# Evaluating the best language models for programming assistance



In a recently published analysis by ZDNet, technology expert David Gewirtz provides an in-depth evaluation of various large language models (LLMs) and their capabilities in the realm of programming assistance. Following the release of OpenAI's ChatGPT, Gewirtz undertook a testing journey, examining a total of 11 LLMs over the last 18 months.

The analysis identifies both promising contenders and underperformers among the myriad of chatbots currently available for coding tasks. Gewirtz notes a marked difference in the abilities of chatbots to execute programming requests, highlighting that, although some LLMs perform adequately on simple tasks, they often fail to create complex applications. In fact, five out of 10 LLMs tested did not successfully produce functional plugins for Gewirtz's initial test.

Among those recommended for programming assistance is Perplexity Pro, which Gewirtz describes as “the best overall AI chatbot for coding.” Perplexity Pro, priced at $20 per month, successfully passed all four of his programming tests and offers the flexibility of running multiple LLMs within its platform. However, Gewirtz expressed some concerns regarding its login method, noting that it relies solely on email-generated pins without standard multi-factor authentication.

Grok, developed by the social media platform formerly known as Twitter, received a favourable mention as well, with Gewirtz asserting its performance surpassed expectations given its recent creation. Although Grok stumbled in one of the programming tests, it managed to fare well in others, marking it as a developing tool worthy of observation.

A free alternative, ChatGPT, available in both a free and a Plus version, also demonstrated notable capabilities, particularly in its GPT-4o mode. Despite having limitations in peak server usage situations, it successfully passed three out of four tests. Gewirtz noted that while the free version is influential in many programming scenarios, it experiences throttling during heavy traffic, which can hinder its usability.

Perplexity's free version, based on GPT-3.5, was similarly acknowledged for its commendable programming performance, especially in research-oriented tasks. Gewirtz emphasised the scope of its research tools, which are considered advantageous for users balancing coding with exploration of broader topics.

On the less favourable side of the evaluation, Gewirtz highlighted several chatbots, including Microsoft’s Copilot, which demonstrated subpar results, failing to pass any of the programming tests. Meta AI and Meta Code Llama, which were also tested, exhibited poor performances, with Gewirtz pointing out their failure to generate functional code despite being designed with programming in mind.

As Gewirtz notes, this technology sector is evolving rapidly, with ongoing advancements likely to improve the reliability and functionality of AI tools for business applications. The results offer a context of both promise and caution, as businesses consider implementing these technologies into their operations.

The analysis concludes by acknowledging that while preferences for specific chatbots may vary among users, those engaged in programming tasks would benefit from being discerning in their choices, particularly when considering budget limitations. The continuous testing and emerging capabilities of chatbots signal a significant shift in integrating AI within business practices, paving the way for future improvements in automation and efficiency.

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

## Bibliography

1. <https://arxiv.org/html/2409.06416v1> - This link supports the capabilities and applications of large language models (LLMs) in software development and testing, including their use in generating code, automated program repair, and code review.
2. <https://direct.mit.edu/dint/article/6/1/240/118046/ChatGPT-is-a-Remarkable-Tool-For-Experts> - This article discusses the capabilities of ChatGPT and other LLMs in various domains, including their performance in structured tasks like code translation and explaining well-known concepts, as well as their limitations.
3. <https://direct.mit.edu/dint/article/6/1/240/118046/ChatGPT-is-a-Remarkable-Tool-For-Experts> - This source provides insights into the cognitive perspective on LLMs, distinguishing between formal and functional linguistic competence, and highlights the challenges and potential avenues for making LLMs more human-like.
4. <https://arxiv.org/html/2409.06416v1> - This study explores the use of LLM agents in test maintenance, including their ability to reason, plan, and interact with environments, and proposes multi-LLM architectures for complex tasks.
5. <https://direct.mit.edu/dint/article/6/1/240/118046/ChatGPT-is-a-Remarkable-Tool-For-Experts> - This article mentions the effectiveness of ChatGPT in supporting software process improvement practices, learning, and research, and the need for caution due to potential inaccuracies.
6. <https://arxiv.org/html/2409.06416v1> - This study discusses the importance of prompt engineering for LLMs, including strategies like few-shot prompting and chain-of-thought prompting to steer the generated output.
7. <https://direct.mit.edu/dint/article/6/1/240/118046/ChatGPT-is-a-Remarkable-Tool-For-Experts> - This source highlights ChatGPT's assistance in drafting and language editing to enhance readability and grammatical accuracy, relevant to its use in programming tasks.
8. <https://arxiv.org/html/2409.06416v1> - This study mentions the use of Retrieval-Augmented Generation (RAG) in LLM agents, allowing them to access external information to better respond to prompts.
9. <https://direct.mit.edu/dint/article/6/1/240/118046/ChatGPT-is-a-Remarkable-Tool-For-Experts> - This article discusses the applicability of ChatGPT across diverse domains, including its performance in tasks such as code translation and explaining well-known concepts.
10. <https://arxiv.org/html/2409.06416v1> - This study emphasizes the importance of considering both technological and organizational concerns when deploying LLMs in industrial development environments.
11. <https://direct.mit.edu/dint/article/6/1/240/118046/ChatGPT-is-a-Remarkable-Tool-For-Experts> - This source underscores the rapid evolution of AI technology and the ongoing advancements that are likely to improve the reliability and functionality of AI tools for business applications.
12. <https://www.zdnet.com/article/the-best-ai-for-coding-in-2025-and-what-not-to-use/> - Please view link - unable to able to access data