Exploring Personalized Learning Support through Retrieval Augmented Generation: A Feasibility Study

Petar Mladenov and **Luis Pinheiro** and **Dino Pelesevic** and **Jasmin Heierli**
Zurich University of Applied Sciences, Winterthur, Switzerland
{mladepet,pinhelui,pelesdin}@students.zhaw.ch, heej@zhaw.ch

Abstract

This paper presents a feasibility study on using language models (LMs) in education to generate and evaluate tasks based on Bloom's taxonomy. We developed a prototype combining retrieval-augmented generation and prompt engineering to assess students' cognitive levels. Initial expert feedback highlights our approach's potential, but it emphasizes the need for broader empirical validation. The study provides a foundation for future research into LMs in personalized education and underscores the importance of real-world testing.

1 Introduction

Knowledge dissemination is essential for advancing societies, yet traditional teaching struggles with teacher shortages and diverse student needs (BFS, 2022). The limited availability of support per student is further compounded by the diverse learning abilities and requirements of individual students.

Consequently, digitalization has produced many e-learning aids[1]. These tools typically offer interactive tasks, explanations, and sample solutions, but often fail to provide personalized learning support that accurately assesses a student's understanding level. Current e-learning aids, despite offering task-based support, do not fully capture the nuances of a student's comprehension, limiting the effectiveness of personalized learning[2].

An effective, personalized learning support requires a sound and valid assessment of the student's state of understanding.

This study explores the feasibility of using language models (LM) with with Retrieval Augmented Generation (RAG) to enhance e-learning personalization via Bloom's Taxonomy (Krathwohl, 2002). We aim to assess student comprehension more accurately by creating and evaluating tasks across two different comprehension levels, using Bloom's taxonomy not just as a framework, but as a personalization tool.

The remainder of this paper is organized as follows: Sect. 2 outlines how we aim to assess a student's level of understanding and the requirements. Sect. 3 details the implementation of our approach, demonstrating its feasiblity. Finally, Sect. 4 discusses the results and potential future work.

2 Requirements

This section details the system's requirements for personalizing the system to assess student understanding of learning materials. We start by describing the desired inputs and outcomes of the system, followed by a list of key requirements.

The objective is to analyze content, identify relevant excerpts, like the one in Figure 1, and generate tasks across two Bloom levels. Using selected


I Beginn der Unternehmenstätigkeit

Das menschliche Leben ist durch **Spannungsfelder** geprägt: Auf der einen Seite stehen unsere vielfältigen und **umfangreichen Bedürfnisse**, auf der anderen Seite die **knappen Mittel**, um diese Bedürfnisse zu befriedigen. Wie jeder Mensch muss auch ein Unterneh-

Figure 1: Example excerpt from an educational resource

excerpts, the LM prompts tasks testing specific Bloom's levels. For instance, one task might assess Bloom's level 1 (Remember), querying a student's ability to recall facts:

```
Wer ist für die strategische Planung in
einem Unternehmen verantwortlich?
A) ...
```

Another task might evaluate Bloom's level 3 (Apply), examining the ability to apply knowledge

[1]For example: https://www.aufgabenfuchs.de, https://www.sofatutor.ch, https://ilearn.ch, http://aufgaben-online.ch, https://lernen-mit-spass.ch, https://www.schlaukopf.ch, https://simpleclub.com, https://brilliant.org

[2]These services build upon the groundwork laid by earlier work in digital education, which we cannot discuss due to space constraints

in new scenarios.

```
Szenario: Sie sind ein Mitglied des
Führungsteams der Schweizer Firma
"Alpine Goods AG", die qualitativ
hochwertige Outdoor-Bekleidung und
Ausrüstung produziert und vertreibt...
Frage: Unter Verwendung der
SWOT-Analyse (Strengths, Weaknesses,
Opportunities, Threats) analysieren Sie
die aktuelle Situation von "Alpine
Goods AG". Was ist eine der wichtigsten
Schwächen, die das Unternehmen beachten
und angehen sollte, und welche
strategischen Entscheidungen könnten
getroffen werden, um diese Schwäche zu
adressieren?
A) ...
```

The generated tasks assess recall of facts (Level 1) and application of knowledge (Level 3). The system corrects responses to determine a student's level of understanding and whether or not they are reaching certain Bloom's levels.

To investigate the feasibility, an implementation must meet these requirements:

- The system shall be able to process different educational resources and use them for assessing the student's learning level for a specific subject.
- The system shall select suitable study material for task generation using a simple keyword.
- The system shall be able to generate tasks that assess different Bloom levels.
- The system shall be able to generate response options for single- or multiple-choice tasks.
- The system shall be able to evaluate the correctness of student responses.

Having defined these requirements, we further restrict our focus to German texts in PDF format and aim to use RAG for these purposes. The next section explains our implementation approach to meet these requirements.

3 Implementation

This section outlines the implementation of a lightweight prototype, designed to be feasible on standard computers and evaluated for potential by an expert. Figure 2 illustrates the system's architecture, featuring four main components.

Figure 2: System architecture.

3.1 Text to vectors

Addressing the need to process diverse educational resources, the prototype uses Chroma[3] vector store for storing text as embedding vectors extracted from PDFs using PyPDF 2[4]. For the prototype, we did not evaluate the performance of different PDF readers. PyPDF is open source and allows the retrieval of text from PDFs.Texts are chunked to align with the embedding model (paraphrase-multilingual-MiniLM-L12-v2[5]), optimized for semantic search in German. The chunking and embedding processes, though practical, lack extensive technical validation at this stage.

As there is currently not much research around chunking for RAG available (Yepes et al., 2024), we decided to take a practical approach respecting the input length of the embedding model we planned to use. To achieve the desired length, the retrieved texts are chunked with Langchain's RecursiveCharacterTextSplitter[6] that splits the text into chunks of 1,000 characters. This roughly correspond to 128 tokens, which is the input size of the embedding model we used.

The chunks are then embedded using the paraphrase-multilingual-MiniLM-L12-v2[7] embedding model that works for short paragraphs of German text (Reimers and Gurevych, 2019). It is small enough to run on a variety of hardware and has been developed for semantic search, specifically. The embedded chunks are then indexed and stored in a Chroma vector store.

Having established the method for converting text to vectors, we next focus on the question generation components for assessing students' understanding levels.

[3] https://docs.trychroma.com/
[4] https://pythonhosted.org/PyPDF2/
[5] https://huggingface.co/cross-encoder/msmarco-MiniLM-L6-en-de-v1
[6] https://python.langchain.com/docs/modules/data_connection/document_transformers/recursive_text_splitter
[7] https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2

3.2 Vector retrieval

To meet the second requirement of selecting study material by keyword we utilized Chroma's vector similarity search.

We enhanced traditional methods by using GPT-4[8] to expand initial terms into five linguistically sound questions, rather than traditionally adding more terms (Carpineto and Romano, 2012).

For example, for an original search term ´Unternehmenstätigkeit´, we used the prompt

```
Du bist Lehrperson für
Betriebswirtschaft. Du bekommst Fragen
zum Lehrmittel über Betriebswirtschaft.
Schlage bis zu 5 zusätzliche verwandte
Fragen vor, um dem Benutzer zu helfen,
die Antworten auf seine Frage zu
finden...
```

yielding five additional questions such as

```
Was sind die verschiedenen Arten von {
    given_topic}?
```

These were validated for relevance and appropriateness by an educational expert.

To further ensure the validity of the expanded queries and their retrieved documents, we applied PCA projection. This visualizes the original and expanded queries alongside their results in a 2-D vector space, confirming their alignment and relevance.

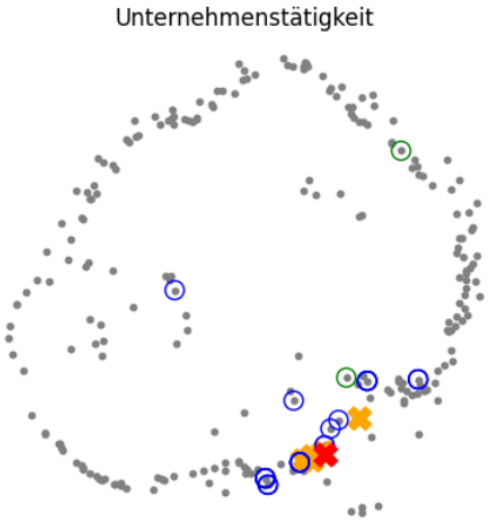

Figure 3: PCA projection of the original (red) and expanded queries (orange) and their respective results (results for original query = green, results for expanded = blue).

The original search term and the five expanded queries then guide the retrieval of the top 5 relevant chunks from the vector store with Euclidian distance. To avoid information overload and redundancies, we refine our selection to the five most relevant chunks. This is achieved with the msmarco-MiniLM-L6-en-de-v1 cross-encoder model, which evaluates chunk relevance regardless of origin.

With these refined chunks, we proceed to the question generation as detailed in the next section.

3.3 Task generation

We opted for RAG combined with prompt engineering rather than fine-tuning a model to enable the dynamic generation of tasks for different subjects and learning materials. Fine-tuning would potentially require separate models for different subjects, which increases resource requirements. Furthermore, the repeated fine-tuning of a single model with new materials could lead to inconsistencies, potentially negatively affecting model performance.

Bloom's taxonomy categorizes educational objectives into six cognitive levels, ranging from simple recall to complex evaluation and creation. As our goal is to evaluate the feasibility of personalization in e-learning, Bloom's taxonomy is a suitable framework to personalise tasks according to students' cognitive levels. For the prototype we focused on level 1 (Remember) and level 3 (Apply) tasks, as this allows us to evaluate whether the LM can generate tasks that are personalized to the students' cognitive levels.

The aim of the task generation is to create tasks aligning with different Bloom's levels. In the scope of this investigation, we focused on single-choice tasks. Consequently, the task generation must also include the generation of response options. Following the idea of RAG, we combine prompt engineering with text injected into prompts.

Our first attempt to obtain tasks assessing different Bloom's levels consisted of specifying the Bloom's level as part of the prompt. While the first results indicated that GPT-4 has inherent knowledge about Bloom's taxonomy, the generated questions for level 1 were overly simplistic, and those for level 3 lacked comprehensive scenario descriptions — as pointed out by our educational expert.

To address this, we crafted prompts containing keywords and phrases indicative of the desired Bloom level (Krathwohl, 2002). For Bloom level 1, our prompts, such as the one shown next, include verbs such as ´list´, ´name´, and ´describe´, en-

[8] https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo

suring simplicity and clarity in the generated tasks.

```
... Deine Aufgabe ist es, eine
Prüfungsfrage auf
Bloom-Niveau 1 zu formulieren, die sich
auf allgemeines Wissen bezieht, das im
Unterricht behandelt wurde. Verwende
einfache Schlüsselwörter
und Verben wie 'sammeln', 'erzählen',
'benennen', 'erinnere', 'was', 'wann',
'wer', 'liste auf', 'zeige', 'gib an'
und ähnliches...
```

The previously retrieved chunks are then injected at the end of such a prompt, followed by an instruction to generate a task solely on this information.

The prompt for the Bloom level 3 task generation also includes instructions to generate a fictious scenario mentioning theories when applicable to solve the task. While we obtained acceptable results instructing the LM to generate questions and response options for single-choice tasks at once, our educational expert concluded that these response options lacked discriminative power.

We obtained better results by prompting the LM separetely for task generation and the generation of answer options. The task generated with the first prompt was included in second prompt. This adds the possibility of specifying separate requirements such as regarding the discriminatory power.

This two-step prompting approach not only enhances the discriminative power of response options but also increases control in tailoring them to specific task requirements, thereby ensuring our prototype's efficacy in engaging students at their personal cognitive level.

3.4 Task correction

The LM evaluated student responses against provided materials, aiming to mimic a teacher's assessment process, as shown in this structured prompt. Accuracy in practical educational settings remains to be tested.

```
Du bist Lehrperson für das Fach
Betriebswirtschaft an einer schweizer
Sekundarschule. Deine Aufgabe ist es,
die Antworten deiner Schüler auf eine
Single-Choice-Frage zu bewerten. Die
Frage lautet: \{question_3\}. Antworte
nur mit 'richtig' oder 'falsch'. Die
Informationen zum Thema sind:
\{retrieved_documents\}. Antwort des
Schülers: \{user_answer\}
```

Note that the expressions in curly brackets are place holders for the full question including answer options, the retrieved documents and the response the student selects.

4 Conclusion

This feasibility study demonstrates the potential of LMs for personalized aducation, specifically through developing a prototype that leverages LMs for task generation and response evaluation, targeting the personalized assessment of students' understanding at different Bloom levels.

In alignment with the identified requirements (Section 2), our retrieval-augmented generation approach, which incorporates a Chroma vector store, effectively processes a range of educational resources, selects relevant content, and seamlessly integrates it into LM-generated prompts. The appropriateness of our tasks, customized for specific Bloom levels, was confirmed by an educational expert. Although our current focus has been on two levels of Bloom's taxonomy, the versatility of our approach suggests potential applicability to other means of personalization.

While we focused on two Bloom levels, our results indicate potential for personalization across other dimensions, such as learning styles, interests, and cognitive abilities. Future work will include systematic validation of our techniques for query expansion and cross-coder reranking. Empirical testing with students will also be crucial to evaluate the accuracy of our system in identifying their respective Bloom's comprehension levels. In conclusion, our study not only confirms the feasibility of using LMs in educational settings but also opens avenues for future research, particularly in enhancing personalized learning experiences and understanding student cognitive levels.

References

BFS. 2022. *Szenarien 2022-2031 für die Lehrkräfte der obligatorischen Schule*. 22806575. Bundesamt für Statistik (BFS), Neuchâtel.

Claudio Carpineto and Giovanni Romano. 2012. A survey of automatic query expansion in information retrieval. *ACM Comput. Surv.*, 44(1).

David R Krathwohl. 2002. A revision of bloom's taxonomy: An overview. *Theory into practice*, 41(4):212–218.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *CoRR*, abs/1908.10084.

Antonio Jimeno Yepes, Yao You, Jan Milczek, Sebastian Laverde, and Renyu Li. 2024. Financial report chunking for effective retrieval augmented generation.