

Erstellung eines auf Large Language Models basierenden Chatbots für die Sekundarstufe I zum Thema Vierecke und dessen Erprobung im Unterricht

Maximilian Mallweger, BEd

Betreuer

Priv.-Doz. Dipl.-Ing. Dr.techn. Martin Ebner
Benedikt Brünner, BEd MEd

Institute of Human-Centred Computing



WISSEN
TECHNIK
LEIDENSCHAFT

Inhalt

- Forschungsfragen
- Theoretischer Hintergrund
- Entwicklungsprozess
- Implementierung und Evaluation
- Unterrichtsversuch
- Zusammenfassung

Forschungsfragen

- Wie kann man mithilfe von einem Large Language Model einen Chatbot für ein spezifisches Thema im Mathematikunterricht erstellen?
- Wie kann ein Chatbot in den Mathematikunterricht eingebunden werden?
- Wie empfinden Schülerinnen und Schüler das Arbeiten mit einem Chatbot im Mathematikunterricht?

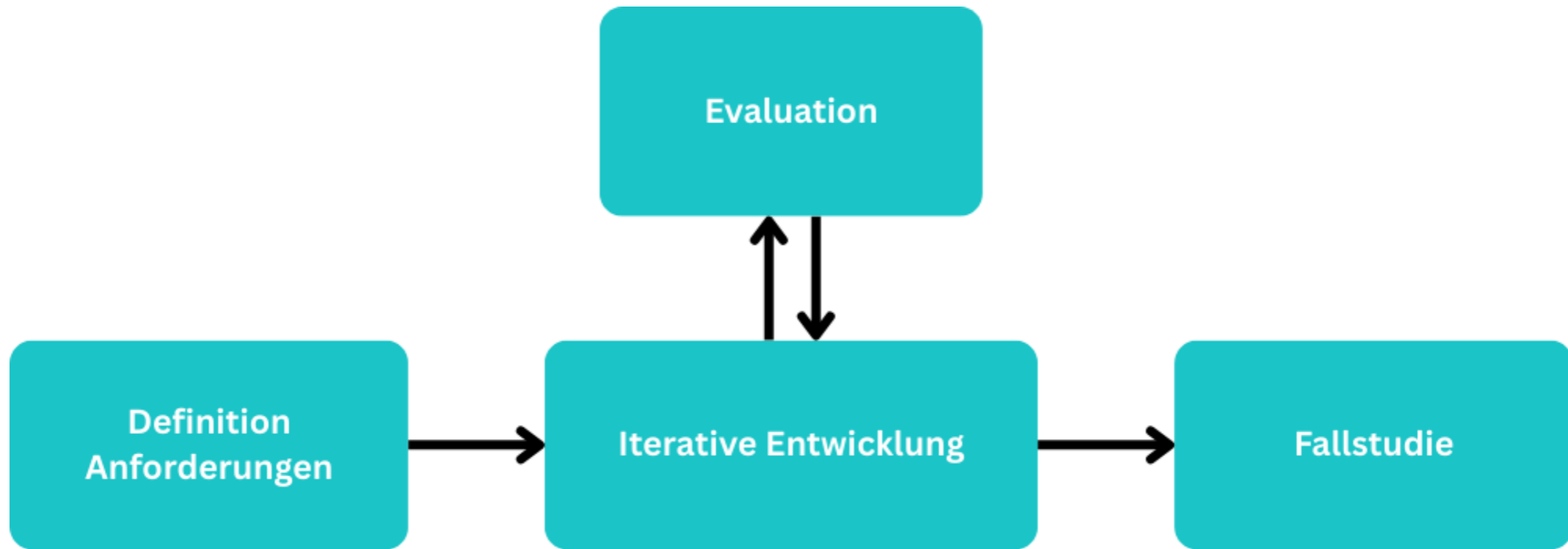
Large Language Models (LLMs) & Fine-Tuning

- Mit sehr großen Textmengen trainierte KI-Systeme
- Können Sprache verstehen und Texte generieren
- Modelle, wie GPT-3 oder GPT-4 basieren auf der Transformer Architektur (Vaswani et al., 2017)
- **Pre-Training:** unüberwachtes Training zur Erlangung allgemeiner sprachlicher Fähigkeiten (Zhao et al., 2023)
- **Fine-Tuning:** Anpassung für spezifische Anwendungsgebiete sowie Unterbindung von Sicherheitsrisiken (Zhao et al., 2023; Dettmers et al., 2023)

Retrieval-Augmented Generation (RAG)

- Einbindung externer Daten zur Verbesserung der Leistung von LLMs (Gao et al., 2023)
- Daten werden in Form von Embeddings, einer Vektorrepräsentation von Daten, in einer Vektordatenbank gespeichert (Malla et al., 2024)
- Im Generierungsprozess werden zur Anfrage passende Informationen abgerufen und dem LLM als zusätzlicher Kontext übergeben (Gao et al., 2023)

Ablauf Entwicklung



Anforderungen an den Chatbot

- Beantwortung fachlicher Fragen
- Lösung von Aufgaben
- Unterstützung bei Aufgaben
- Erstellung von Karteikarten
- Erstellung von Aufgaben

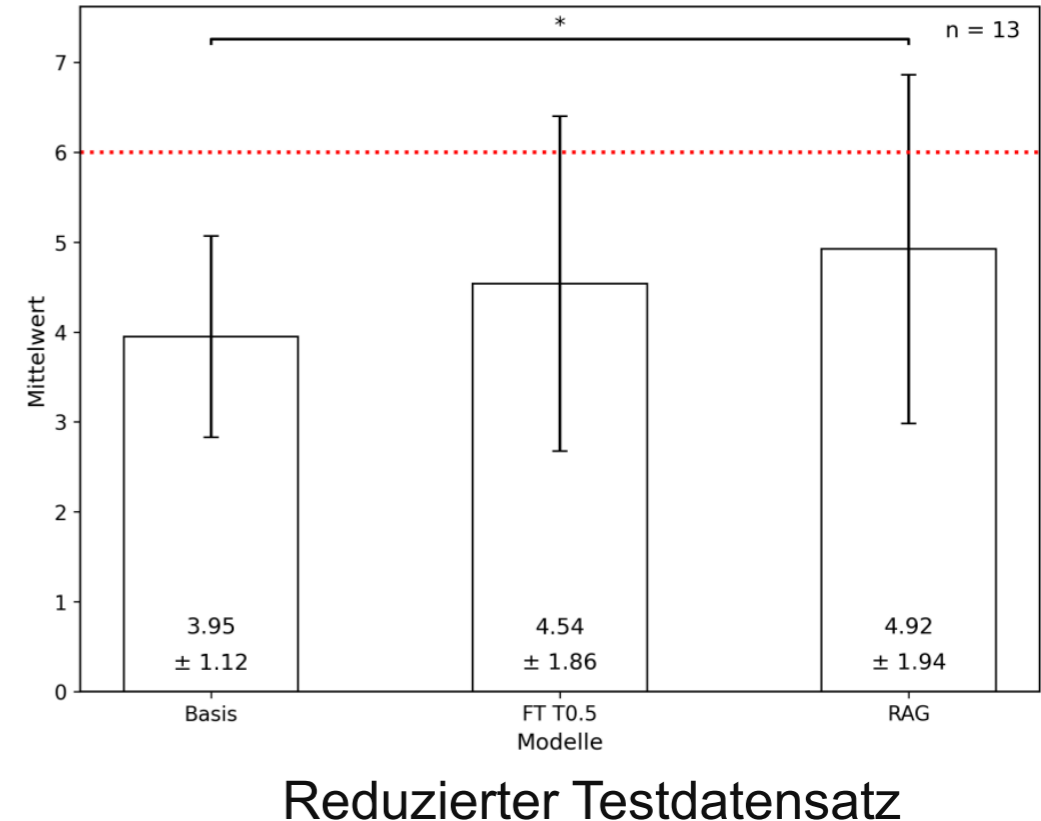
Evaluierungsstrategie

- Leistungsbewertung mit Testdatensätzen (13 / 29 Prompts)
- Zu jedem Prompt werden 3 Antworten generiert
- Antworten werden nach festgelegten Bewertungskriterien ausgewertet
- Mittelwert der 3 Bewertungen wird für die weitere statistische Auswertung herangezogen

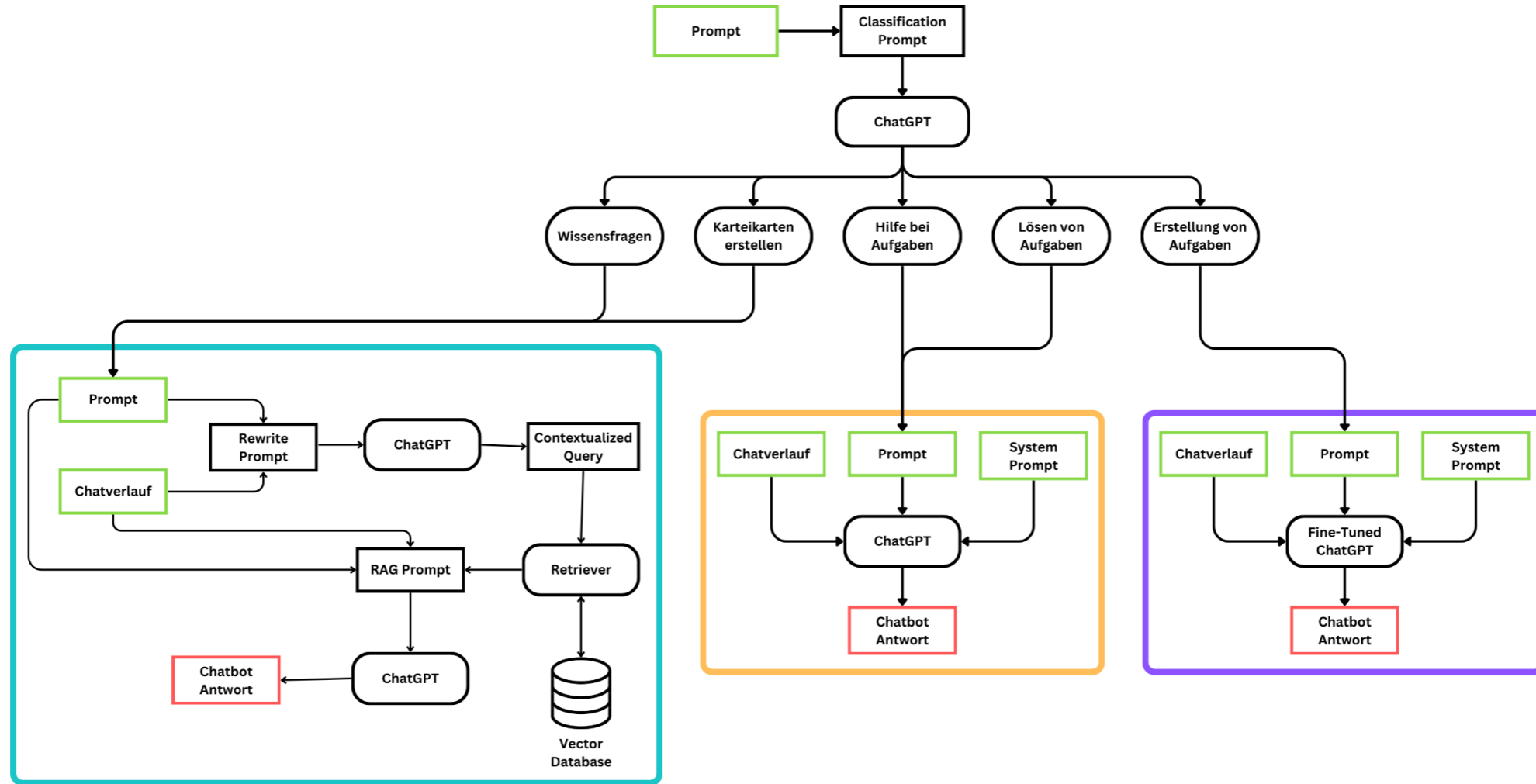
Bewertungskriterium	Erfüllt	Nicht erfüllt
Ausschluss Kriterium	1	0
Korrektheit der Sprache	1	0
Fachliche Korrektheit	1	0
Angemessenheit der Sprache	1	0
Komplexität der Antwort	1	0
Vollständigkeit der Antwort	1	0

Iterative Entwicklung

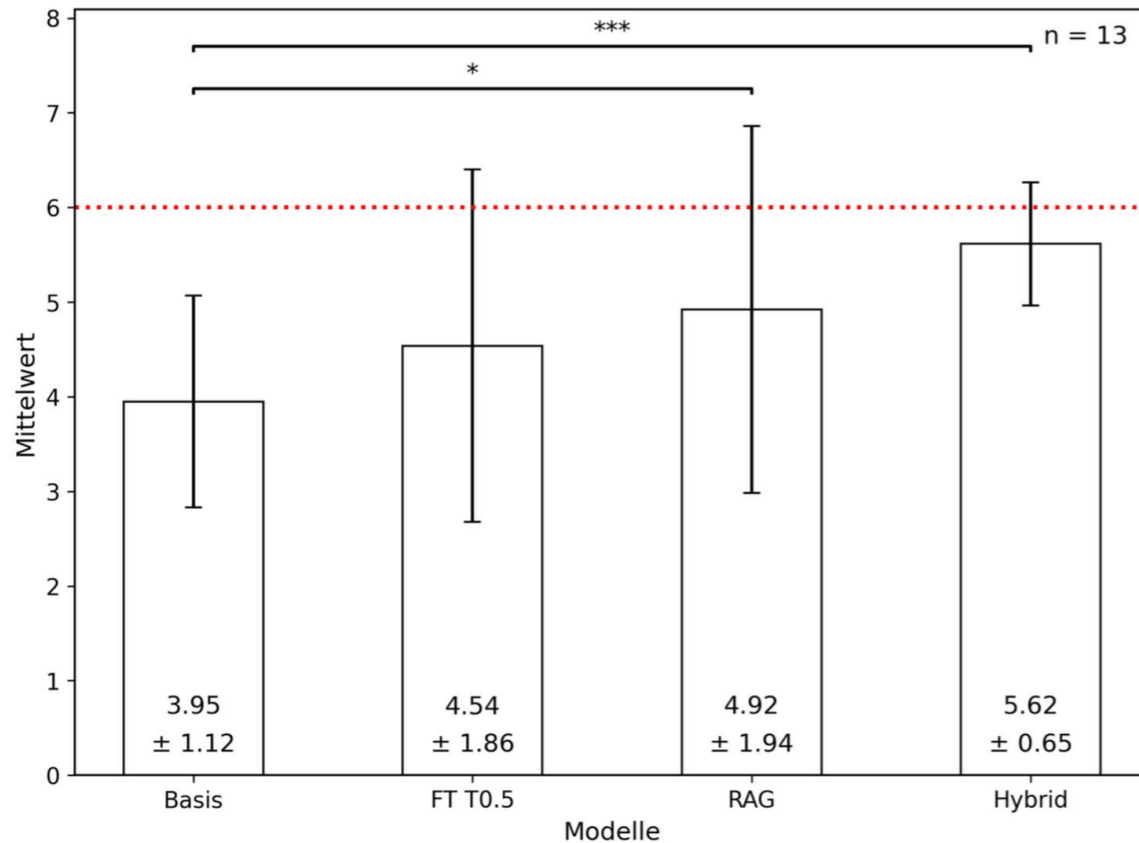
- Basis: gpt-3.5-turbo-1106
- Fine-Tuning
 - Probleme mit strukturierten Daten
 - Gute Leistung bei der Erstellung von Aufgaben
- Retrieval-Augmented Generation (RAG)
 - Gute Leistung bei inhaltlichen Fragestellungen und strukturierten Daten
 - Probleme bei der Aufgabenerstellung



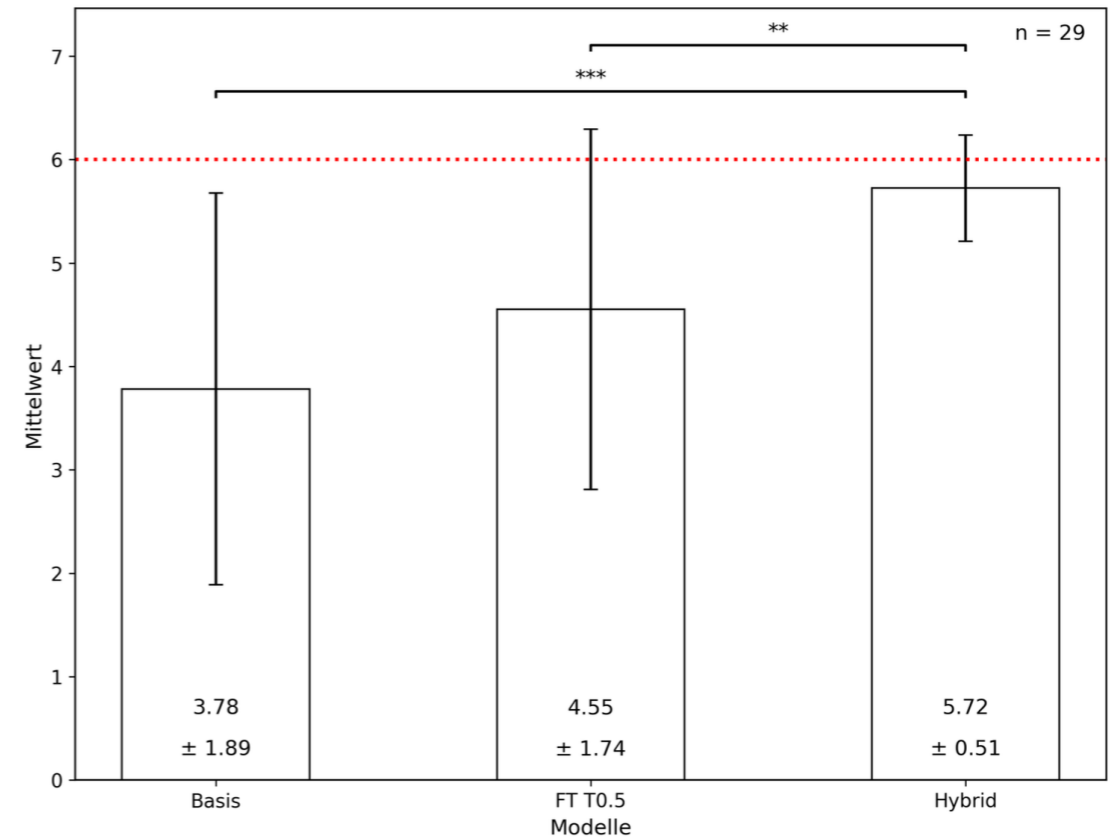
Architektur Hybrid-Implementation



Leistungsvergleich der Implementierungen



Reduzierter Testdatensatz

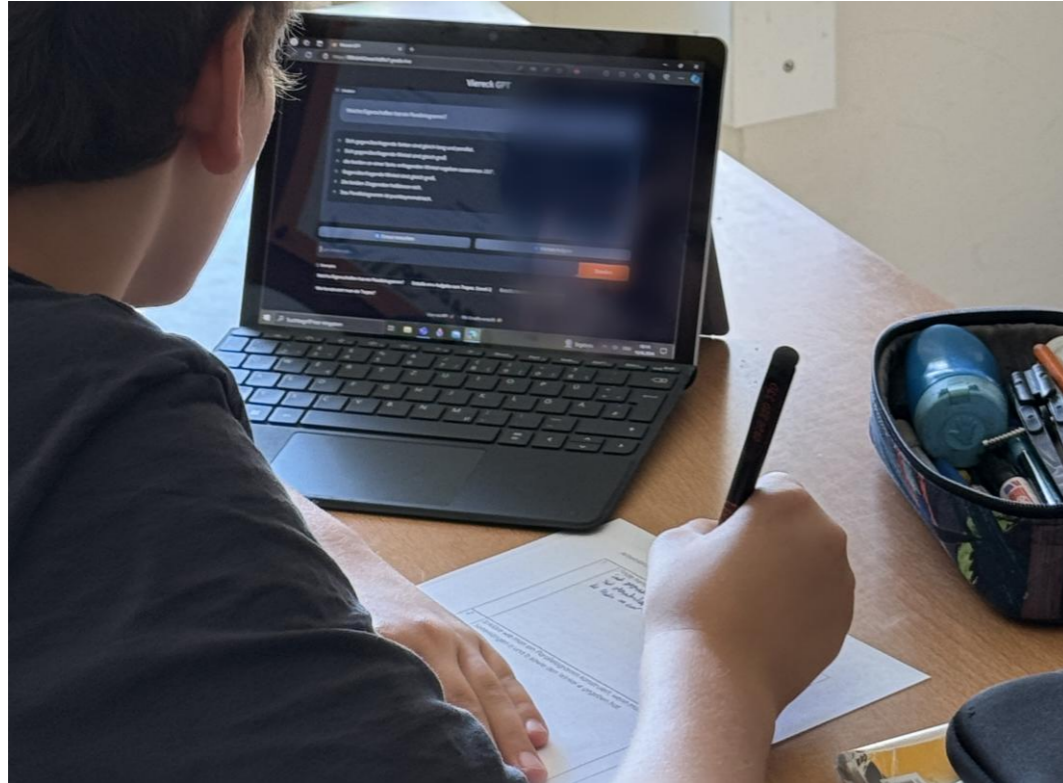


Erweiterter Testdatensatz

Rahmenbedingungen Unterrichtsversuch

- Doppelstunde in einer 2. Klasse Mittelschule (n = 20)
- Ablauf der Einheit:
 - Wissensüberprüfung 1
 - Arbeitsblatt
 - Erstellung Übersichtsblatt
 - Wissensüberprüfung 2
 - Fragebogen
- Unterrichtsbeobachtung

Einsatz des Chatbots im Unterricht

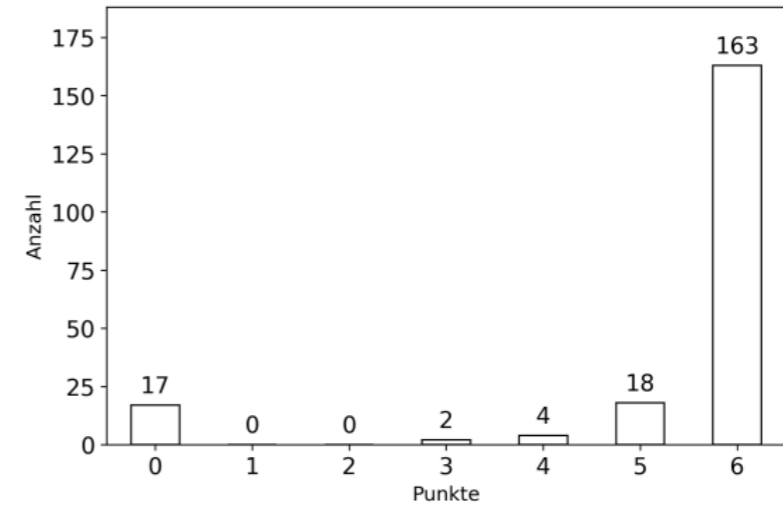
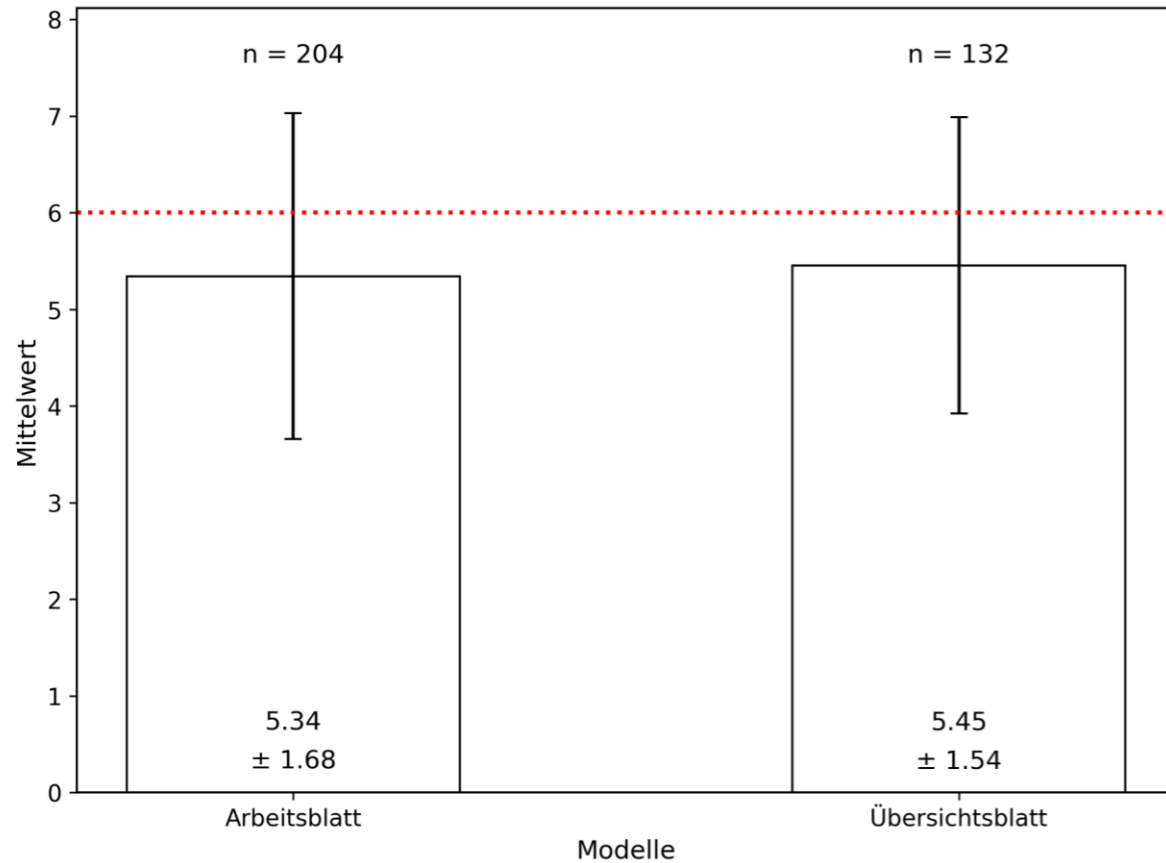


Bearbeitung des Arbeitsblattes

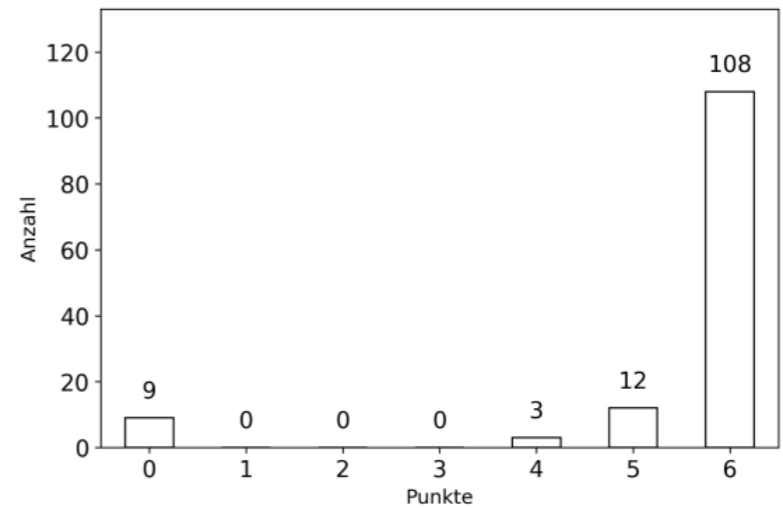


Erstellung der Übersichtsblätter

Leistung des Chatbots



Punkteverteilung Arbeitsblatt



Punkteverteilung Übersichtsblatt

Ergebnisse der Arbeitsphase

- Wissensüberprüfungen:
 - Wissensüberprüfung 1: $\bar{x} = 6,8 / 8,5$ Punkte, zwei negative Leistungen
 - Wissensüberprüfung 2: $\bar{x} = 6,1 / 8,5$ Punkte, zwei negative Leistungen
 - Unterschied ist nicht signifikant
- Arbeitsblatt: $\bar{x} = 11,93 / 15$ Punkte, eine negative Leistung
- Übersichtsblatt: $\bar{x} = 2,25 / 3$ Punkte, vier negative Leistungen

Allgemeine Erkenntnisse und Ergebnisse des Fragebogens

- Problematische Muster konnten identifiziert werden:
 - Abschreiben von Chatbot Antworten
 - Übermäßiges Vertrauen in Chatbot Antworten
 - Unklar formulierte Anfragen
- Feedback aus dem Fragebogen war größtenteils positiv:
 - Verwendung war einfach
 - Chatbot war hilfreich
 - Lernende würden Chatbot gerne öfter, auch zu Hause, verwenden

Zusammenfassung

- Der Chatbot wurde mithilfe einer Kombination aus Fine-Tuning, RAG und einem Standard-LLM implementiert.
- Der Chatbot wurde zur Unterstützung bei Arbeitsaufträgen sowie zur Erstellung von Lernmaterialien in den Unterricht eingebunden.
- Lernende empfinden das Arbeiten mit dem Chatbot sehr positiv und würden ihn gerne als Unterstützungsmaßnahme verwenden.

Literatur

- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 6000–6010.
- 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., ... Wen, J.-R. (2023). A Survey of Large Language Models. *arXiv preprint arXiv:2303.18223*. <https://arxiv.org/abs/2303.18223>
- Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLoRA: Efficient Finetuning of Quantized LLMs. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt & S. Levine (Hrsg.), *Advances in Neural Information Processing Systems* (S. 10088–10115, Bd. 36). Curran Associates, Inc. <https://proceedings.neurips.cc/paperfiles/paper/2023/file/1feb87871436031bdc0f2beaa62a049b-Paper-Conference.pdf>
- Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2023). Retrieval-Augmented Generation for Large Language Models: A Survey. *arXiv preprint arXiv:2312.10997*. <https://arxiv.org/abs/2312.10997>
- Malla, A., Omwenga, M. M., & Kumar Bera, P. (2024). Exploring Image Similarity through Generative Language Models: A Comparative Study of GPT-4 with Word Embeddings and Traditional Approaches. *2024 IEEE International Conference on Electro Information Technology (eIT)*, 275–279. <https://doi.org/10.1109/eIT60633.2024.10609905>


Bedienoberfläche


Viereck GPT

Chatbot

Wie berechnet man den Umfang von einem Parallelogramm?

Der Umfang des Parallelogramms kann mit $U = 2 \cdot a + 2 \cdot b$ berechnet werden, wobei a und b die beiden Seitenlängen des Parallelogramms sind.

 Erneut versuchen

 Nächste Aufgabe

Type a message... Senden

Examples

Welche Eigenschaften hat ein Parallelogramm? Erstelle eine Aufgabe zum Trapez. (Level 1) Erstelle Karteikarten zum Quadrat Wie konstruiert man ein Trapez?