Nils Amelung[1]
Christopher Kralisch[1]
Annette Marohn[1]

[1]Universität Münster

easy.plAnnIng – KI-gestützte Unterrichtsplanung

Ausgangslage und Zielsetzung

Die weitläufige Implementierung von evidenzbasiert entwickeltem Unterrichtsmaterial in die Schulpraxis stellt für die Forschung eine ebenso große Herausforderung dar wie das Übertragen dieser Materialien auf neue Kontexte für die Lehrkräfte. „Für eine Verbreitung der Evidenzorientierung – und damit für den Transfer wissenschaftlichen (Interventions-)Wissens in die Praxis – wäre es […] von großer Bedeutung, Wissen leichter zugänglich zu machen“ (Gräsel, 2019, S. 7). Durch die Veröffentlichung von ChatGPT durch OpenAI im Jahr 2022, wodurch die ersten auf großen Sprachmodellen (LLMs) basierenden Chatbots frei verfügbar wurden, haben sich in diesem Bereich neue, bislang ungenutzte Möglichkeiten ergeben. Viele Lehrkräfte nutzen zwar bereits LLMs in der Planung ihres Unterrichts, allerdings neigen diese Modelle nach wie vor zu Konfabulationen (häufig auch als Halluzinationen bezeichnet (Smith et al., 2023)) und verfügen (vermutlich) über eine geringe fachdidaktische Datengrundlage. Für den Bereich der Chemiedidaktik könnten optimierte Chatbots hier in allen Aspekten Fortschritte ermöglichen.

Aus diesem Grund wird im Rahmen des Projektes easy.plAnnIng der Chatbot AnnI entwickelt, der (Chemie-)Lehrkräfte in der Planung ihres Unterrichts zeiteffizient unterstützen und zusätzlich fachdidaktisches Wissen zu spezifischen Unterrichtskonzepten vermitteln soll. Die Unterrichtsplanungen sollen dabei auf den im Institut für Didaktik der Chemie der Universität Münster entwickelten Konzepten basieren, wobei zunächst ausschließlich mit dem Konzept *choice²learn* (Marohn, 2008, 2021) gestartet wird. Durch das algorithmische Ergänzen von fachdidaktischem und konzeptspezifischem Wissen sowie der Verwendung von dem Stand der Technik entsprechenden Open-Weight Modellen, soll der Chatbot weniger Konfabulationen produzieren und eine gesteigerte didaktische Qualität in der Unterrichtsplanung erzielen.

Ziel ist es, den Chatbot am Ende des Projektes interessierten Lehrkräften kostenfrei zur Verfügung zu stellen, wobei die Ergebnisse der Nutzung möglichst unabhängig von der AI-Literacy der Nutzenden sein sollen.

Methodik

Die Entwicklung des Chatbots erfolgt im Sinne des *Design-Based Research* Ansatzes (McKenney & Reeves, 2012) mit einem dualen Fokus auf Theorie und Praxis in iterativen Mesozyklen. Diese setzen sich jeweils aus unterschiedlichen Kombinationen der Microzyklen Konstruktion/Design, Durchführung/Erprobung und Analyse/Reflexion zusammen. Ziel ist dabei stets die (Weiter-)Entwicklung der Anwendung für die Praxis.

Die Bewertung der Qualität der mit Hilfe von AnnI geplanten Unterrichtsstunden und des Chatbots selbst wird durch eine qualitative Inhaltsanalyse der Chatverläufe, der Unterrichtsplanungen sowie Lehrkräfteinterviews erfolgen. Außerdem ist eine Integration in das Praxissemester (5-monatige Praxisphase im Master in NRW) geplant, um auch die praktische Umsetzbarkeit der Stunden zu überprüfen sowie den Nutzen für Lehrkräfte in frühen Ausbildungsphasen ermitteln zu können. Nachfolgend wird der theoriebasierte

Entwicklungsstand des Chatbots aufgezeigt und das für die praktische Umsetzung gewählte Konzept beschrieben.

Entwicklung des Chatbots AnnI

Der Chatbot AnnI basiert auf einem variablen Open-Weight Modell (aktuell gpt-oss-120b (OpenAI et al., 2025)), das durch UniGPT (Radas et al., 2025) bereitgestellt wird. Als Open-Weight Modelle werden alle LLMs bezeichnet, deren interne Parameter öffentlich zugänglich sind und die dementsprechend von Nutzer*innen heruntergeladen und beliebig verändert werden können. Die Trainingsdaten und der Quellcode werden dabei in der Regel allerdings nicht veröffentlicht. Von Open-Source Modellen kann nur dann gesprochen werden, wenn sowohl das Training inkl. Daten als auch der vollständige Code und die Struktur frei zugänglich gemacht werden (Sapkota et al., 2025).

Zur „Anpassung" eines LLMs waren zwei geläufige Methoden denkbar: *Fine-Tuning* und *Retrieval-Augmented Generation (RAG)* (Balaguer et al., 2024). Unter *Fine-Tuning* wird das nachträgliche Anpassen eines bereits trainierten Modells für einen bestimmten Bereich bezeichnet, indem das Training mit ausschließlich bereichsspezifischen Daten weitergeführt wird. Diese Methode hat allerdings den Nachteil, dass es sich um einen sehr rechenintensiven Prozess handelt und für jede Aktualisierung des zugrunde liegenden LLMs ein erneutes Fine-Tuning dieses Modells benötigt wird (Balaguer et al., 2024). Auf Grund der rasanten Entwicklungen im Bereich der LLMs schien dies im Rahmen dieses Projektes nicht zielführend. Das Konzept des *RAG* hingegen lässt das LLM unverändert, hat allerdings den Nachteil, dass hier für jede Anfrage mehr Tokens (Ansammlung von Zeichen, in die ein LLM alles unterteilt) benötigt werden, was im Regelfall die Antwortzeit und ggf. anfallende Kosten des LLMs erhöht (Balaguer et al., 2024). Die Idee des RAG ist es, indirekt das „Wissen" des LLMs zu erweitern, indem externe, kontextrelevante Daten zu Prompts hinzugefügt werden (Zhao et al., 2024). So werden z. B. beim query-based RAG (siehe Abbildung 1), welches die Basis von AnnI bildet, die Eingaben der Nutzer*innen unbemerkt und automatisiert durch den *Retriever* um möglichst relevante Kontextinformationen angereichert. Der in Abbildung 1 dargestellte Ablauf der Methode wird im Folgenden näher erläutert.

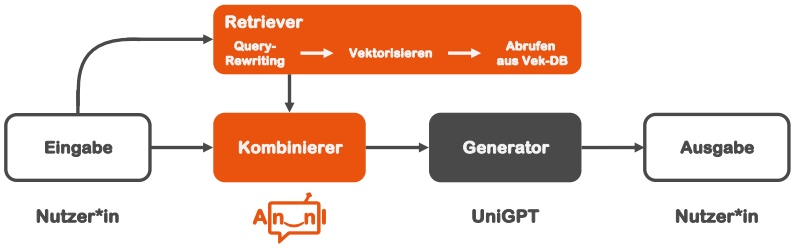

Abb. 1: Ablauf des query-based RAG nach Zhao et al., 2024

Bevor mit der RAG begonnen werden kann, sind allerdings zunächst einige Vorbereitung zu treffen. Um die Datengrundlage zu schaffen, wurden neben Artikeln zum Konzept choice^{2}learn auch zahlreiche Experimentiervorschriften zu Schulexperimenten mit Hilfe von PyMuPDF (Adhikari & Agarwal, 2025) eingelesen und anschließend bestmöglich durch Algorithmen und LLMs bereinigt. Die dadurch erhaltenen Texte wurden anschließend in Abschnitte unterteilt, was in der Literatur als *Chunking* bezeichnet wird (Wang et al., 2024). Hierbei wird versucht, die Informationen möglichst fein zu granulieren, dabei aber trotzdem noch ausreichend Kontext zu erhalten, damit ein LLM die Information sinnvoll verarbeiten

kann. Als letzter Schritt der Vorbereitung werden die erhaltenen Texte von einem sogenannten Embedding-Modell vektorisiert. Diese Modelle wurden darauf trainiert, Wörter, Sätze oder Abschnitte als Vektoren (hier mit 384 Dimensionen) abzubilden. Die daraus erhaltenen Vektoren werden dann in einer speziellen Vektordatenbank gespeichert, auf die der *Retriever* dann zugreifen kann.

Bevor der *Retriever* nun über die Eingabe der Nutzer*innen auf die Datenbank zugreift, wird diese von einem LLM mehrfach umformuliert, um die Chance auf relevante Informationen zu erhöhen. Dieser Prozess wird auch als *Query-Rewriting* bezeichnet (Ma et al., 2023). Das hier erhaltene Ergebnis wird dann vom gleichen Embedding-Modell vektorisiert, sodass nun die ähnlichsten Vektoren durch die Berechnung der Kosinus-Ähnlichkeit aus der Datenbank gesucht werden können, da diese mit einer hohen Wahrscheinlichkeit die sinnvollsten Informationen zu der Frage der Nutzer*in enthalten. Die Texte, die den hier erhaltenen Vektoren entsprechen, werden dann im *Kombinierer* mit der ursprünglichen Eingabe kombiniert. Der dadurch erhaltene Text wird anschließend an das LLM *(Generator)* übergeben, welches die entsprechende für die Nutzer*innen sichtbare Ausgabe erzeugt. Alle vorherigen Schritte laufen, abgesehen von einer leicht erhöhten Antwortzeit, für Nutzer*innen unbemerkt ab.

Durch diese Technik ergibt sich für Lehrkräfte eine Möglichkeit, ihre Unterrichtsplanungen an evidenzbasiert entwickelten Konzepten zu orientieren, die der Chatbot literaturbasiert präsentieren und erklären kann. AnnI sollte dabei keine Quellen konfabulieren, da der Verweis auf die Quellen algorithmisch und nicht durch das LLM probabilistisch erfolgt. Damit soll neben einer Steigerung der Unterrichtsqualität sowie einer Entlastung der Lehrkraft auch eine Überprüfung der KI-generierten Antworten erleichtert werden. Diese sind weiterhin stets kritisch zu hinterfragen, da eine Verarbeitung der Literaturausschnitte durch das LLM erfolgt und die Ausschnitte nicht wörtlich wiedergegeben werden.

Abschließend werden weitere im Rahmen dieses Projekts verwendete Modelle und Softwarelösungen kurz erwähnt. Zur Implementierung wird ergänzend zu den üblichen SQL-Datenbanken auf Vektordatenbanken zurückgegriffen, die über eine lokal gehostete Instanz von Weaviate (Dilocker et al., 2016/2025) betrieben werden. Als Embedding Modell wird die mehrsprachige Version des MiniLM-L12-v2 verwendet (Reimers & Gurevych, 2019; *sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2 · Hugging Face*, o. J.), die ebenfalls lokal gehostet wird. Es werden also ausschließlich Open-Source Software sowie lokal betriebene Open-Weight Modelle verwendet, wodurch die laufenden Kosten minimiert und der Datenschutz maximiert werden.

Ausblick

Im weiteren Verlauf des Projektes wird der Chatbot zunächst über weitere Systemprompts und die Implementierung weiterer Funktionen zur Unterstützung bei der Planung von Chemieunterricht nach dem Konzept choice^{2}learn befähigt. Anhand einer Beurteilung der daraus resultierenden Planungen sowie Befragungen der mit AnnI arbeitenden (angehenden) Lehrkräfte wird dann eine iterative Weiterentwicklung erfolgen. Abschließend soll AnnI durch eine Erweiterung der Wissensbasis, also der Vektordatenbanken, auch bei der Unterrichtsplanung im Rahmen weiterer Konzepte unterstützen können und final als kostenfreie Webanwendung zugänglich gemacht werden. An dieser Stelle wird betont, dass der entwickelte Chatbot keinesfalls das Ziel verfolgt, die Lehrkraft im Planungsprozess zu ersetzen, sondern vielmehr als entlastende und kompetente Assistenz verstanden werden soll,

die ergänzend zu Lehrkräftefortbildungen einen niedrigschwelligeren Zugang zu evidenzbasiert entwickelten Materialien und Konzepten ermöglicht.

Literatur

Adhikari, N. S. & Agarwal, S. (2025). A comparative study of PDF parsing tools across diverse document categories (No. arXiv:2410.09871). arXiv. https://doi.org/10.48550/arXiv.2410.09871

Balaguer, A., Benara, V., Cunha, R. L. de F., Filho, R. de M. E., Hendry, T., Holstein, D., Marsman, J., Mecklenburg, N., Malvar, S., Nunes, L. O., Padilha, R., Sharp, M., Silva, B., Sharma, S., Aski, V. & Chandra, R. (2024). RAG vs Fine-tuning: Pipelines, Tradeoffs, and a Case Study on Agriculture (No. arXiv:2401.08406). arXiv. https://doi.org/10.48550/arXiv.2401.08406

Dilocker, E., van Luijt, B., Voorbach, B., Hasan, M. S., Rodriguez, A., Kulawiak, D. A., Antas, M. & Duckworth, P. (2025). *Weaviate [Go]*. https://github.com/weaviate/weaviate (Ursprünglich erschienen 2016)

Gräsel, C. (2019). Transfer von Forschungsergebnissen in die Praxis. In C. Donie, F. Foerster, M. Obermayr, A. Deckwerth, G. Kammermeyer, G. Lenske, M. Leuchter, & A. Wildemann (Hrsg.), *Grundschulpädagogik zwischen Wissenschaft und Transfer* (S. 2–11). Springer Fachmedien Wiesbaden.

Ma, X., Gong, Y., He, P., Zhao, H. & Duan, N. (2023). Query rewriting for retrieval-augmented large language models (No. arXiv:2305.14283). arXiv. https://doi.org/10.48550/arXiv.2305.14283

Marohn, A. (2008). „Choice2learn"—Eine Konzeption zur Exploration und Veränderung von Lernervorstellungen im naturwissenschaftlichen Unterricht. *ZfDN, 14*, 57–83.

Marohn, A. (2021). Umgang mit Vielfalt: Das Unterrichtskonzept choice2learn. *MNU-Journal, 1*, 85–92.

McKenney, S. & Reeves, T. C. (2012). *Conducting educational design research*. Routledge.

OpenAI, Agarwal, S., Ahmad, L., Ai, J., Altman, S., Applebaum, A., Arbus, E., Arora, R. K., Bai, Y., Baker, B., Bao, H., Barak, B., Bennett, A., Bertao, T., Brett, N., Brevdo, E., Brockman, G., Bubeck, S., Chang, C., … Zhao, S. (2025). Gpt-oss-120b & gpt-oss-20b Model Card (No. arXiv:2508.10925). arXiv. https://doi.org/10.48550/arXiv.2508.10925

Radas, J., Risse, B. & Vogl, R. (2025). Building UniGPT: A customizable on-premise LLM-solution for universities. 108–198. https://doi.org/10.29007/jv1l

Reimers, N. & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using siamese BERT-networks (No. arXiv:1908.10084). arXiv. https://doi.org/10.48550/arXiv.1908.10084

Sapkota, R., Raza, S. & Karkee, M. (2025). Comprehensive analysis of transparency and accessibility of ChatGPT, DeepSeek, and other SoTA large language models. https://doi.org/10.48550/ARXIV.2502.18505

Sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2 · Hugging Face. (o. J.). Abgerufen 23. September 2025, von https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2

Smith, A. L., Greaves, F. & Panch, T. (2023). Hallucination or confabulation? Neuroanatomy as metaphor in large language models. *PLOS Digital Health, 2*(11), e0000388. https://doi.org/10.1371/journal.pdig.0000388

Wang, X., Wang, Z., Gao, X., Zhang, F., Wu, Y., Xu, Z., Shi, T., Wang, Z., Li, S., Qian, Q., Yin, R., Lv, C., Zheng, X. & Huang, X. (2024). Searching for best practices in retrieval-augmented generation. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Hrsg.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (S. 17716–17736). Association for Computational Linguistics.

Zhao, P., Zhang, H., Yu, Q., Wang, Z., Geng, Y., Fu, F., Yang, L., Zhang, W., Jiang, J. & Cui, B. (2024). Retrieval-augmented generation for AI-generated content: A survey (No. arXiv:2402.19473). arXiv. https://doi.org/10.48550/arXiv.2402.19473