Statistical approximation is not general intelligence

Walter Quattrociocchi[1], Valerio Capraro[2], Gary Marcus[3]

[1]Sapienza University of Rome, Rome, Italy

[2]University of Milan–Bicocca, Milan, Italy

[3]New York University, New York, NY, USA

Contact author: walterquattrociocchi@gmail.com

A shorter version of this piece appeared as:

Quattrociocchi, W., Capraro, V., Marcus, G. (2026). Statistical approximation is not general intelligence. *Nature*, 659, 792.

Rumors that humanity has already achieved artificial general intelligence (AGI) have been greatly exaggerated. Such rumors are often fueled by recent advances in large language models (LLMs), whose outputs show strong benchmark performance, high fluency across domains, and, in some cases, correct solutions to open problems in mathematics. These developments are often taken as evidence that general intelligence has been achieved.

Such interpretations rest on a fundamental confusion between performance on individual, often well-known tasks and intelligence writ large. Task-level performance, even when impressive, is not sufficient evidence of general intelligence. Here, we argue that recent claims of achieving AGI rest on a conceptual error: conflating increasingly sophisticated statistical approximations with intelligence itself. We also show that recent claims (Chen et al., 2026) about putative success on AGI hinge on redefining what AGI has historically meant.

Definitions of artificial general intelligence

The term artificial general intelligence was originally introduced to denote systems capable of robust, flexible competence across a wide range of environments and tasks, emphasizing generality, flexibility, adaptability, and transfer under novelty rather than success on fixed or curated task batteries.

A widely cited formal definition by Legg and Hutter (2007) characterized intelligence as an agent's ability to achieve goals across a broad range of environments, with robustness and generalization at its core. Related accounts, including those by Goertzel (2014), similarly emphasized open-ended learning and domain-general problem-solving over performance on predefined tasks.

This understanding remained broadly stable for many years. Benchmark-driven progress was widely recognized as valuable but insufficient to establish abstraction, reliability, or genuine generalization. More recently, Hendrycks et al. (2023) reaffirmed the importance of robustness, dependable generalization beyond curated benchmarks, and resistance to systematic failure as essential criteria for AGI—criteria on which current systems fall short.

In recent years, however, some have tried to redefine AGI, weakening the term. Some accounts implicitly redefined AGI in terms of broad behavioral performance across benchmarks, often in parallel with growing commercial and strategic incentives. Other redefinitions attempted to reframe AGI in economic rather than cognitive terms, equating it with the ability to perform a large share of economically useful human work.[1]

Most recently, AGI has at times been treated as synonymous with strong performance on standardized benchmarks. By focusing on specific benchmarks, these definitions miss the generality that was central to the original definitions, as discussed in the next section, thereby reflecting a fundamental misunderstanding of how intelligence should be evaluated. As such, they conflate increasingly sophisticated approximations of intelligence with intelligence itself.

[1] https://garymarcus.substack.com/p/the-five-stages-of-agi-grief?utm_source=publication-search

Benchmark success per se does not entail general intelligence

Much of the argument that artificial general intelligence has already been achieved rests on benchmark performance (e.g., Chen et al., 2026). Benchmarks evaluate specific capabilities under controlled conditions and have been useful for tracking progress. For example, Chen and colleagues argue that success on the Turing Test constitutes evidence of AGI.

However, benchmark success is a limited indicator of general intelligence. By design, benchmarks isolate narrow competencies and abstract away real-world context, making it difficult to distinguish genuine generalization from pattern recognition. Strong benchmark performance often provides little evidence of robustness under novelty, uncertainty, or shifting objectives.

Moreover, benchmarks are often easily gamed. The Turing Test itself illustrates this point: systems have been reported to "pass" it by exploiting superficial conversational cues and the gullibility of non-expert evaluators, rather than by exhibiting genuine understanding or reasoning (e.g., Eugene Goostman). In recent years, as the financial stakes of benchmark performance have increased, systems have been trained more often directly on benchmarks ("teaching to the test") or on synthetic data closely resembling them. As a result, model development and data curation are optimized for benchmark success, producing systems that perform well under test conditions but degrade in real-world settings that differ only modestly from them, as shown in recent medical studies, where models remain accurate despite missing key inputs yet become unstable under minor distribution shifts, generating fluent but flawed reasoning (Gu et al. 2025).

Economic evidence underscores this gap. At the task level, analyses of workplace automation show that even state-of-the-art systems can reliably perform only a small fraction of the tasks required in typical human occupations (Eloundou et al., 2023; Felten et al., 2023), despite high scores on idealized benchmarks. At the aggregate level, macroeconomic analyses suggest similarly modest effects: (Acemoglu, 2025) estimates that AI-driven automation would increase total factor productivity by no more than 0.66% over a ten-year horizon. Additional evidence suggests that only a minority of firms report economically meaningful returns from AI deployment (Nanda, 2025), a finding difficult to reconcile with claims of general intelligence.

Taken together, these considerations indicate that benchmark performance, even across many tasks, is not sufficient evidence of general intelligence. General intelligence would presumably entail strong benchmark performance, but available real-world evidence suggests that the core properties of general intelligence – flexibility, generality, and reliability – remain elusive.

The limits of behavior-based definitions

The overindexing on often gameable benchmarks reflects a broader and recurring conceptual error: a tendency to treat observable behavior as decisive evidence of intelligence, without sufficient attention to underlying mechanisms. In psychology, behaviorism offered a pragmatic way to study cognition without reference to internal mental states; in artificial intelligence, operational tests such as the Turing Test similarly prioritize external performance over underlying mechanisms.

These approaches were useful, but ultimately insufficient. A central insight from cognitive psychology is that similar behaviors can be produced by fundamentally different processes, and that producing the right output does not imply the same cognitive capacities (Anderson, 1976). A classic illustration comes from behavioral psychology: pigeons trained to discriminate complex visual stimuli—such as photographs of people versus those without people—were able to extend this discrimination to previously unseen images (Herrnstein & Loveland, 1964). Yet such performance does not imply abstraction, transfer, or flexible reasoning beyond the trained context.

The same lesson applies to artificial systems. Large language models increasingly approximate human behavior across many tasks, often producing outputs indistinguishable from those of humans in controlled settings. However, as in classic cases from behavioral psychology, similar outputs can arise from fundamentally different underlying processes. Behavioral similarity alone therefore provides no insight into whether the underlying processes support core components of general intelligence, such as judgment, error correction, or reliable generalization.

Concrete evidence of this divergence comes from controlled comparisons of human and LLM judgments. In a direct evaluation of news source reliability conducted under identical conditions, large language models often matched human evaluators in their final classifications while relying on systematically different evaluative criteria. For example, when assessing politically salient or controversial news sources with mixed or incomplete evidence, human evaluators frequently downgraded confidence or withheld firm judgments, explicitly citing uncertainty, lack of corroboration, or potential costs of error. Language models, by contrast, tended to issue confident classifications in the same cases, even when uncertainty was explicitly signaled in the input. In particular, models exhibited high confidence even when evidence was sparse, conflicting, or explicitly flagged as uncertain, whereas human judgments were strongly modulated by uncertainty, justification, and perceived costs of error. As a result, agreement at the level of outputs concealed substantial divergence in the underlying judgment processes, reflecting a systematic substitution of epistemic evaluation with generative plausibility—a phenomenon we term *epistemia* (Loru et al., 2025).

More broadly, this pattern points to deeper epistemological divergences between human and machine judgment. Across domains involving causality, value trade-offs, justification, and error monitoring, humans attend to reasons, stakes, and responsibility, whereas language models optimize for linguistic plausibility—structural fault lines between performance alignment and genuine evaluative competence (Quattrociocchi et al., 2025). Related distortions are also observed in behavioral simulations, where surface-level alignment masks systematic

exaggeration of salient traits (Nudo et al., 2025). These divergences become clear when current systems are evaluated against the original criteria for artificial general intelligence.

Where current systems fall short

By the standards articulated in the original definitions of artificial general intelligence—robustness across environments, reliable generalization under novelty, and autonomous goal-directed behavior—current AI systems remain limited. Despite impressive gains in narrow competence and fluency, today's large language models lack persistent goals, struggle with long-horizon reasoning, and depend extensively on human scaffolding for task formulation, evaluation, and correction. Reports that language models have produced correct proofs for isolated open problems in mathematics, including specific Erdős problems, do not alter this assessment. As noted by mathematicians such as Terence Tao[2], these results primarily reflect the ability to rapidly search, recombine, and iterate over existing techniques, rather than the emergence of genuinely novel or domain-general problem-solving strategies. Moreover, inclusion in the Erdős list does not by itself imply exceptional conceptual difficulty, as some problems remain unsolved due to relative obscurity rather than depth.

These limitations are central rather than peripheral. They directly concern reliability under uncertainty, resistance to systematic failure, and cross-domain transfer without task-specific tuning. On these dimensions, current systems remain brittle, sensitive to prompt framing, and inconsistent outside curated evaluation settings. Recognizing these constraints does not diminish recent progress; it clarifies its scope. Current AI systems are powerful and increasingly useful tools, but they do not exhibit the flexible, self-directed competence that the original concept of artificial general intelligence was intended to capture. As AI systems become embedded in scientific and institutional decision-making (NIST, 2023; European Commission, 2025), overestimating their cognitive capacities risks misallocating trust, responsibility, and authority. Confusing increasingly sophisticated statistical approximation with general intelligence is therefore not only a conceptual error, but a strategic misjudgment.

[2] https://github.com/teorth/erdosproblems/wiki/AI-contributions-to-Erdős-problems

References

Chen, E. K., Belkin, M., Bergen, L., & Danks, D. (2026). Does AI already have human-level intelligence? The evidence is clear. *Nature*, *650*(8100), 36-40.

Legg, S., & Hutter, M. (2007). Universal intelligence: A definition of machine intelligence. *Minds and machines*, *17*(4), 391-444.

Goertzel, B. *Artificial General Intelligence: Concept, State of the Art, and Future Prospects.* Springer (2014).

Hendrycks, D., Song, D., Szegedy, C., Lee, H., Gal, Y., Brynjolfsson, E., ... & Bengio, Y. (2025). A definition of AGI. *arXiv preprint arXiv:2510.18212*.

Gu, Y., Fu, J., Liu, X., Valanarasu, J. M. J., Codella, N. C., Tan, R., ... & Vozila, P. (2025). The Illusion of Readiness in Health AI. *arXiv preprint arXiv:2509.18234*.

Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2024). GPTs are GPTs: Labor market impact potential of LLMs. *Science*, *384*(6702), 1306-1308.

Felten, E., Raj, M., & Seamans, R. (2023). How will language modelers like ChatGPT affect occupations and industries?. *arXiv preprint arXiv:2303.01157*.

Acemoglu, D. (2025). The simple macroeconomics of AI. Economic Policy, 40(121), 13-58

Nanda, M. (2025). State of AI in business 2025. *Preprint at https://www. artificialintelligence-news. com/wp-content/uploads/2025/08/ai_report_2025. Pdf*.

Anderson, J. R. (2013). *Language, memory, and thought*. Psychology Press.

Herrnstein, R. J., & Loveland, D. H. (1964). Complex visual concept in the pigeon. *Science*, *146*(3643), 549-551.

Loru, E., Nudo, J., Di Marco, N., Santirocchi, A., Atzeni, R., Cinelli, M., ... & Quattrociocchi, W. (2025). The simulation of judgment in LLMs. *Proceedings of the National Academy of Sciences*, *122*(42), e2518443122.

Nudo, J., Pandolfo, M. E., Loru, E., Samory, M., Cinelli, M., & Quattrociocchi, W. (2026). Generative exaggeration in LLM social agents: Consistency, bias, and toxicity. *Online Social Networks and Media*, *51*, 100344.

Quattrociocchi, W., Capraro, V., & Perc, M. (2025). Epistemological fault lines between human and artificial intelligence. *arXiv preprint arXiv:2512.19466*.

NIST. *AI Risk Management Framework.* (2023). https://www.nist.gov/itl/ai-risk-management-framework

European Commission. Guidelines for providers of general-purpose AI models. (2025). https://digital-strategy.ec.europa.eu/en/policies/guidelines-gpai-providers