FermentAI: Large Language Models in Chemical Engineering Education for Learning Fermentation Processes

\*

aPROSYS, Dept. of Chemical and Biochemical Engineering, Technical University of Denmark, Søltofts Plads, Building 228 A, 2800 Kgs. Lyngby, Denmark.

ulkr@kt.dtu.dk

Abstract

Recent developments in artificial intelligence (AI), leading to the release of continuously improving large language models (LLMs) provide the opportunity for educators to automate repetitive tasks with stimulant experiences for the students. In fact, the ability of LLMs to extract content and key information from text offers a powerful tool for enhancing the learning experience. In this work, we present an example of how LLMs can be used to automate educational processes. We implement *FermentAI*, a virtual tutor (VT) to answer students’ questions about fermentation for a Master’s Degree course taught at the Technical University of Denmark. The model used is a pre-trained sequence-to-sequence model. The prompt provided to the LLM is composed of a question and its context, the curated information for the model to generate the answer. The context is retrieved through semantic search by calculating the cosine similarity between the query question and a dataset of historical questions. The primary objective of this work is to create an interactive, freely available, and accurate tool that students can use to ask questions about fermentation. The VT is integrated into BioVL, an online educational platform for (bio)chemical processes developed by the authors. The code and data are open-source on GitHub.

**Keywords**: Large Language Models, AI in Education, Chatbots in Engineering Curriculum, Education 4.0, Prompt Engineering.

* 1. Introduction

Education 4.0 represents a significant shift in the education sector through the incorporation of cutting-edge technology, including virtual reality, artificial intelligence (AI), and online platforms, into the learning process. Its primary objective is to prepare students to face the challenges resulting from this digital transformation and to thrive in a digital and automated world. This new paradigm prioritizes interactive and personalized learning experiences that promote digital literacy, critical thinking, and creativity. In addition, it aims to produce tech-savvy professionals who can adapt to rapidly changing landscapes and drive progress.

If applied correctly and ethically, AI could enable educators to automate repetitive tasks with stimulant experiences for the students. Particularly, natural language processing (NLP) systems have gained considerable interest in educational research and practice, due to their ability to learn and output natural language. Among others, these models have been used in applications for essay scoring, discourse analysis (e.g., between students), intelligent tutoring systems, and tools that support collaborative learning activities (Ferreira et al., 2019). Recent years have witnessed the rise of large language models (LLM), which are NLP models trained on massive amounts of text, mostly available online (Brown et al., 2020). The result of this large-scale training is that these models obtain knowledge about a wide range of topics and, hence are suitable for being applied in many different contexts and fields. This ability to extract content and key information from text offers a powerful tool for enhancing the learning experience. They could be used to identify knowledge gaps, aid formative assessment, provide personalized feedback, and facilitate the grading process, resulting in reducing educators' workload and providing more accurate and consistent evaluations (Hopfenbeck et al., 2023; Kasneci et al., 2023). Other AI-driven tools that could highly benefit students’ learning are chatbots and virtual tutors (VT). The benefits of these applications on learning include the fact that, if deployed online, students could continually have access to them to clarify easily answered questions, not needing the mediation of a teacher, and therefore and therefore reducing the workload related to some of the teacher’s responsibilities. Given all these new trends in the field, it follows that AI is becoming an integral part of new educational platforms in an effort to successfully deploy Education 4.0.

This work presents an initial investigation of how LLMs can be used to automate educational processes. We implemented *FermentAI*, a VT for answering students’ questions about fermentation. This aims to provide students with an open-access interactive tool that is trained on curated data and therefore returns high-quality and relevant responses. *FermentAI* is integrated into BioVL (Caño de las Heras et al., 2022), users can chat with it at: www.biovl.com/fermentAI. To foster transparency and knowledge transfer, code and data are available on GitHub at: https://github.com/FiammettaC/FermentAI.

* 1. Data

The data used for the model development is from the “Process Adaptation in Fermentation Based Biomanufacturing” (28455 course, 2023) Master’s Degree course taught at the Technical University of Denmark (DTU). The main purpose of the course is to introduce the used methods and tools when transferring a cell from a lab or research environment to pilot and full scale. Students should therefore acquire a better understanding of the interactions between the cells and the reactor, and learn how to aim for optimal conditions for the cells to grow in large scale. Some of the learning objectives of the course include: (i) simulate different reactor operating modes and interpret the simulation results; (ii) describe the effect of introducing basic process control on a process in a bioreactor; (iii) design and implement a basic strategy for data collection and handling on a fermentation process; and (iv) distinguish between basic AI tools for fermentation data processing.

In this work, a dataset composed of question-answer pairs is collected. The data is retrieved from past exams administered in the aforementioned course, where the answers are curated by the teachers. The set of questions and answers have a conversational nature and are therefore deemed suitable for this type of application.

* 1. Methods

The model implemented in this work, *FermentAI*, is based on a pre-trained LLM, FLAN-T5 (Chung et al., 2022), and is used to perform a question-answering task. The Hugging Face library is used for the pre-trained model implementation. The pre-trained model, FLAN-T5, is chosen because of the extensive fine-tuning on multiple downstream tasks performed, especially question-answering, and its manageable size.

For this task, we perform zero-shot learning, meaning that we do not provide additional examples to the LLM to be able to perform the given task. The prompt provided to the *FermentAI* is composed of a question (asked by the student) and its context, containing the curated information for the model to answer the question. This is done because in-context learning, which means enriching the prompt with some context, usually improves the quality of the generation. The context is retrieved through semantic search by calculating the cosine similarity, measuring the cosine of the angle between two vectors, of the query question and the set of historical questions populating the dataset. The algorithm then returns the most similar historical question (similarity closest to 1) and returns its context. This means that for each question the students ask, the algorithm calculates the most similar historical question and retrieves its context, which is the answer that was previously given to the historical question. Then the question and the context are given as prompt, providing the VT the necessary information to answer the questions asked. A schematic of the input given to the model is presented in Figure 1.

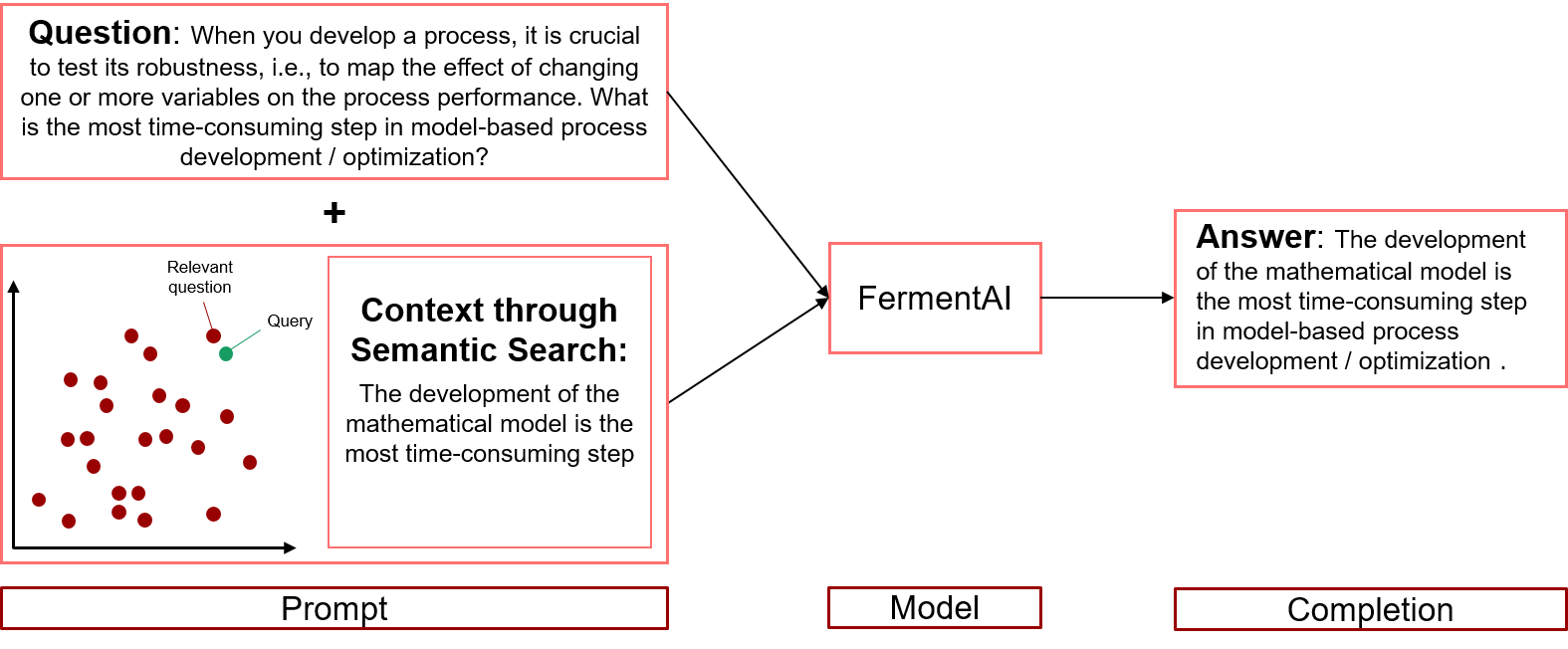


Figure 1: Schematics of the *FermentAI* pipeline. The context is retrieved through semantic search by calculating the cosine similarity between the query question and all historical questions stored in the dataset. The prompt is prefixed by the following instruction: “*Based on the HISTORICAL ANSWER, please answer this QUESTION*”.

We benchmark the results of the small model (google/flan-t5-small with 77M parameters), the base model (google/flan-t5-base with 248M parameters), and the xl model (google/flan-t5-xl with 2.85B parameters) to investigate whether a larger model, and therefore more parameters, can be beneficial in the specific domain and task performed. We then highlight the trade-off between performance and runtime speed, to facilitate other researchers working on similar applications. The model can be used with a CPU, although the deployed version makes use of GPUs.

At this stage of the project, we chose not to fine-tune the model but to provide the context to enrich the prompt of the LLM to save computing resources. The two approaches will be further assessed and benchmarked in the future.

To evaluate the responses generated by the model, we calculated two metrics widely used to evaluate QA models, the F1 score and cosine similarity. The F1 score is the harmonic mean of precision and recall score; it is implemented by taking into account the set of predicted tokens and therefore has the drawback that the order in which the tokens are predicted does not count. The cosine similarity is a metric often used to calculate document similarity in the NLP field.

* 1. Results and Discussion

*3.1 Quantitative evaluation: model performance*

The results of the model evaluation of the model on a subset of data are shown in Table 1. The cosine similarity is generally quite high, meaning that the generated output is similar to the original answers. Table 1 also shows that the F1 scores are not as high as the cosine similarity: however, this does not necessarily mean that the model is not able to perform the task. The metric compares the set of tokens produced to the ones in the original answer, and therefore possible explanations for the low F1 scores could include that the model tries to exclude redundant or not meaningful information, generate synonyms of words, or paraphrase the content of the original answer. Moreover, the used metrics, especially the F1 score, are affected by the length of the completion compared to the original answer.

Table 1: Average results of *FermentAI* model given the prompt (student’s question and context, retrieved through semantic search given historical answers). The results are calculated over a subset of questions (N=10) with a single high-performance GPU (NVIDIA GeForce RTX 2060).

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **F1 score** | **Cosine similarity** | **Runtime/completion** |
| Flan-T5 small | 0.62 | 0.83 | 3.3 s |
| Flan-T5 base | 0.74 | 0.94 | 9.3 s |
| Flan-T5 xl | 0.64 | 0.82 | 82.0 s |

One of the major issues of the model is the length of the output, which could be a challenging parameter to tune. In fact, if the minimum number of characters expected for the generation is too high, there is the risk that the model will repeat itself if the answer is straightforward. On the other hand, a fixed parameter that is too short would also mean that some sentences are cut and left unfinished. This issue will be further investigated in the future.

The fourth column in Table 1 compares the runtime necessary for one completion. It is important to notice that, generally, larger models have a longer runtime, meaning that they will output the answer more slowly. This has to be taken into consideration and addressed properly to ensure a fast and smooth experience for students.

At this stage of the implementation, to expert evaluation, the results are considered satisfactory. We plan to benchmark various approaches in the future, such as comparing prompt engineering and instruction fine-tuning, to investigate what is the best strategy to adopt in this use case and whether it can be generalized to different applications. Fine-tuning could be beneficial in fields that use language in an idiosyncratic manner, such as in our specific case. Finally, more quantitative metrics will also be used.

*3.2 Qualitative evaluation: further analyses*

An initial qualitative evaluation of the model revealed that the answers generated are coherent with the question asked and generally complete. For example, to the given prompt: *“QUESTION: When processing time series data, stationarity is important. How do you define stationarity? HISTORICAL ANSWER: Stationary data is time series data that keeps the same statistical properties. Non-stationary data will have properties that change as a function of time”*, the three models generated, respectively: (small) “non-stationary data will have properties that change as a function of time”, (base) “stationary data is time series data that keeps the same statistical properties. Non-stationary data will have properties that change”, and (xl) “stationary data is data that has the same statistical properties over a long period of time”. We can observe that all generated answers are correct, although they are formulated differently, where the most complete is arguably the completion generated by the base model, supporting the results previously discussed.

However, rigorously evaluating the models implemented in educational systems is very important; therefore, we plan to further evaluate the performance of the models and assess the fit of the VT for the given task. We plan to perform two tests with students: first, a small-scale test where only a few students are asked to get familiar with the platform and extensively interact with it. This is needed to establish whether the model can sustain a long QA session or if it gets repetitive after a few answers concerning the same topic (suggesting that the model might have limited knowledge regarding a topic). This small-scale test would also allow us to gain insights into the platform itself, whether it is user-friendly and intuitive whether further improvements to the interface are needed. Afterward, we aim to test the model with a larger set of volunteering students from DTU, as well as experts in the subject, such as the teachers of the course.

*3.3 Deployment of the model online and code availability*

*FermentAI* is integrated into BioVL (Caño de las Heras et al., 2022), an online educational platform dedicated to teaching (bio)chemical processes. Anybody can freely chat with the implemented VT at: www.biovl.com/fermentAI. The VT is currently in the form of a chatbot, however, future developments envisage an avatar or human-like agent, where students can utter the questions and the VT will voice an answer.

To foster transparency and knowledge transfer, the code and the pre-processed data are open-source, stored in a public repository and shared publicly on GitHub at: https://github.com/FiammettaC/FermentAI.

*3.4 Transferability to* *other domains of chemical and biochemical engineering*

Although the project presented has a quite specific application, i. e. answering questions related to the Process adaptation in Fermentation Based Biomanufacturing course taught at DTU, the approach discussed in this work could be applied to other domains within chemical and biochemical engineering. Similar models could be used, for example, to answer questions related to process control or thermodynamics, given that the dataset is curated to include discursive answers. At this stage the model could not, in fact, return or correct equations. However, provided the availability of high-quality domain data, it could explain the theory and answer simple questions, which could relieve some of the pressure from the teachers and allow them to focus on less trivial tasks.

*3.5 Reflection and possible implications regarding the use of AI in Education*

Finally, a reflection regarding the ethical and fair use of AI is indispensable. Emerging research in the field suggests that the use of AI in Education has the potential to support teaching and learning and improve student performance; however, its misuse, which may result from algorithmic bias and lack of clear regulations, could inhibit human rights and result in the reinforcement of existing inequalities (Prinsloo, 2020; Yang et al., 2021). Thus, although the expectations and promise of how AI can contribute to advancing education are substantial, it must be remembered that AI is not infallible, and attention should be put into its implementation to avoid unintended consequences. Therefore, developing these solutions requires understanding the possible issues and limitations, as well as the time and effort it takes to safely develop, test and use these types of systems. Moreover, even though an AI-powered system in Education could be used fairly and ethically, we should also reflect on whether these systems could benefit students and improve their education on a case-by-case basis. Extensive testing and model validation should be performed before rolling out these models in an educational context.

* 1. Conclusion

This work presented *FermentAI*, an educational virtual tutor to answer questions about fermentation processes. The data used to test the model is composed of the exam questions (and answers) from the “Process Adaptation in Fermentation Based Biomanufacturing” Master’s Degree course taught at DTU.

The model used is a pre-trained LLM, FLAN-T5, fine-tuned on multiple downstream tasks, including question-answering. Instead of fine-tuning the model on our domain-specific data, we performed prompt engineering, enriching the prompt given to the model with the context retrieved through semantic search from a database of historical questions. The performance of the model, calculated using F1 score and cosine similarity, is considered to be satisfactory. However, more qualitative assessment is needed, therefore the model will be extensively evaluated in the future, including tests with students.

Moreover, some considerations that researchers and developers should make to ensure fair and ethical application of AI in Education were discussed. Finally, we discuss some of the implications of using AI in Education, by providing some advantages and limitations. We hope that this effort will encourage researchers and practitioners working in this field to contribute with reflections and resources to ensure a more ethical and safe transition to a reality where AI models are fully embedded in the educational system, to safeguard learners and where every aspect of the ’digital contract’ is transparent.

References

T. Brown, B. Mann, N. Ryder, M. Subbiah, J.D Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, and S. Agarwal, 2020, Language models are few-shot learners. Advances in neural information processing systems, 33, pp.1877-1901.

S. Caño de las Heras, C.L. Gargalo, F. Caccavale, B. Kensington-Miller, K.V. Gernaey, S. Baroutian, and U. Krühne, 2022, From Paper to web: Students as partners for virtual laboratories in (Bio) chemical engineering education. Frontiers in Chemical Engineering, 4, p.959188.

H.W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, E. Li, X. Wang, M. Dehghani, S. Brahma, and A. Webson, 2022, Scaling instruction-finetuned language models. arXiv preprint arXiv:2210.11416.

R. Ferreira-Mello, M. André, A. Pinheiro, E. Costa, C. Romero, 2019, Text mining in education, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 9, e1332.

T.N. Hopfenbeck, 2023, The future of educational assessment: self-assessment, grit and ChatGTP?. Assessment in Education: Principles, Policy & Practice, 30(2), pp.99-103.

E. Kasneci, K. Seßler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. Hüllermeier, and S. Krusche, 2023, ChatGPT for good? On opportunities and challenges of large language models for education. Learning and individual differences, 103, p.102274.

P. Prinsloo, 2020, Of ‘black boxes’ and algorithmic decision-making in (higher) education–a commentary, Big Data & Society 7, 2053951720933994.

S.J. Yang, H. Ogata, T. Matsui, N.-S. Chen, 2021, Human-centered artificial intelligence in education: Seeing the invisible through the visible, Computers and Education: Artificial Intelligence 2, 100008.

28455 course: Process adaptation in Fermentation Based Biomanufacturing (28455), available at: https://kurser.dtu.dk/course/2023-2024/28455?menulanguage=en, Accessed: 27/11/2023.

Hugging Face Flan-T5 implementation, available at: https://huggingface.co/docs/transformers/model\_doc/flan-t5, Accessed: 27/11/2023.