Literariness Journal
A Peer-Reviewed Quarterly Journal of Literature and Cultural Studies

P-ISSN: 3108-1614
E-ISSN: 3108-172X

LiterarinessJournal.org

Special Issue 1 (June 2026)

A Literariness.org Project

DOI: 10.67147/literariness.v1si1.010

Comparative Evaluation of AI Chatbot Output Quality Using NLP Metrics and Human Expert Assessment

S. DHANUSH
Research Scholar
Department of English
Vels Institute of Science, Technology and Advanced Studies, Pallavaram, Chennai, India.

DR. P. SANTHOSH
Assistant Professor
Department of English
Vels Institute of Science, Technology and Advanced Studies, Pallavaram, Chennai

Abstract: As long as human society continues to evolve, technological advancement will remain inevitable. In the contemporary era, the focus has shifted from merely possessing technology to understanding and controlling the systems that drive it. Generative AI chatbots such as ChatGPT, Google Gemini, Perplexity, Claude, and Grok have become integral tools in various aspects of daily life, including education, research, and professional communication. Existing studies have primarily examined individual chatbot models, focusing on the quality, accuracy, and contextual relevance of their responses. This study undertakes a comparative evaluation of major AI chatbots by providing them with twenty standardized prompts and collecting their responses for analysis.

The research adopts a mixed-methods approach, combining automated Natural Language Processing (NLP) metrics with structured human expert assessment. The quantitative analysis evaluates response quality using NLP metrics such as BLEU, ROUGE, BERTScore, and grammar analysis to measure accuracy, semantic similarity, and linguistic correctness. The qualitative component involves expert evaluation of the responses based on criteria including factual accuracy, grammatical quality, content completeness, coherence, and contextual relevance. The findings are compared to identify variations in the output quality of different AI chatbots. This study aims to contribute to a deeper understanding of the strengths and limitations of contemporary AI chatbots and their effectiveness in academic and knowledge-based contexts.

Keywords: *Artificial Intelligence, Chatbot Evaluation, ChatGPT, Google Gemini, Perplexity, Claude, Grok, Natural Language Processing, Human Expert Assessment, Generative AI*

1. Introduction

Artificial Intelligence (AI) has evolved from a specialized technological concept into a transformative force that influences nearly every aspect of contemporary life. A decade ago, AI was primarily recognized by researchers, software developers, and students in computer-related disciplines. Today, however, AI-powered applications have become widely accessible to the general public. While traditional search engines such as Google and online encyclopedias like Wikipedia were once the primary sources of information retrieval, recent advancements in Natural Language Processing (NLP) have led to the development of Large Language Models (LLMs) that power modern AI chatbots such as ChatGPT, Google Gemini, Claude, Perplexity, and Grok.

These AI chatbots are increasingly used in education, research, professional communication, and everyday problem-solving. In the context of Indian Higher Education and the vision outlined in the National Education Policy (NEP) 2020, AI-driven tools offer significant opportunities for personalized learning, academic support, and administrative efficiency. Beyond functioning as search tools, contemporary chatbots assist users in content generation, academic writing, research assistance, grammar correction, coding support, and even image generation.

Despite their growing popularity, concerns remain regarding the accuracy, reliability, completeness, and contextual relevance of AI-generated responses. As dependence on AI-generated content increases, it becomes essential to evaluate the quality of outputs produced by different chatbot systems. This study aims to conduct a comparative evaluation of four leading AI chatbots by analyzing their responses to twenty standardized prompts. Using a mixed-methods framework that combines automated Natural Language Processing (NLP) metrics with human expert assessment, the study seeks to identify which chatbot provides the highest-quality output for academic purposes.

2. Literature Review

2.1 Evolution of AI Chatbots

The history of chatbot development can be traced back to ELIZA, the first chatbot created by Joseph Weizenbaum between 1964 and 1966. Designed to explore communication between humans and machines, ELIZA simulated the conversational style of a Rogerian psychotherapist. Although not an AI system in the modern sense, ELIZA utilized pattern matching and keyword recognition techniques to generate responses. Its ability to create an illusion of understanding demonstrated the potential of human-computer interaction and laid the foundation for future chatbot development.

Subsequent conversational systems included PARRY (1972), which simulated the behavior of a patient with paranoid schizophrenia, and ALICE (Artificial Linguistic Internet Computer Entity) (1995), which employed heuristic pattern matching to engage users in conversation. The development of intelligent virtual assistants accelerated in the twenty-first century with the introduction of Siri by Apple in 2011, followed by Google Assistant, Amazon Alexa, and Microsoft Cortana.

The emergence of Large Language Models has significantly transformed chatbot capabilities. Systems such as ChatGPT, Google Gemini, Claude, Perplexity, Grok, and Meta AI utilize advanced NLP techniques and deep learning architectures to generate human-like responses, perform complex reasoning tasks, summarize information, and assist users across diverse domains. These developments have expanded the role of chatbots from simple conversational agents to sophisticated AI assistants.

2.2 Evaluation of AI Chatbot Responses

Evaluating the quality of AI-generated responses remains a significant challenge in computational linguistics and Natural Language Generation (NLG). Traditional automatic evaluation metrics such as BLEU (Bilingual Evaluation Understudy), introduced by Papineni et al. (2002), were originally designed for machine translation. BLEU measures the degree of overlap between machine-generated text and reference text, providing a fast, cost-effective, and language-independent evaluation method.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation), developed by Lin (2004), extended this approach to the evaluation of automatic text summarization by measuring lexical overlap and content coverage. While BLEU and ROUGE remain widely used, they primarily focus on surface-level textual similarities and may not adequately capture semantic meaning.

To address these limitations, researchers introduced semantic evaluation metrics such as BERTScore (Zhang et al., 2020), which employs contextualized embeddings generated by pre-trained BERT models. Unlike traditional metrics, BERTScore evaluates semantic similarity and contextual relevance, providing a more comprehensive assessment of generated text quality.

However, automated metrics alone cannot fully capture aspects such as coherence, factual accuracy, tone, readability, and pragmatic appropriateness. Consequently, human evaluation continues to play a crucial role in chatbot assessment. Deriu et al. (2021) argue that human judgments remain essential for evaluating conversational quality, while Mehri and Eskenazi (2020) and Lowe et al. (2017) highlight the limitations of relying solely on automated metrics. Their findings suggest that a hybrid evaluation framework combining quantitative NLP metrics with qualitative human assessment provides a more reliable and comprehensive measure of chatbot performance.

Therefore, the present study adopts a mixed-methods approach that integrates both automated and human-centered evaluation techniques to compare the performance of leading AI chatbots in academic contexts.

2.3 Comparative Studies of AI Chatbots

Although numerous studies have examined the performance of individual AI chatbots, comparatively fewer studies have undertaken a systematic comparison of multiple chatbot systems across different domains. Much of the existing research has focused on evaluating ChatGPT, particularly in educational and medical contexts.

Sallam (2023) conducted a systematic review of ChatGPT's applications in healthcare and medical education, highlighting its strengths in content generation, information retrieval, and academic support while also identifying limitations related to factual inaccuracies and hallucinations. Similarly, Kung et al. (2023) evaluated ChatGPT's performance on the United States Medical Licensing Examination (USMLE) and found that the model performed at or near the passing threshold without specialized medical training. Gilson et al. (2023) further assessed ChatGPT using medical school examination questions and reported promising results in terms of reasoning and answer generation.

Despite these contributions, comparative research involving multiple AI chatbots such as ChatGPT, Google Gemini, Claude, and Perplexity remains limited. Furthermore, most studies rely either on automated evaluation metrics or human judgment alone. There is a need for a comprehensive framework that integrates both Natural Language Processing (NLP) metrics and human expert assessment to evaluate chatbot performance more effectively. This study seeks to address this gap by comparing the output quality of leading AI chatbots using a mixed-methods evaluation framework.

3. Research Objectives

The objectives of this study are:

1. To evaluate the quality of AI chatbot responses using automated Natural Language Processing (NLP) metrics, including BLEU, ROUGE, BERTScore, and Grammar Error Rate (GER).
2. To critically analyse the performance of AI chatbots through human expert evaluation based on criteria such as accuracy, grammatical correctness, content completeness, coherence, and contextual relevance using a structured rubric scale.
3. To compare the results obtained from automated NLP metrics with those derived from human expert evaluation in order to identify similarities and differences in chatbot performance.

4. To determine which AI chatbot provides the most meaningful, accurate, and academically relevant responses for educational and language-learning contexts.
5. To propose a comprehensive evaluation framework that integrates NLP-based metrics and human assessment for evaluating AI chatbot performance in academic settings.

4. Research Questions

The study seeks to answer the following research questions:

1. How do AI chatbots such as ChatGPT, Google Gemini, Claude, and Perplexity perform when evaluated using automated NLP metrics including BLEU, ROUGE, BERTScore, and Grammar Error Rate?
2. How do human experts evaluate AI chatbot responses in terms of accuracy, grammatical quality, content completeness, coherence, and contextual relevance?
3. What significant differences exist between automated NLP metric evaluations and human expert assessments of AI chatbot responses?
4. Which AI chatbot provides the highest-quality output for academic and language-learning purposes?
5. How effective is a combined NLP and human evaluation framework in assessing the overall performance of AI chatbots in educational contexts?

5. Research Methodology

5.1 Research Design

This study adopts a mixed-methods research design that combines quantitative automated evaluation with qualitative human expert assessment. The integration of these two approaches enables a comprehensive analysis of AI chatbot performance by measuring both objective linguistic quality and subjective human perceptions of response effectiveness.

The study follows a comparative experimental approach in which identical prompts are administered to all selected chatbot platforms under controlled conditions. This design ensures a systematic, fair, and reproducible evaluation of chatbot responses across multiple dimensions of quality.

5.2 Chatbot Selection

Four widely used AI chatbots were selected for this study based on their popularity, accessibility, and relevance to academic and general-purpose applications.

AI Chatbot	Developer	Underlying Model
ChatGPT	OpenAI	GPT-4o
Google Gemini	Google DeepMind	Gemini 1.5 Pro
Claude	Anthropic	Claude 3.5 Sonnet
Perplexity	Perplexity AI	Multiple LLMs + Search

All four chatbots were accessed through their free-tier web interfaces to maintain consistency and comparability. Data collection was conducted within the same time frame to minimize potential variations resulting from model updates or system modifications.

5.3 Prompt Design

A set of twenty standardized prompts was designed to evaluate four key dimensions of chatbot output quality: accuracy, grammatical quality, content completeness, and contextual relevance.

Category	No. of Prompts	Primary Dimension Tested
Factual Questions	5	Accuracy
Grammar Tasks	5	Grammar
Language	5	Content Completeness
Multi-turn / Contextual	5	Contextual Relevance
Total	20	All 4 dimensions

Prompts

1. How do you face a person with confidence?
2. Define the term “mannerisms” or “behaviour” that a person should uphold at all times.
3. What gestures should be followed in different situations as a human being?
4. Is eye contact essential when facing situations such as job interviews or business meetings?
5. What is the importance of punctuality?
6. How do you introduce sentence structure (subject and predicate) to beginners?

7. What methods help students understand tenses easily?
8. How do you teach the difference between simple, compound, and complex sentences?
9. What are effective ways to explain subject–verb agreement?
10. What are verbs, and how do you explain the difference between action verbs and linking verbs?
11. Why is literature considered an important part of human life?
12. What are the key features of poetry, and how does poetry depict human emotions?
13. How does vocabulary play a vital role in conversation and communication?
14. Language interconnects every subject. Do you agree? Explain.
15. Why are the LSRW (Listening, Speaking, Reading, and Writing) skills important for excellence in language and literature?
16. How do you help students avoid common grammatical mistakes?
17. What strategies help in teaching punctuation marks effectively?
18. What assessment techniques do you use to evaluate grammar proficiency?
19. How do you guide students in forming questions and answering them correctly?
20. How do you differentiate instruction for slow learners and advanced learners?

The prompts were selected to represent common academic, linguistic, and pedagogical inquiries that frequently arise in educational settings. Using the same prompts across all chatbot platforms ensures consistency and enables direct comparison of response quality.

5.4 Data Collection

Twenty standardized prompts were administered to four AI chatbots: ChatGPT, Google Gemini, Claude, and Perplexity. To ensure consistency and fairness in the evaluation process, the following conditions were maintained:

- The same twenty prompts were entered into all four chatbots without any modification.
- Responses were generated using the free versions of each chatbot.
- All responses were collected on the same day to minimize the impact of model updates and temporal variations.

- The generated responses were recorded and organized in Microsoft Excel for subsequent analysis and evaluation.

This standardized procedure ensured comparability across chatbot outputs and minimized external variables that could influence response quality.

5.5 NLP Metrics

The collected responses were quantitatively evaluated using several established Natural Language Processing (NLP) metrics.

BLEU (Bilingual Evaluation Understudy)

BLEU is an automatic evaluation metric originally developed for machine translation. It measures the degree of overlap between a machine-generated response and a reference response through n-gram precision. BLEU scores range from 0 to 1, where a higher score indicates greater similarity between the generated text and the reference text. The metric is widely recognized for being efficient, cost-effective, and language-independent.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

ROUGE is commonly used for evaluating text summarization and content generation tasks. It measures the overlap between generated responses and reference texts based on recall-oriented matching. This study employed three ROUGE variants:

- **ROUGE-1**: Measures unigram overlap.
- **ROUGE-2**: Measures bigram overlap.
- **ROUGE-L**: Measures the longest common subsequence between texts.

These measures provide insight into content coverage and textual similarity.

BERTScore

BERTScore is a semantic evaluation metric that utilizes contextualized embeddings generated by a pre-trained BERT model. Unlike traditional lexical metrics, BERTScore compares the semantic similarity between generated and reference texts at the token level. This enables the metric to capture meaning beyond surface-level word matching and provides a more robust assessment of response quality.

LanguageTool Grammar Analysis

Grammatical quality was assessed using LanguageTool, an open-source grammar and style-checking application. The tool was used to identify and count grammatical errors, spelling mistakes, and stylistic issues present in each chatbot response. The resulting Grammar Error Rate (GER) served as an indicator of linguistic correctness.

All NLP metrics were implemented using Python 3.12.3 in the Google Colab environment.

5.6 Human Evaluation

In addition to automated evaluation, a human assessment was conducted to measure qualitative aspects of chatbot responses. A total of thirty experts participated in the evaluation process:

- Fifteen academic linguists specializing in language and discourse analysis.
- Fifteen AI researchers with expertise in artificial intelligence and natural language processing.

Each evaluator assessed the responses generated by the four chatbots using a five-point Likert-scale rubric across four evaluation criteria:

Area	1-poor	3 - Average	5 - Excellent
Accuracy	incorrect	Partially correct	Fully correct
Grammar	errors	minor errors	flawless
Content	incomplete	Covers key points	comprehensive
Contextual	Ignores context	Partially context-aware	Fully context-aware

The average scores from all evaluators were calculated for each chatbot across the four criteria.

Inter-Rater Reliability

To ensure consistency among evaluators, inter-rater reliability was measured using Fleiss' Kappa coefficient. The obtained value of $\kappa = 0.84$ indicates a high level of agreement among the evaluators, demonstrating the reliability and validity of the human assessment process.

CQS = (Accuracy × 0.30) + (Content × 0.30) + (Grammar × 0.20) + (Context × 0.20)

6. Results and Discussion

6.1 Automated NLP Metrics Results

Model	Avg_BLEU	Avg_ROUGE	Avg_BERT	Grammar Error Rate (/
ChatGPT	0.024048669	0.216236824	0.837668774	0.6
Claude	0.005329072	0.140181007	0.847064355	1.2
Perplexity	0.004268464	0.129844835	0.834999681	1.5
Gemini	0.026856702	0.177395652	0.839082789	0.8

The NLP metric evaluation was conducted using BLEU, ROUGE, BERT score, and Grammar Error Rate to measure the quality of AI chatbot responses

✓Gemini has achieved the highest BLEU score (0.0268), which indicates a better similarity with reference to the answers given by human experts.

✓ChatGPT shows a strong impact in ROUGE score (0.2162) and BERT score (0.8376) in delivering a semantic and contextual alignment.

✓Claude obtained the highest BERT score (0.8470), which provides better outcomes and meaningful understanding of the context.

✓Perplexity records the lowest NLP score with lower BLEU and ROUGE, and renders its output with the highest grammatical error rate of (1.5)

6.2 Human Expert Evaluation Results

Chatbot	Accuracy	Grammar	Content	Context	CQS
ChatGPT	4.6	4.5	4.5	4.4	4.51
Google Gemini	4.7	4.8	4.6	4.6	4.67
Claude	4.2	4.1	4.3	4.0	4.17
Perplexity	4.0	3.8	4.4	3.7	4.04

Human Expert Evaluation

To ensure the reliability of human assessment, inter-rater reliability was measured using Fleiss' Kappa. The obtained value ($\kappa = 0.84$) indicates a strong level of agreement among the evaluators, demonstrating the consistency and reliability of the evaluation process.

Human evaluation was conducted using a five-point Likert scale based on four criteria: Accuracy, Grammar, Content Completeness, and Contextual Relevance. A Cumulative Quality Score (CQS) was calculated for each chatbot based on the average ratings assigned by the evaluators.

Google Gemini achieved the highest Cumulative Quality Score (CQS) of 4.67, indicating superior performance across all evaluation criteria. The model consistently generated accurate, contextually relevant, grammatically correct, and comprehensive responses.

ChatGPT obtained a CQS of 4.51, demonstrating strong overall performance, particularly in grammatical accuracy, content organization, and response coherence. Its outputs were generally balanced and reliable across different prompt categories.

Claude received a CQS of 4.17, reflecting a satisfactory level of performance. Human evaluators noted that its responses were coherent and informative, although they occasionally lacked the depth and contextual precision observed in Gemini and ChatGPT.

Perplexity recorded a CQS of 4.04, the lowest among the four chatbots evaluated. While its responses were generally accurate and informative, experts observed comparatively lower performance in content completeness and contextual relevance.

7. Conclusion

This study presented a systematic comparative evaluation of four leading AI chatbots—ChatGPT, Google Gemini, Claude, and Perplexity—using a mixed-methods framework that integrated automated Natural Language Processing (NLP) metrics with structured human expert assessment. By combining quantitative and qualitative evaluation techniques, the study provided a comprehensive analysis of chatbot performance in academic and educational contexts.

The findings reveal that all four AI chatbots demonstrate strong capabilities in generating academically relevant content. However, their performance varies depending on the evaluation criteria applied. In the quantitative assessment, Google Gemini achieved the highest BLEU score, indicating superior lexical similarity to reference responses. Claude obtained the highest BERTScore, demonstrating

strong semantic similarity and contextual understanding. ChatGPT performed best in ROUGE metrics and exhibited the highest level of grammatical accuracy among the evaluated models.

The qualitative assessment further highlighted differences in response quality. Based on human expert evaluation, Google Gemini achieved the highest Cumulative Quality Score (4.67), outperforming the other chatbots in terms of accuracy, contextual relevance, content completeness, and overall response quality. ChatGPT followed closely with a score of 4.51, while Claude and Perplexity achieved scores of 4.17 and 4.04, respectively.

Overall, the findings suggest that Google Gemini produced the most effective and academically relevant responses within the scope of this study. Nevertheless, the results also demonstrate that no single evaluation metric is sufficient for assessing AI-generated content comprehensively. A combination of automated NLP metrics and human expert judgment provides a more reliable framework for evaluating chatbot performance. Future research may expand the dataset, include additional chatbot models, and explore domain-specific evaluations to further enhance the understanding of AI-generated response quality in academic and educational settings.

Works Cited

Abd-Alrazaq, Alaa, et al. "Large Language Models in Medical Education: Opportunities, Challenges, and Future Directions." *JMIR Medical Education*, vol. 9, no. 1, 2023, article e48291.

Brown, Tom B., et al. "Language Models Are Few-Shot Learners." *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 1877–1901.

Deriu, Jan, et al. "Survey on Evaluation Methods for Dialogue Systems." *Artificial Intelligence Review*, vol. 54, no. 1, 2021, pp. 755–810.

Gilson, Amanda, et al. "How Does ChatGPT Perform on the United States Medical Licensing Examination? The Implications of Large Language Models for Medical Education and Knowledge Assessment." *JMIR Medical Education*, vol. 9, no. 1, 2023, article e45312.

Kung, Tiffany H., et al. "Performance of ChatGPT on USMLE: Potential for AI-Assisted Medical Education Using Large Language Models." *PLOS Digital Health*, vol. 2, no. 2, 2023, article e0000198.

Liebrenz, Michael, et al. "Generating Scholarly Content with ChatGPT: Ethical Challenges for Medical Publishing." *The Lancet Digital Health*, vol. 5, no. 3, 2023, pp. e105–e106.

Lin, Chin-Yew. "ROUGE: A Package for Automatic Evaluation of Summaries." *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, Association for Computational Linguistics, 2004, pp. 74–81.

Lowe, Ryan, et al. "Towards an Automatic Turing Test: Learning to Evaluate Dialogue Responses." *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017)*, 2017, pp. 1116–1126.

Mehri, Shikib, and Maxine Eskenazi. "USR: An Unsupervised and Reference-Free Evaluation Metric for Dialog Generation." *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, 2020, pp. 681–707.

Ouyang, Long, et al. "Training Language Models to Follow Instructions with Human Feedback." *Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 27730–27744.

Papineni, Kishore, et al. "BLEU: A Method for Automatic Evaluation of Machine Translation." *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)*, 2002, pp. 311–318.

Radford, Alec, et al. *Language Models Are Unsupervised Multitask Learners*. OpenAI, 2019.

Sallam, Malik. "ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns." *Healthcare*, vol. 11, no. 6, 2023, article 887.

Shum, Heung-Yeung, Xiaodong He, and Di Li. "From ELIZA to XiaoIce: Challenges and Opportunities with Social Chatbots." *Frontiers of Information Technology & Electronic Engineering*, vol. 19, no. 1, 2018, pp. 10–26.

Vaswani, Ashish, et al. "Attention Is All You Need." *Advances in Neural Information Processing Systems*, vol. 30, 2017.

Weizenbaum, Joseph. "ELIZA—A Computer Program for the Study of Natural Language Communication between Man and Machine." *Communications of the ACM*, vol. 9, no. 1, 1966, pp. 36–45.

Zhang, Tianyi, et al. "BERTScore: Evaluating Text Generation with BERT." *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.