GfA, Sankt Augustin (Hrsg.): Frühjahrskongress 2025, Aachen
Arbeit 5.0: Menschzentrierte Innovationen für die Zukunft der Arbeit

„Hey KI, wie funktioniert ChatGPT?“ – Interaktive Lernerfahrung mittels Web-basierter Anwendung zur Vermittlung der Funktionsweise von LLMs

Lena KÖLMEL[1], Lina KLUY[1], Yu CAO[2], Jun MA[2], Sinan SEN[2], Barbara DEML[1]

[1] *Institut für Arbeitswissenschaft* und Betriebsorganisation, Karlsruher Institut für Technologie,
Engler-Bunte-Ring 4, D-76131 Karlsruhe
[2] Datalyxt GmbH, Ludwig-Erhard-Allee 10, D-76131 Karlsruhe

Kurzfassung: Large Language Models (LLMs) werden zunehmend in privaten, beruflichen und schulischen Kontexten eingesetzt und unterstützen eine heterogene Gruppe an Nutzenden bei sprachbezogenen Aufgaben wie dem Formulieren von E-Mails oder der Erstellung von Gliederungen für Präsentationen. Dabei ist die technische Funktionsweise von LLMs für einen relevanten Teil der Nutzenden unklar, sodass naive Erklärungsansätze und falsche Annahmen über die technischen Hintergründe, Möglichkeiten und Grenzen dieser Technologie existieren. Um diese Wissenslücke zu schließen, wurde eine Webanwendung entwickelt. Ziel dieser Anwendung ist es, durch interaktives Lernen die technische Funktionsweise von LLMs für Nutzende ohne IT-Kenntnisse niederschwellig und anschaulich zu erklären und erfahrbar zu machen. Das Lern- und Interaktionskonzept basiert auf dem Vergleich von je zwei Versionen desselben Sprachmodells (GTP 3.5), deren Konfigurationen sich in einem Konfigurations-Hyperparameter unterscheiden und infolge dessen verschiedene Textausgaben erzeugen. Durch eine bewertende Einordnung der Ausgaben der Modellversionen werden sowohl das Konzept der Parametrisierung als auch technische Grundlagen und die Auswirkung spezifischer Hyperparameter anschaulich erklärt. Übergeordnetes Lernziel ist es, interessierten Nutzenden ein laiengerechtes technisches Verständnis von LLMs zu vermitteln, um eine realistische Einschätzung der Technologie vornehmen zu können und bei der zukünftigen Nutzung unterstützt zu werden. Die Anwendung wurde unter Laborbedingungen mit Fokus auf Usability und Lernerfahrung evaluiert. Um den intendierten Wissenszuwachs und ein gesteigertes Verständnis von LLMs zu quantifizieren wurde ein Leistungstest entwickelt und zu zwei Messzeitpunkten (Prä-/Post-Messung) präsentiert. Die Ergebnisse werden berichtet.

Schlüsselwörter: Large Language Models, KI-basierte Bildung, AI Literacy, Hyperparamter

1. KI-Kompetenz und Fachwissen als Nutzungsvoraussetzung von LLMs

Generative KI-Anwendungen und große Sprachmodelle (engl. Large Language Models, kurz LLMs) haben seit der Veröffentlichung von ChatGPT durch OpenAI im

November 2022 eine rasante Verbreitung erfahren und werden gegenwärtig im privaten, schulischen als auch beruflichen Kontext für verschiedene Aufgaben eingesetzt (Friha et al., 2024; Kalla et al., 2023): LLMs unterstützen Programmierer:innen beim Schreiben von Code (Joel et al., 2024), erzeugen personalisierte Anschreiben im Bewerbungskontext (Zinjad et al., 2024) und übersetzen umfangreiche Texte in Sekunden (Naveen, 2024). Insbesondere der Einsatz im schulischen und akademischen Kontext wird kritisch diskutiert und ist von sowohl Chancen als auch Risiken geprägt (Sharma, & Yadav, 2022; Yu, 2023). Im Zentrum des Diskurses steht die Forderung nach dem Ausbau KI-bezogener Kompetenzen, um einen selbstbestimmten und reflektierten Umgang mit generativen KI-Technologien zu ermöglichen. Die Steuerung durch natürlichsprachliche Eingabeaufforderungen („Prompts“) reduziert zwar die notwendigen technischen Fähigkeiten, allerdings führt dies auch dazu, dass Nutzende die Systeme einerseits bedienen können, andererseits die zugrundeliegende technische Funktionsweise nicht verstehen (Wienrich & Carolus, 2021). Dabei handelt es sich beim Grundverständnis um eine relevante Dimension im Umgang mit KI-Systemen: KI-Literacy-Kompetenzmodelle (zum Beispiel Carolus et al., 2023; Laupichler et al., 2023; Long & Magerko, 2020) weisen abstrahiert stets die folgenden drei Kerndimensionen auf: 1) Grundverständnis über KI erlangen, 2) KI-Systeme kritisch einordnen und 3) KI-Systeme kompetent und sachkundig einsetzen. Mit der neu entwickelten Hey-KI-Anwendung sollen bezogen auf den Anwendungsbereich von LLM-basierten Chatapplikationen jene Kerndimensionen abgebildet werden.

2. Hey-KI-Anwendung: LLMs verstehen durch den Einfluss von Konfigurations-Hyperparametern

Das übergeordnete Lernziel der Hey-KI-Anwendung besteht darin, Nutzenden die grundlegende technische Funktionsweise von LLMs zu vermitteln, um so eine eigenverantwortliche und informierte Nutzung von textbasierten generativen KI-Anwendungen zu ermöglichen. Das Lern- und Interaktionskonzept basiert auf dem Vergleich von je zwei Versionen desselben Sprachmodells (GTP-3.5 Turbo; OpenAI, 2022), deren Konfigurationen sich in einem Hyperparameter unterscheiden und infolge dessen verschiedene Textausgaben erzeugen. Konfigurations-Hyperparameter werden genutzt, um das grundsätzliche Verhalten von LLMs zu beeinflussen, beispielsweise wie deterministisch versus zufällig die erzeugten Ausgaben sind.

In der Hey-KI-Anwendung werden fünf Konfigurations-Hyperparameter abgebildet: 1) *Prompts*: Nutzenden wird erklärt, dass es mehrere Varianten an Prompts gibt. Mit dem Begriff Prompt, wie er im alltäglichen Sprachgebrauch genutzt wird, sind üblicherweise User Prompts gemeint. Sie stellen die herkömmliche Eingabemodalität zur Interaktion mit einem LLM-basierten Chatbot dar. In Abgrenzung dazu werden *System Prompts* eingeführt, die eine übergeordnete grundsätzliche Konfiguration von Modelleigenschaften darstellen. Die definierten Eigenschaften sind meist für die Dauer einer Sitzung beständig und werden für gewöhnlich vom Modell selbst oder Entwickler:innen vergeben. Anhand dieses Beispiels sollen Nutzende lernen, dass Prompts auf verschiedenen Modellebenen relevant sind und LLMs bereits vor der ersten Interaktion mit Nutzenden mit einer Art Persönlichkeit versehen werden können.

Anhand der Konfigurations-Hyperparameter 2) *Temperature* und 3) *Nucleus Sampling* (auch *top_p*) wird erläutert, dass LLMs Textausgaben durch die Berechnung

von Wahrscheinlichkeiten zwischen aufeinanderfolgenden Worten beziehungsweise Tokens (Verarbeitungseinheit von LLMs; Tokens können Wörter, Silben oder einzelne Zeichen sein) Textausgaben erzeugen. Die *Temperature* eines Modells basiert auf einer Skalierung der Wahrscheinlichkeitsverteilung des nächst auszuwählenden Tokens, wohingegen bei *Nucleus Sampling* nachfolgende Tokens basierend auf einem probabilistischen Wahrscheinlichkeitsintervall ausgewählt werden. Durch diese Hyperparameter sollen Nutzende ein Verständnis für die grundlegende auf Wahrscheinlichkeiten basierende Funktionsweise von LLMs erlangen und zudem einen Eindruck davon gewinnen, wie Wahrscheinlichkeitswerte die Wirkung von Texten hinsichtlich Kreativität bzw. Diversität versus Monotonie beeinflussen und so mitunter das Auftreten von Halluzinationen begünstigen.

Mit der 4) *Frequency Penalty* (Frequenzstrafe) und 5) *Presence Penalty* (Anwesenheitsstrafe) werden Parameter zur Beeinflussung von Wort-Wiederholungen erläutert. Das Wiederholen von bereits genutzten Worten wird dabei absolut (*Frequency Penalty*) oder relativierend an der Gesamtlänge einer Textausgabe (*Presence Penalty*) bestraft. Nutzende lernen so eine direkte Möglichkeit zur systemseitigen Beeinflussung von LLMs kennen und erlangen ein Verständnis darüber, dass restriktive Einstellungen den Lesefluss und die Kohärenz eines Textes negativ beeinflussen können.

3. Interaktionskonzept der Anwendung

Neben einem Einführungstext zu Hintergrund und Funktionsweise der Anwendung befindet sich auf der Startseite ein Informationstext (Abbildung 1(1)), der relevante technische Details und Hintergrundinformationen aufbereitet für ein nicht-technisches Publikum beinhaltet, um ein grundlegendes Verständnis über LLMs zu vermitteln. Zentral auf der Startseite ist ein Menü zur Auswahl der fünf Konfigurations-Hyperparameter (Abbildung 1(2); s. Beschreibung der Parameter in Absatz 2 Hey-KI-Anwendung) positioniert.

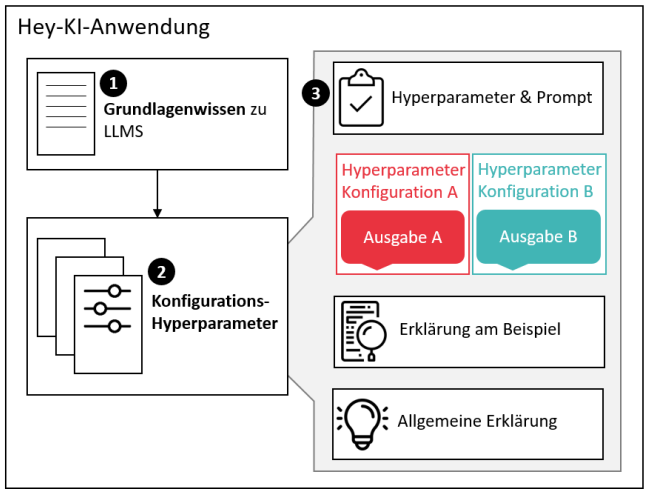

***Abbildung 1**: Visualisierung des Interaktionskonzepts der Hey-KI-Anwendung.*

Durch Auswahl eines Hyperparameters gelangen Nutzende auf eine Unterseite, die für alle Hyperparameter identisch aufgebaut ist (Abbildung 1(3)). Dort wird das Grundkonzept des jeweiligen Hyperparameters beschrieben, gefolgt von einer Prompt-basierten Aufgabenstellung, beispielsweise das Fortführen eines Satzanfangs

oder das Schreiben einer Textzusammenfassung. Durch Button-Klick haben Nutzende die Möglichkeit, für zwei kontrastierende Hyperparameter-Konfigurationen (hoch versus niedrig) die entsprechende LLM-basierte Textausgabe (Ausgabe A und B) zu erzeugen. Durch eine anschließende bewertende Einordnung werden die Funktionsweise des Hyperparameters und mögliche Nutzungsszenarien ausführlich erklärt. Die Erläuterung erfolgt im ersten Schritt unter Rückbezug auf die zu vergleichenden Textausgaben, und im zweiten Schritt auf einer grundlegenderen Abstraktionsebene. Für die in dieser Studie evaluierte Version der Hey-KI-Anwendung wurden die angezeigten LLM-Textausgaben mit den entsprechenden Hyperparameter-Konfigurationen zuvor erzeugt, um für die Durchführung verlässliche und einheitliche Ausgaben gewährleisten zu können.

4. Evaluationsstudie

Die Evaluation wurde als Laboruntersuchung mit Studierenden konzipiert und umfasst drei Erkenntnisziele: 1) Wissenszuwachs, 2) Bewertung der Lernerfahrung und 3) Usability der Anwendung. Zur Quantifizierung des Wissenszuwachses wurde ein Leistungstest bestehend aus 19 wahr/falsch-Fragen entwickelt. Der Test bildet grundsätzliche Kenntnisse über die technische Funktionsweise von LLMs ab, ohne Bezug auf Hyperparameter zu nehmen. Die Fragen wurden messwiederholt vor und nach Interaktion mit der Anwendung präsentiert, sodass sowohl das individuelle Vorwissen der Probanden als auch ein Zuwachs an Fachwissen quantifiziert werden kann. Ergänzend wurden elf Single-Choice Fragen (je drei Distraktoren) entwickelt, die sich auf die Funktionsweise von Konfigurations-Hyperparametern im Kontext von LLMs beziehen. Die Fragen werden nicht-messwiederholt nach der Interaktionsphase präsentiert, da insbesondere bei einer nicht-technischen Zielgruppe davon auszugehen ist, dass kein Vorwissen bezüglich Hyperparameter bei LLMs besteht.

Zur Bewertung der Lernerfahrung wurden sieben Items formuliert. Neben einer subjektiven Einschätzung des Wissenszuwachses wurde die Angemessenheit der Informationsdarbietung und Verständlichkeit erfragt. Um die Nutzungserfahrung der Anwendung zu erfassen, wurde die deutsche Kurzfassung des User Experience Questionnaire genutzt (Laugewitz et al., 2006). Zudem wurden die Proband:innen um eine Schätzung der selbstbestimmten Interaktionsdauer mit der Anwendung (in Minuten) außerhalb des Studienkontexts gebeten.

4.1 Stichprobe

Proband:innen wurden durch lokale offline und online Werbung am Campus rekrutiert. Die Studie dauerte eine Stunde und wurde mit einer Versuchspersonenstunde oder 12€ vergütet. Insgesamt nahmen *N*=37 Personen an der Evaluationsstudie teil, knapp die Hälfte davon war weiblich (*n*=17). Das durchschnittliche Alter betrug 22 Jahre (*M*=22,8 Jahre, *SD*=4,6 Jahre) und *n*=36 Personen gaben an gegenwärtig zu studieren, davon *n*=35 in einem technischen oder naturwissenschaftlichen Studiengang. Durchschnittlich gaben die Proband:innen an LLM-basierte Anwendungen häufig zu nutzen (*M*=4,1, *SD*=1.1; fünfstufige Häufigkeitsskale von *nie* bis *sehr oft*). Das LLM-bezogene Vorwissen wurde durchschnittlich als grundlegend bis mittel eingeschätzt (*M*=2,4, *SD*=0,8; fünfstufige Skala von *keine Kenntnisse* bis *sehr genaue Kenntnisse*).

4.2 Ergebnisse

In Tabelle 1 werden die deskriptiven Statistiken und Korrelationen der Konstrukte berichtet. Ein *t*-Test für gepaarte Stichproben zeigt, dass die Leistung im LLM-Quiz zu Messzeitpunkt zwei nach der Interaktion signifikant höher ist als zu Messzeitpunkt eins vor der Interaktionsphase ($t(36)=-6.91$, $p<.01$, $d=1,14$).

Tabelle 1: *Deskriptive Statistik und Korrelation der Konstrukte (N=37). Dauer wird in Minuten angegeben. Möglicher Maximalwert im LLM-Quiz (pre und post) ist 19. Möglicher Maximalwert im Hyperparameter-Quiz ist 11. Siebenstufige Likertskala zur Erfassung der subjektiven Lernerfahrung. Skala zur Erfassung der Nutzungserfahrung ist -3 bis +3.*

Konstrukt	***M***	***SD***	**a.**	**b.**	**c.**	**d.**	**e.**	**f.**
a. LLM-Quiz pre	11.27	1.93						
b. LLM-Quiz post	13.62	2.16	.49**					
c. Hyperparameter Quiz	9.13	1.78	.18	.47**				
d. Interaktionsdauer real	19.25	3.62	-.20	.04	.25			
e. Interaktionsdauer geschätzt	24.89	23.66	-.04	-.29	-.03	.19		
f. Subjektive Lernerfahrung	6.22	.50	-.23	.00	.27	.17	.-10	
g. Nutzungserfahrung	1.71	.57	-.13	-.03	.31	.37*	-.01	.57**

** Die Korrelation ist auf dem Niveau von .01 (2-seitig) signifikant.
* Die Korrelation ist auf dem Niveau von .05 (2-seitig) signifikant.

5. Diskussion

Die Ergebnisse der Evaluation zeigen, dass die Hey-KI-Anwendung dazu in der Lage ist, LLM-bezogenes Wissen auf geeignete Weise zu vermitteln. Vor allem die inhaltlich sehr spezifischen Fragen zu Konfigurations-Hyperparametern und deren Auswirkungen auf Wahrscheinlichkeitsverteilungen und Ausgabetexte konnten im Durchschnitt mit absoluter Mehrheit (neun von elf Fragen) korrekt beantwortet werden. Im Bereich LLM-bezogenem Grundlagenwissen ergab der Prä-/Post-Vergleich einen signifikanten Wissenszuwachs, allerdings konnten auch zu Messzeitpunkt zwei durchschnittlich nur 13 Fragen von 19 Fragen korrekt beantwortet werden. Mit Hinblick darauf, dass die Ergebnisse aus einer studentischen Stichprobe mit technischem beziehungsweise naturwissenschaftlichem Hintergrund stammen, wird die Notwendigkeit für Bildungsangebote im Bereich LLM verdeutlicht. Eine häufige Nutzung LLM-basierter Anwendungen und Interesse an der Thematik resultieren nicht automatisch in adäquaten Kenntnissen und damit einer Befähigung zur kritischen Reflexion von Textausgaben. Im Kontext von LLM-basierten Anwendungen sind Fähigkeiten dieser Art von besonderer Relevanz, denn Menschen neigen dazu LLM-erzeugten Inhalten selbst in sensiblen Lebensbereichen wie Medizin oder Recht zu stark zu vertrauen (Schneiders et al., 2024; Shekar et al., 2024). In Erklärungsansätzen wird dies darauf zurückgeführt, dass große Sprachmodelle falsche Informationen meist mit absoluter Sicherheit kommunizieren – die verwendete Sprache lässt selten Rückschlüsse auf die zugrundeliegende Souveränität zu (Kim et al., 2024). Dass Bereitschaft zu Weiterbildung in diesem Bereich da ist, zeigt die positive Bewertung der subjektiven Lern- sowie Nutzungserfahrung der Hey-KI-Anwendung, deren Fokus eher auf Informationsvermittlung statt spielerischem

Ausprobieren liegt. Zudem gaben die Proband:innen eine geschätzte Nutzungsdauer von durchschnittlich knapp 25 Minuten außerhalb des Studienkontextes an. In der Gesamtheit betrachtet ebenen diese Ergebnisse den Weg für eine Erweiterung der Hey-KI-Anwendung um ergänzende Bestandteile und eine Erprobung in außeruniversitären Kontexten.

6. Literatur

Carolus, A., Augustin, Y., Markus, A., & Wienrich, C. (2023). Digital interaction literacy model–Conceptualizing competencies for literate interactions with voice-based AI systems. *Computers and Education: Artificial Intelligence*, *4*, 100-114. doi: 10.1016/j.caeai.2022.100114

Friha, O., Ferrag, M. A., Kantarci, B., Cakmak, B., Ozgun, A., & Ghoualmi-Zine, N. (2024). Llm-based edge intelligence: A comprehensive survey on architectures, applications, security and trustworthiness. *IEEE Open Journal of the Communications Society*. *5*, 5799-5856. doi: 10.1109/OJCOMS.2024.3456549

Joel, S., Wu, J. J., & Fard, F. H. (2024). A Survey on LLM-based Code Generation for Low-Resource and Domain-Specific Programming Languages. *arXiv preprint arXiv:2410.03981*.

Kalla, D., Smith, N., Samaah, F., & Kuraku, S. (2023). Study and analysis of chat GPT and its impact on different fields of study. *International journal of innovative science and research technology*, *8*, 827-833.

Kim, S. S., Liao, Q. V., Vorvoreanu, M., Ballard, S., & Vaughan, J. W. (2024). "I'm Not Sure, But...": Examining the Impact of Large Language Models' Uncertainty Expression on User Reliance and Trust. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, *4*, 822-835. doi: 10.1145/3630106.3658941

Laupichler, M. C., Aster, A., Haverkamp, N., & Raupach, T. (2023). Development of the “Scale for the assessment of non-experts’ AI literacy”–An exploratory factor analysis. *Computers in Human Behavior Reports*, *12*, 100338. doi: 10.1016/j.chbr.2023.100338

Naveen, P. (2024). The rise of AI in job applications: a generative adversarial tug-of-war. *AI & SOCIETY*, 1-2. doi: 10.1007/s00146-024-02054-3

OpenAI. (2022). *ChatGPT-3.5 Turbo* [Large language model, API]. https://platform.openai.com

Schneiders, E., Seabrooke, T., Krook, J., Hyde, R., Leesakul, N., Clos, J., & Fischer, J. (2024). Objection Overruled! Lay People can Distinguish Large Language Models from Lawyers, but still Favour Advice from an LLM. *arXiv preprint arXiv:2409.07871*.

Laugwitz, B., Schrepp, M., & Held, T. (2006). Konstruktion eines Fragebogens zur Messung der User Experience von Softwareprodukten. *Mensch und Computer 2006: Mensch und Computer im Strukturwandel*, 125-134.

Sharma, S., & Yadav, R. (2022). Chat GPT–A technological remedy or challenge for education system. *Global Journal of Enterprise Information System*, *14*, 46-51.

Shekar, S., Pataranutaporn, P., Sarabu, C., Cecchi, G. A., & Maes, P. (2024). People over trust AI-generated medical responses and view them to be as valid as doctors, despite low accuracy. *arXiv preprint arXiv:2408.15266*.

Yu, H. (2023). Reflection on whether Chat GPT should be banned by academia from the perspective of education and teaching. *Frontiers in Psychology*, *14*, 1181712. doi: 10.3389/fpsyg.2023.1181712

Zinjad, S. B., Bhattacharjee, A., Bhilegaonkar, A., & Liu, H. (2024). Resumeflow: An llm-facilitated pipeline for personalized resume generation and refinement. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2781-2785. doi: 10.1145/3626772.3657680

Danksagung: Ein besonderer Dank gilt Fabian Bohnacker für die mehrjährige Mitarbeit im Projekt und Léa Riffel für die Unterstützung bei der Datenerhebung. Die Anwendung wurde im Rahmen des Kompetenzzentrums KARL – Künstliche Intelligenz für Arbeit und Lernen in der Region Karlsruhe entwickelt und evaluiert. Dieses Forschungs- und Entwicklungsprojekt wird durch das Bundesministerium für Bildung und Forschung (BMBF) im Programm „Zukunft der Arbeit: Regionale Kompetenzzentren der Arbeitsforschung“ (02L19C250) gefördert und vom Projektträger Karlsruhe (PTKA) betreut. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt bei den Autor:innen.