wofür allerdings auch noch weitere Forschung erforderlich ist. Dafür kann das Concept Image als Ausgangspunkt dienen.

Zukünftige Studien sollten zudem das Potential von LLMs, für automatisiertes Lernendenfeedback und Lehrendenassistenz verwendet zu werden, mit Finetuning erforschen. Ein klassischer Ansatz mithilfe von Encoder-Modellen wie der BERT-Familie könnten unter der Voraussetzung fruchtbringend sein, dass genügend Trainingsdaten vorhanden sind. Hybride Modelle, wie Retrieval Augmented Generation (RAG) kombinieren die Stärken der impliziten Wissensgrundlage vortrainierter Modelle mit retrievalbasiertem Wissen (Lewis et al., 2020). RAG-Modelle halluzinieren weniger und der generierte Output ist eher faktenbasiert (Lewis et al., 2020). Daher wären RAG-Modelle ein Ansatz, der für die Erfassung des Concept Images von Lernenden vielversprechend sein kann.

Unsere Ergebnisse können dazu dienen, weiter zu untersuchen, wie Concept Images automatisiert erfasst werden können. Auch wenn die Schlagwortsuche, wie wir sie bei der manuellen Annotation verwendet haben, zunächst einen hilfreichen Ansatz darstellen kann, um Rückschlüsse über das Concept Image zu ziehen, so ist die Erstellung entsprechender Listen aufwändig. Entsprechend ist die Verwendung traditioneller Korpusanalysen zwar theoretisch eine methodische Alternative, um das gleiche Ziel zu erreichen, für die Vorlesungspraxis allerdings mit zu viel Aufwand verbunden. Wenn die automatische Annotation gute Ergebnisse liefert, kann dieser Schritt wegfallen. Auch die Erstellung möglicher Kategorien könnte mit einer Weiterentwicklung der Modelle obsolet werden, sodass das simple Prompting mit der Bitte nach der Zuordnung von passenden Concept Images bereits hilfreiche Ergebnisse liefern würde.

In einem möglichen Anwendungsszenarien könnten Lehrenden in Vorlesungen die Lernenden um spontane schriftliche Assoziationen zu einem Begriff, einem Bild oder einer anderen Art von Vignette bitten. Diese Assoziationen können dann nach direkter Eingabe der Studierenden in einer entsprechenden Umgebung innerhalb weniger Sekunden annotiert werden, sodass die Lehrenden sowohl qualitativ als auch quantitativ einen besseren Einblick in den Lernstand der Lernenden haben.

Zusammenfassend können wir unsere Forschungsfrage wie folgt beantworten: Die Qualität der Klassifikation ist stark modellabhängig. Basierend auf unseren Ergebnissen zeigen GPT und Mistral Potenziale für Anwendungen in der Hochschullehre. Deren konkrete Ausgestaltung hängt jedoch auch von der weiteren Entwicklung dieser Modelle ab. Insgesamt können LLMs teilweise bei der Ermittlung des Concept Images in der Lehre eingesetzt werden. Voraussetzung ist jedoch, dass sich das Concept Image für den jeweiligen Gegenstand auch manuell eindeutig klassifizieren lässt und entsprechend schriftliche Daten vorliegen. Potenziale können durch extensives Finetuning ausgebaut werden, was allerdings die Zugänglichkeit für Dozierende ohne Programmierkenntnisse stark einschränkt.

Danksagungen

Wir bedanken uns bei den Studierenden, die an der Erhebung teilgenommen haben. Außerdem danken wir den anonymen Reviewer:innen für ihre hilfreichen Kommentare.

Endnoten

[1] Formal werden Bäume als kreisfreie Graphen definiert. Wir definieren sie hier so, da dies anschaulicher ist. Eine formale Einführung als kreisfreie Graphen würde den Rahmen des vorliegenden Beitrags übersteigen.

[2] GPT wird über die Weboberfläche uhiki.de verwendet.

Literatur

Bewersdorff, A., Seßler, K., Baur, A., Kasneci, E. & Nerdel, C. (2023). Assessing student errors in experimentation using artificial intelligence and large language models: A comparative study with human raters. *Computers and Education: Artificial Intelligence, 5*, 100177. https://doi.org/10.1016/j.caeai.2023.100177

Bingolbali, E. & Monaghan, J. (2008). Concept image revisited. *Educational Studies in Mathematics, 68*(1), 19–35. https://doi.org/10.1007/s10649-007-9112-2

Bondy, A. & Murty, U.S.R. (2025). *Graph Theory*. Springer London. https://doi.org/10.1007-978-1-84628-970-5

Buchholtz, N., & Huget, J. (2024). *ChatGPT as a reflection tool to promote lesson planning competencies of pre-service teachers* (S. 129–136). In E. Faggiano, A. Clark-Wilson, M. Tabach, & H.-G. Weigand (Hrsg.), *Proceedings of the 17th ERME Topic Conference MEDA 4*. MEDA 4. https://hal.science/hal-05022073

Chen, S., Wong, S., Chen, L., & Tian, Y. (2023). *Extending context window of large language models via positional interpolation*. arXiv. https://arxiv.org/abs/2306.15595

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of deep bidirectional transformers for language understanding*. In J. Burstein, C. Doran, & T. Solorio (Hrsg.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Volume 1 (Long and Short Papers)* (S. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Diestel, R. (2017). *Graphentheorie* (5. Auflage). Spektrum.

Dilling, F. (2024). Large language models as formative assessment and feedback tools? – A systematic report. In P. Iannone, F. Moons, C. Drüke-Noe, E. Geraniou, F. Morselli, K. Klingbeil, M. Veldhuis, S. Olsher, H. Corinna, G. Peter, L. Edith, R. Benjamin, A. Bilge Yilmaz, Ö. Mehmet Fatih, K. Bilge Kuşdemir, A. Catarina, P. Torulf, A. Miglena, Z. Agnese Del, … G. Martina (Hrsg.), *Proceedings of FAME 1 – Feedback & Assessment in Mathematics Education* (ETC 14), 5-7 June 2024, Utrecht (The Netherlands). (S. 98–105). Utrecht University & ERME. https://doi.org/10.5281/zenodo.14231455

Ferrarello, D., Gionfriddo, M., Grasso, F. & Mammana, M. F. (2022). Graph theory and combinatorial calculus: an early approach to enhance robust understanding. *ZDM – Mathematics Education, 54*(4), 847–864. https://doi.org/10.1007/s11858-022-01407-w

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines, 30*(4), 681–694. https://doi.org/10.1007/s11023-020-09548-1

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., … Ma, Z. (2024). The Llama 3 Herd of Models. https://arxiv.org/abs/2407.21783

Greefrath, G., Siller, H.-S., Vorhölter, K. & Kaiser, G. (2022). Mathematical modelling and discrete mathematics: opportunities for modern mathematics teaching. *ZDM – Mathematics Education, 54*(4), 865–879. https://doi.org/10.1007/s11858-022-01339-5

Hankeln, C. (2024). Challenges in Using ChatGPT for Assessing Conceptual Understanding in Mathematics Education. *Journal of Mathematics Education, 17*(1), 1–15. https://doi.org/10.26711/007577152790171

Harries, J. P. & Wetter, T. (o. J.). Repository for the EM German Model. https://github.com/jphme/EM_German

Hiebert, J. & Carpenter, T. P. (1992). Learning and teaching with understanding. In Grouws, D. A. (Hrsg.), *Handbook of research on mathematics teaching and learning: A project of the National Council of Teachers of Mathematics* (S. 65–97).

Hou, C., Zhu, G., Zheng, J., Zhang, L., Huang, X., Zhong, T., Li, S., Du, H., & Ker, C. L. (2024). Prompt-based and fine-tuned GPT models for context-dependent and -independent deductive coding in social annotation. In *Proceedings of the 14th Learning Analytics and Knowledge Conference (LAK '24)* (pp. 518–528). Association for Computing Machinery. https://doi.org/10.1145/3636555.3636910

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems, 43*(2), Article 42. https://doi.org/10.1145/3703155

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Kiesler, N., Lohr, D. & Keuning, H. (2023). Exploring the Potential of Large Language Models to Generate Formative Programming Feedback. In *2023 IEEE Frontiers in Education Conference (FIE)*, 1–5. https://doi.org/10.1109/FIE58773.2023.10343457

Kontorovich, I. (2018). Why Johnny struggles when familiar concepts are taken to a new mathematical domain: towards a polysemous approach. *Educational Studies in Mathematics, 97*(1), 5–20. https://doi.org/10.1007/s10649-017-9778-z

Kruse, T. (2025). *Ein fachsprachliches elektronisches Lernwörterbuch*. De Gruyter. https://doi.org/10.1515/9783111591353

Kruse-Kurbach, T. & Knierim, A. (2025). Replication Data for: Das Concept Image in Studierendenantworten mit Large Language Models klassifizieren (Version V1). GRO.data. https://doi.org/10.25625/M27U4K

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Hrsg.), *Advances in Neural Information Processing Systems (NeurIPS 33)* (pp. 9459–9474). Curran Associates, Inc. https://doi.org/10.1145/3703155

Li, J., Consul, S., Zhou, E., Wong, J., Farooqui, N., Ye, Y., Manohar, N., Wei, Z., Wu, T., Echols, B., Zhou, S. & Diamos, G. (2024). Banishing LLM Hallucinations Requires Rethinking Generalization. https://arxiv.org/abs/2406.17642

Ling, C., Zhao, X., Lu, J., Deng, C., Zheng, C., Wang, J., Chowdhury, T., Li, Y., Cui, H., Zhang, X., Zhao, T., Panalkar, A., Mehta, D., Pasquali, S., Cheng, W., Wang, H., Liu, Y., Chen, Z., Chen, H., White, C., Gu, Q., Pei, J., Yang, C., & Zhao, L. (2024). Domain

Specialization as the Key to Make Large Language Models Disruptive: A Comprehensive Survey. https://arxiv.org/abs/2305.18703

McCarty, T. & Sealey, V. (2025). Mathematicians' Conceptualizations of Differentials in Calculus and Differential Equations. *International Journal of Research in Undergraduate Mathematics Education, 11*(2), 343–363. https://doi.org/10.1007/s40753-024-00254-2

Meyer, J., Jansen, T., Schiller, R., Liebenow, L. W., Steinbach, M., Horbach, A. & Fleckenstein, J. (2024). Using LLMs to bring evidence-based feedback into the classroom: AI-generated feedback increases secondary students' text revision, motivation, and positive emotions. *Computers and Education: Artificial Intelligence, 6*, 100199. https://doi.org/10.1016/j.caeai.2023.100199

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N. & Mian, A. (2024). A Comprehensive Overview of Large Language Models. https://arxiv.org/abs/2307.06435

Nitzsche, M. (2009). *Graphen für Einsteiger: Rund um das Haus vom Nikolaus* (3. Auflage). Vieweg+Teubner Verlag. https://doi.org/10.1007/978-3-8348-9968-2

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., … Zoph, B. (2024). GPT-4 Technical Report. https://arxiv.org/abs/2303.08774

Panaoura, A., Michael-Chrysanthou, P., Gagatsis, A., Elia, I. & Philippou, A. (2017). A Structural Model Related to the Understanding of the Concept of Function: Definition and Problem Solving. *International Journal of Science and Mathematics Education, 15*(4), 723–740. https://doi.org/10.1007/s10763-016-9714-1

Parnami, A. & Lee, M. (2022). Learning from Few Examples: A Summary of Approaches to Few-Shot Learning. https://arxiv.org/abs/2203.04291

Pepin, B., Buchholtz, N. & Salinas-Hernández, U. (2025). A Scoping Survey of ChatGPT in Mathematics Education. *Digital Experiences in Mathematics Education, 11*(1), 9–41. https://doi.org/10.1007/s40751-025-00172-1

Renkl, A., Eitel, A. & Glogger-Frey, I. (2020). Die Vorlesung – nur schlecht, wenn schlecht vorgelesen: Warum eine gut gemachte Vorlesung einen Platz im Methodenrepertoire verdient. In R. Egger & B. Eugster (Hrsg.), *Lob der Vorlesung* (S. 113–136). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-29049-8_6

Sabra, H. & Tabchi, T. (2024). Connectivity in resources for teaching graph theory in engineering education. *ZDM – Mathematics Education, 56*(7), 1473–1487. https://doi.org/10.1007/s11858-024-01613-8

Sandefur, J., Lockwood, E., Hart, E. & Greefrath, G. (2022). Teaching and learning discrete mathematics. *ZDM – Mathematics Education, 54*(4), 753–775. https://doi.org/10.1007/s11858-022-01399-7

Schiller, R., Fleckenstein, J., Mertens, U., Horbach, A. & Meyer, J. (2024). Understanding the effectiveness of automated feedback: Using process data to uncover the role of behavioral engagement. *Computers & Education, 223*, 1–16. https://doi.org/10.1016/j.compedu.2024.105163

Schorcht, S., Peters, F. & Kriegel, J. (2025). Communicative AI Agents in Mathematical Task Design: A Qualitative Study of GPT Network Acting as a Multi-professional Team. *Digital Experiences in Mathematics Education, 11*(1), 77–113. https://doi.org/10.1007/s40751-024-00161-w

Suresh, S., Mukherjee, K., Yu, X., Huang, W.-C., Padua, L. & Rogers, T. T. (2023). Conceptual structure coheres in human cognition but not in large language models. https://arxiv.org/abs/2304.02754

Tall, D. & Vinner, S. (1981). Concept image and concept definition in mathematics with particular reference to limits and continuity. *Educational Studies in Mathematics, 12*(2), 151–169. https://doi.org/10.1007/BF00305619

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E. & Lample, G. (2023). LLaMA: Open and Efficient Foundation Language Models. https://arxiv.org/abs/2302.13971

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS 2017)* (pp. 6000–6010). Curran Associates. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Vinner, S. (1991). The Role of Definitions in the Teaching and Learning of Mathematics. In D. Tall (Hrsg.), Advanced mathematical thinking (Bd. 11, S. 65–81). Kluwer Academics

Vygotsky, L. (1962). *Thought and language*. (E. Hanfmann & G. Vakar, Hrsg.). The MIT Press.

Waldo, J. & Boussard, S. (2024). GPTs and Hallucination: Why do large language models hallucinate? *Queue, 22*(4), 19–33. https://doi.org/10.1145/3688007

Wang, C., Liu, X., Yue, Y., Tang, X., Zhang, T., Jiayang, C., Yao, Y., Gao, W., Hu, X., Qi, Z., Wang, Y., Yang, L., Wang, J., Xie, X., Zhang, Z. & Zhang, Y. (2023). Survey on Factuality in Large Language Models: Knowledge, Retrieval and Domain-Specificity. https://arxiv.org/abs/2310.07521

Wang, W., Zheng, V. W., Yu, H. & Miao, C. (2019). A Survey of Zero-Shot Learning: Settings, Methods, and Applications. *ACM Trans. Intell. Syst. Technol., 10*(2). https://doi.org/10.1145/3293318

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V. & Zhou, D. (2022). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho & A. Oh (Hrsg.), *Advances in Neural Information Processing Systems 35* (NeurIPS 2022). https://proceedings.neurips.cc/paper_files/paper/2022

Windler, M. (2018). *Der Einfluss graphentheoretischer Konzepte im Mathematikunterricht der Grundschule auf psychologische Schülerinnen- und Schülermerkmale* [PhD Thesis, Stiftung Universität Hildesheim]. https://hildok.bsz-bw.de/frontdoor/index/index/docId/858

Wu, T., He, S., Liu, J., Sun, S., Liu, K., Han, Q.-L. & Tang, Y. (2023). A Brief Overview of ChatGPT: The History, Status Quo and Potential Future Development. *IEEE/CAA Journal of Automatica Sinica, 10*(5), 1122–1136. https://doi.org/10.1109/JAS.2023.123618

Wu, X. K., Chen, M., Li, W., Wang, R., Lu, L., Liu, J., Hwang, K., Hao, Y., Pan, Y., Meng, Q., Huang, K., Hu, L., Guizani, M., Chao, N., Fortino, G., Lin, F., Tian, Y., Niyato, D., & Wang, F. Y. (2025). LLM Fine-Tuning: Concepts, Opportunities, and

Challenges. Big Data and Cognitive Computing, *9*(4), Article 87. https://doi.org/10.3390/bdcc9040087

Anschrift der Verfasser:innen

Aenne Knierim
Universität Hildesheim
Institut für Informationswissenschaft und Sprachtechnologie
Universitätsplatz 1
31141 Hildesheim
knierim@uni-hildesheim.de

Theresa Kruse
Universität Hildesheim
Institut für Informationswissenschaft und Sprachtechnologie
Universitätsplatz 1
31141 Hildesheim
theresa.kruse@uni-hildesheim.de