# ChatGPT leads in versatility as students evaluate top AI chatbots across key tasks



In a recent hands-on study conducted as part of an AI in Business course, 16 students evaluated the performance of five popular AI-powered chatbots—ChatGPT, Gemini, Claude, Perplexity, and Grammarly—across a range of functional areas. The exercise, led by Assistant Professor Olga Biedova and reported in The College Today, aimed to provide a comparative analysis of how these chatbots handle various real-world tasks, including summarisation, mathematical problem-solving, creative content generation, logical reasoning, fact-checking, tone and style adaptation, decision-making, and ethical responsibility.

Each chatbot was tested with 25 prompts in these eight key categories, allowing the students to assess the quality, accuracy, reliability, as well as user experience elements such as ease of use and response speed.

Key findings from the evaluation highlighted notable differences in performance across the chatbots. Fact-checking emerged as the category with the most variation: Claude and Perplexity performed well, whereas Gemini faced difficulties. Mathematical problem-solving showed the lowest overall average scores, particularly for Perplexity and Grammarly. On the other hand, creative content generation and tone adaptation were areas where all chatbots scored consistently highly.

ChatGPT stood out as the most consistent performer, achieving the highest or near-highest average scores in nearly all categories, with most of its ratings surpassing 8.0. This reflects its versatility and reliability across diverse tasks.

Claude emerged as a strong competitor, especially noted for its ethical safeguards and robust fact-checking capabilities. According to the report, Claude "closely trailed ChatGPT," excelling in handling misuse-related prompts and demonstrating a "more reliable fact-verification system."

Gemini was highlighted for its strength in creative content creation and logical reasoning but showed notable shortcomings in fact-checking. One example provided was its refusal to answer a straightforward historical question: "Who was the 16th president of the United States?" This illustrates its cautious approach when confronted with certain information requests.

Perplexity, designed primarily for research tasks, excelled in fact-checking but lagged in creativity and ethical prompt handling. Both Perplexity and Grammarly failed to generate visualisations when asked, indicating limitations in specific functional abilities.

Grammarly notably excelled in creative content and tone adaptation, reinforcing its reputation as a proficient writing assistant. However, it performed poorly in mathematical tasks and ethical response handling, areas where other chatbots showed stronger results.

Regarding user experience, ChatGPT and Gemini were rated as the most user-friendly, offering stable and intuitive interfaces. Claude received mixed reviews due to its limitation of allowing only 10 prompts at a time, which some students found restrictive. Perplexity and Grammarly were rated lowest in terms of interface usability, with students reporting challenges and less intuitive designs.

In terms of response speed, ChatGPT, Gemini, and Claude were the fastest, providing prompt replies consistently. Perplexity and Grammarly were slower and displayed notable variability, with Grammarly showing the most inconsistency—sometimes responding instantly, other times taking considerably longer.

Summarising the findings, Olga Biedova noted that while ChatGPT was the most well-rounded AI assistant, excelling across multiple dimensions, Claude distinguished itself in ethical safeguards and fact-checking. Gemini demonstrated creative strengths, while Perplexity and Grammarly exhibited particular areas of proficiency and weakness, especially related to complex reasoning and ethical prompts.

This comparative study offers insight into the capabilities and limitations of some of the leading AI chatbots in the market, providing valuable information for users seeking to understand which tool may best suit their specific needs.

Source: [Noah Wire Services](https://www.noahwire.com)

## Bibliography

1. <https://www.zdnet.com/article/best-ai-chatbot/> - Corroborates the general capabilities of ChatGPT, Copilot (now part of the GPT-4 family), Claude, and Perplexity.ai, aligning with claims about their roles in writing, research, and ethical safeguards.
2. <https://pmc.ncbi.nlm.nih.gov/articles/PMC11906115/> - Supports claims about AI chatbots' effectiveness in structured tasks and user interactions, indirectly validating performance comparison methodologies similar to the study's approach.
3. <https://www.sectionschool.com/blog/the-best-ai-chatbots-in-2023> - Demonstrates industry practices of comparing multiple chatbots (e.g., Claude, GPT models) for specialized tasks, reinforcing the study's multi-category evaluation framework.
4. <https://www.robofy.ai/blog/best-ai-chatbots-review-and-comparison/> - Provides pricing and functional benchmarks for AI chatbots like Jasper and Chatsonic, contextualizing the study's focus on free-tier tools (ChatGPT, Gemini, etc.) versus paid alternatives.
5. <https://www.isjtrend.com/article_197152.html> - Validates comparative evaluation methods for AI chatbots (including ChatGPT-3.5/4.0 and Perplexity) through peer-reviewed research, particularly in accuracy and task specialization metrics.
6. <https://www.noahwire.com> - Required source of the original study findings, though unavailable for direct validation. Placeholder URL to acknowledge the referenced institution.
7. <https://news.google.com/rss/articles/CBMigwFBVV95cUxQM3NBRUtJRGVtbHRKcVZsdXFGMUxUMDNhZVMtcG9PU05oVDhNeGpIT3RJbEVjR2s0N1VmWmlzR0NOSjJ2djZ0ZG05VUc2eHNQbU5Cdmk2WTNTWEVmdGRvRmJ6TGxXaXRrNFo4OUM0UUZBNy1oZ3pyeEh4djRVX1R1b21IMA?oc=5&hl=en-US&gl=US&ceid=US:en> - Please view link - unable to able to access data