LLM Evaluation

Version 2.0

Kazi Rahimu Islam

Prepared for Prof David Smith

Date [04-17-2025]

Abstract

This report presents a comparative evaluation of three prominent AI language models—ChatGPT, DeepSeek, and Claude AI—focusing on their ability to summarize complex topics from diverse domains such as technology, economy, and politics. The evaluation is based on summaries generated from five selected articles, each limited to 300 words. Key criteria used for assessment include text summarization, data extraction, factual accuracy, completeness, usefulness, and overall strengths and weaknesses. Findings reveal that Claude AI consistently outperforms the other models by offering structured, accurate, and comprehensive summaries suitable for a wide audience. ChatGPT shows strength in synthesizing information but struggles with depth and nuanced content. DeepSeek, while accessible for general readers, lacks robustness in dealing with politically sensitive topics and tends to oversimplify complex materials. This analysis offers insights into the strengths and limitations of current AI models and highlights the importance of transparency and contextual sensitivity in language model deployment.

Objective:

This document compares the performance of three AI language models—ChatGPT, DeepSeek, and Claude AI—in summarizing articles related to technology, economy, and politics. With a 300-word limit for each summary, the models were evaluated based on criteria such as text summarization, data extraction, accuracy, completeness, and usefulness. The goal was to assess how effectively each model condenses complex topics, extracts key information, and presents it clearly and accurately. Based on my evaluation (Kazi Islam), Claude AI emerged as the most balanced and effective model, excelling in all areas, particularly in providing well-structured and accurate summaries. In comparison, GPT showed inconsistencies in-depth and occasional omissions, while DeepSeek oversimplified content and struggled with politically sensitive topics, especially those critical of the Chinese government. This analysis draws from a range of sources, including articles from Forbes, Nvidia, USC Dornsife, Webisoft, and the U.S. Department of State, ensuring a thorough evaluation.

Literature Review

This evaluation draws on summaries generated by three major AI language models—ChatGPT, DeepSeek, and Claude AI—based on five selected articles covering topics in technology, economy, music and math, the metaverse, and diplomacy. The sources include articles from Forbes, Nvidia, USC Dornsife, Webisoft, and the U.S. Department of State. These articles served as consistent benchmarks to test how each model performs under identical conditions.

Each model was assessed using a uniform set of criteria: text summarization, data extraction, accuracy, completeness, usefulness, and an analysis of strengths and weaknesses. The evaluation revealed notable differences in how each model handles complex content. Claude AI stood out for its ability to condense detailed information while maintaining accuracy and structure. ChatGPT showed strength in synthesizing information from multiple sources but lacked depth in more technical areas. DeepSeek performed adequately with general content but failed to summarize politically sensitive topics, highlighting limitations in its response capabilities.

This internal comparison provides a structured, hands-on assessment of LLM capabilities and limitations based entirely on controlled testing using real-world articles, without relying on external frameworks or prior studies.

1. Article Selection:

The first step involved gathering five diverse articles that cover topics related to technology, economy, and politics. These articles served as the input for the AI language models. The sources selected include:

- **Economy**: An article from Forbes on the economic impact of Biden-Harris vs. Trump.
- **Technology**: A blog from Nvidia on enabling quantum computing with AI.
- **Music and Math**: A study on the relationship between music and math from USC Dornsife.
- **Metaverse**: A comparison of the multiverse vs. metaverse from Webisoft.
- **Diplomacy**: An article from the U.S. Department of State about diplomacy.

2. Setting the Criteria:

Once the articles are chosen, a set of evaluation criteria is established. The models are assessed based on:

- **Text Summarization**: How well the model structures and condenses the content.
- **Accuracy**: Accuracy of key points and data.
- **Completeness**: How factually complete the summaries are, with attention to omissions.
- **Reasoning:** Ability to identify relationships, patterns, and logical structures.
- **Analysis:** Depth of interpretation and ability to extract meaning, significance, and implications.
- **Hallucination:** Creation of false information and fabricated details.

3. AI Model Summary Generation:

Three AI models—ChatGPT, DeepSeek, and Claude AI—are used to generate summaries for each article. Each summary was capped at 300 words. These models are asked to produce summaries without additional instructions or bias, ensuring a fair comparison.

4. Evaluation and Grading:

Once the summaries are generated, each AI model is evaluated according to the set criteria. The evaluation involves assessing how well each model:

- Condensed the information into the word limit while maintaining clarity and structure.
- Extracted relevant data and concepts from the articles.
- Ensured the summaries are factually accurate and comprehensive.
- Provided a clear, organized, and useful summary for a general audience.

Criteria	ChatGPT	DeepSeek	Claude AI
Text Summarization	Structured and clear, but occasional oversimplification.	Strong overall, but occasional inconsistency in depth	Excellent condensation of complex topics, maintaining key points and structure
Accuracy	Factually sound with minor omissions.	Generally aligns well with sources, minor omissions	Summaries closely align with original content, no apparent errors
Completeness	Consistent coverage but lacks depth in niche areas.	Covers main points, sometimes misses nuanced details	Covers main points comprehensively, minor omissions in some areas
Analysis	Struggles with deeper interdisciplinary connections.	Shows better depth than ChatGPT but varies by domain.	Exceptional analytical depth, identifying subtle connections and maintaining contextual awareness
Reasoning	Occasionally makes leaps in logic. Can conflate correlation with causation	Strong technical reasoning though sometimes skips steps in explanations	Exhibits superior reasoning with clear, evidence-based justifications.
Hallucination	More prone to minor fabrications or misattributions, especially in detailed scenarios. Requires fact-checking.	Rarely fabricates information, showing good factual discipline.	Low hallucination rates with only occasional minor inaccuracies.

5. Final Analysis and Conclusion:

According to me (Kazi Islam), Claud AI is the best of them all as it is more balanced in all the criterias because it excels across the board, particularly in summarizing text, extracting key information, and maintaining accuracy. Claude is highly effective at organizing content, presenting data clearly, and providing a balanced perspective. GPT can be inconsistent in its depth of coverage, occasionally omitting important details. DeepSeek, on the other hand, tends to oversimplify and lacks consistency. Deepseek has a bigger issue, it doesn't provide any summary or response for political and controversial topics which seem to criticize the Chinese government.

6. Version History

Version	Created on	Created by
1.0	3/19/2025	Kazi Islam
2.0	4/17/2025	Kazi Islam