

Framework for Comparison of Different LLM System Providers

Course: Generative AI, Spring Semester 2024

Christine
Ansam
Adrian
Michael
Albin
Grzegorz

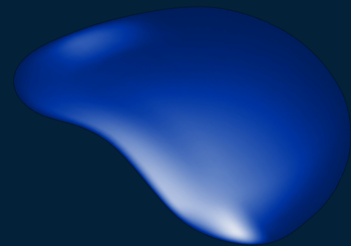
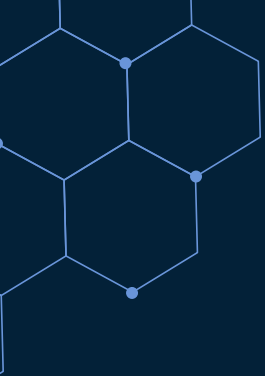


TABLE OF CONTENTS

01	SYSTEM PROVIDERS
02	TOPICS & PROMPTS
03	GRADING SCALE
04	FRAMEWORK
05	QUANTITATIVE RESULTS
06	QUALITATIVE RESULTS
07	INSIGHTS
08	CONCLUSION & OUTLOOK

01 SYSTEM PROVIDERS



**Chat
GPT 3.5**



**Chat
GPT 4**



**Gemini
(free)**



**Gemini
Advanced**



**LLaMA
70b**

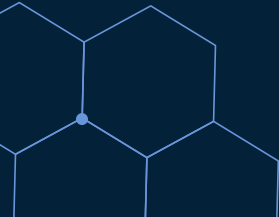


02

TOPICS & PROMPTS



CREATIVITY
MATH
THEORY OF MIND
INTUITIVE PHYSICS
IDEA GENERATION
TRANSLATION



...Write a Haiku about a fork

... Find the eigenvalues and eigenvectors of a given matrix:
[[13, 2, -18], [14, 1, -18], [10, 2, -15]]

... Andreas secretly likes Isabelle. Isabelle owns a very jealous dog. The dog's owner is easily upset. One day, Andreas sees Isabelle on a date with Peter. Who is upset, Peter, Andreas, Isabelle, or the dog? Explain your reasoning.

03 GRADING SCALE

1. SCORE OF 1-10

2. SHORT WRITTEN REVIEW

Taking the following structure into account:

- Correctness of the answer
- Completeness of the answer
- Logical coherence of the reasoning
- If there are additional assumptions made, are they reasonable or not
- Are there any hallucinations or unnecessary statements in the response

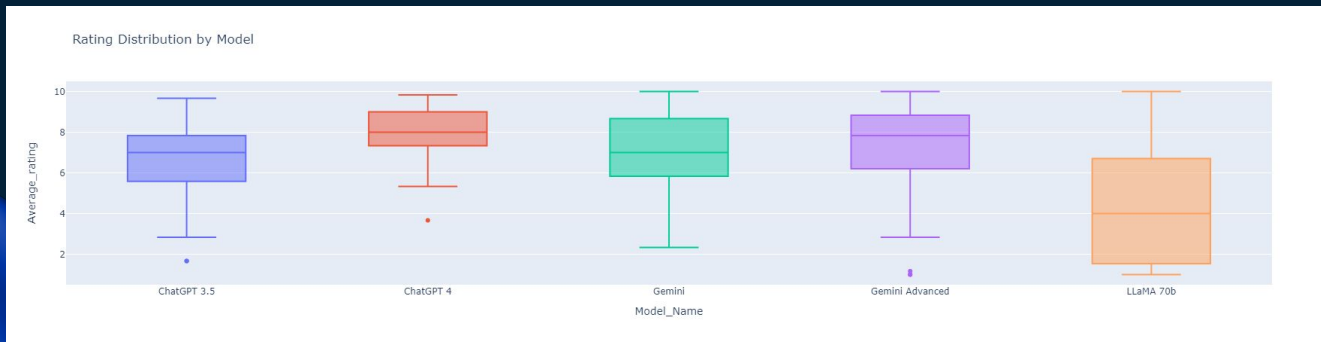
04

FRAMEWORK



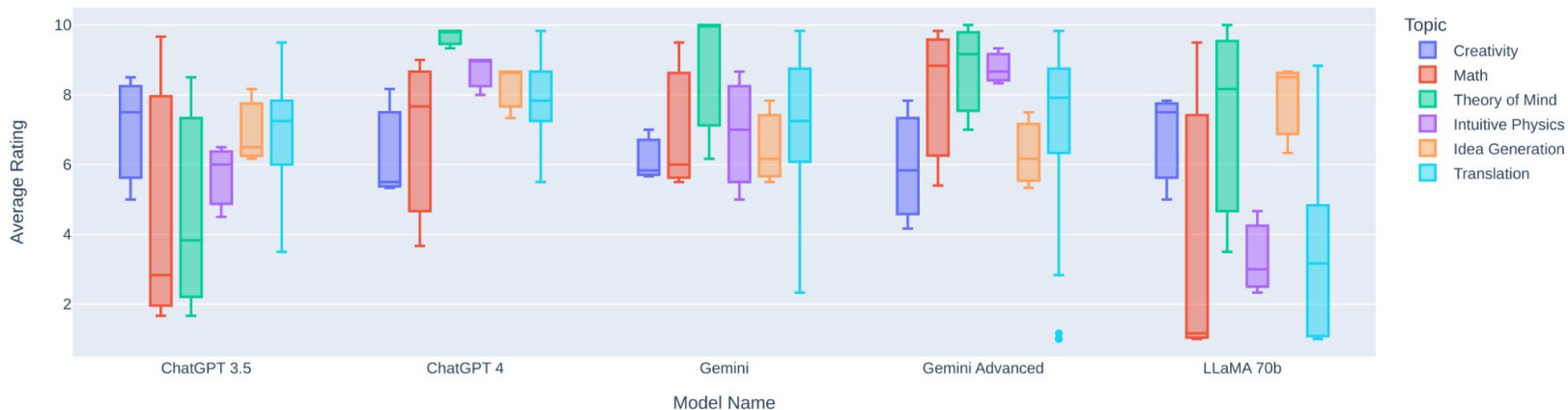
Prompt	Model	Actual Model (Fill in after Rating is complete)	Response	Rating Christine	Comments Christine
Write a Haiku about a fork	1	ChatGPT 3.5	Silver prongs reach out, Guiding food from plate to mouth, A dance on the tongue.		8 Simple yet efficient
	2	ChatGPT 4	In hands it dances, Tines whisper through the air's song, Feast's silent partner.		4 Confusing
	3	Gemini	Curved metal tines gleam, Separates bites, a single path, Shared meal, two become one.		4 Confusing
	4	Gemini Advanced	Metal tines curve up, Ready to capture and lift, Tiny spear for food.		8 Cute
	5	LLaMA 70b	Tines raised high Food's best friend, a humble fork Brings nourishment		6 Confusing

05 QUANTITATIVE RESULTS

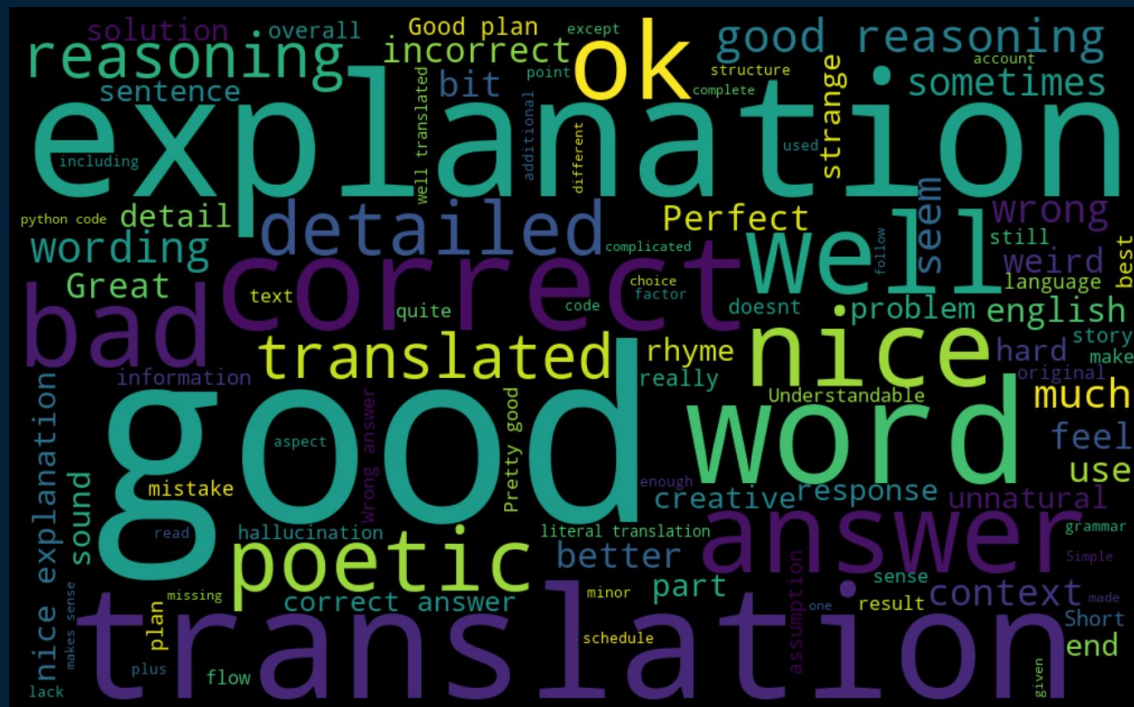


05 QUANTITATIVE RESULTS

Model Performance by Topic



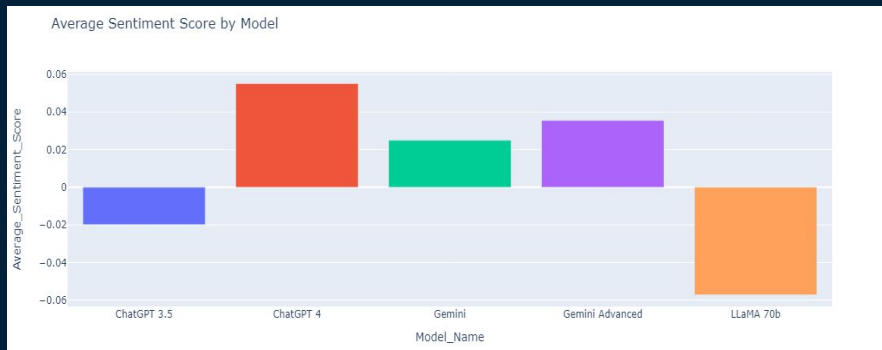
QUALITATIVE RESULT



06

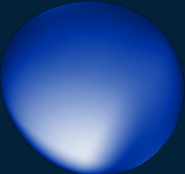
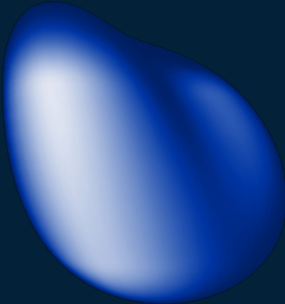
QUALITATIVE RESULT

x



07 INSIGHTS

Model strengths & weaknesses

- 
- ChatGPT demonstrates its ability to handle complex prompts that require deep understanding and creative output
 - The Gemini models excel at providing factual information and leveraging real-time data
 - LLaMA 70b, although smaller in scale, is notable for its performance in specific domains such as idea generation.
 - lags behind larger scale models in broader applications and depth of knowledge, particularly in translation for languages such as Arabic and Hebrew.
- 

08

CONCLUSION & OUTLOOK

- Broader evaluation criteria, Larger sample size
- Multimodal and Specialized Models
- benefits of hybrid approaches that combine the strengths of different models to achieve superior performance across a broader range of tasks.
 - Users looking for creative content generation may prefer ChatGPT 4, while those looking for up-to-date factual information may prefer Gemini Advanced.



QUESTIONS?