AI-assisted student grading system in geotechnics

Enrico Soranzo[1)]

[1)] Institute of Geotechnical Engineering, Department of Landscape, Water and Infrastructure, BOKU University, Austria

Introduction

This study investigates the use of automated grading in geotechnics by leveraging large language models (LLMs). Key objectives include:

- generating educational content by fine-tuning LLMs with geotechnics textbooks,
- developing ML grading systems validated on synthetic and real student data and
- enhance student learning through self-assessment tools.

Methodology

Using 26 geotechnics textbooks data were scraped and processed via a Retrieval-Augmented Generation (RAG) approach (Lewis et al. 2020). 30 geotechnics-related questions were generated and reviewed.

Corresponding correct answers were generated as references. ChatGPT-4.0 was used to create synthetic "student" answers. These answers varied in quality. A total of 100 synthetic answers for each of the questions were generated (3000 answers in total). By training the open-source DistilBERT LLM (Sanh et al. 2019) on the data, the process achieves independence from third-party applications. Synthetic "student" answers were graded on a scale of 1 ("very good") to 5 ("fail") (Table 1).

An interactive tool was developed using the Streamlit framework to evaluate student responses to geotechnical questions, accessible at https://soranz84-carlo.hf.space (Figure 1). The tool dynamically retrieves questions and correct answers from a Google Sheet. When a user submits an answer, the system calculates the cosine similarity between the input and the correct answer. Based on predefined thresholds, responses are graded according to the Austrian grading system. User responses are logged in a separate Google Sheet for further analysis.

Figure 2 illustrates the tool in action, showing a student response graded as "good" with a cosine similarity score of 0.76.

Table 1. Question and answers with 5 levels of accuracy and grades.

Question	"Student" answers	Grade
Explain the difference between cohesive soils and cohesionless soils. Provide examples of each and discuss their engineering significance.	Cohesive soils are fine-grained and retain water, making them suitable for applications like clay liners, while cohesionless soils are coarse-grained and allow water to drain freely	1
	Cohesive soils are sticky and retain water, while cohesionless soils are loose and drain well	2
	Cohesive soils are impermeable, cohesionless soils are highly permeable.	3
	Cohesive soils are sticky and cohesionless soils are loose.	4
	I don't know.	5

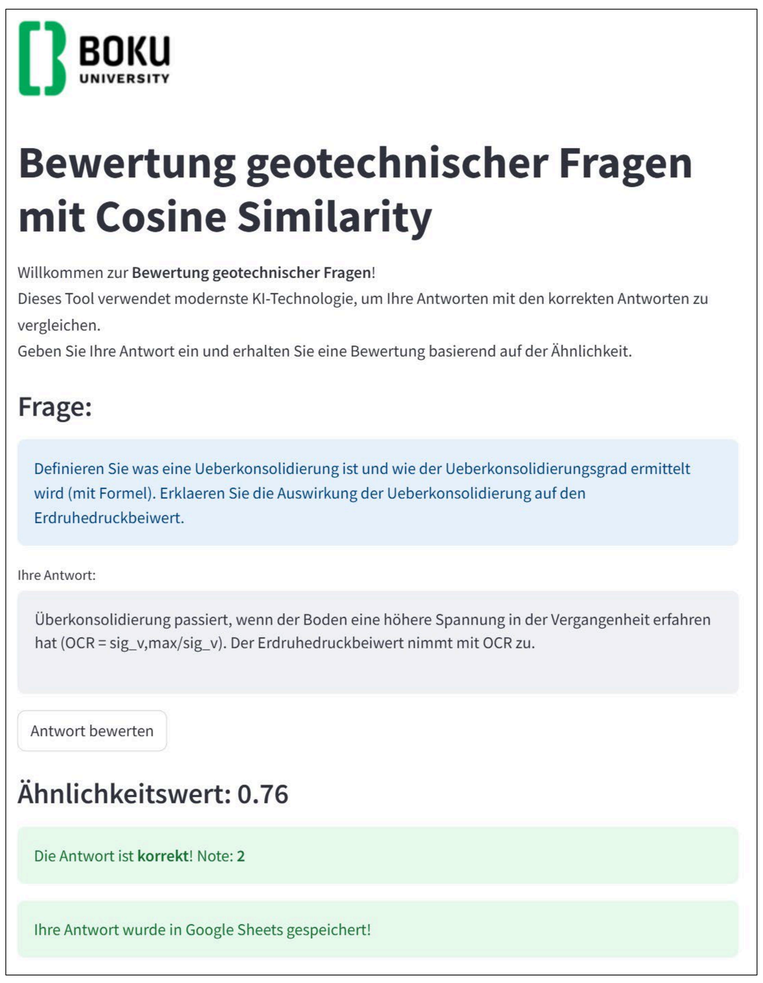

Figure 1. Screenshot of the web application for dynamic feedback.

Results

Fine-tuning the LLM on ground truth labels yielded strong performance. The confusion matrix for the test set (Figure 2) highlights this: grade 2 achieved 186 true positives with only 3 errors, and grade 5 achieved 22 true positives with no errors. The metrics further underscore the model's strong performance. On the testing dataset, both accuracy and F_1-score were 97.5%. A small gap between training and testing performance indicates good generalization.

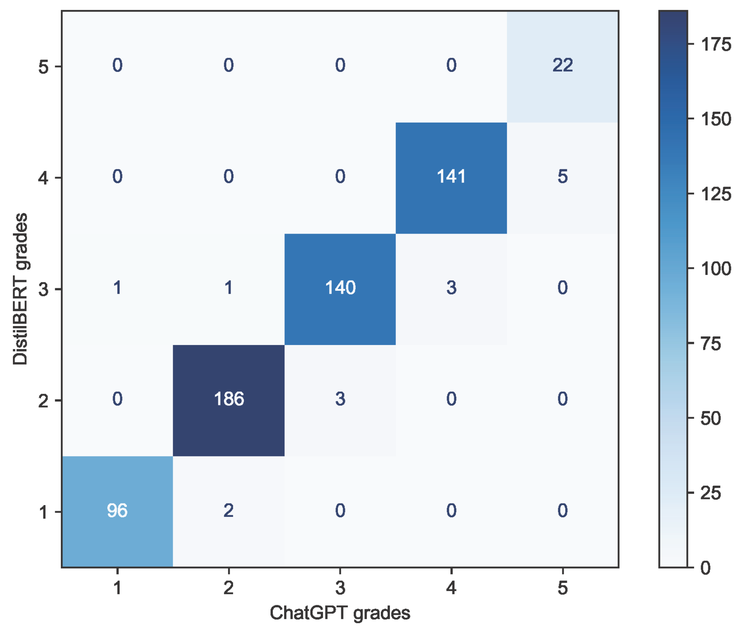

Figure 2. Confusion matrix of the grading based on LLM (test set).

Discussion and conclusions

The study highlights both the potential and challenges of automated grading systems for open-ended questions. In retrieval-augmented generation, the model generated relevant Q&A pairs for geotechnics by leveraging a curated corpus of 26 textbooks. This ensured accuracy and alignment with best practices, while synthetic student answers of varying quality enriched the dataset, providing a strong foundation for training the grading system.

The LLM-based grading system demonstrated significant performance, achieving around 98% across various metrics. The model effectively differentiated between relevant, detailed answers and irrelevant or incorrect ones, showcasing its potential for educational use.

Key contributions include:

1. Generating high-quality, relevant Q&A pairs tailored to geotechnics using LLMs and RAG.
2. Developing an ML-driven grading system using cosine similarity and LLMs, achieving grades comparable to human evaluators.
3. Enabling immediate feedback for students on open-ended questions, improving learning outcomes.

Acknowledgements

This study was funded by the European Union under the MSCA Staff Exchanges project 101182689 Geotechnical Resilience through Intelligent Design (GRID). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Research Executive Agency (REA). Neither the European Union nor the granting authority can be held responsible for them. The author would like to thank the bachelor students of the Soil Mechanics and Geotechnical Engineering course at BOKU for answering the test questions. Because of GDPR, they must remain incognito, but their contributions will rock on in our hearts.

References

Lewis, P.S.H., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., and Kiela, D. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *CoRR* [Online] Available at: https://arxiv.org/abs/2005.11401 [Accessed 2nd July 2025].

Sanh, V., Debut, L, Chaumond, J. and Wolf, T. 2019. "Distilbert, a distilled version of Bert: smaller, faster, cheaper and lighter", https://arxiv.org/abs/1910.01108.

Soranzo, E. 2025. Large language models for automated grading in geotechnics. *Machine Learning and Data Science in Geotechnics* 1 (1): 124–144. https://doi.org/10.1108/MLAG-01-2025-0001

Contact information for the presenting author:

enrico.soranzo@boku.ac.at