

AI's Secrets

What Students and Educators Need to Know About Chatbots

Professor Ken Purnell

Head of Educational Neuroscience, CQUniversity Australia

February 2026

Author note: Since July 2025, my work on AI and the brain has attracted sustained national media coverage across television, radio, and print, with more than 400 mentions and an estimated cumulative audience of over 13 million.

You may wish to view the short video overviewing key aspects of this paper at <https://share.descript.com/view/3v6ZfAUEuEh>

Generative AI, particularly large language model chatbots, is now embedded in students' study practices. These systems offer genuine benefits for brainstorming, drafting, and language refinement, yet they pose new risks to the quality of learning, academic integrity, and institutional responsibility. Understanding what chatbots can and cannot do is essential for educators and students.

Yann LeCun, then Meta's chief AI scientist, offers a revealing perspective: a four-year-old has encountered roughly 50 times as much real-world data as the largest language models today (LeCun, 2024a). Over 16,000 waking hours, a four-year-old absorbs continuous sensory experiences, including touch, sight, sound, and spatial relationships, that shape understanding in ways text alone cannot replicate. By contrast, current chatbots learn only from static text patterns, divorced from embodied experience, causality, and the grounded world models that enable genuine understanding (LeCun, 2024b). Figure 1 illustrates this fundamental limitation.

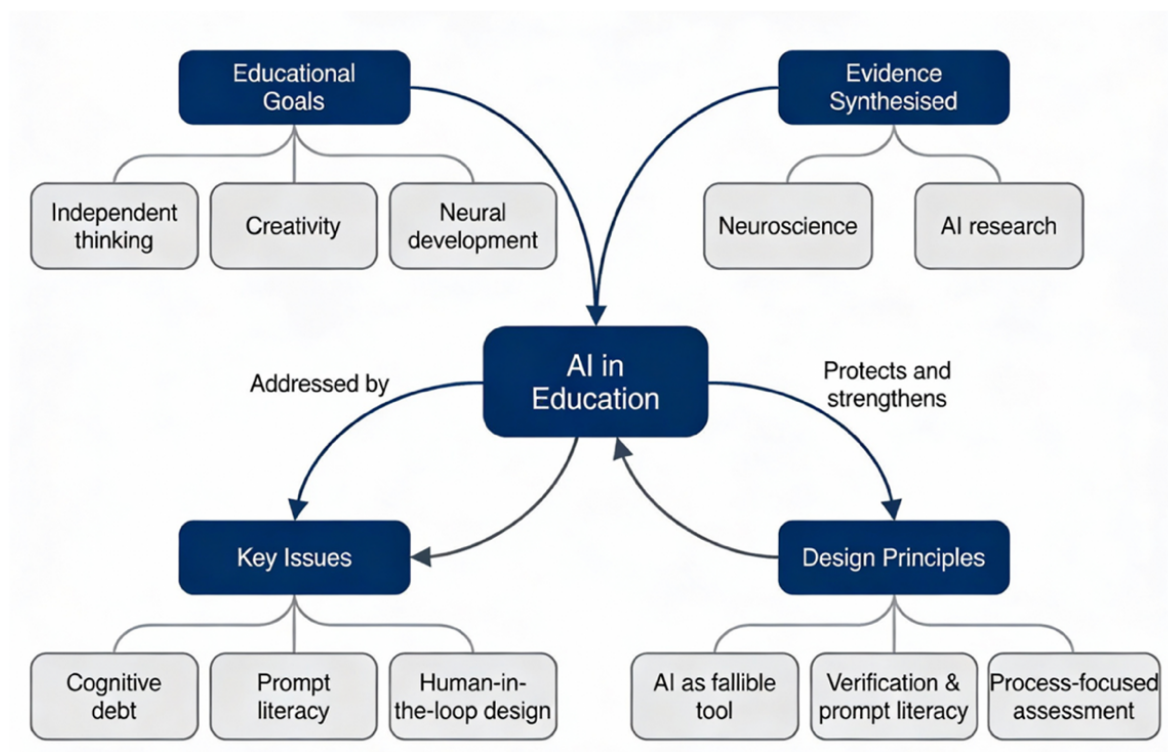


Figure 1. The architectural limitations of current chatbots: human learning, embodied experience, and AI's text-only training constraint

This contrast reveals the central truth: generative AI can support specific tasks, but it cannot replace human expertise, critical thinking, or deep learning. This paper outlines what students and educators need

to understand about how chatbots work, where they systematically fail, and why uncritical adoption carries significant cognitive, ethical, and environmental consequences.

The scale of infrastructure investment is staggering. Reinsel et al. (2018) projected that global data volumes would reach about 175 zettabytes by 2025, with an increasing share processed by large-scale data centres, raising pressing concerns about energy use, carbon emissions, and water consumption. Researchers are experimenting with **five-dimensional optical data storage** in nanostructured glass, using femtosecond lasers to create permanent nanoscale structures capable of storing hundreds of terabytes on a single disc, with lifespans measured in billions of years (Wang et al., 2022). Although 5D optical storage remains experimental and faces substantial technical barriers, it raises a provocative question: might the massive data-centre infrastructure buildout underway become technologically obsolete before its 20-year lifespan ends? If ultra-dense, ultra-durable storage were eventually commercialised at scale, it could reduce the physical footprint, cooling demands, and electronic waste associated with conventional data centres (Wang et al., 2022), potentially calling into question the long-term design assumptions underlying current infrastructure models. Indeed, some industry leaders now speculate that **future AI infrastructure may migrate off-planet**, with proponents arguing that solar-powered, space-based data centres could use the cold vacuum of space as a natural heat sink, thereby easing terrestrial cooling, land, and water pressures. However, these ideas remain largely theoretical and face formidable engineering, economic, and regulatory hurdles (EurekAlert, 2025; Musk, cited in *pv magazine Australia*, 2026; Wen et al., 2025).

The Fundamental Limitation: What Chatbots Are

Current large language models function primarily as probabilistic text predictors, with their capacity for genuine reasoning remaining limited and contested in AI research. They excel at predicting the next word from patterns in massive datasets, but prediction is not understanding. At their core, they operate as **statistical algorithms that generate text most likely to fit a prompt**, based on patterns in training data, rather than as systems explicitly designed to verify truth. This distinction is crucial because it clarifies a central paradox: outputs that sound confident, coherent, and persuasive can still be entirely false. When a language model encounters ambiguous prompts or novel situations outside its training distribution, it defaults to pattern-matching, which can produce fluent fiction (McKenna et al., 2023). LeCun and others argue that progress beyond this approach will require AI systems that learn through interaction, causal relationships, and embodied experience, rather than merely scaling model size and adding more text (LeCun, 2024b; Metz, 2026).

The Hallucination Problem: Scale and Scope

Researchers use the term *AI hallucinations* to describe outputs that appear fluent and contextually appropriate yet contain false claims, invented references, fabricated evidence, or logically incoherent assertions (Shao, 2025; Wang et al., 2024). These errors are particularly problematic in educational contexts because they are difficult to detect. Unlike an obviously incorrect or nonsensical response, a hallucinated claim can sound credible, especially to learners who lack the domain knowledge to identify the error.

This risk is not hypothetical. Research shows that conversational AI systems frequently generate non-existent references, fabricate evidence, and produce internally inconsistent arguments while maintaining polished, confident prose (see, for example, Google FACTS Team, 2025; McKenna et al., 2023). This is not a trivial technical limitation. It strikes at the core of academic integrity. Students who uncritically reproduce AI-generated work risk allegations of misconduct, even without intent to deceive. Institutions that deploy or endorse such tools therefore have a responsibility to implement robust verification processes, human oversight, and explicit guidance frameworks to safeguard academic standards and public trust.

Large-scale benchmarking underscores the seriousness of this issue. A comprehensive factuality benchmark released in late 2025 found that the highest-performing model achieved approximately 68.8 per cent accuracy across diverse knowledge tasks, with no system exceeding 70 per cent (Google FACTS Team, 2025). In practical terms, roughly one in three responses contained factual errors despite high

linguistic confidence. Independent newsroom testing by the BBC and the European Broadcasting Union reached similar conclusions, reporting accuracy issues in approximately 48 per cent of responses, with 17 per cent classified as serious factual or sourcing errors (Gaudiaut, 2025).

The Cognitive Cost: Learning and Memory

Beyond factual accuracy lies a deeper concern rooted in cognitive science. When students outsource analysis, synthesis, and argumentation to AI, they bypass the cognitive processes that build durable memories and deeper understanding. Recent neuroscientific research lends weight to this concern. In a brain-imaging study, students who used conversational AI to generate essays showed substantial reductions in functional connectivity during writing tasks and struggled to recall key details from their own essays just minutes after writing them (Kosmyna et al., 2025). These are preliminary findings that require cautious interpretation, yet the pattern is troubling: when AI generates the work, the brain engages less deeply because the neural pathways necessary for learning are not activated. Popular media outlets have amplified these concerns, with Chapados (2025) arguing that an MIT study shows AI tools can impair memory, critical thinking, and brain activity when overused.

To understand why, consider how the brain allocates its limited cognitive resources. Working memory, the mental workspace that temporarily holds and manipulates information, has strict capacity limits; on average, adults can hold approximately 4 to 7 items at a time (Cowan, 2014). This constraint is not a flaw; it is a feature. The peripheral sensory systems gather data at approximately one billion bits per second, yet the human brain's conscious processing operates at only around 10 bits per second (Zheng & Meister, 2024). This severe bottleneck explains why task-relevant prior knowledge is so consequential: when individuals retrieve organised knowledge structures, called schemas, from long-term memory, complex, familiar information is processed as single units rather than as multiple discrete elements, thereby reducing cognitive load (Ericsson & Kintsch, 1995; Purnell et al., 1991; Sweller, 1994). The more task-relevant knowledge stored in long-term memory, the less demand is placed on working memory, as prior learning provides ready-made frameworks for understanding new material. When students think deeply about a problem, grapple with complexity, and construct their own explanations, they actively build these durable schemas and strengthen long-term learning.

Cognitive load theory explains the mechanism. When tasks are well designed, the mental effort required to build meaningful schemas - what researchers call *germane cognitive load* - promotes deep learning (Cowan, 2014; Dubinsky & Hamid, 2024; Purnell, 2026). However, when AI significantly reduces the cognitive struggle, it may also diminish the productive cognitive challenge necessary for effective memory encoding and concept formation. Students who receive polished AI drafts experience ease at the expense of comprehension, precisely the opposite of what learning requires.

What Chatbots Actually Do Well

The picture is not entirely bleak. Generative AI excels at specific, bounded tasks when used intentionally and with sustained human oversight. For example:

Ideation and conceptual clarification. Chatbots are excellent thinking partners for brainstorming and exploring alternative perspectives. A student struggling with an abstract idea can request simple explanations, analogies, or worked examples and often receives genuinely useful responses. The key is intentionality: AI should expand thinking, not displace it.

Accessibility and genuine support. For students with dyslexia, language-based learning disabilities, or hearing impairments, AI tools can be transformative. Chatbots can summarise dense readings, generate alternative explanations, transcribe audio notes, and adapt text for accessibility. In these contexts, AI is a scaffold for equity, not a shortcut.

Scaffolded feedback and low-stakes iteration. Students can draft responses, request feedback, revise, and iterate, provided they retain ownership of their thinking and verify content independently. This

approach, in which AI serves as a feedback partner rather than an authority or author, builds skills through practice with human accountability.

Research synthesis with verification. AI can help students locate and synthesise information from multiple sources, provided human judgment remains central and verification is non-negotiable. The student must independently verify claims against authoritative databases and conduct their own critical evaluation.

The common thread is clear: AI works best as a thinking amplifier within a framework of human oversight and critical judgment. It is most problematic when students use it to replace rather than enhance their own thinking, and it may impede learning.

The Hidden Costs: Infrastructure, Environment, and Governance

The widespread adoption of AI systems entails substantial hidden costs that institutions and individuals often overlook. For example,

Energy and climate impacts. Data centres account for just over 4 per cent of total electricity consumption in the United States, and demand is projected to more than double by around 2030, largely driven by AI workloads (Gabbatiss, 2025; O'Donnell & Crownhart, 2025). When schools, universities, and businesses adopt AI at scale, they engage in broader climate and energy trade-offs that warrant explicit consideration in institutional planning. Beyond energy, data centre operations compete for scarce local resources - land and water - in the communities where they are sited, often exacerbating existing environmental pressures and creating tensions with local water supplies and agricultural or urban needs.

Financial unsustainability and the infrastructure timescale problem. The financial sustainability of this infrastructure is increasingly scrutinised in industry analyses. Recent industry forecasts project that the global data-centre market will reach US\$500 billion to over US\$1 trillion by the early to mid-2030s (Alaamer, 2025; Knight Frank, 2025; Maintworld, 2025). Data centres lock in energy demand for 20 years or more, long before technological obsolescence might render them unnecessary. Yet internal corporate communications reveal a critical mismatch: institutions struggle to integrate advanced AI systems into established workflows in complex, regulated environments (Weiss, 2026). This gap between infrastructure investment and operational capacity highlights potential implementation challenges when technology adoption outpaces organisational readiness and workflow integration.

Copyright, compliance, and data governance. Copyright and intellectual property concerns compound these challenges. Large-scale data acquisition initiatives have involved purchasing and digitising substantial numbers of physical books to train language models (Schaeffer et al., 2026). For universities, these initiatives pose significant compliance risks. Educational providers are increasingly advised to implement rigorous internal AI governance frameworks that include mandatory human review of AI-generated content, explicit copyright checks, and clear expectations for student use of AI in assessments (MIT Sloan Teaching & Learning Technologies, 2024; Sandhu, 2024). These frameworks need to address fundamental questions, such as: What student data is sent to third-party AI services? How long is it retained, and is it used to train commercial models? Have algorithms been audited for bias? Can students opt out of AI-integrated features?

Towards pragmatic institutional practice. A sustainable approach recognises these realities: **deploy AI where it demonstrably improves learning outcomes; avoid adoption driven solely by novelty or vendor pressure.** Institutions that take their responsibility to protect academic integrity, student data, and environmental stewardship seriously will make deliberate, evidence-based decisions about which tools to adopt and how to govern their use.

Emerging Capabilities and Technical Limitations

Multimodal large language models, which process text, images, audio, and sometimes real-time data, bring AI closer to what LeCun calls “grounded” learning. Students can now upload lecture slides for explanation, analyse visual sources, or access real-time data. This creates genuine pedagogical

opportunities but also introduces new risks: AI systems can misinterpret images, generate deepfakes, and reinforce biases in training data.

However, recent research reveals sobering technical constraints that temper these prospects. Pure large language model-based agents face fundamental mathematical limitations in completing tasks beyond modest complexity (Landymore, 2026). Empirically, currently marketed agentic features cannot reliably run long, open-ended tasks in the background after a conversation ends, despite marketing claims to the contrary (Barlow, 2026). Independent verification of AI system capabilities is essential, as performance may vary significantly across use cases and contexts. Institutions and individual learners should evaluate AI tools through rigorous, evidence-based testing rather than relying solely on vendor claims.

Toward Responsible Implementation

Rather than banning or uncritically adopting AI, institutions and educators can adopt deliberate, evidence-based approaches that align with educational goals. The path forward rests on four complementary strategies:

AI literacy as a core curriculum. Students must be taught to recognise how AI works and where it fails. Specifically, they should: document AI use transparently in all work; cross-check every claim and citation independently before citing them; distinguish between AI-generated text and their own analysis; use AI to expand, not displace, their reasoning; and treat all AI outputs as rough first drafts requiring rigorous verification. These practices shift AI from a shortcut to a genuine thinking tool.

Retrieval-Augmented Generation (RAG) systems. Some institutions are pioneering more trustworthy alternatives that use RAG to search curated, vetted repositories, including course materials, institutional databases, and authoritative academic sources, before generating responses. By grounding outputs in verified information sources, these systems are more transparent and defensible than general-purpose chatbots, substantially reducing the risk of fabrication and hallucination.

Assessment redesign. Many educational institutions are moving towards methods that resist AI shortcuts. These include in-class, invigilated assessments (essays under exam conditions, oral presentations, real-time problem-solving); portfolios that document students' thinking across multiple drafts and reflections; iterative assessment with explicit human feedback between submissions; and explicit instruction that rewards students for showing their working. Together, these redesigns prioritise learning over efficiency.

Institutional data governance. As AI integrates into learning management systems and grading platforms, institutions must answer critical questions transparently. These include: Which student data is sent to third-party AI services, and for how long it is retained. Is that data used to train commercial models? Have the algorithms underpinning these systems been independently audited for bias? Can students meaningfully opt out of AI-integrated features? These questions should be answered publicly and regularly to build confidence among students and educators and to benefit the community.

These approaches preserve academic integrity while preparing students for a world where AI is embedded in professional practice. **The key is intentionality: adopting AI where evidence shows it improves learning, and refusing to adopt it when driven by vendor pressure or the mere novelty of the technology.**

Bringing It Together

Generative AI is neither magic nor the enemy. It is a probabilistic tool that excels at ideation, concept explanation, and drafting, yet falls short in reasoning, verification, and ethical judgment. The same technology can support a student with dyslexia's learning, streamline a researcher's literature review, or enable wholesale academic misconduct if used uncritically.

The path forward requires educators and students to have:

- **Clear understanding of how AI works and where it fails**, grounded in technical literacy and neuroscience
- **Explicit instruction in AI literacy**: prompting, verification, source evaluation, and ethical use.
- **Institutional governance** that balances innovation with integrity, equity with caution, and pedagogical benefit with environmental cost.
- **Ongoing research** into the cognitive, neural, and social impacts of AI-assisted learning.

As LeCun reminds us, the most powerful learning comes from rich, embodied interaction with the world—something no chatbot can yet replace. **The secret that educators and students must grasp is this: the tool is only as good as the human judgment that guides it.** Do not outsource your thinking to a machine. Instead, use AI to amplify what you can think, verify, and decide. Your human agency and responsibility to engage critically with the world are irreplaceable; the tool is not.

References

- Alaamer, K. (2025, April 22). *This is the state of play in the global data centre gold rush*. World Economic Forum. <https://www.weforum.org/stories/2025/04/data-centre-gold-rush-ai/>
- Barlow, G. (2026, January 26). I can't actually keep working on a long, manual task like this in the background once a message turn ends: ChatGPT has a major limitation that needs to be addressed and fast. *TechRadar*. <https://www.techradar.com/ai-platforms-assistants/chatgpt/i-cant-actually-keep-working-on-a-long-manual-task-like-this-in-the-background-once-a-message-turn-ends-chatgpt-has-a-major-limitation-that-needs-to-be-addressed-and-fast>
- Chapados, A. (2025, June 20). MIT studied the effects of using AI on the human brain—the results are not good. *Blaze Media*. <https://www.theblaze.com/return/mit-chatgpt-ai-brain-study>
- Cowan, N. (2014). Working memory underpins cognitive development, learning, and education. *Educational Psychology Review*, 26(2), 197–223. <https://doi.org/10.1007/s10648-013-9246-y>
- Dubinsky, J. M., & Hamid, A. A. (2024). The neuroscience of active learning and direct instruction. *Neuroscience & Biobehavioral Reviews*, 163, Article 105737. <https://doi.org/10.1016/j.neubiorev.2024.105737>
- Ericsson, K. A., & Kintsch, W. (1995). Long-term working memory. *Psychological Review*, 102(2), 211–245. <https://doi.org/10.1037/0033-295X.102.2.211>
- EurekAlert!. (2025, October 26). NTU Singapore scientists propose carbon-neutral data centres powered from space. *EurekAlert*. <https://www.eurekalert.org/news-releases/1103424>
- Gabbatiss, J. (2025, September 15). AI: Five charts that put data-centre energy use and emissions into context. *Carbon Brief*. <https://www.carbonbrief.org/ai-five-charts-that-put-data-centre-energy-use-and-emissions-into-context/>
- Gaudiaut, T. (2025, November 29). How accurate are AI chatbots? *Statista*. <https://www.statista.com/chart/35534/ai-chatbots-accuracy-rate-of-inaccurate-responses-2025/>
- Google FACTS Team. (2025, December 11). The FACTS Leaderboard: A comprehensive benchmark for large language model factuality. *Google*. https://storage.googleapis.com/deepmind-media/FACTS/FACTS_benchmark_suite_paper.pdf
- Knight Frank. (2025, April). Global data centre market is projected to reach US\$4 trillion by 2030. *Knight Frank*. <https://www.knightfrank.com.au/blog/2025/04/17/global-data-centre-market-is-projected-to-reach-us4-trillion-by-2030>
- Kosmyna, N., Hauptmann, E., Yuan, Y. T., Situ, J., Liao, X.-H., Beresnitzky, A. V., Braunstein, I., & Maes, P. (2025, June 9). Your brain on ChatGPT: Accumulation of cognitive debt when using an AI assistant for essay writing task. *arXiv*. <https://doi.org/10.48550/arXiv.2506.08872>
- Landymore, F. (2026, January 26). AI agents are mathematically incapable of doing functional work, paper finds. *Futurism*. <https://futurism.com/artificial-intelligence/ai-agents-incapable-math>
- LeCun, Y. [@ylecun]. (2024a, January 25). A child has seen 50 times more data than the biggest LLMs [Post]. X. <https://x.com/ylecun/status/1750614681209983231>
- LeCun, Y. (2024b, March 7). Yann LeCun: Meta AI, open source, limits of LLMs, & the future of AI (Lex Fridman Podcast #418) [Interview transcript]. *Lex Fridman Podcast*. <https://lexfridman.com/yann-lecun-3-transcript/>

- Maintworld. (2025, November 29). Data center market to exceed \$1 trillion by 2035. *Maintworld*. <https://www.maintworld.com/News/Data-Center-Market-to-Exceed-1-Trillion-by-2035>
- McKenna, N., Li, T., Cheng, L., Hosseini, M. J., Johnson, M., & Steedman, M. (2023). Sources of hallucination by large language models on inference tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 2758–2774). Association for Computational Linguistics. <https://aclanthology.org/2023.findings-emnlp.182.pdf>
- Metz, C. (2026, January 26). An AI pioneer warns the tech 'herd' is marching into a dead end. *The New York Times*. <https://www.nytimes.com/2026/01/26/technology/an-ai-pioneer-warns-the-tech-herd-is-marching-into-a-dead-end.html>
- MIT Sloan Teaching & Learning Technologies. (2024). When AI gets it wrong: Addressing AI hallucinations and bias. *MIT Sloan Teaching & Learning Technologies*. <https://mitsloanedtech.mit.edu/ai/basics/addressing-ai-hallucinations-and-bias/>
- O'Donnell, J., & Crownhart, C. (2025, May 20). We did the math on AI's energy footprint. Here's the story you haven't heard. *MIT Technology Review*. <https://www.technologyreview.com/2025/05/20/1116327/ai-energy-usage-climate-footprint-big-tech/>
- Purnell, K. N., Solman, R. T., & Sweller, J. (1991). The effects of technical illustrations on cognitive load. *Instructional Science*, 20, 443–462. <https://doi.org/10.1007/BF00116358>
- Purnell, K. (2026). *Cognitive Load Theory informed by educational neuroscience and artificial intelligence: Implications for preservice teachers and teacher educators*. Zenodo. <https://doi.org/10.5281/zenodo.18370535>
- pv magazine Australia. (2026, January 27). Musk says solar space-based AI data centres possible in two to three years. *pv magazine Australia*. <https://www.pv-magazine-australia.com/2026/01/28/musk-says-solar-space-based-ai-data-centres-possible-in-two-to-three-years/>
- Reinsel, D., Gantz, J., & Rydning, J. (2018). *The digitization of the world: From edge to core (Data Age 2025)*. Seagate Technology; International Data Corporation. <https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf>
- Sandhu, S. (2024, October 17). Why AI should not be used for course development: The looming threat of copyright infringement. CAQA. <https://caqa.com.au/blogs/news/why-ai-should-not-be-used-for-course-development-the-looming-threat-of-copyright-infringement>
- Schaeffer, A., Oremus, W., & Tiku, N. (2026, January 27). Inside an AI start-up's plan to scan and dispose of millions of books. *The Washington Post*. <https://www.washingtonpost.com/technology/2026/01/27/anthropic-ai-scan-destroy-books/>
- Shao, A. (2025, August 27). New sources of inaccuracy? A conceptual framework for AI hallucinations and misinformation. *Harvard Kennedy School Misinformation Review*. <https://misinfoeview.hks.harvard.edu/article/new-sources-of-inaccuracy-a-conceptual-framework-for-studying-ai-hallucinations/>
- Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction*, 4(4), 295–312. [https://doi.org/10.1016/0959-4752\(94\)90003-5](https://doi.org/10.1016/0959-4752(94)90003-5)
- Wang, H., Lei, Y., Wang, L., Sakakura, M., Yu, Y., Shayeganrad, G., & Kazansky, P. (2022). 100-layer error-free 5D optical data storage by ultrafast laser nanostructuring in glass. *Laser & Photonics Reviews*, 16(4), Article 2100563. <https://doi.org/10.1002/lpor.202100563>
- Wang, Y., Wang, M., Manzoor, M. A., Liu, F., Georgiev, G., Das, R. J., & Nakov, P. (2024). Factuality of large language models in the year 2024. *arXiv*. <https://arxiv.org/html/2402.02420v2>
- Weiss, G. (2026, January 26). Emails show Bank of America's struggles with Nvidia AI: 'You have to help us as local car mechanics drive the race car!' *Business Insider*. <https://www.businessinsider.com/bank-of-america-nvidia-ai-internal-emails-2026-1>
- Wen, S., Hsu, C., & Al-Falasi, M. (2026, January 16). How data centres in space sustainably enable the AI age. *World Economic Forum*. <https://www.weforum.org/stories/2026/01/data-centres-space-ai-revolution/>
- Zheng, J., & Meister, M. (2024). The unbearable slowness of being: Why do we live at 10 bits/s? *Neuron*, 113(1), 1–20. <https://doi.org/10.1016/j.neuron.2024.11.008>