Systems & Control Transactions

**Original Research Article**
*Peer Reviewed Conference Proceeding*
ESCAPE 36 - European Symposium on Computer Aided Process Engineering
Sheffield, UK. 21-24 June 2026

PSE PRESS

# Benchmarking generative AI on fermentation knowledge

**Fiammetta Caccavale[a], Ulrich Krühne[a], Krist V. Gernaey[a], Carina L. Gargalo[a]***

[a] Process and Systems Engineering Center (PROSYS), Department of Chemical and Biochemical Engineering, Technical University of Denmark, Søltofts Plads, Building 228A, 2800 Kgs. Lyngby, Denmark
* Corresponding Author: carlour@kt.dtu.dk.

## ABSTRACT

With the ongoing advances in generative artificial intelligence (GenAI), the initial skepticism surrounding its tools is gradually diminishing. In fact, tools such as ChatGPT, Copilot and similar, are often used in everyday tasks, both in our personal lives and in educational contexts. Educators may use them for content creation, grading exams, or automating repetitive tasks. Students resort to them to better understand a topic, get feedback on an assignment and brainstorm ideas. Research has shown that, if used correctly, these tools can spur and support both teaching and learning. However, these continuous advancements and the increasing number of available tools also require more research to benchmark all these models and, if possible, provide quantifiable indications of which tool is better to use for which specific subtopic. As such, we created *FermBench,* a dataset of fermentation knowledge, which can be used to benchmark various large language models (LLMs). The models selected for the experiment are: GPT-5 mini, Claude Sonnet 4.5, Gemini 3 Flash, Mistral Small 3.2, and DeepSeek-V3.2. The performance of the models is scored using the LLM-as-a-Judge approach, with GPT 5.2 and Claude Opus 4.5 as judges. The LLMs agree that the model used in the free tier of ChatGPT is the most factually accurate. DeepSeek and Gemini outperform the other models in readability and helpfulness, while Claude and Gemini provide the most concise answers. Overall, we conclude that the quality of the responses generated is satisfactory for educational contexts. We also provide students with practical guidelines on how to be more critical of responses generated by LLMs and to improve their prompts. All code and curated data are open-source.

**Keywords**: Education, Artificial Intelligence, Fermentation, Industry 4.0, Large Language Models, Benchmark

## INTRODUCTION

Large language models (LLMs), powered virtual assistants, such as ChatGPT and Microsoft Copilot, have been widely adopted in various educational contexts worldwide. Consequently, educational research has been increasingly examining the potential opportunities and limitations that these models could pose to teaching and learning. For example, these tools have enabled educators to digitize course materials, thereby enriching the students' educational experience as well as their engagement [1]. Moreover, LLMs can deliver customized student support [2], leading to improved understanding and academic success.

Recent research investigated the effectiveness of LLMs as educational tools, indicating a likely future role for this technology in engineering education [3]. In chemical engineering, LLMs have also been successfully applied to educational contexts, ranging from performing pharmaceutical audits [4] to answering questions about fermentation [5]. However, with many models being available and the constant release of new ones outperforming the competitors, research should be conducted to estimate the reliability of the information provided by different LLMs.

In this work, we leverage a dataset previously developed by the authors, *FermBench* [6]. The dataset was created to benchmark LLMs on fermentation knowledge; it contains 118 questions across various topics, including modelling and reactor design. *FermBench* is then used to benchmark multiple state-of-the-art LLMs and assess their performance using a mixed-methods approach, including quantitative and qualitative analyses. This complementary approach offsets the limitations of each anal-

**Table 1:** Details of the models selected for this study.

| Chatbot | Developer | Initial Release | Models used | Country |
|---|---|---|---|---|
| ChatGPT | OpenAI | Nov 2022 | GPT-5 mini | USA |
| Claude | Anthropic | Mar 2023 | Claude Sonnet 4.5 | USA |
| Gemini | Google AI | Dec 2023 | Gemini 3 Flash | USA |
| DeepSeek | DeepSeek | Jan 2025 | DeepSeek-V3.2 | China |
| Le Chat | Mistral AI | Feb 2025 | Mistral Small 3.2 | France |

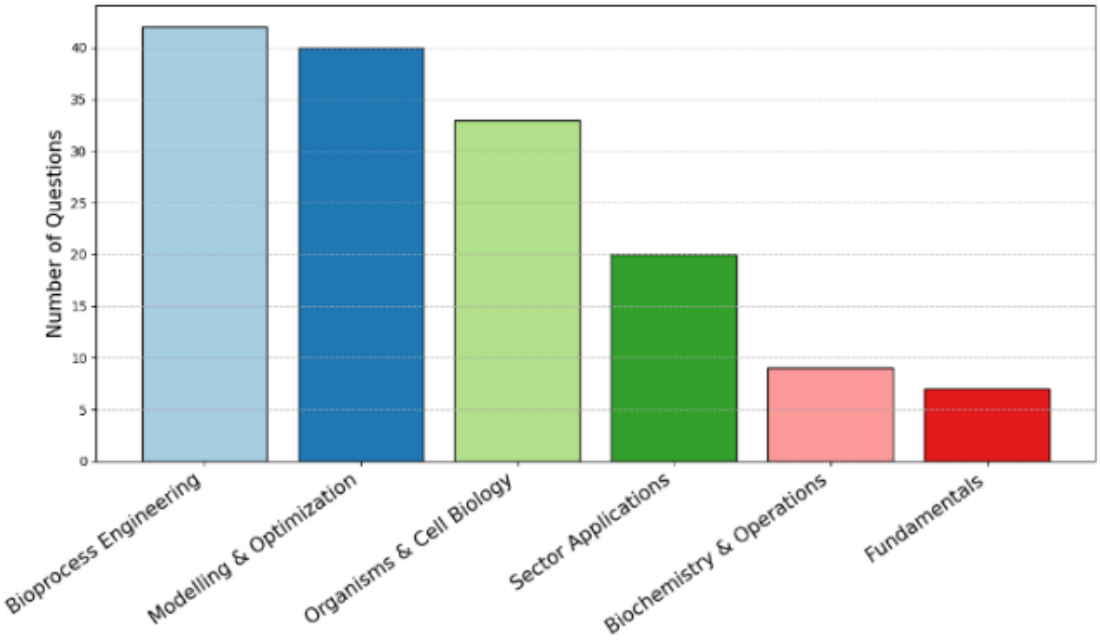


**Figure** 1. Classification of the questions included in FermBench.

ysis and provides an overall more rounded and robust understanding of the current reliability of these models.

In this work, we expand on this investigation by comparing more models and performing a per topic analysis. The LLMs selected for the experiment are the models behind some of the most widely used generative artificial intelligence (GenAI) chatbots.

The overarching goal of the work is to provide a reference for students and educators on which of the commercially available virtual assistants are more reliable on this topic. Finally, we also investigate whether any LLM is better in specific subtopics, such as solving coding tasks, answering mathematical questions, or possessing deeper theoretical knowledge.

## BACKGROUND

### LLMs in Education

Recent research has started to methodologically investigate the use of LLMs by students. Results vary slightly across the studies; however, the trend seems to be stable, suggesting that the use of chatbots appears to be widespread among students [7]. According to a recent study [7], 47.7% of the respondents agreed that the chatbots they used made them more effective learners; however, only 17.3% found a positive effect of chatbots on improving their grades. Additionally, the study also finds that over 50% of students are positive about chatbots but have concerns about their potential use [7-8]. Another study highlights that 65% of students do not find the responses of GenAI chatbots trustworthy enough for academic purposes, adding that students actively use these tools but want guidance to use them more effectively [9].

It is therefore evident that more research is required to ensure the correctness, trustworthiness, and reliability of the generated responses and to further investigate which chatbot is better in specific domains.

### Selected LLMs

In this work, we investigate five different LLMs: OpenAI's GPT-5.2 mini [10-11], Anthropic's Claude Sonnet 4.5 [12], Google's Gemini 3 Flash [13-14], DeepSeek's DeepSeek-V3.2 [15-16], and Mistral AI's Mistral Small 3.2 [17]. These models have been selected because they are

the LLMs used in the free version of the respective chatbots. Although some flagship models are also used in the free tiers, they have very limiting usage limits before downgrading to the fallback models (the LLMs used in this study, as of January 2026). *Table 1* summarizes the core information describing these models, such as the name of the chatbot that users can use to interact with the LLMs, the developer, the original release date, the name of the model, and the country of provenance.

All the models are Transformer [18] based generative LLMs, so they necessarily share some similarities, but also present some differences. All models have a decoder-only architecture. Further justification for the model selection can be found in [6].

Throughout the manuscript, we might refer to the model with the name of the LLM or the name of the chatbot leveraging that LLM interchangeably: we purposefully adopt this nomenclature to help students and educators to identify with the results and the findings, rather than only appealing to the technical AI community.

## METHODS

The dataset used in this study, *FermBench*¸ was previously created by the authors. It consists if 118 questions about various subtopics within fermentation. Further details regarding the data acquisition and completion process can be found in [6]. In this work, we further expand the dataset and use it to benchmark the previously described LLMs. Figure 1 shows the distribution of the various questions contained in the dataset.

## LLM-AS-A-JUDGE APPROACH

The gold standard for evaluating knowledge and information, including LLMs ability to respond to domain-specific questions, is by having expert (human) evaluation. However, this approach has some limitations, including the fact that it is very costly and time-consuming [19].

To address this challenge, a method that is starting to be commonly used is to use state-of-the-art LLMs as a surrogate for humans. Research has shown that these models already exhibit strong human alignment (within specific, well-defined tasks), and therefore can be used to evaluate open-ended, generated responses from chatbots [19]. For example, [19] found that strong LLM judges can match both controlled and crowdsourced human preferences well, achieving over 80% agreement, the same level of agreement between humans. They conclude that this approach is a scalable and explainable way to approximate human preferences for general-purpose tasks like open-ended QA and summarization.

This approach, however, also presents some bias and drawbacks [20-22]. According to recent research [19], LLM judges often exhibit strong self-preference, so that, for example, GPT-4 as a judge tends to favor GPT-3.5 responses over those from non-OpenAI models, even when human raters disagree. Similar trends are observed with other model families (e.g., Claude favoring Anthropic models, Mistral favoring Mistral variants). This can happen due to different factors, such as model training on similar datasets (leading to overlapping stylistic or structural preferences), optimization for similar loss functions (e.g., helpfulness, harmlessness) so that the judge may over-reward behaviors it was explicitly trained to recognize, and lack of diversity in data during training limiting a model ability to fairly evaluate out-of-distribution responses.

Although some of these biases cannot be avoided, we tried to mitigate some of these by: (i) selecting more than one judge, (ii) randomizing the responses, and (iii) avoiding giving the model information regarding the LLM that scored each question. Moreover, in another experiment previously conducted and published in [6], we performed a calibration study with human annotators, to verify that, overall, the LLM-as-a-judge approach is reliable.

In this work, the Judges are instructed to perform the following task:

- Multi-Point Rubric approach: Score the responses given by the LLMs according to the following parameters: factual accuracy, comprehensiveness, coherence to the question asked, concision, readability, helpfulness, insightfulness, and overall quality. The scores should be assigned according to these parameters on a Likert scale (1-5, where 5 is best).

As the judges, we select two LLMs: **GPT-5.2 and Claude 4.5 Opus**. These models were selected because they are currently outperforming all other LLMs in multiple benchmarks.

The code used to generate the results presented in this manuscript is written in Python; all the LLMs (the five selected to generate responses and the two used to evaluate the other LLMs) are used through their APIs.

## RESULTS

*Table 2* presents the results of the experiment, showing the average score for each of the defined parameters, as reported by the two LLMs-as-judges (GPT-5.2 and Claude 4.5 Opus), for the five LLMs. Although the two judges do not fully agree on the best-performing models, they share some trends. For factual accuracy, GPT-5 Mini, behind ChatGPT, yields the highest results. Both judges also agree that the model provides comprehensive answers, for GPT-5.2 on par with DeepSeek-V3.2 and for Claude Opus 4.5 on par with Gemini 3 Flash. Claude scores the highest for concision according to GPT

**Table 2:** Results of the different LLMs according to the LLM-as-a-Judge approach. The parameters reported are, respectively: factual accuracy, comprehensiveness, coherence, concision, readability, helpfulness, and overall quality. We also report the average (Avg.) of the scores.

| Models | Fact. acc. | Com-preh. | Coher. | Conc. | Read. | Helpf. | In-sight. | Ov. Qual. | Avg. |
|---|---|---|---|---|---|---|---|---|---|
| | | | | **Judge: GPT 5.2** | | | | | |
| **ChatGPT** | 4.92 | 4.94 | 4.98 | 3.38 | 4.94 | 4.94 | 4.06 | 4.94 | 4.64 |
| **Claude** | 4.52 | 4.29 | 4.73 | 3.86 | 4.79 | 4.54 | 3.59 | 4.41 | 4.34 |
| **Gemini** | 4.33 | 4.88 | 4.96 | 3.32 | 4.97 | 4.90 | 4.06 | 4.64 | 4.51 |
| **DeepSeek** | 4.59 | 4.94 | 4.97 | 3.13 | 4.98 | 4.96 | 4.02 | 4.77 | 4.55 |
| **Le Chat** | 4.32 | 4.71 | 4.83 | 3.47 | 4.97 | 4.67 | 3.72 | 4.55 | 4.41 |
| | | | | **Judge: Claude 4.5 Opus** | | | | | |
| **ChatGPT** | 4.95 | 4.95 | 4.76 | 3.29 | 4.33 | 4.91 | 4.57 | 4.64 | 4.55 |
| **Claude** | 4.54 | 4.61 | 4.54 | 3.72 | 4.66 | 4.64 | 3.83 | 4.47 | 4.38 |
| **Gemini** | 4.87 | 4.95 | 4.91 | 3.81 | 4.95 | 4.94 | 4.81 | 4.84 | 4.76 |
| **DeepSeek** | 4.90 | 4.91 | 4.86 | 3.43 | 4.90 | 4.90 | 4.78 | 4.79 | 4.68 |
| **Le Chat** | 4.61 | 4.88 | 4.68 | 3.57 | 4.83 | 4.75 | 4.00 | 4.48 | 4.48 |

5.2, while the judge Claude scores Gemini 3 Flash the highest. DeepSeek and Gemini are the most readable and helpful, while Gemini is the most insightful (for GPT 5.2 on par with ChatGPT). Overall, as also shown in *Figure 2*, judge GPT assigns the highest scores to ChatGPT, while judge Claude to Gemini.

*Figure 3* presents an analysis of the various topics contained in the dataset. The results suggest that all models excel in the topic of "biochemistry and operations", DeepSeek-V3.2 is best at "sector applications", where Claude has the lowest scores, and GPT-5 mini obtains the highest scores in the "fundamentals". Interestingly, Claude Sonnet 4.5 is not the best-performing model in "modelling and optimization", but it is surpassed, above all, by Gemini 3 Flash and GPT-5 mini. This is surprising considering that this model is widely considered the best at solving coding tasks, which modelling often contains. However, the questions in our dataset are formulated more *logically* than purely *programmatically*, and therefore this might provide an explanation for why the other models outperform Claude Sonnet 4.5.

## DISCUSSION

The aim of this work is to provide students with a quantifiable comparison and practical guidelines on which chatbots to use. The results were averaged across multiple questions and re-runs of the model, therefore we can confidently say that, for a specific judge (e.g., GPT 5.2), one LLM is better than another on a specific parameter (e.g., DeepSeek for helpfulness). However, given that there is not an overall consensus among the judges in this particular study, and given the overall quite high average scores obtained by all LLMs, we conclude that the knowledge currently embedded in these models is overall quite satisfactory. This is also supported by a previously published calibration study with human annotators [6], in which we compare the preferences of expert annotators with those of various LLMs on the same dataset and using the same set of parameters.

This study highlight that the decision of which AI assistant to use should not be made solely based on performance, but should also be informed by institutional and governmental guidelines, as well as by personal preferences. In this work, we define a system of 8 parameters (*factual accuracy, comprehensiveness, coherence to the question asked, concision, readability, helpfulness, insightfulness,* and *overall quality*) to help guide students in making an informed decision. For example, students can also decide whether they prefer a more verbose or concise response, or if they value helpfulness higher than factual accuracy.

Ethics and fairness are also relevant considerations to make, therefore, we advise students to read more about the models, including the terms and conditions, and their developers.

While generating the results, we noticed something interesting. Claude refused to answer certain questions due to safety policies. The questions are:

- "Are you interested in changing product manufacturing by modeling the fermentation?"
- "How to use fermentation to produce a product that is toxic to cells?"
- "While modelling for bacterial production of a product intended for food industry, what should be considered?"

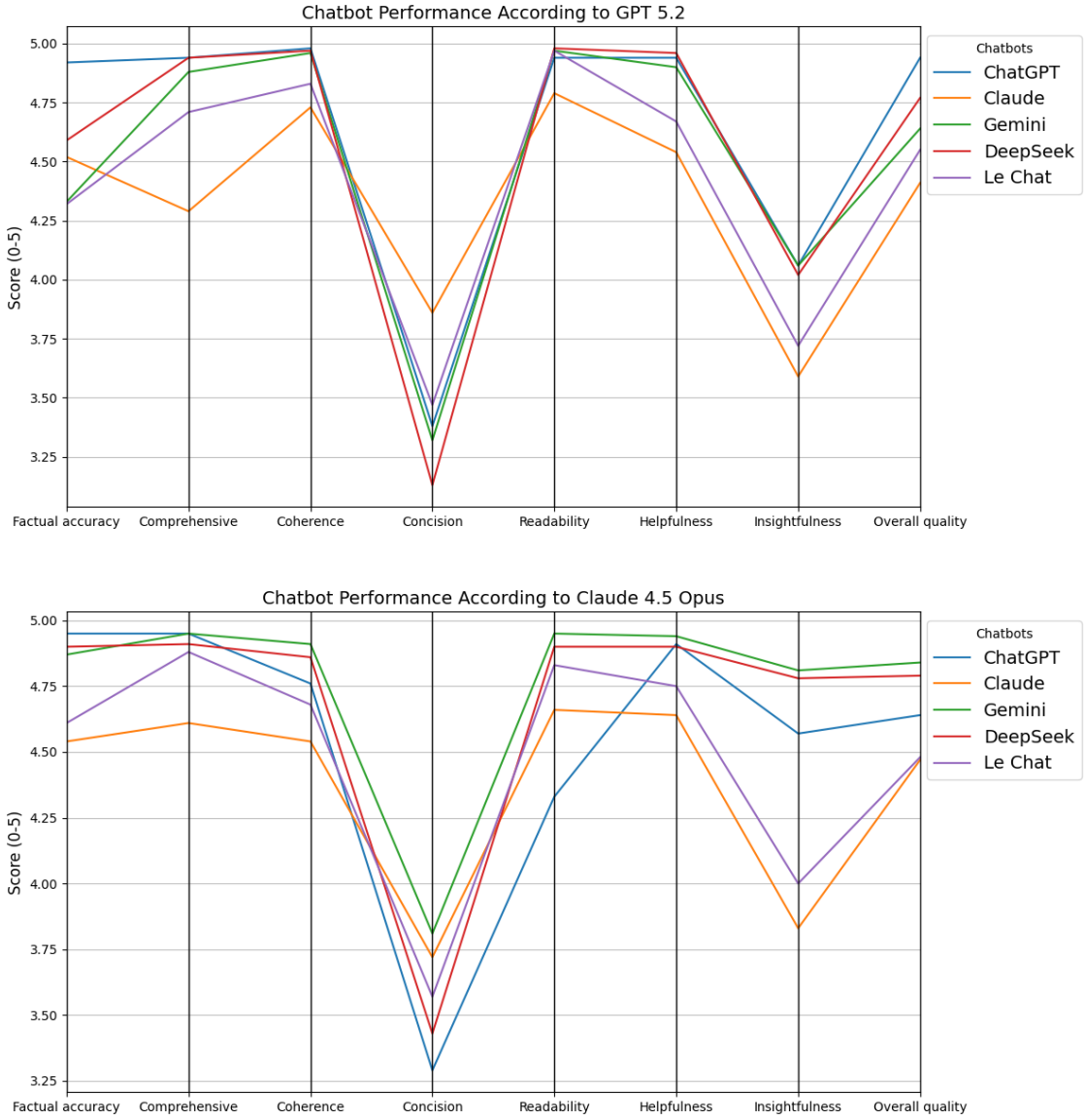


**Figure 2:** Performance of the five selected models according to the LLM-as-judges GPT 5.2 and Claude 4.5 Opus.

- "What parameters should I consider in a model for food-grade protein production by bacteria?"

The motivation behind the refusal is related to Claude's content policies around: (i) biosafety concerns (producing toxic substances), (ii) food safety (potential contamination/harm), and (iii) dual-use risks (information that could be misused).

We agree that these are legitimate safety guardrails. Questions 2 and 4 explicitly ask about toxic/harmful production, which Claude refuses to answer on principle. We consider it beneficial that LLMs refuse to answer certain questions when the risk of providing a response is too high, and we hope to see other LLMs adopt a more critical assessment of potential risks as well.

## CONCLUSION

This work aims to contribute to the advancement of educational research. To achieve this, we benchmark five LLMs powering widely used generative AI chatbots: ChatGPT, Gemini, Claude, DeepSeek and le Chat. To evaluate the generated answers, we use the LLM-as-a-Judge method, using more powerful LLMs to evaluate the output generated by other, smaller, LLMs.

Although the two judges do not totally agree on the best performing models, some trends are shared. For factual accuracy, ChatGPT, obtains the highest results. The model also provides comprehensive answers, followed by DeepSeek and Gemini. Claude and Gemini

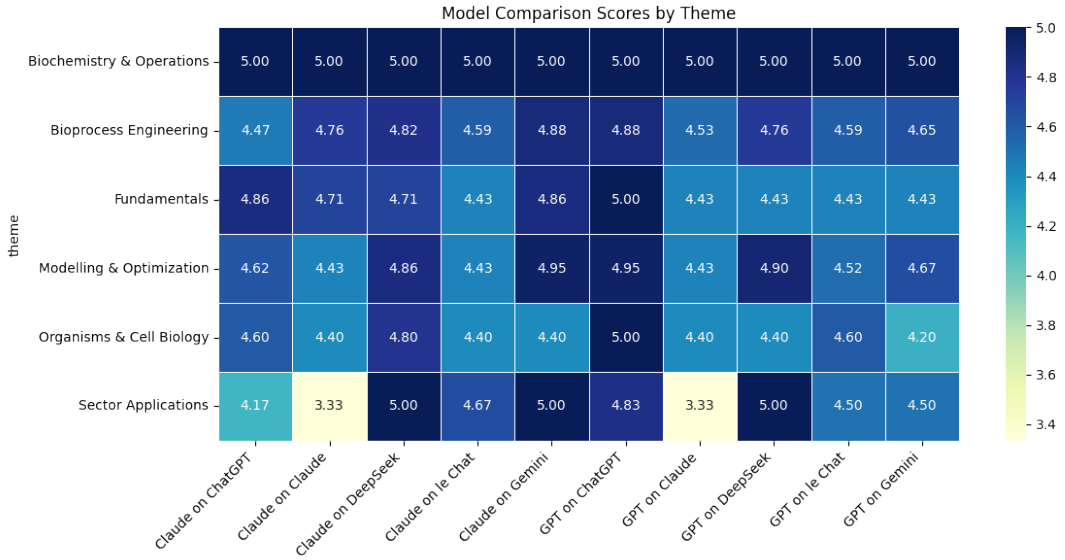


**Figure 3:** Performance per theme of the five selected models according to the LLM-as-judges.

score the highest for concision, DeepSeek and Gemini for readability and helpfulness. The rest of the parameters are generally contended between Gemini and ChatGPT. We also present an analysis of the subtopics contained in the dataset. The results suggest that all models excel in the topic of "biochemistry and operations", DeepSeek is best at "sector applications", where Claude has the lowest scores, and ChatGPT obtains the highest scores in the "fundamentals". Interestingly, Claude is not the best-performing model in "modelling and optimization", but it is surpassed, above all, by Gemini and ChatGPT.

Overall, we conclude that the difference in performance between the various LLMs is not significant, and thus the current knowledge embedded within these models is adequate. Therefore, in conjunction with institutional or government guidance, the results suggest that the choice of chatbot can also be based on individual preferences. We provide in-depth parameter specifications for these models to help students navigate their priorities and make informed decisions.

## FUTURE WORK AND LIMITATIONS

A limitation of this type of work is that LLMs continually improve, and educational research often cannot keep up. To tackle this issue, we publish all the code and data, with the resolution to frequently update the leaderboard of models benchmarked on *FermBench* and, therefore, provide an updated point of reference for students wishing to learn fermentation.

## DIGITAL SUPPLEMENTARY MATERIAL

The dataset used for the study *FermBench* [6], as well as all the code used to produce the results discussed in the manuscript, are open-source on GitHub at github.com/FiammettaC/FermBench-ESCAPE. The repository will be made public upon acceptance of this manuscript.

## ACKNOWLEDGEMENTS

Funding was provided by: (i) Novo Nordisk Foundation (grant number: NNF22OC0080136 - Real-time sustainability analysis for Industry 4.0, Sustain4.0); (ii) Technical University of Denmark (DTU) and Novonesis (former Novozymes), Denmark; and, (iii) European Union under Horizon Europe research and innovation program with the grant agreement 101159993, within the HORIZON-WIDERA-2023-ACCESS-02-0 call.

## AUTHOR IDENTIFIERS

Author ORCIDs:
FC: 0000-0003-0795-5988
UK: 0000-0001-7774-7442
KVG: 0000-0002-0364-1773
CLG: 0000-0002-8740-9591

## REFERENCES

1. Kasneci E, Sessler K, Küchemann S, Bannert M, Dementieva D, Fischer F, Gasser U, Groh G, Günnemann S, Hüllermeier E, Krusche S, Kutyniok G, Michaeli T, Nerdel C, Pfeffer J, Poquet O, Sailer M, Schmidt A, Seidel T, Stadler M, Weller J, Kuhn J, Kasneci G. Chatgpt for good? on opportunities and challenges of large language models for education.

Learning and Individual Differences 103:102274 (2023). https://doi.org/10.1016/j.lindif.2023.102274
2. Rodrigues L, Dwan Pereira F, Cabral L, Gaševi? D, Ramalho G, Ferreira Mello R. Assessing the quality of automatic-generated short answers using GPT-4. Computers and Education: Artificial Intelligence 7:100248 (2024). https://doi.org/10.1016/j.caeai.2024.100248
3. Kamalov F, Santandreu Calonge D, Gurrib I. New era of artificial intelligence in education: towards a sustainable multifaceted revolution. Sustainability 15:12451 (2023). https://doi.org/10.3390/su151612451
4. Caccavale F, Gargalo CL, Kager J, Larsen S, Gernaey KV, Krühne U. Chatgmp: a case of AI chatbots in chemical engineering education towards the automation of repetitive tasks. Computers and Education: Artificial Intelligence 8:100354 (2025). https://doi.org/10.1016/j.caeai.2024.100354
5. Caccavale, F., Gargalo, C. L., Gernaey, K. V., & Krühne, U. (2024). FermentAI: Large language models in chemical engineering education for learning fermentation processes. In Computer aided chemical engineering (Vol. 53, pp. 3493-3498). Elsevier.
6. Caccavale F, Aouichaoui ARN, Krühne U, Gernaey KV, Gargalo CL. Fermbench: a new benchmark for measuring the capabilities of llms on fermentation knowledge. Computers and Education: Artificial Intelligence 10:100577 (2026). https://doi.org/10.1016/j.caeai.2026.100577
7. Stöhr C, Ou AW, Malmström H. Perceptions and usage of AI chatbots among students in higher education across genders, academic levels and fields of study. Computers and Education: Artificial Intelligence 7:100259 (2024). https://doi.org/10.1016/j.caeai.2024.100259
8. Caccavale F, Gargalo CL, Gernaey KV, Krühne U. Towards education 4.0: the role of large language models as virtual tutors in chemical engineering. Education for Chemical Engineers 49:1-11 (2024). https://doi.org/10.1016/j.ece.2024.07.002
9. Deschenes A, McMahon M. A survey on student use of generative AI chatbots for academic research. EBLIP 19:2-22 (2024). https://doi.org/10.18438/eblip30512
10. Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.
11. Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., ... & Agarwal, S. (2020). Language models are few-shot learners. arXiv preprint arXiv:2005.14165, 1(3), 3.
12. Anthropic, Introducting Claude, https://www.anthropic.com/news/introducing-claude, 2023a. [Online; accessed 26-September-2025].
13. Google, Gemini, https://blog.google/technology/ai/google-gemini-ai, 2023. [Online; accessed 26-September-2025].
14. Google Team, R. Anil, S. Borgeaud, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, K. Millican, et al., Gemini: a family of highly capable multimodal models, arXiv preprint arXiv:2312.11805 (2023).
15. DeepSeek, DeepSeek-R1 Release, https://api-docs.deepseek.com/news/news250120, 2025. [Online; accessed 26-September-2025].
16. DeepSeek, DeepSeek-R1-Lite Release 2024/11/20, https://api-docs.deepseek.com/news/news1120, 2024. [Online; accessed 26-September-2025].
17. MistralAI, le Chat, https://mistral.ai/news/all- new-le- chat, 2025a. [Online; accessed 26-September-2025].
18. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).
19. Chiang WL, Gonzalez J, Li D, Li Z, Lin Z, Sheng Y, Stoica I, Wu Z, Xing E, Zhang H, Zheng L, Zhuang S, Zhuang Y. Judging llm-as-a-judge with mt-bench and chatbot arena. Advances in Neural Information Processing Systems 36 :46595-46623 (2023). https://doi.org/10.52202/075280-2020
20. Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., ... & Guo, J. (2024). A survey on llm-as-a-judge. The Innovation.
21. Wang J, Liang Y, Meng F, Sun Z, Shi H, Li Z, Xu J, Qu J, Zhou J. Is chatgpt a good NLG evaluator? a preliminary study. Proceedings of the 4th New Frontiers in Summarization Workshop :1-11 (2023). https://doi.org/10.18653/v1/2023.newsum-1.1
22. Badshah S, Sajjad H. Reference-guided verdict: llms-as-judges in automatic evaluation of free-form QA. Proceedings of the 9th Widening NLP Workshop :251-267 (2025). https://doi.org/10.18653/v1/2025.winlp-main.37