CAN AI PASS THE CZECH UNIVERSITY ENTRANCE TEST?

TOMASZEK Lukas – MIKLOŠÍKOVÁ Miroslava – ŠALOUN Petr, CZ

MŮŽE AI ÚSPĚŠNĚ SLOŽIT PŘIJÍMACÍ TEST NA ČESKÉ VYSOKÉ ŠKOLY?

Abstract
This article explores the ability of large language models to successfully complete the general academic prerequisites test, which is part of the national comparative examinations conducted by SCIO and used as an entrance exam for Czech universities. For the experiment, a 2024 test was adapted for processing by reformatting questions, mathematical expressions, and graphical information. Several language models were tested, including GPT-3.5, GPT-4, Gemini, Claude 3.5, LLaMA 3.1, and Mistral Large 2. The results show that all models outperformed the median student score, with stronger performance in the verbal section compared to the analytical section. Additionally, the models demonstrated the ability to clearly explain their reasoning processes, offering potential support for students preparing for these exams.

Keywords: Large Language Models, SCIO Exam, Entrance Test

Úvod

Artificial intelligence (AI), especially in the form of large language models (LLMs), is becoming an increasingly integral part of many areas of human activity — education included (Jeon & Lee, 2023; Kasneci et al., 2023). These models can generate text, answer questions, analyze complex information, and explain academic topics with remarkable accuracy. Given their ability to understand and process language and data, the natural question arises: can AI become a reliable partner in learning, exam preparation, or even knowledge assessment?
In recent years, it has become clear that LLMs offer real potential in education. Teachers use them to create lesson materials, while students rely on them for practicing and generating questions. They help overcome language barriers (Fitas, 2025), support critical thinking (Yuxian, 2025), and provide alternative perspectives on problems. When used effectively, AI doesn't replace traditional teaching — it acts as a smart supplement, offering both inspiration and support. However, before fully integrating AI into the education system, we need to ask an important question: what is it truly capable of? If we want to use AI for preparing students for entrance exams or assessing their skills, we must first test its performance in real exam conditions. The SCIO test, widely used for university admissions in the Czech Republic, presents an ideal opportunity to practically evaluate AI's capabilities.
In this article, we evaluate how well selected LLMs can complete the general academic prerequisites test. The article is organized as follows: first, we introduce the datasets used to assess the knowledge and skills of these LLMs and we
provide an overview of the SCIO test itself. Next, we describe the experiment conducted, present the results, and offer a discussion..

1 LLMs datasets (Times New Roman, 12 pt, bold type, alignment: left, headline in bold)

There are several LLMs available online (Naveed et al., 2023). Each model has its own strengths and limitations—some perform better in specific domains, while others excel in different areas. As such, it is difficult to identify a single model as the best overall. One common method for

evaluating model quality is user-based testing, where users submit the same prompt to two models and compare the responses to determine which one is better. However, this approach is time-intensive and requires a significant amount of human effort.

An alternative method for evaluating LLMs is through the use of datasets. (Rein et al., 2024) introduced a dataset containing 448 challenging questions created by experts in biology, chemistry, and physics. According to the authors, these questions are difficult even for PhD-level experts, who achieve an average accuracy of around 65 %. A similar, yet significantly larger dataset was published by (Li et al., 2024), comprising 12,000 questions across 14 academic domains. Each question includes 10 possible answers, with multiple correct options possible. In contrast, the dataset introduced by (Zhou et al., 2023) focuses on testing LLMs in terms of task execution and response verification. It features prompts such as "Write a text of at least 300 words about..." or "Write two paragraphs about...".

All the datasets discussed so far are in English. However, there are also several valuable datasets available in the Czech language. One example is the dataset by (Fajcik et al., 2024), which comprises 50 challenging tasks written in Czech. Another notable resource is the dataset by (Kapsa et al., 2024), which contains questions from recent 6th and 9th grade primary school exams, as well as final high school graduation tests. Czech-language datasets are particularly important, as language differences can significantly impact the performance of LLMs. A model that performs well in English may not achieve the same level of accuracy in Czech, due to variations in linguistic structure, vocabulary, and the availability of training data.

Multilingual datasets, which contain sets of questions in various languages, are also worth mentioning. One example is the Belebele dataset (Bandarkar et al., 2024), which includes questions in Czech. It consists of 900 questions, each featuring a short text passage followed by a comprehension question and four answer choices. The dataset is designed to evaluate reading comprehension across different languages, offering valuable insights into model performance in various linguistic settings. Thanks to its structure, it is particularly useful for assessing a model's ability to generalize across language contexts.

A wide variety of datasets are available online (Joshi et al., 2017; Kwiatkowski et al., 2019; Liu et al., 2024), intended for both training LLMs and evaluating their performance. In addition to these specialized resources, there are also tests originally designed to assess human knowledge rather than AI. A notable example is the SCIO tests, which serve as a national student assessment in the Czech Republic and often replace or complement university entrance exams. Despite their original purpose, these tests can also be effectively used to evaluate the capabilities of LLMs.

2 SCIO test

The National Comparative Exams by SCIO offer a range of tests across various subjects, including Mathematics, Biology, Chemistry, Foreign Languages, and more. Among these, the most frequently taken and widely recognized is the general academic prerequisites test, which serves as a crucial indicator of a candidate's readiness for university-level studies. Rather than testing specific subject knowledge, this exam evaluates general skills and abilities essential for academic success in higher education.

The General academic prerequisites test is divided into two main sections (each has 33 questions) that together provide a comprehensive assessment of a candidate's academic potential. The first section focuses on verbal skills, such as reading comprehension, information processing, and critical thinking. The second section evaluates analytical abilities, including logical reasoning, data interpretation, and working with quantitative information. Both sections are timed, ensuring an objective and balanced measure of a candidate's preparedness for university study.

3 Experiment design

As part of the experiment, a 2024 general academic prerequisites test in the Czech language was selected and adapted for evaluation by LLMs. The test was broken down into individual questions and reformatted to ensure compatibility with text-based model inputs. Blank spaces, where the correct answer was expected, were replaced with placeholder symbols, accompanied by prompts such as "replace the symbol with the correct answer." Mathematical expressions, equations, and tables were converted into TeX format for clarity and consistency. A single graph included in the original test was also transformed into a table to allow for easier interpretation by the models.
A total of 66 questions, covering both the verbal and analytical sections of the test, were presented in Czech to the following LLMs:

- GPT-3.5,
- GPT-4,
- Gemini,
- Claude 3.5,
- LLama 3.1 and
- Mistral Large 2.

Each model was queried using a standardized prompt format through its public interface, without any custom parameter adjustments. The test format consistently included four answer options per question, with exactly one correct choice. Models were instructed to select a single, definitive answer. If a response was missing or ambiguous, the prompt was repeated to obtain a valid answer.
Model outputs were recorded and evaluated against the official answer key. For each correct answer, the model earned 1 point; for each incorrect answer, 0.33 points were deducted. No points were awarded or subtracted for unanswered items. Based on these rules, the number of correct and incorrect answers was tallied for each model, and the final scores were calculated accordingly. This methodology enables a fair, standardized comparison of model performance under realistic exam conditions.

4 Results

Table 1 shows the number of correct and incorrect answers provided by each LLM in both sections of the test. Figure 1 presents the recalculated scores, illustrating the performance of each model in the verbal and analytical sections. Figure 2 then compares the total scores of the models with the median score achieved by students.

Table 1: The number of correct and incorrect answers in the verbal and analytical sections

	GPT 3.5	GPT 4	Gemini	Claude 3.5	LLama 3.1	Mistral Large 2
Correct – verbal part	25	28	28	29	29	31
Incorrect – verbal part	8	5	5	4	4	2
Correct – analytical part	23	21	21	22	22	23
Incorrect – analytical part	10	12	12	11	11	10

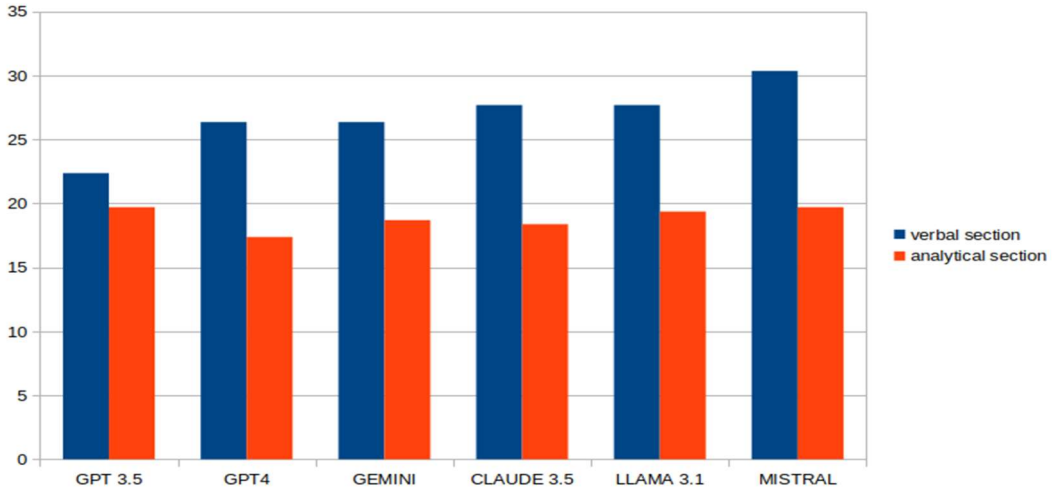

Figure 1 – Score comparison of LLMs across test sections

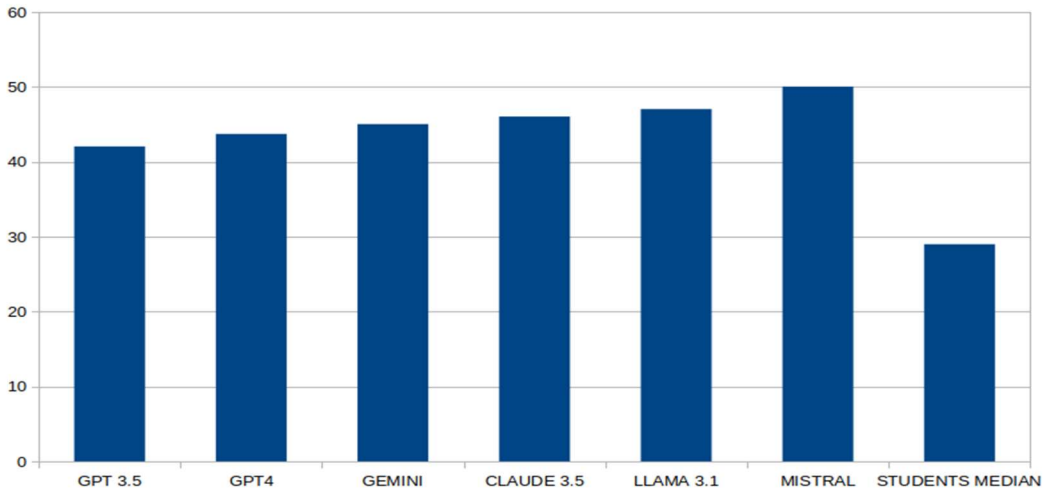

Figure 2 – The number of points each LLM would have obtained in comparison to the median score of student participants

5 Discussion

The results indicate that all tested LLMs outperformed the average student. The median score among students was 29 points, while the highest score achieved by a student was 63.3 points. In comparison, the LLMs scored between 42.06 points (lowest, ChatGPT 3.5) and 50.04 points (highest, Mistral Large 2).

Overall, it can be said that all language models performed better in the verbal section compared to the analytical section. This outcome is expected, as language models are highly proficient in working with text. However, they may face challenges in tasks requiring complex reasoning or quantitative analysis. A contributing factor could be the use of zero-shot prompting (Kojima et al., 2022 in this experiment, where models were simply given the question without any additional guidance. It is likely that using techniques such as chain-of-thought prompting (Wei et al., 2022) would have led to improved performance, particularly in the analytical section.

On the other hand, it is worth noting that when the models selected the correct answers, most of them were also able to clearly and thoroughly explain the reasoning behind their choices. The models demonstrated strong capabilities in eliminating incorrect options and articulating why a particular answer was right or wrong. This ability can be highly beneficial for students preparing for the test, as language models can provide detailed explanations of the problemsolving process. However, students must always critically evaluate the responses provided by the models.
A recommended strategy in this context (especially for understanding and verifying solution steps) is self-consistency prompting (Wang et al., 2022). This involves asking the same question multiple times or querying several different models and comparing the results. This approach can help students identify consistent reasoning patterns and gain deeper insight into the logic behind correct answers. However, it is important to approach this method with caution, as it was often observed that when one model failed to answer a question correctly, others tended to fail as well. Likewise, when one model gave the right answer, most of the others did too. This consistency suggests a shared reasoning structure among models, which, while useful, may also lead to collective blind spots.
Another interesting finding emerged from comparing different versions of ChatGPT. Although version 4 is the newer model, it outperformed version 3.5 only in the verbal section of the test. In contrast, version 3.5 achieved better results in the analytical section. On the other hand, models from other companies generally delivered better overall performance than ChatGPT. However, to obtain more reliable and comprehensive insights, it would be advisable to expand the testing to include multiple versions of the test and to create a more robust dataset.

Conclusion

In conclusion, we would like to emphasize that LLMs are capable of successfully completing the general academic prerequisites test by SCIO with a high level of accuracy. Additionally, they can effectively explain the reasoning process behind their answers, which can be highly beneficial for students preparing for these exams.
On the other hand, it is important to acknowledge that these models are not infallible and do make mistakes. In future research, we aim to explore ways to improve the accuracy of their responses so that these tools can be more effectively used in student preparation. Additionally, we plan to investigate methods to intentionally alter questions or prompts to cause the LLMs to respond incorrectly. This line of research could be valuable for security purposes, such as preventing students from misusing LLMs during their entrance exams.

Bibliography

Bandarkar, L., Liang, D., Muller, B., Artetxe, M., Shukla, S. N., Husa, D., ... & Khabsa, M. (2023). The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. arXiv preprint arXiv:2308.16884.
Fajcik, M., Docekal, M., Dolezal, J., Ondrej, K., Beneš, K., Kapsa, J., ... & Kydlicek, H. (2024). BenCzechMark: A Czech-centric Multitask and Multimetric Benchmark for Large Language Models with Duel Scoring Mechanism. arXiv preprint arXiv:2412.17933.
Fitas, R. (2025). Inclusive Education with AI: Supporting Special Needs and Tackling Language Barriers. arXiv preprint arXiv:2504.14120.
Jeon, J., & Lee, S. (2023). Large language models in education: A focus on the complementary relationship between human teachers and ChatGPT. Education and Information Technologies, 28(12), 15873-15892.

Joshi, M., Choi, E., Weld, D. S., & Zettlemoyer, L. (2017). Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. arXiv preprint arXiv:1705.03551.
Kapsa, J., Fajčík, M., Smrž, P., & Hradiš, M. (2024). Cermatqa: A benchmark for czech language understanding and for reasoning on czech math assignments. https://huggingface.co/datasets/CZLC/cermat czech open, https://huggingface.co/datasets/CZLC/cermat math open, https://huggingface.co/datasets/CZLC/cermat czech tf, https://huggingface.co/datasets/CZLC/cermat czech mc, https://huggingface.co/datasets/CZLC/cermat math mc. (Dataset published on Hugging Face)
Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., ... & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. Learning and individual differences, 103, 102274.
Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. Advances in neural information processing systems, 35, 22199-22213.
Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., ... & Petrov, S. (2019). Natural questions: a benchmark for question answering research. Transactions of the Association for Computational Linguistics, 7, 453-466.
Li, T., Chiang, W. L., Frick, E., Dunlap, L., Wu, T., Zhu, B., ... & Stoica, I. (2024). From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline. arXiv preprint arXiv:2406.11939.
Liu, Y., Cao, J., Liu, C., Ding, K., & Jin, L. (2024). Datasets for large language models: A comprehensive survey. arXiv preprint arXiv:2402.18041.
Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., ... & Mian, A. (2023). A comprehensive overview of large language models. arXiv preprint arXiv:2307.06435.
Rein, D., Hou, B. L., Stickland, A. C., Petty, J., Pang, R. Y., Dirani, J., ... & Bowman, S. R. (2024). Gpqa: A graduate-level google-proof q&a benchmark. In First Conference on Language Modeling.
Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., ... & Zhou, D. (2022). Self-consistency improves chain of thought reasoning in language models. arXiv preprint arXiv:2203.11171.
Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems, 35, 24824-24837.
Yuxian, J. (2025). Bridging the knowledge-skill gap: The role of large language model and critical thinking in education. Computers & Education, 105357.
Zhou, J., Lu, T., Mishra, S., Brahma, S., Basu, S., Luan, Y., ... & Hou, L. (2023). Instruction-following evaluation for large language models. arXiv preprint arXiv:2311.07911.

Acknowledgement:

I would like to thank SCIO for allowing me to use their tests in the creation of this article.
During the preparation of this work the author(s) used GPT-4 in order to improve the language and readability of this work. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Contact address:
Lukas Tomaszek,
Palacky University Olomouc, Křížkovského 8, 77900 Olomouc, Czech Republic,
lukas.tomaszek01@upol.cz

Miroslava Miklošíková
VSB - Technical University of Ostrava, 17. listopadu 15, 70800 Ostrava, Czech Republic
miroslava.miklosikova@vsb.cz

Petr Šaloun
Palacky University Olomouc, Křížkovského 8, 77900 Olomouc, Czech Republic,
petr.saloun@upol.cz