



Mikko Haavisto

The author is a doctoral researcher at the University of Turku, focusing on the use of educational technologies such as videos, artificial intelligence, and virtual reality in teaching and learning.

Kuva: Liisa Kallio

It might be time to reconsider AI-chatbots

Imagine it is November 2022, and ChatGPT has just launched. Visionaries quickly declare the arrival of intelligent tutoring systems for everyone, promising patient, intelligent, and affordable private tutors for all learners. But soon enough, problems arise. ChatGPT confidently produces false information, lacks knowledge about current events, and using it requires sending personal data across the world. It is slow, power-hungry, and cannot even do basic arithmetic. These criticisms were justified in 2022. However, in artificial intelligence (AI) years, that was ages ago, so let us see how things have changed.

KEYWORDS: Artificial intelligence, Large language models, Educational technology

Increases in capability and efficiency

Multiple factors have contributed to the rapid increase in AI capabilities, such as better hardware, more advanced AI models, and new tools integrated into the AI chatbots. These improvements have led to significant efficiency gains and a drop in prices. For example, one of today's smartest models, Deepseek V3, costs about \$1.50 to generate one million words (roughly the length of all seven Harry Potter books combined).

Although the original ChatGPT already exceeded average human performance in broad cross-disciplinary knowledge (Hendrycks et al., 2021), many people still found it lacking, since asking the chatbot was compared to searching for the answer online. A test that directly reflects this comparison is the Google-proof graduate-level benchmark (GPQA, Rein et al., 2023), which evaluates AI models without web access against humans who can use web search. Indeed, early ChatGPT answered only around 30% of the GPQA questions correctly, while non-expert and expert humans with web search scored around 34%

and 70%, respectively, using 30 minutes per question. However, current best models outperform even expert humans. For example, Claude 3.7 Sonnet (reasoning) gets around 80% correct, using just 2.5 minutes per question.

Example question from the GPQA benchmark

In a given population, 1 out of every 400 people has a cancer caused by a completely recessive allele, *b*. Assuming the population is in Hardy-Weinberg equilibrium, which of the following is the expected proportion of individuals who carry the *b* allele but are not expected to develop the cancer?

- a. 1/400
- b. 19/400
- c. 20/400
- d. 38/400



Early ChatGPT struggled with multi-step reasoning and even basic common sense. Modern reasoning models, such as Deepseek R1, address this by “thinking aloud”: they first generate a detailed, step-by-step chain of thought before producing a final response. These models can solve complex tasks more reliably by breaking problems into smaller, manageable parts and carefully considering each one. This shift from simply recognizing patterns in massive datasets to task-specific reasoning is a major step forward in AI, with much of its potential still unexplored. An added benefit of this approach is greater explainability. Since the model outputs its reasoning process, users can follow along and better understand how conclusions are reached. This makes it easier to assess whether the final response is correct.

There have also been significant hardware advancements. The original ChatGPT was trained and deployed on NVIDIA’s A100 chips. Since then, newer chips like the H100 have been developed, which are several times more efficient. Combined with model optimization techniques, these improvements enable operating models with just a fraction of the power consumption. As a result, today’s AI models are multiple times faster and cheaper to run. Remarkably, the original ChatGPT was already highly efficient: human writers are estimated to produce 130 to 1100 times more carbon dioxide emissions per page of text than the AI (Tomlinson et al., 2024).

Local models solve copyright and data privacy

Originally, ChatGPT posed serious copyright and data privacy concerns, since writing or pasting something into the chat box was sent on servers around the world without guarantee of confidentiality. In 2025, there is a solution to this problem: locally running, freely available AI language models. Because everything happens on the user’s

own device, sensitive and proprietary information is fully secure.

The original ChatGPT was so resource-intensive that it required multiple powerful computers just to operate. In contrast, today’s local models with better performance require no specialized hardware, just a standard laptop. Benchmarks like Chatbot Arena (lmarena.ai, Chiang et al., 2024) demonstrate this. On the website, visitors are shown responses from two anonymized models and asked to choose which one they prefer. The models are then ranked based on these preferences. Currently, an improved version of the original ChatGPT (GPT-3.5-turbo-1106) ranks 135th on the leaderboard, behind even Gemma 2 2b, a tiny local model that can run on nearly any modern computer.

Many institutions now use enterprise AI services, such as Microsoft Copilot. While they cannot guarantee absolute data privacy like local models, they can still be used with proprietary information, depending on company policy. This approach offers some of the benefits of local models without requiring personal hardware.

Your first chat with a local model in 5 minutes, for free, no login required

1. Download, install, and open **LM Studio** on your laptop or desktop.
2. Click **Skip onboarding** in the top right corner.
3. Click the **magnifying glass icon**.
4. Click **Gemma 3 1B**, then click **Download**.
5. Once it is downloaded, click **Load Model**.
6. That’s it! Start chatting with your fully private AI chatbot.

Bonus: Try **Gemma 3 4B**: it is smarter, multilingual, and you can even input images.



Overcoming the hallucination problem

Perhaps the biggest failing of the original ChatGPT was hallucination — confidently stating incorrect information. The root cause was that it relied entirely on its pretraining, which is like cramming the entire internet into a size small enough to fit on a smartphone. The problem is that detailed information is lost in the process, especially information that appeared only a few times online. Another limitation was that the original ChatGPT had no knowledge of anything that happened after its training data was collected.

Today, there are multiple solutions to these problems. First, chatbots now have access to web search, allowing them to retrieve up-to-date information and include references in their responses. Furthermore, AIs are increasingly integrated with real-time data sources like weather services and news feeds. In short, just like humans, having access to the internet allows AI to respond more accurately when discussing current events and specific details.

Second, the original ChatGPT was notoriously reluctant to respond with a simple “I don’t know.” Today, cross-checking an AI’s answers is quite straightforward. Two different models can be asked simultaneously, and if their answers differ, the chatbot does not know. A third model can even be used to compare the answers and automatically detect discrepancies. Thanks to increases in efficiency, running a chain of multiple models is now cheaper than asking a single model in 2022.

Third, the original ChatGPT’s context size was around 3,000 words, meaning users could not provide much reference material for it to search for facts. In contrast, current models like Qwen-2.5 (open, local) and Gemini 2.5 (proprietary, online) can handle over 700,000 words. This allows users to attach multiple books’ worth of information to the chat. When the AI has the relevant

content in its “working memory”, it hallucinates far less. The reason is straightforward: imagine having to answer a question based on a book you read some time ago, versus answering it while being able to browse the book.

It is important to note that when attaching copyrighted or confidential materials, a local model should be used to ensure data privacy. Additionally, it is best practice to ask the AI to produce word-for-word quotes from the source material, making the information easier to verify.

AI chatbots as instructional tutors

Recent developments have introduced several features that make AI especially useful as a tutor. One infamous issue with the original ChatGPT was its poor numerical ability: it often gave incorrect answers even for the most basic math problems (e.g., $11 - 8 + 24 - 57 = ?$). Luckily, today’s AI models can provide reliable answers by using calculators, much like humans would. With built-in reasoning and Python code execution, current models can reliably solve arithmetic and algebraic problems. Specialized systems, such as AlphaGeometry 2, can even solve most of Olympiad-level problems in their domain.

Another past limitation was the slow pace of written communication during learning. Current chatbots like GPT-4o now support voice interaction, image sharing, and even video calling, which enables a fluent learning experience. There are also powerful, freely available solutions that can run locally on most laptops; for example, Koko-ro for text-to-speech, Whisper for speech-to-text, and Gemma 3 4b, which can understand image input.

AI chatbots can sometimes be too helpful: they give the answer away easily, rather than guiding students to discover the solution themselves. Their explanations might also be too complicated for younger students. These issues



can be addressed in both local AIs as well as proprietary services by writing a system prompt – a predefined instruction that guides the AI's behavior. System prompts are pieces of text in the AI's memory and are not visible or editable by the user, allowing educators to shape how the AI responds. With this customization, teachers can prevent the chatbot from simply giving away answers, keep it focused on the lesson topic, and make it tailor its tone and complexity to the target audience and context.

Conclusion

AI chatbots have made remarkable progress in just 2.5 years, improving significantly across nearly every aspect. If you have not tried AI recently, now might be a great time to give it another go. That said, integrating all the discussed advances into a single system still requires a high-end computer and technical expertise. At present, there is no plug-and-play solution that offers a top-performing AI capable of running quickly and locally on a basic laptop, while also supporting web search, calculators, voice conversations, and less commonly spoken languages like Finnish. But with all the key components already in place, that may soon change.

References

- Chiang, W.-L., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., Zhang, H., Zhu, B., Jordan, M., Gonzalez, J. E., & Stoica, I. (2024). Chatbot arena: An open platform for evaluating llms by human preference (arXiv:2403.04132). arXiv. <https://doi.org/10.48550/arXiv.2403.04132>
- Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2021). Measuring massive multitask language understanding (arXiv:2009.03300). arXiv. <https://doi.org/10.48550/arXiv.2009.03300>
- Rein, D., Hou, B. L., Stickland, A. C., Petty, J., Pang, R. Y., Dirani, J., Michael, J., & Bowman, S. R. (2023). GPQA: A graduate-level Google-proof Q&A benchmark (arXiv:2311.12022). arXiv. <https://doi.org/10.48550/arXiv.2311.12022>
- Tomlinson, B., Black, R. W., Patterson, D. J., & Torrance, A. W. (2024). The carbon emissions of writing and illustrating are lower for AI than for humans. *Scientific Reports*, 14(1), 3732. <https://doi.org/10.1038/s41598-024-54271-x>

How I used AI when writing this article

1. I gathered my past texts, presentations, and relevant articles on AI into a single folder.
2. I passed those files, along with the topic and ideas for this article, to **Qwen-2.5 14b Instruct 1M** (running locally), asking it to find relevant points I have made and suggest new angles.
3. Based on that, I created a bullet-point draft covering the arguments, points, and article structure. I gave this draft to **Chat-GPT o3 Deep Research** multiple times, asking it to research relevant topics and write reports covering all bullet points.
4. I reviewed the reports, decided on the final content, and wrote the first full draft. While writing, I passed awkward or clunky phrases to **GPT-4o** for better alternatives.
5. Finally, **Gemma 3 27b** (running locally) reviewed the entire manuscript, for example pointing out grammatical and stylistic errors as well as detecting redundancy.