Comparative Analysis of Generative AI Systems Across Diverse Scenarios

Written by group Nr. 8:

Adrian Imfeld - adrian.imfeld@stud.hslu.ch
Albin Plathottathil - albin.plathottathil@stud.hslu.ch
Ansam Zedan - ansam.zedan@stud.hslu.ch
Christine Kortemeyer - christine.kortemeyer@stud.hslu.ch
Grzegorz Jaczech - grzegorz.jaczech@stud.hslu.ch
Michael Jung - michael.jung@stud.hslu.ch

Link to Github Repository

Abstract

The study aims to compare the performance and output quality of several generative AI systems, including ChatGPT 3.5, ChatGPT 4, Gemini, Gemini Advanced, and LLaMA 70b. The comparison framework is based on responses to prompts in categories such as creativity, mathematics, theory of mind, intuitive physics, idea generation, and translation into multiple languages. Results are scored by six reviewers on a scale of 1-10, supplemented by qualitative reviews. The analysis includes both quantitative scoring and qualitative NLP analysis of the reviews and responses to provide insight into the strengths and weaknesses of each model. Results indicate ChatGPT 4's lead in creative and theoretical tasks, Gemini's strength in factual accuracy, and LLaMA 70b's niche effectiveness despite lower overall performance. The study also underscores the importance of model selection based on task specificity and highlights the rapid evolution of AI capabilities, suggesting the need for continuous benchmarking and model development aligned with ethical AI practices.

1. Introduction

Generative AI has rapidly evolved in recent years, building on advancements in natural language processing (NLP) and machine learning. Here's a brief overview of its evolution:

Early Stages (1950s-1970s):

- Initial research in computational creativity and language generation.
- Notable figures like Alan Turing explore the possibility of machines exhibiting intelligent behavior.
- Early prototypes capable of simple text generation and rule-based creativity.

The Rise of Neural Networks (1980s-1990s):

- Introduction of neural networks and deep learning architectures.
- Early neural language models (NLMs) capable of basic language understanding and generation.
- However, these models lacked the sophistication and coherence of modern generative AI.

The Transformer Revolution (2017-Present):

- Introduction of the Transformer architecture, a neural network model revolutionizing NLP.
- Transformers enabled more efficient and contextually aware language processing.
- Generative AI models built on Transformers achieved significant improvements in text generation, translation, and summarization tasks.

Contemporary Advancements (2020-Present):

- Development of large language models (LLMs) with billions or trillions of parameters.
- LLMs like ChatGPT, Gemini, LLaMA, and others exhibit impressive language comprehension and generation capabilities.
- Generative AI applications expand into various domains, including code generation, image creation, and music composition.

Current Trends and Future Directions:

- Ongoing efforts to improve the quality, diversity, and ethics of generative AI systems.
- Exploration of multimodal generative models that combine text, image, and audio generation.
- Integration of generative AI into creative industries, customer service, and other real-world applications.

Comparing generative AI models is important for several reasons:

- **Evaluate model performance:** Developers and researchers can evaluate their relative performance on different tasks and identify areas where one model excels over others. This evaluation helps select the most suitable model for specific use cases.
- **Identify strengths and weaknesses:** Highlight their individual strengths and weaknesses. This information guides developers in improving existing models and addressing their limitations. It also enables users to make informed decisions when selecting a model for their projects.
- **Benchmark progress:** Comparison over time allows developers and researchers to track the progress and improvements in the field. This benchmarking helps measure the effectiveness of new techniques and algorithms and assess the overall advancement of generative AI technology.
- **Foster healthy competition:** Comparison fosters healthy competition among generative AI developers, encouraging them to continually improve their models and

push the boundaries of AI capabilities. This competition ultimately benefits the field by driving innovation and progress.

- **Inform decision making:** Provide valuable insights for users and organizations considering deploying these models in real-world applications. By understanding the relative performance, strengths, and weaknesses of different models, users can make informed decisions that are aligned with their specific needs and goals.

2. Methodology

2.1 Framework Definition

The prompt categories used in this study were carefully selected to evaluate the performance of generative AI systems across a wide range of tasks and domains. Each category represents a different aspect of language understanding and generation, allowing us to assess the capabilities of models in different scenarios.

1. **Creativity**: Prompts in this category are designed to test the models' ability to generate novel and original ideas, including writing short stories, poems, and song lyrics.
    1.1. ex. “Write rap song lyrics about an annoyingly loud geothermal construction site in your backyard.”
2. **Math**: These problems are used to assess the models' numerical reasoning and problem-solving skills.
    2.1. ex. “Find the eigenvalues and eigenvectors of a given matrix: [[13, 2, -18], [14, 1, -18], [10, 2, -15]]”
3. **Theory of Mind**: These prompts evaluate the models' understanding of human thoughts, beliefs, and intentions.
    3.1. ex. “Carl, Gustav, and Jung go to the same school. One day the teacher tells the students that the earth is 6000 years old. Gustav is sick that day and Carl is late. The next day, the teacher asks the students about the age of earth. What will Carl, Gustav, and Jung answer? Explain your reasoning.”
4. **Intuitive Physics**: Prompts in this category test the models' understanding of basic physical principles, such as gravity and motion.
    4.1. ex. “You have a rectangular piece of cardboard that you are holding up by all four corners. Now you let go of one corner, what happens? Explain your reasoning.”
5. **Idea Generation**: Prompts in this category are open-ended and designed to elicit a variety of creative ideas and solutions to problems.
6. **Translation**: Translation prompts evaluate the models' ability to translate text from English to German, Polish, Arabic and Hebrew while preserving meaning and context.

The rationale for including these categories is to provide a comprehensive evaluation of the performance of generative AI systems across different tasks and domains. By covering a wide range of scenarios, we aimed to gain insights into the models' strengths and weaknesses and to identify areas where they excel or need improvement.

2.2 Evaluation Method

Blind review process explanation:
In order to ensure an unbiased evaluation of generative AI systems, a blind review process was established involving six reviewers who were blinded to the specific models being evaluated. This approach was designed to mitigate any potential biases or preconceptions that the reviewers might have about particular models, thus ensuring a fair and unbiased evaluation process.

Grading scale and qualitative review guidelines:
The blind review process used a 1-10 rating scale, ensuring consistency through qualitative guidelines that emphasized creativity, coherence, accuracy, and response quality. The scale assessed correctness, completeness, logical reasoning, reasonableness of assumptions, and the presence of hallucinations or redundant statements in the responses.

NLP analysis approach for quantitative and qualitative data:
A comparative analysis of model performance across different prompt categories was performed, revealing the strengths and weaknesses of each model in specific areas. In addition, correlation analysis between scores and qualitative review comments, enhanced by sentiment analysis using the SentimentIntensityAnalyzer from the nltk library, helped identify the interplay between quantitative scores and the sentiment expressed in feedback. This analysis leveraged the SentimentIntensityAnalyzer's ability to assign sentiment scores to textual data, quantitatively assessing the positive, negative or Neutral sentiment within the reviewers' comments and providing a more nuanced understanding of the qualitative aspects of the review.

Further analysis:
A comparative analysis of the models' performance across different prompt categories was conducted, revealing the strengths and weaknesses of each model in specific domains or tasks. To further refine our findings, an average of all provided scores was calculated to allow for a more nuanced comparison. In addition, a correlation analysis was performed between these averaged scores and the qualitative review comments. The purpose of this analysis was to uncover the relationships between the quantitative scores and the qualitative feedback, thereby enhancing our understanding of the interplay between these two critical aspects of evaluation.

3. AI Models Overview

ChatGPT 3.5 & ChatGPT 4
OpenAI's flagship conversational AI models, known for their ability to generate human-quality text, translate languages, and answer questions in an informative manner. ChatGPT is based on the Generative Pre-trained Transformer (GPT) architecture. ChatGPT 3.5 (175 billion parameters) was a significant improvement, while ChatGPT 4 (1.76 trillion parameters) pushes the boundaries with greater contextual understanding and creativity. The model features improved conversational skills, better ability to follow directions, and more fact-based responses.

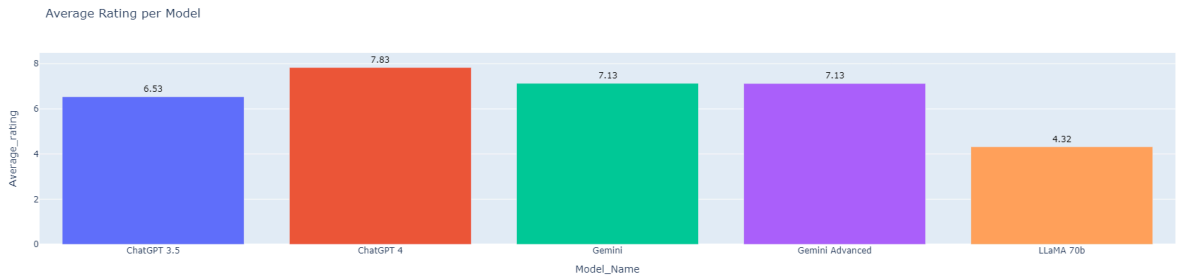
Gemini & Gemini Advanced
Google AI's conversational language models are designed to be informative and comprehensive. Gemini Advanced is at the forefront of Google's AI capabilities. It's built on Google's extensive research into language models and the massive datasets they're trained on. Leverages Google Search for real-time information. Its features focus on factual accuracy, the ability to access up-to-date information through Google Search, and comprehensive summaries. While Google hasn't publicly disclosed the exact parameter counts for its Gemini models some sources claim that Ultra has 175 trillion parameters, Pro has 50 trillion parameters, and Nano 10 trillion parameters.

LLaMA 70B
A basic language model from Meta AI (Facebook). Small in size compared to the others, but very efficient. It belongs to the class of large language models (LLMs) and is inspired by similar architectures. Its features are strong code generation capabilities and potential for fine-tuning for a wide range of tasks. Despite its relatively small size, it is remarkably competitive.

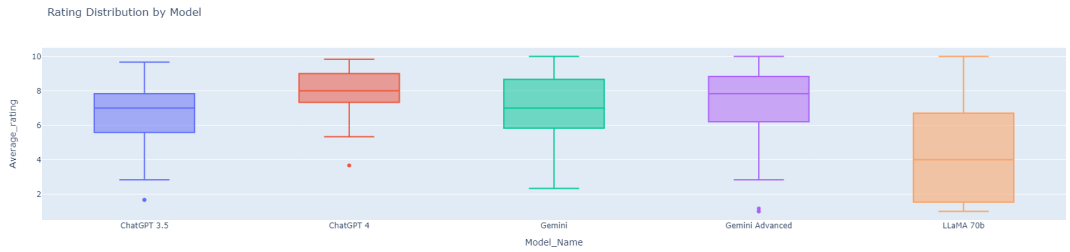
4. Results

4.1 Quantitative Analysis

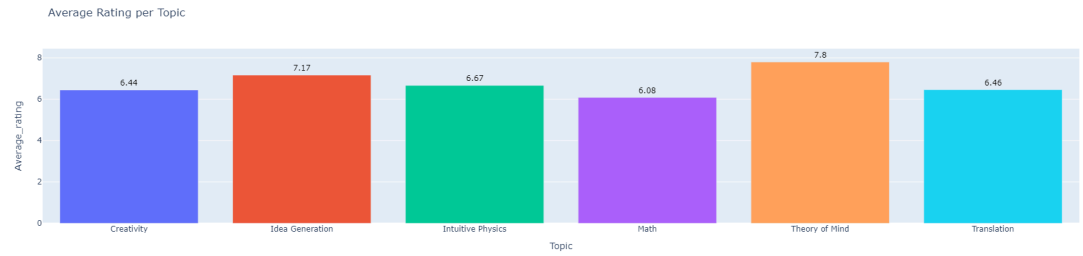
Average Rating per Model:
ChatGPT 4 is the highest rated AI model with an average rating of 7.83.
Gemini and Gemini Advanced are tied for second with an average rating of 7.13.
ChatGPT 3.5 follows with a rating of 6.53.
LLaMA 70b has the lowest average rating with 4.32.



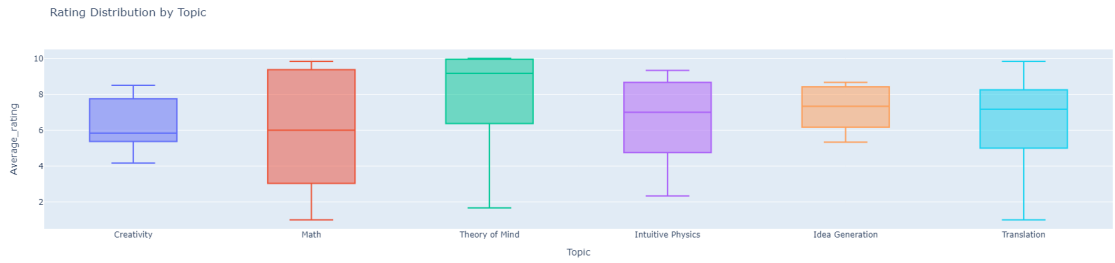
Rating Distribution by Model:
The box plots show the range and central tendency of the ratings for each model.
ChatGPT 4 has a narrower IQR and a higher median, indicating consistently high ratings. The Gemini models show a moderate spread with no extreme outliers.
LLaMA 70b has a wider spread and lower median, indicating more variability and generally lower ratings.



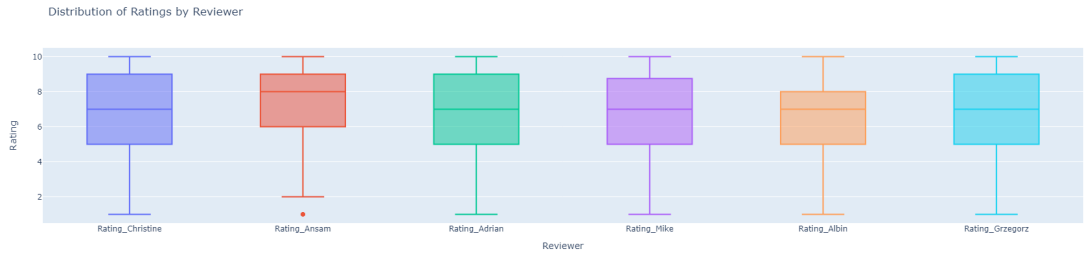
Average rating per topic:
The Theory of Mind topic has the highest average rating of 7.8.
Idea generation has a rating of 7.17.
Creativity has a rating of 6.44.
Intuitive Physics and Translation are at the bottom with ratings of 6.67 and 6.46 respectively.



Rating Distribution by Topic:
The box plot for each topic shows how the ratings are distributed. Theory of Mind and Translation appear to have higher medians and narrower spreads, indicating consistent and favorable ratings. Math and Intuitive Physics have wider spreads and lower medians, indicating less consistent ratings.

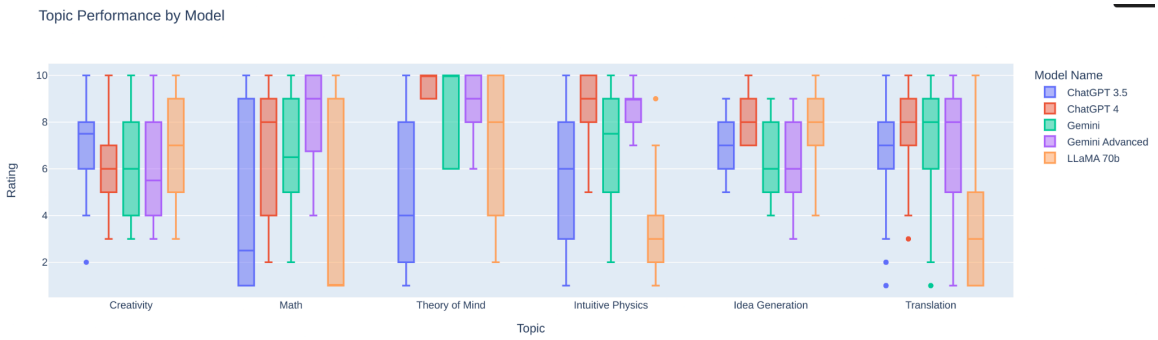


Distribution of ratings by reviewer:
The box plots show the distribution of the ratings given by each reviewer. For example, Christine has a median rating of about 8, with the majority of ratings falling between 6 and 9. Ansam has a slightly lower median and a smaller interquartile range (IQR), indicating more variability in the ratings. Adrian and Mike appear to have similar distributions. Albin and Grzegorz show higher variability with ratings spread across the scale from 2 to 10.

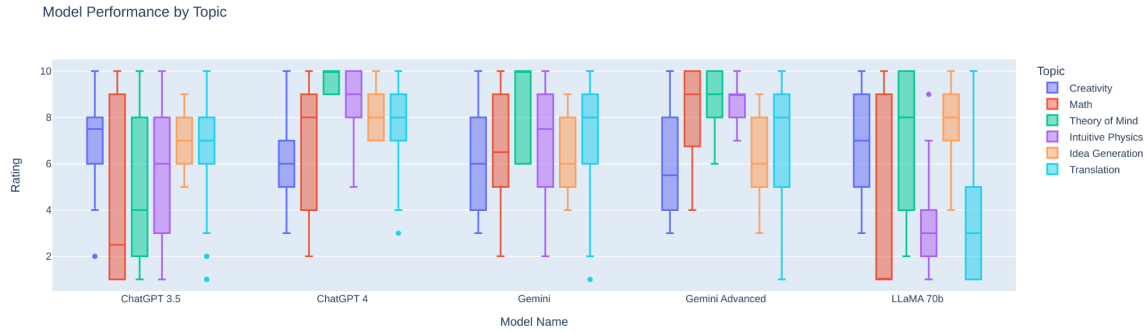


Model performance by topic:

These grouped box plots show the individual scores for each AI model across different topics. ChatGPT 4 appears to excel in Theory of Mind and Idea Generation. Gemini Advanced has high ratings in Theory of Mind and Intuitive Physics. LLaMA 70b has its strengths in Theory of Mind and Idea Generation, but lacks considerably in translation capabilities. Each model shows varying performance in different topics, indicating specialization in certain areas.



By looking at the same data grouped by models, the high ratings and low spread (high consistency) for ChatGPT 4 across topics becomes apparent. While ChatGPT 4 outperforms 3.5, Gemini Advanced is surprisingly close to Gemini. By contrast, LLaMA 70 b's performance varies highly across topics.



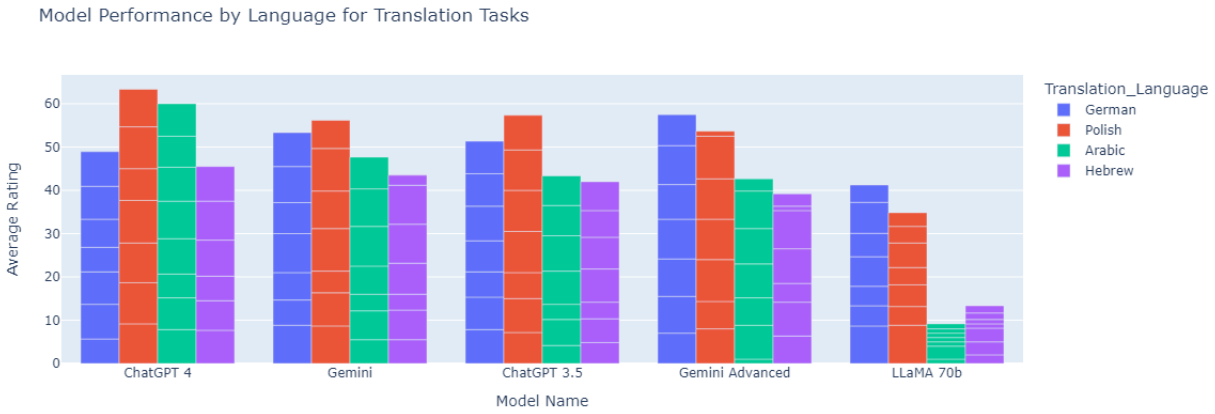
We performed a two-way ANOVA (6 topics by 5 models) which showed highly significant main and interaction effects ($p < 0.001$).

	df	sum_sq	mean_sq	F	PR(>F)
Model_Name	4.0	1876.583287	469.145822	88.004377	5.705176e-66
Topic	5.0	201.192764	40.238553	7.548120	5.388631e-07
Model_Name:Topic	20.0	948.137353	47.406868	8.892783	2.634155e-25
Residual	1256.0	6695.657358	5.330937	NaN	NaN

Models therefore differ in overall performance across topics, and show individual strengths and weaknesses per topic.

Model Performance by Language for Translation Tasks

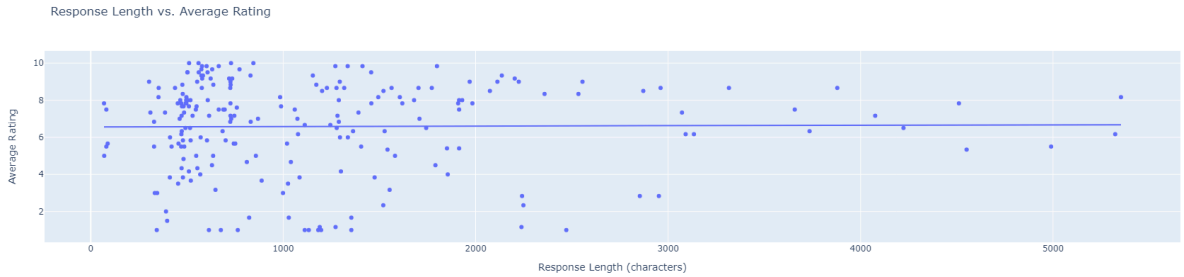
We observe a pattern in the average scores that describe the effectiveness of each model per language. In particular, ChatGPT 4 excels in Arabic translation compared to German, which is a departure from the trend of most models performing better in Latin-derived languages. In sharp contrast to these results, LLaMA 70b's translation skills are distinctively poor, particularly in Semitic languages such as Arabic and Hebrew, where it received the lowest score of 1, reflecting a complete failure to translate to these languages. These data suggest that while some models have made progress in handling the complexities of different language families, others, such as LLaMA 70b, struggle significantly, highlighting the importance of specialized training for different language structures.



Model Performance by Translation Type:

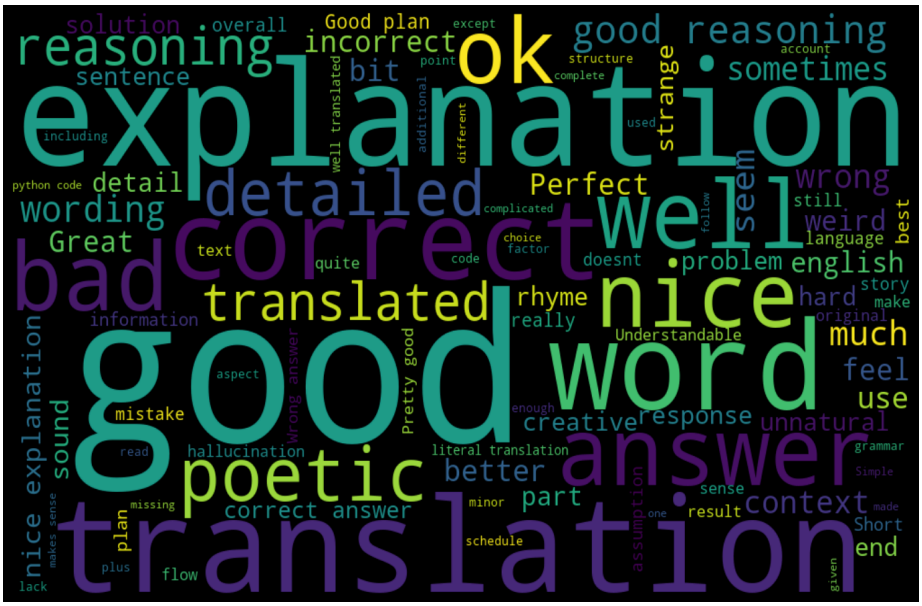
The prompts were assessed by analyzing three brief paragraphs, which were classified into three categories: English C1 level vocabulary, poetry, and scientific writing. In addition, all sorts of paragraphs were consolidated into a single prompt for a comprehensive translation exam (referred to as "Translating all together"). ChatGPT 4 and 3.5 demonstrate the strongest performance in the "Translating all together" category, suggesting they possess a comprehensive translation capability across a variety of text types. Gemini maintains a steady performance across all individual categories. In contrast, Gemini Advanced excels in translating poetry and scientific text. However, Gemini Advanced's performance dips significantly in the "Translating all together" category, possibly due to completely failing to translate some texts. Meanwhile, LLama's performance is generally lower when compared to the other models.

Response Length vs. Average Rating:
The scatter plot shows the relationship between the length of responses and the average rating given. There is a horizontal line indicating the average rating around which most of the data points are clustered. There doesn't seem to be a clear trend or correlation between response length and average rating. The responses are spread across the range of lengths and ratings.



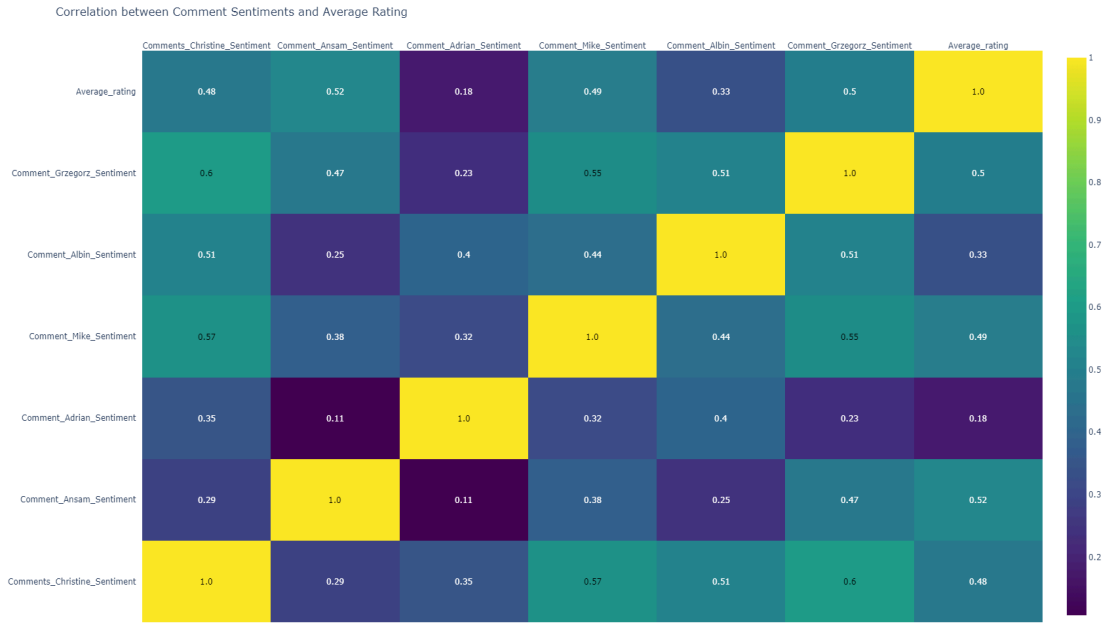
4.2 Qualitative Analysis

Word cloud:
The word cloud visualizes the frequency of words used in reviewers' comments. Larger words such as "detailed," "good," and "incorrect" indicate that these terms were used more frequently. This variety of words suggests a range of feedback, from positive (e.g., "perfect," "great," "nice") to critical (e.g., "wrong," "bad," "mistake").

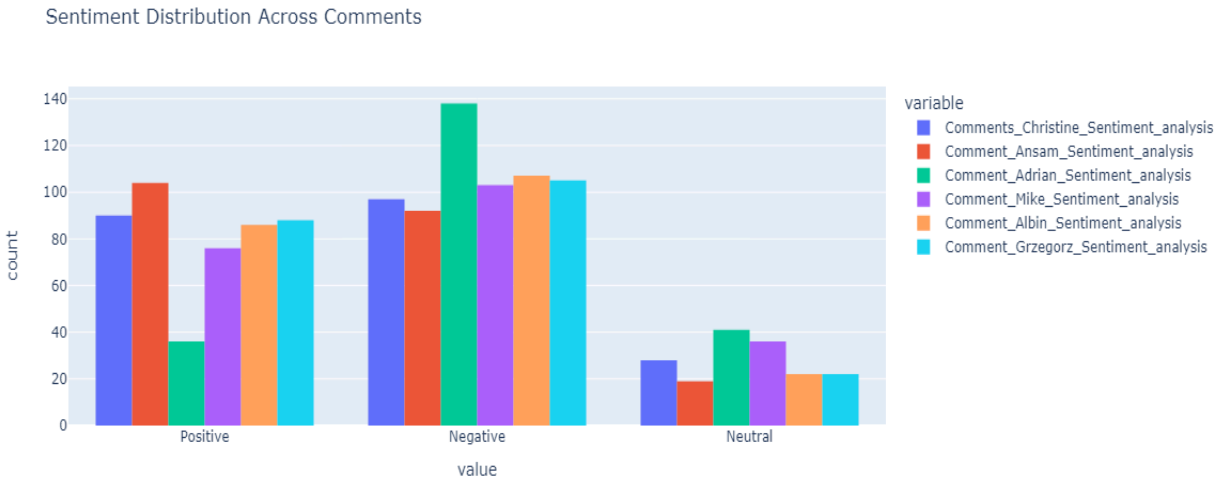


Correlation between comment sentiment and average rating:
The heat map shows the correlation between the sentiment scores of each reviewer's comments and the average ratings given. There are moderate positive correlations in some cases, such as between Christine's sentiment scores and average ratings (0.48). Ansam's sentiment scores have a strong correlation with average ratings (0.52), indicating that more positive sentiment is often

associated with higher ratings



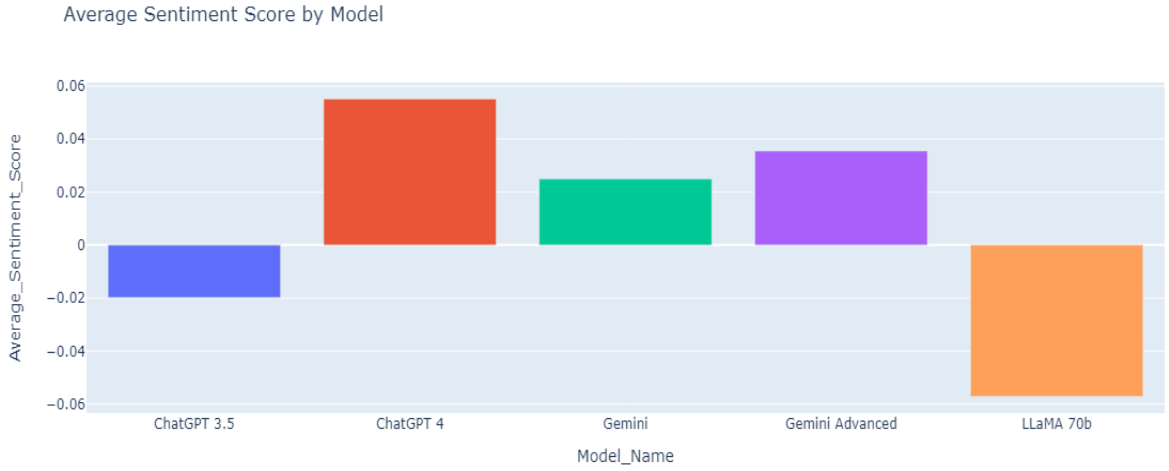
Sentiment distribution across comments:
The bar chart shows the number of positive, negative, and neutral comments for each reviewer. Ansam has the highest number of positive sentiment comments. Adrian has a notably high number of negative sentiment comments as well as fewer positive comments than the other reviewers. Neutral sentiments are fairly evenly distributed among the reviewers.



Sentiment Score Distribution:
This graph shows the distribution and density of the sentiment scores from each reviewer's comments. Comments from Christine, Ansam, Adrian, Mike, Albin, and Grzegorz show different distributions, with Christine's comments tending to have a more positive sentiment. The width

of each violin indicates the density of comments at different sentiment levels, and most seem to be centered around zero.

Average Sentiment Score by Model:
ChatGPT 4 has the highest average sentiment score, just above zero, indicating a generally positive reception. Gemini has a positive sentiment score, but lower than ChatGPT 4. Gemini Advanced has average sentiment scores close to zero, but on the positive side. ChatGPT 3.5 and LLaMA 70b have a negative average sentiment score, indicating a more critical perception by reviewers.



5. Discussion

Interpretation of results:
The study evaluated the capabilities and limitations of today's generative AI models, including ChatGPT 3.5, ChatGPT 4, Gemini, Gemini Advanced, and LLaMA 70b. ChatGPT 4 in particular

demonstrated its capabilities, achieving the highest scores in theory of mind and idea generation. These results indicate a refined understanding of complex human concepts and a flair for creativity, possibly due to its sophisticated training and extensive data ingestion. The Gemini suite, which includes standard and advanced variants, demonstrated robust performance, particularly in harnessing current information and generating comprehensive responses which reflects Google AI's ability to leverage large datasets. LLaMA 70b, while trailing in overall scores and showing marked weaknesses in translation tasks, demonstrated proficiency in select areas such as idea generation and theory of mind. This highlights the benefits of efficiency and tailored training, even within the constraints of a more compact model.

Model strengths and weaknesses in various categories:

- ChatGPT demonstrates its ability to handle complex prompts that require deep understanding and creative output, due to its advanced training and large dataset. However, this extensive knowledge can occasionally produce responses that appear generic when applied to specialized tasks that require a more tailored approach.
- The Gemini models excel at providing factual information and leveraging real-time data, making them invaluable for applications that require up-to-date knowledge. Their performance may be limited by the inherent challenge of maintaining creative and nuanced responses compared to models such as ChatGPT 4.
- LLaMA 70b, although smaller in scale, is notable for its performance in specific domains such as idea generation. This model demonstrates that targeted efficiency and focused training can be highly effective. However, it lags behind larger scale models in broader applications and depth of knowledge, particularly in translation for languages such as Arabic and Hebrew. This underscores the need for specialized training of models to effectively handle the intricacies of different linguistic structures.

Implications of findings for developers and users:

For developers, these findings emphasize the importance of selecting models based on the specific needs of their application. A balance between model size, efficiency, and task specificity must be considered to optimize performance and user satisfaction. The findings also underscore the potential benefits of hybrid approaches that combine the strengths of different models to achieve superior performance across a broader range of tasks.

For users, understanding the different capabilities of each model can lead to more informed choices, especially when integrating AI into workflows, products, or services. Users looking for creative content generation may prefer ChatGPT 4, while those looking for up-to-date factual information may prefer Gemini Advanced.

Limitations of the study and evaluation framework:

While this study provides valuable insights into the comparative performance of AI models across different prompt categories, it acknowledges certain limitations in its evaluation framework and methodology. The blind review process is intended to reduce bias, but cannot completely eliminate subjective interpretations of the models' results. In addition, the evaluation's focus on specific categories may not fully capture the comprehensive capabilities or shortcomings of individual models, particularly in emerging or niche areas.

Another notable limitation is the inherently dynamic nature of AI development. As models continually evolve, the introduction of new versions can significantly alter the competitive landscape, making this study a temporal snapshot in a rapidly advancing field. Ongoing research and evaluation are critical to keeping abreast of these advances.

In addition to these considerations, it's important to recognize that each prompt category places different demands on the rating scale. For example, the criteria for scoring a two-week travel itinerary are very different from those for scoring a haiku. This variation in scoring requirements could have been more finely tuned at the outset of the study to account for the nuanced differences between tasks, ensuring a more tailored and accurate assessment of model performance across different applications.

6. Conclusion

Summary of Key Findings

Our comparative analysis of leading generative AI systems-ChatGPT 3.5, ChatGPT 4, Gemini, Gemini Advanced, and LLaMA 70b-illuminates the current AI landscape, revealing distinct capabilities across different prompts. ChatGPT 4 excels at generating creative responses and understanding complex human concepts, earning the distinction of being the top-ranked model in our study. Both Gemini and Gemini Advanced excel at delivering accurate information, a testament to the power of extensive data usage and live data integration. Meanwhile, despite a modest overall score, LLaMA 70b is recognized for its efficiency and effectiveness in specialized tasks, particularly in idea generation and theory of mind.

Overall model ranking based on performance

- ChatGPT 4 - Leader with the highest average score, particularly strong in Theory of Mind and Idea Generation.
- Gemini and Gemini Advanced (Tied) - Strong performance in providing factual, up-to-date information, with a slight edge for Gemini Advanced in certain tasks.
- ChatGPT 3.5 - Solid performance, but surpassed by its successor in handling complex and creative prompts.
- LLaMA 70b - while lower in the average ranking, shows pronounced translation difficulties, but also demonstrates the effectiveness of a focused and streamlined model architecture.

Recommendations for future research and comparative frameworks

- Rather than relying solely on average scores, future research can provide a more nuanced and comprehensive assessment that better captures the complex performance landscape of AI models.
- Broader evaluation criteria: Future research should consider expanding the evaluation criteria to include emerging AI capabilities and application-specific performance metrics to provide a more comprehensive understanding of each model's utility across a broader range of tasks.

- Dynamic Benchmarking: As AI models rapidly evolve, establishing a dynamic benchmarking system that regularly updates evaluation data will provide more timely insights into model performance and advances in the field.
- Multimodal and Specialized Models: With the emergence of multimodal AI systems and models that specialize in specific domains (e.g., medical, legal), future comparisons should include these categories to account for the growing diversity and specialization within the AI landscape.
- Ethical and societal impacts: Include assessments of the ethical considerations and societal impacts of models, particularly with respect to bias, fairness, and misinformation, to ensure responsible AI development and deployment.
- Open collaboration and benchmarking: Encourage open collaboration between academia, industry, and independent researchers to develop standardized benchmarking frameworks that facilitate fair, transparent, and reproducible comparisons of AI models.

7. References

1. nltk: https://www.nltk.org/
2. Textblob: https://textblob.readthedocs.io/en/dev/
3. wordcloud: https://amueller.github.io/word_cloud/
4. ChatGPT: https://openai.com/blog
5. Google AI's language models: https://blog.research.google/
6. Meta AI's LLaMA: https://ai.meta.com/blog/code-llama-large-language-model-coding/
7. Evaluating Large Language Models: A Comprehensive Survey: https://arxiv.org/pdf/2310.19736.pdf