Project Synopsis – Group 10

Karl Johan Murphy Mogensen, s234813
Mikel Taotao Yu s234812

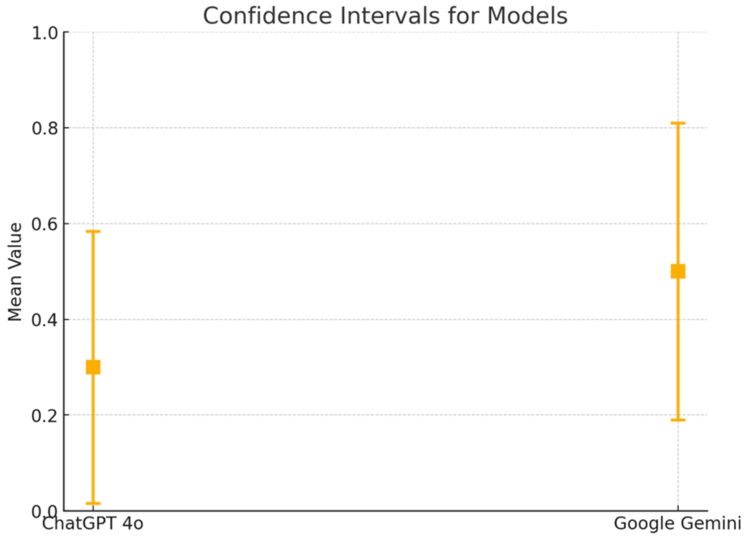
This project aims to evaluate ChatGPT 4o and Google Gemini in their capabilities to solve math problems, namely problems from the DTU mathematics 1 course. Instilling mathematic 'sense' in LLMs has been a challenge in previous years and therefore, we would like to investigate the current mathematical capabilities of ChatGPT and Gemini. Since GPT 4o is commonly regarded as the best public LLM, we hypothesize that it will perform better than Gemini and will thus carry out a comparison.

In order to simplify the problem, the model will only be evaluated on its amount of correct answers and not its mathematical reasoning as it introduces an element of subjectivity and complexity to the problem.

To set a direction for the project, a small pilot sample was retrieved from both GPT 4o and Gemini by using the exam set from December 2023. In the pilot sample, the following accuracies and their 95% confidence intervals were obtained for ChatGPT 4o and Google Gemini, respectively:

ChatGPT 4o: 0.3
CI: [0.01597423, 0.58402577]

Google Gemini: 0.5
CI: [0.1901025, 0.8098975]



Since the sample size for the pilot study was fairly low, we expect these accuracies to change and narrower confidence intervals as more exam sets are sampled.

An apparent challenge with this project is the risk of overwhelming the LLMs with too many tasks at once. By uploading the whole exam set at once, there is a risk exceeding the token limit of the model, resulting in ‘hallucinations’. The solution to this would be to prompt questions individually which is, however, more time-consuming.

Another issue is that potential differences might be caused by the different difficulties of separate exam sets instead of different performances in the models. To examine this, a two-way ANOVA could potentially be employed in order to ascertain it.