RESEARCH ARTICLE

An NLP-Based Evaluation of LLMs Across Creativity, Factual Accuracy, Open-Ended and Technical Explanations

Qazi Novera Tansue Nasa[1,*] **and Ashik Chandra Das**[1]

[1] Institute of Computer Science, University of Potsdam, Potsdam 14476, Germany

Abstract

The rapid advancement of AI-based language models has transformed the field of Natural Language Processing (NLP) into a powerful tool for text generation. This study evaluates the performance of models in different categories such as factual accuracy, creative writing, open-ended writing, and technical explanation. We have considered three popular and advanced large language models (LLMs) for this analysis. To quantify their performance, we have applied a combination of statistical and linguistic metrics. We have used Dale-Chall to analyze the readability score of the responses. For lexical diversity, we have used the type-token ratio technique. In addition, a cosine similarity with TF-IDF is used for semantic similarity. Furthermore, sentiment polarity and grammatical correctness are also analyzed. Moreover, we have conducted an F-test to determine whether the differences in performance among the LLMs are statistically significant ($p < 0.05$). We have found minimal differences between LLMs, with ChatGPT showing slightly better performance compared to the others.

Keywords: LLMs evaluation, NLP, ChatGPT, Gemini, DeepSeek, ANOVA.

1 Introduction

Artificial intelligence is one of the fastest growing terms in the current time. Users employ these techniques for various purposes, including creative writing and academic research, with text generation being one of the key applications. People normally use large language models (LLMs) like ChatGPT, DeepSeek, Gemini etc. for generating text. Each of these models has its own strengths in areas such as creativity, factual accuracy, open-ended reasoning, and technical explanation. Although these types of models are used for generating human-like responses, there is a question on their outputs about accuracy [1]. These models perform differently depending on how they are tested. This makes questions about how dependable they are and how useful they are. In addition, it asks questions about how they generate responses.

From an applied perspective, the inconsistent performance of LLMs makes it difficult to use them confidently in real world situations. As an example,

Submitted: 10 November 2025
Accepted: 01 December 2025
Published: 01 February 2026

Vol. 3, **No.** 2, 2026.
10.62762/TETAI.2025.264517

***Corresponding author:**
✉ Qazi Novera Tansue Nasa
nasa.novera@gmail.com

Citation
Nasa, Q. N. T., & Das, A. C. (2026). An NLP-Based Evaluation of LLMs Across Creativity, Factual Accuracy, Open-Ended and Technical Explanations. *ICCK Transactions on Emerging Topics in Artificial Intelligence*, 3(2), 76–85.

ChatGPT is known for its smooth writing and creative generation. But studies show that it sometimes gives wrong information or 'hallucinates' [2]. DeepSeek is a new model that says that it is better for finding exact information [3]. But we do not know much about its creativity as empirical studies are less focused. On the other hand, Gemini is designed to understand different types of information and is claimed to be both creative and accurate [4]. However, its performance with various prompts has not been fully studied. These variations in model performance directly impact industries that depend on AI-generated content such as journalism, education, and policymaking [5].

LLMs can create creative, factual, technical explanation and open-ended responses. Focusing on these response mode enables us the adaptability and flexibility of LLMs. Also we can identify their thinking patterns. But the important questions are about how they generate, how we can trust them and whether they can truly be creative [6]. As the models are trying to provide original results, making sense, and fitting with the situations, it is very difficult to measure these responses [7]. The accuracy of these contents depend on its training data, how it finds information and how it fits human knowledge [8]. Moreover, handling open-ended prompts requires accurate thinking, understanding the context which challenges how we understand [19]. For these challenges, we need to compare LLMs to find how well they handle these different tasks.

Our objective for this study is to find a comparative analysis of the popular LLMs particularly ChatGPT, DeepSeek and Gemini. We analyze the text, generated from these three models of three different categories. These categories are described as: Creativity: Uniqueness, logical flow and innovation of AI content [9]. Factual Accuracy: Evaluating accuracy and correctness from various areas of knowledge [2]. Open-Endedness: Analyzing the ability of AI models to handle abstract prompts with contextually appropriate outcomes [10]. Technical Explanation: Ability analysis of AI models in terms of explaining technical knowledge that connect technology to reality.

To achieve these objectives, we have reviewed existing literature on AI-generated text, including creativity metrics, factual accuracy evaluations, and assessments of open-ended reasoning. Then our methodology explains the detailed dataset, evaluation criteria, and experimental setup. Including this, we explain the metric of the parameters to analyze. The result and discussion section provides the outcome of the analysis with a proper discussion. After that, we conclude this with our future steps. By studying the results from the three models, this research helps to understand the level of LLMs in text generation. Researchers, developers and users can get knowledge to decide which model performs better in which case.

2 Theoritical Background

In the field of Natural Language Processing (NLP), artificial intelligence is growing very fast. By using NLP, one can make computers for understanding and generating human language. Recently, some large language models are very popular which mainly provide text-based responses [11]. From them ChatGPT, DeepSeek, Gemini is the most favorite. These types of models are built on transformer-based architecture. Also, these models have significantly improved machine understanding of language. LLMs are trained on large datasets including books, articles, websites, and code. This training model helps a user to learn grammar, logic, coding etc. These models are called "large" because of using billions or even trillions of parameters. These parameters together with weights are used for training the neural network. Parameters help the model "remember" patterns in the training data. The model's performance depends on the complexity of the model. If the model has more parameters, it can handle more complex languages [1]. Transformer architecture enables these models to capture long-range dependencies in sequential data [11].

OpenAI's creation, ChatGPT, is a family of models distinguished by its conversational abilities and adaptability to various tasks. According to the documentation, it can perform data analysis, document understanding, and integrating with external tools etc. The latest model GPT-4 is trained on huge datasets and refined for instruction by following general language processing [12]. Gemini is developed by Google. It is a multimodal model which is capable of processing text, images, audio, and video. From its documentation, it has advanced reasoning capabilities ("thinking models"), document understanding, and integration with Google Cloud services. The Gemini family includes various models which are used for different needs, from speed to complex reasoning [4]. DeepSeek is developed by DeepSeek AI. It offers models like DeepSeek-V3 (a generalist model) and DeepSeek-R1 (focused on reasoning and coding). Its efficiency, long context

window, and context catching are the main features. These models are designed to handle complex reasoning and technical tasks effectively [3].

The quality of the output of a language model can be analyzed based on several theoretical frameworks. Models like GPT-4 often score highly in grammar, coherence, and fluency. Meanwhile, open models such as LLaMA have shown strong performance in similar dimensions while being more accessible for research [13]. Task Alignment is used to measure how well the model's output matches the intent of user. This refers to how effectively each model can perform tasks based on given instructions [14]. LLMs can take over biases from their training data by leading to unfair or discriminatory outputs [15]. Response analysis can help to identify and mitigate these biases. Besides, evaluating responses for harmful content, misinformation, or hallucinations is necessary for ensuring the safe and reliable deployment of LLMs [16]. Some researchers also focus on information quality. Like factuality is the degree to which the response is grounded in evidence and avoids hallucinations [17]. There may be the presence of unfair or discriminatory outputs. These are related to protected attributes like race, gender, religion [18]. By studying these models comparatively, a researcher can better understand their strengths and weaknesses. These types of analysis help a user to choose the best model for their respective work.

3 Methodology

This research compares how well ChatGPT, DeepSeek and Gemini provide responses for the same prompts with the factors like creativity, accuracy, readability, and grammar. To conduct this study, first we collect responses (generated text from each) as data then evaluate these responses by using statistics and qualitative analysis. This analysis indicates which model is the best option for various purposes of prompts.

3.1 Data Collection

We have used 2 different stages. Each stage has different prompts. These prompts are categorized into four types. We have set our first stage prompts as: Creativity: "Write a short science fiction story about Mars colonization." Factual Accuracy: "Write the key difference between AC and DC current." Open-ended Discussion: "What are the ethical implications of AI in journalism?" Technical Explanations: "Explain the concept of blockchain technology in simple term." And our second stage prompts are as follows: Creativity: "Write a poem about space exploration." Factual Accuracy: "Write the benefits of renewable energy." Open-ended Discussion: "What are the problems associated to design a futuristic city?" Technical Explanations: "Describe the step-by-step process through which transactions are validated and recorded on a blockchain, highlighting key technical components involved." These prompts are designed to simulate real world tasks. They require imagination, precision, critical thinking and technical understanding to find responses.

3.2 Data Analysis

We have submitted each prompt simultaneously to ChatGPT 4.0, DeepSeek and Gemini to ensure the proper testing condition. The same machine is used for three different models. We have recorded the response time and the responses in a structured way for further analysis. Both quantitative and qualitative metrics are used to find an acceptable outcome. The quantitative metrices are in Table 1.

Table 1. Short description of the metric and corresponding tools in Python.

Metric	Description	Tool Used
Readability	Dale-Chall scores measure response clarity and readability.	textstat library
Vocabulary Richness	Type-Token Ratio (TTR) evaluates vocabulary diversity.	Python (NLTK)
Coherence	TF-IDF Similarity measures logical flow and relevance.	sklearn library
Sentiment Polarity	Determines whether the response is positive, negative, or neutral.	TextBlob
Grammar and Spelling	Counts the number of grammatical errors.	Language Tool

To conduct this analysis, we have used different Python libraries such as: NLTK, TextBlob, Scikit-learn, Textstat, LanguageTool. For inferential statistics, we have used F-test to determine statistically significance in performance differences.

4 Result

In this section, we discuss the findings from the analysis. First, we discuss the time taken by each

Table 2. Response generation time (in seconds) of ChatGPT 4.0, DeepSeek-R1, and Gemini across various tasks and stages.

Stage	Category	ChatGPT 4.0	DeepSeek-R1	Gemini
Stage 1	Creativity	29.11	44.53	8.40
	Factual Accuracy	22.34	53.90	6.32
	Open Ended	14.14	18.34	7.99
	Technical	9.67	80.72	5.38
Stage 2	Creativity	18.06	52.67	4.35
	Factual Accuracy	18.45	62.42	13.49
	Open Ended	24.72	67.41	16.96
	Technical	66.90	109.77	8.06

Table 3. Creativity analysis scores of LLMs with different parameters.

Phase	LLM	Readability	Vocab. Richness	Coherence / Similarity	Sentiment Polarity	Grammar / Spelling
Stage 1	ChatGPT	70.09	0.71	0.63	0.068	9
	DeepSeek	70.26	0.74		0.000	28
	Gemini	75.10	0.63		0.002	0
Stage 2	ChatGPT	66.67	0.89	0.28	0.031	0
	DeepSeek	76.15	0.83		-0.025	0
	Gemini	84.47	0.77		0.106	0

model. This indicates which model is faster. In case of fast response rather than accurate response, we can choose the faster model. The times needed for each model are described in Table 2.

From Table 2, Gemini takes less time to provide the response. In all cases, Gemini takes less than 10 seconds. But the other two models take more time. Among them, DeepSeek takes the longest time. ChatGPT takes a manageable time duration to give the response.

Creativity Analysis: In this section we have presented the scores of generated texts from the LLM. We have applied different methods to find the scores of Readabilities, Vocabulary Richness, Coherence and similarity, Sentiment Polarity, and, Grammar and Spelling. As we have conducted this research in two phases, we are presenting them as Stage 1 and Stage 2.

The Table 3 shows the comparative scores among three different LLMs. In terms of Readability (Dale-Chall score), Gemini obtained the highest scores in both stages. Note that higher Dale-Chall scores indicate more difficult text (due to complex vocabulary and sentence structure), so Gemini produced the most challenging-to-read responses in creativity tasks. We can see the opposite picture in terms of Vocabulary richness where Gemini has the lowest one. Similarly, Coherence and Sentiment Polarity scores are also displayed. On the other hand, in stage 1, DeepSeek had the highest number of grammatical and spelling errors (28 errors), while Gemini showed the best performance with the fewest errors (0 errors).

Factual Accuracy: As we have categorized our questions in different parts, factual accuracy is one of them. In this case, the answers are known to us. But for the analytics of their solutions, we have applied different parameters to find their scores like Creativity Analysis. The detailed results are presented in Table 4.

For the Factual Accuracy Analysis, we can see a different outcome compared to the Creativity Analysis. Here, DeepSeek shows the highest value for readability in both stages where ChatGPT shows better vocabulary richness. Due to incomplete TF-IDF cosine similarity values for DeepSeek and Gemini, cross-model coherence comparison is not possible; only ChatGPT values are available. But for the sentiment polarity score, there is a mixed score, where ChatGPT scores least among all. Moreover, in stage 2, we have found that ChatGPT produces more grammatical mistakes compared to the other two models. The other two models do not have any grammatical mistakes.

Open Ended: In this section, we have discussed the generated text of our tested LLMs focusing on the open-ended topic. Assessing LLMs, open-ended

Table 4. Factual accuracy analysis scores of LLMs with different parameters.

Phase	LLM	Readability	Vocab. Richness	Coherence / Similarity	Sentiment Polarity	Grammar / Spelling
Stage 1	ChatGPT	46.47	0.82	0.48	0.064	0
	DeepSeek	52.76	0.74		0.086	0
	Gemini	38.82	0.65		0.047	0
Stage 2	ChatGPT	17.43	0.81	0.54	0.026	4
	DeepSeek	19.57	0.72		0.056	0
	Gemini	14.70	0.66		0.123	1

questions provide a clear picture since they directly obtain their creativity and generative talents. Unlike closed-ended formats, these questions demand LLMs to make responses freely, displaying their capacity to develop new and contextually suitable text. This may be significant for activities such as content creation and open-domain question answering. Open-ended questions can demonstrate the model's ability to be creative and to do more than just retrieve or recognize information by not limiting the response to a predetermined set of alternatives. The quantitative evaluation results for open-ended responses are summarized in Table 5.

For open-ended question, comparing to all, Gemini performs best. Where in stage 2, answers from ChatGPT fail to perform well. Though for Vocabulary richness, ChatGPT always performs well for open ended question answers. Depending on the question, it also works better with a score of 0.127 for stage 2. No significant difference in grammar and spelling performance among the three LLMs.

Technical Explanation: For most professional applications, technical questions are significant. It can clarify details and propose solutions. It can also assist people in technical ways. Besides, LLM says that it is beneficial to work and decreases the workload for humans. So, our analysis is given through Table 6.

5 Discussion

In this section we discuss the results from our analysis. Here, we have drawn a Boxplot from the obtained data from our analysis. Also, we have used a line graph to see the trends between the performance of two stages among the LLMs. Then we used the F-statistics to reach our final decision. With this variation between these three LLM is compared.

The comparison of the box plot of Figure 1, indicates a comparison among the three LLMs. Considering most of the metrics, ChatGPT shows a strong and consistent performance compared to the other two. ChatGPT showed strong performance in vocabulary richness and sentiment polarity. For grammar and spelling (error count, higher = worse), Gemini had the fewest errors overall, indicating superior grammatical correctness. On the other hand, DeepSeek shows better readability scores, but shows an underperformance in sentiment polarity and grammar spelling. The reason behind this is that it suggests a simpler but less refined output. Gemini showed high grammatical accuracy (fewest errors) but lower vocabulary richness and more variable sentiment polarity, potentially indicating less consistent stylistic output in some tasks.

Overall, ChatGPT shows the most consistent performance across all metrics with smaller spreads and fewer outliers. But Gemini sacrifices Vocabulary Richness for higher grammar spelling. Meanwhile, DeepSeek shows room for improvement in Grammar, in spite of having a satisfiable performance in other metrics. Now, in the next step, we have broken down this performance, indicating two stages for better understanding.

The pair plot in Figure 2 shows clear differences in how the three LLMs perform among various writing quality metrics. ChatGPT forms consistent clusters across all metric combinations. This means that its output is steady and balanced. It consistently performs well, especially in vocabulary richness and coherence. Gemini's points are more spread in the grammar and readability plots. This shows that when Gemini produces easy to read responses, it can vary in accuracy and structure. DeepSeek shows the most scattered distribution. There are visible outliers and broad spread in grammar and sentiment polarity. It indicates inconsistent behavior across different cases. Overall, ChatGPT illustrates the most stable and reliable performance, while Gemini and DeepSeek show greater variation depending on the metric.

Table 5. Open-Ended question analysis scores of LLMs with different parameters.

Phase	LLM	Readability	Vocab. Richness	Coherence / Similarity	Sentiment Polarity	Grammar / Spelling
Stage 1	ChatGPT	17.34	0.73	0.651	0.058	0
	DeepSeek	22.00	0.58		0.075	13
	Gemini	44.03	0.52		0.046	1
Stage 2	ChatGPT	16.35	0.73	0.481	0.127	1
	DeepSeek	1.86	0.78		0.060	1
	Gemini	18.86	0.58		0.114	0

Table 6. Technical explanation analysis scores of LLMs with different parameters.

Phase	LLM	Readability	Vocab. Richness	Coherence / Similarity	Sentiment Polarity	Grammar / Spelling
Stage 1	ChatGPT	64.00	0.77	0.615	0.230	0
	DeepSeek	68.06	0.62		0.144	18
	Gemini	55.34	0.64		0.080	1
Stage 2	ChatGPT	31.38	0.57	0.732	0.075	28
	DeepSeek	33.92	0.65		0.035	21
	Gemini	38.72	0.51		0.024	13

Figure 3 illustrates the trends of the evaluated metrics across two stages using a line graph. The metric values are normalized to a range of 0 to 1, whereas the total score section reports values obtained by summing the individual metric scores. Coherence/similarity scores were only fully computed for ChatGPT; values for DeepSeek and Gemini were incomplete, preventing direct comparison across models. It shows the behavioral patterns of each LLM. ChatGPT stands out for its consistent trajectory, maintaining strong performance across readability, coherence, and language correctness. DeepSeek exhibits high readability in both stages but there is a noticeable change in grammar and vocabulary richness. This is an unstable trend while two stages show different patterns. Gemini presents the most interesting behavior in grammar and sentiment polarity. It suggests an unstable response style. From a standpoint of deployment, the trends show the robustness of ChatGPT through different phases. Initially, it is the safest choice for tasks where consistent quality requires.

To find the best model, we have calculated the one -way ANOVA (F-test). We have used this test as it compares multiple group means simultaneously. Moreover, it determines if there is any statistically significant difference between the group of two or more models. The F-test in ANAOVA refers to a test comparing two statistical models. As a result, this test can identify which one has a significant impact on the variable. For applying this test technique, we have made a dataset containing the test result from our result section. Then we have applied this technique to find the best output.

Table 7. One-way ANOVA (F-test) results by metric.

Metric	F-value	p-value	Significance
Readability	0.15	0.8629	Not Significant
Vocabulary Richness	4.78	0.0195	Significant
Coherence Similarity	0.00	1.0000	Not Significant
Sentiment Polarity	0.63	0.5399	Not Significant
Grammar Spelling	1.65	0.2156	Not Significant

Table 8. Tukey HSD multiple comparison results for Readability (FWER = 0.05).

Group 1	Group 2	Mean Diff	p-adj	Lower	Upper	Reject
ChatGPT	DeepSeek	0.047	0.951	-0.348	0.442	False
ChatGPT	Gemini	0.085	0.851	-0.310	0.480	False
DeepSeek	Gemini	0.038	0.968	-0.357	0.433	False

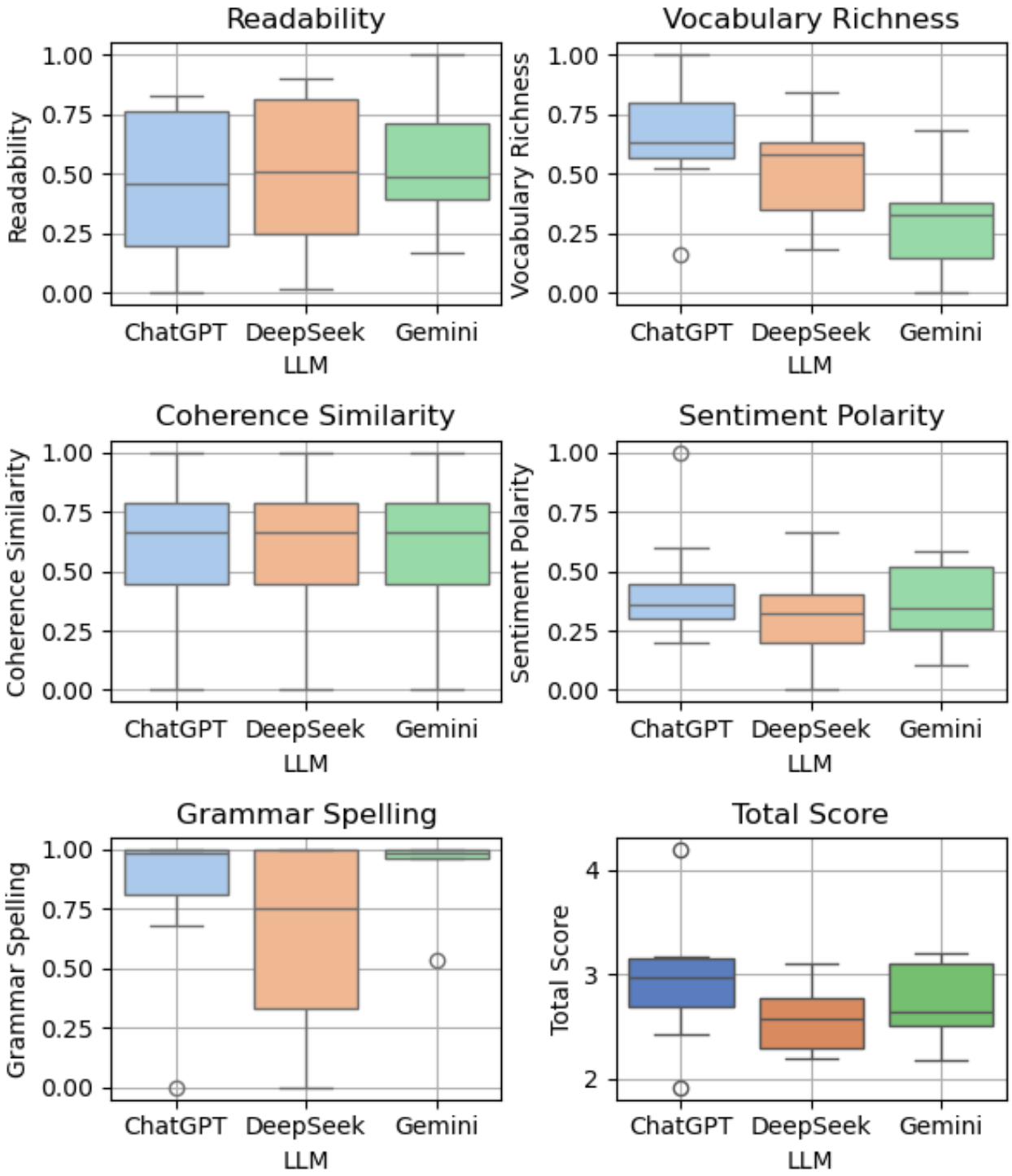

Figure 1. Distribution of the metrics for different LLM's.

Table 9. Tukey HSD multiple comparison results for Coherence Similarity (FWER = 0.05).

Group 1	Group 2	Mean Diff	p-adj	Lower	Upper	Reject
ChatGPT	DeepSeek	0.0	1.0	-0.391	0.391	False
ChatGPT	Gemini	0.0	1.0	-0.391	0.391	False
DeepSeek	Gemini	0.0	1.0	-0.391	0.391	False

Table 10. Tukey HSD multiple comparison results for Sentiment Polarity (FWER = 0.05).

Group 1	Group 2	Mean Diff	p-adj	Lower	Upper	Reject
ChatGPT	DeepSeek	-0.122	0.510	-0.394	0.151	False
ChatGPT	Gemini	-0.067	0.810	-0.340	0.205	False
DeepSeek	Gemini	0.054	0.871	-0.218	0.327	False

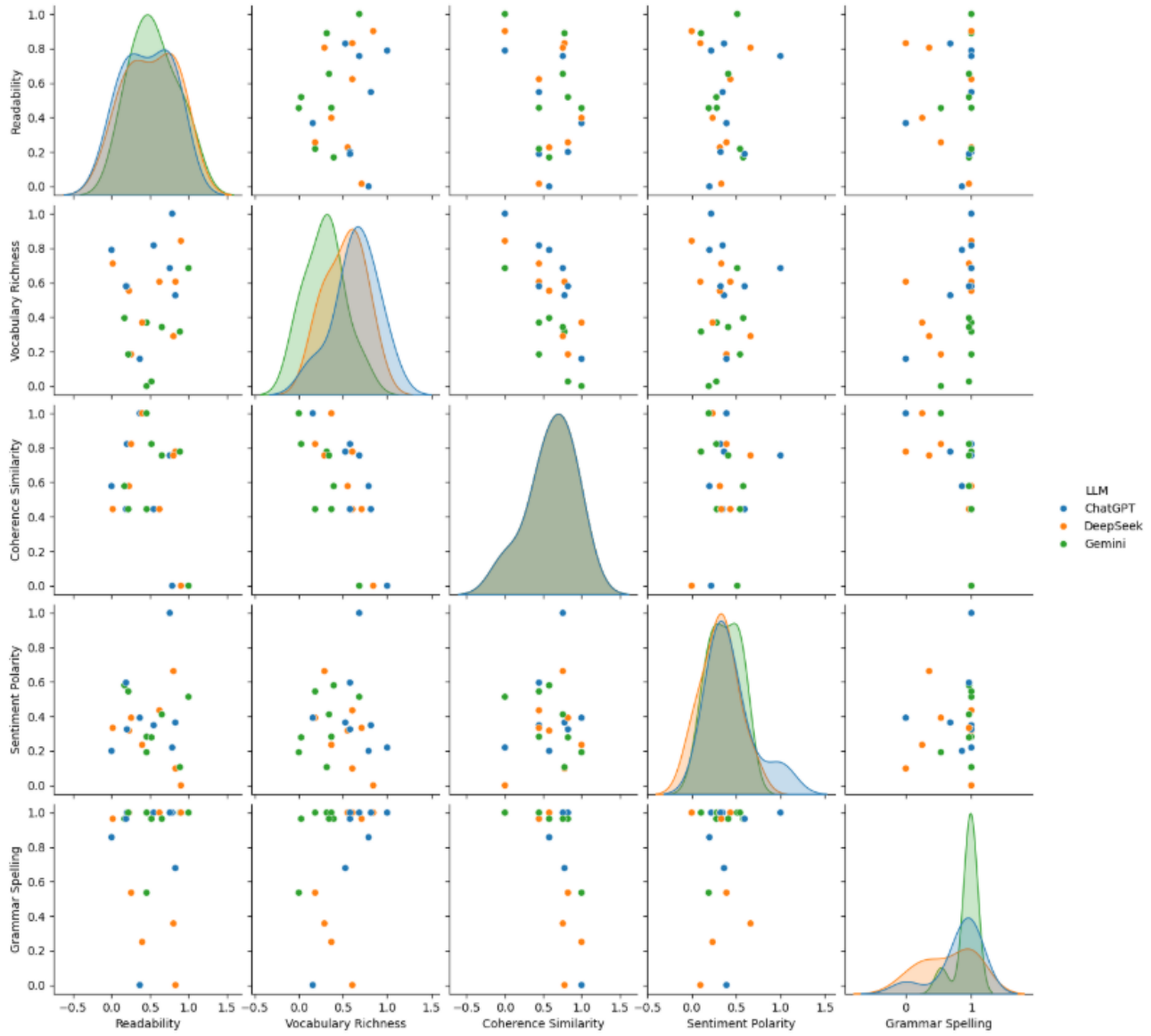

Figure 2. Performances of LLMs among various writing quality metrics.

Table 11. Tukey HSD multiple comparison results for Grammar Spelling (FWER = 0.05).

Group 1	Group 2	Mean Diff	p-adj	Lower	Upper	Reject
ChatGPT	DeepSeek	-0.174	0.534	-0.579	0.231	False
ChatGPT	Gemini	0.116	0.753	-0.289	0.521	False
DeepSeek	Gemini	0.290	0.192	-0.115	0.695	False

The results are shown in Table 7 as a screenshot. Here we can find that only vocabulary richness is statistically significant. This means that there is at least one model that performs differently on vocabulary richness. But for the other metrics, they are not statistically significant. So, from this test we have not achieved the best results. For these cases, we need to find another test which provides pairwise test scores so that we can compare them easily.

We have applied Post-hoc Test (Tukey) to find the pairwise test scores. It is a statistical method which is used after a significant ANOVA test. This method finds specific groups that have a statistically significant difference in their means. It shows a pairwise comparison between all possible groups.

Since other metrics are not statistically significant, we have to find exactly which model differs. For this, we have chosen Post-hoc Test (Tukey). This is a powerful tool to find the scores of each pair.

Tables 8, 9, 10 and 11 show the Post-hoc Test (Tukey) scores of our used metrics. In this technique, we have the pairwise scores among the three models. In this comparison, it is also seen that the models are not statistically significant. But we can get the mean scores

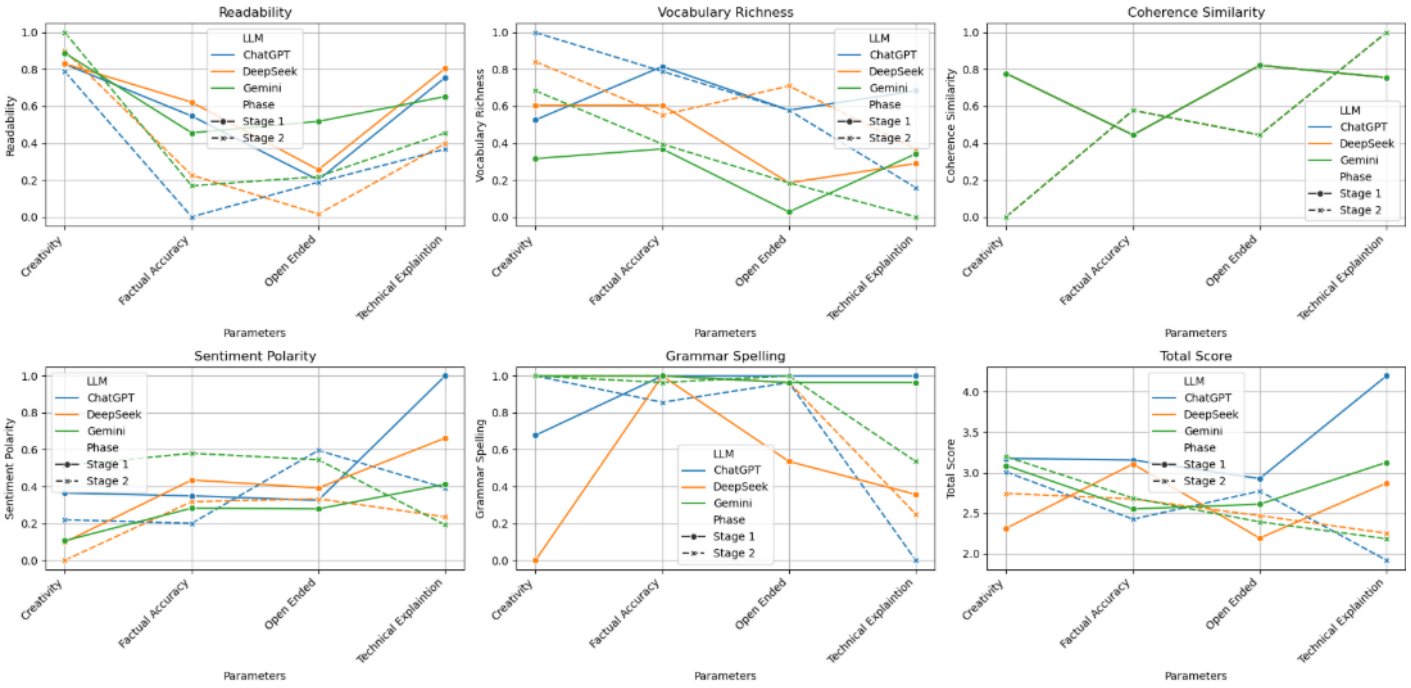

Figure 3. Distribution of scores in different tests.

Table 12. Mean evaluation scores of LLMs across all metrics.

LLM	Readability	Vocabulary Richness	Coherence Similarity	Sentiment Polarity	Grammar Spelling	Total Score
ChatGPT	0.459292	0.641447	0.602778	0.430882	0.187500	0.464380
DeepSeek	0.506526	0.519737	0.602778	0.309314	0.361607	0.459992
Gemini	0.544463	0.289474	0.602778	0.363725	0.071429	0.374374

in terms of the metrics and the total scores. From there, we can finalize our discussion properly.

As we have discussed before, the above figure shows how the metrics are not significant. So, our next step is to find the mean scores. The largest mean score indicates the highest preferable model.

Table 12 shows the mean scores of different LLMs. It is seen that all the models have almost similar scores. Among them, ChatGPT scores comparatively better. But DeepSeek has a big difference in scores though it stands at the top for two metrics. Gemini has only a top score in Readability.

6 Conclusion

This paper shows the performance analysis of three popular LLMs to get a comparative overview among them. First, we have categorized our evaluation process into four different areas. These are Creativity, Factual accuracy, Open-ended, Technical explanation. We have recorded the response with their response time that makes our dataset. Then we have analyzed them by considering different metrics. For this we have used some tools and packages in python. After recording all the scores, we then have analyzed them and showed them graphically. The graphical representation explains why ChatGPT performs better among the others. Also, for statistical analysis, we first performed an F-test on the obtained results to find which is more significant and found the p-value. For showing which pairs of groups differ significantly, we have looked through the Tukey HSD (Honestly Significant Difference). Finally, we have calculated the mean scores of the models. From this analysis, we have come to the final decision that among the three LLMs, ChatGPT shows slightly better than the other two models in different metrics.

Data Availability Statement

The data and code supporting the findings of this study are publicly available at the following repository: https://github.com/acdas10/NLP-based-LLM-analysis

Funding

This work was supported without any funding.

Conflicts of Interest

The authors declare no conflicts of interest.

AI Use Statement

The authors declare that no generative AI was used in the preparation of this manuscript.

Ethical Approval and Consent to Participate

Not applicable.

References

[1] Bommasani, R. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

[2] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ... & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM computing surveys, 55*(12), 1-38. [CrossRef]

[3] Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., ... & He, Y. (2025). Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

[4] Google DeepMind. (2023). Introducing Gemini: our largest and most capable AI model. Retrieved from https://www.deepmind.com/blog/

[5] Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P. S., ... & Gabriel, I. (2021). Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

[6] Boden, M. A. (2004). *The creative mind: Myths and mechanisms*. Routledge.

[7] Oltețeanu, A. M. (2020). *Cognition and the Creative Machine: Cognitive AI for Creative Problem Solving*. Springer Nature. [CrossRef]

[8] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems, 33*, 9459-9474.

[9] Das, S. (2022). The meaning of creativity through the ages: from inspiration to artificial intelligence. In *Creative business education: exploring the contours of pedagogical praxis* (pp. 27-53). Cham: Springer International Publishing. [CrossRef]

[10] Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W. T., Koh, P., ... & Hajishirzi, H. (2023, December). Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 12076-12100). [CrossRef]

[11] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems, 30*.

[12] OpenAI. (2024). OpenAI API Documentation. Retrieved from https://platform.openai.com/docs/overview

[13] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., ... & Lample, G. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

[14] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in neural information processing systems, 35*, 27730-27744.

[15] Sheng, E., Chang, K. W., Natarajan, P., & Peng, N. (2019, November). The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)* (pp. 3407-3412). [CrossRef]

[16] Weidinger, L., Uesato, J., Bielecki, J., van den Driessche, G., Chrzanowski, M., Krasheninnikov, D., ... & Tréger, R. (2021). Ethical and social risks of Large Language Models. *arXiv preprint arXiv:2112.04359*.

[17] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ... & Fung, G. (2023). Survey of Hallucination in Large Language Models. *ACM Computing Surveys, 56*(2), 1-38.

[18] Mehrabi, N., Morstatter, B., Saxena, V., Lerman, K., & Narayanan, M. G. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6), 1-35. [CrossRef]

[19] Marcus, G., & Davis, E. (2019). *Rebooting AI: Building artificial intelligence we can trust*. Pantheon.

Qazi Novera Tansue Nasa working as a research assistant at the University of Potsdam. Masters in Data Science. (Email: nasa.novera@gmail.com)

Ashik Chandra Das received the B.Sc. degree in Mathematics from Jagannath University and M.Sc. in Data Science. He is doing research on application of AI. (Email: ashikdas624@gmail.com)