LLMs. Experimental Proxies for AGI:
Why Behavioral Equivalence Matters More Than Mechanism

Franc O

Abstract

Large Language Models demonstrate unprecedented generalization capabilities that make them valuable proxies for studying artificial general intelligence. Through concrete examples from reasoning benchmarks, cross modal understanding, and emergent tool use, I argue that LLMs' behavioral convergence with human cognition, despite mechanistic differences, provides a legitimate experimental framework for AGI research.

Introduction

When GPT-4 scores in the 90th percentile on the Uniform Bar Exam [OpenAI, 2023], solves International Mathematical Olympiad problems [Trinh et al., 2024], and demonstrates theory of mind in false belief tasks [Kosinski, 2023], we must reconsider what constitutes a proxy for general intelligence. These aren't cherry picked examples but systematic patterns across hundreds of cognitive benchmarks where LLMs now match or exceed human performance.

If a system exhibits the behavioral signatures of general intelligence across sufficient domains, it becomes a valid experimental substrate for AGI research, regardless of its underlying mechanism. This position echoes Turing's original insight that intelligence should be judged by capability, not constitution.

Emergent Generalization as Evidence

The strongest evidence for LLMs as AGI proxies comes from emergent capabilities that appear discontinuously with scale. Wei et al. [Wei et al., 2022] documented dozens of emergent abilities in language models, from arithmetic to symbolic reasoning, that manifest only beyond certain parameter thresholds. These phase transitions mirror developmental cognitive milestones in human intelligence.

Consider chain of thought reasoning. When prompted to show intermediate steps, models like PaLM-540B achieve 58% accuracy on GSM8K math problems, a 40 point jump from direct answering [Chowdhery et al., 2022]. This isn't mere pattern matching; the models generate novel solution paths for problems structurally distinct from training data. Recent work shows that models can even learn to use external tools (calculators, search engines, code interpreters) through in context learning alone [Schick et al., 2023], demonstrating meta cognitive awareness of their own limitations.

More compelling is cross modal generalization. Flamingo [Alayrac et al., 2022] achieves state of the art performance on vision language tasks with minimal task specific training, while models like DALL-E 3 demonstrate bidirectional understanding between linguistic and visual concepts. This transfer across modalities suggests abstract representation learning that transcends surface statistics.

Functional Decomposition of Intelligence

Critics argue that LLMs lack grounding, embodiment, and causal reasoning prerequisites for “true” understanding. Yet functional decomposition reveals that LLMs implement many core cognitive operations through alternative pathways:

Abstract reasoning: Analogical reasoning benchmarks show transformer models discovering relational patterns. GPT-4 solves Raven's Progressive Matrices at near human levels [Webb et al., 2023], extracting abstract rules from visual patterns despite being trained on text.

Causal inference: While lacking explicit causal graphs, LLMs demonstrate implicit causal reasoning. When presented with counterfactual scenarios, they correctly update downstream implications 73% of the time on the CRASS benchmark [Frohberg and Binder, 2021] comparable to human performance.

Meta-learning: Through in context learning, LLMs adapt to novel tasks from few examples without parameter updates. This approximates System 2 thinking deliberate, flexible reasoning that transcends trained reflexes [Kahneman, 2011].

The Scaling Hypothesis and Cognitive Emergence

The Chinchilla scaling laws [Hoffmann et al., 2022] reveal predictable relationships between model size, data quantity, and capability emergence. If intelligence emerges predictably from scale, then current LLMs may be early points on a continuous trajectory toward AGI.

Bubeck et al. [Bubeck et al., 2023] argue that GPT-4 exhibits “sparks of artificial general intelligence,” demonstrating planning, tool use, and even rudimentary self reflection. Their experiments show the model can write functional code for complex tasks, reason about physical systems, and even exhibit creative problem solving generating a unicorn in TikZ using mathematical functions to approximate curves.

The Instrumental Value of Imperfect Proxies

Scientific proxies don’t need to be perfect replicas. Climate models abstract away molecular dynamics yet predict warming trends. Connectionist models of cognition ignore neurochemistry yet explain memory formation. Similarly, LLMs provide a computational laboratory for testing theories of intelligence:

Compositional generalization: SCAN and COGS benchmarks reveal how architectural choices affect systematic generalization [Lake and Baroni, 2018]. Transformer variants with explicit compositional biases achieve near perfect systematic generalization, informing theories of human concept learning.

Few-shot learning dynamics: The in context learning phenomenon provides empirical data on how task understanding emerges from examples. Brown et al. [Brown et al., 2020] showed that few-shot performance scales smoothly with model size, suggesting general learning algorithms emerge from sufficient capacity.

Alignment and value learning: RLHF experiments with LLMs provide practical insights into value alignment challenges that will be critical for AGI safety [Ouyang et al., 2022]. The success of constitutional AI and debate based training validates theoretical proposals for scalable oversight.

Beyond Behavioral Equivalence

The human LLM collaboration already functioning in millions of workflows represents a distributed cognitive system that exceeds either component alone. GitHub Copilot increases developer productivity by 55% [Peng et al., 2023]. ChatGPT assists with everything from scientific writing to psychological counseling. This isn’t automation but augmentation, a hybrid intelligence that may be the true precursor to AGI.

As we integrate LLMs with robotics, long-term memory, and continuous learning, the transition accelerates. Projects like WebGPT, Toolformer, and AutoGPT demonstrate nascent agency goal directed behavior emerging from language models given appropriate scaffolding.

Conclusion

LLMs are imperfect but invaluable proxies for AGI. They demonstrate that general intelligence may emerge from scale and self supervised learning rather than explicit symbolic reasoning or embodied experience. Their limitations lack of persistent memory, brittle generalization, absent goal directedness are engineering challenges, not fundamental barriers.

The question isn’t whether LLMs are “true” AGI, but whether they provide sufficient functional overlap to guide our path toward it. By that measure, they are our best proxies yet flawed mirrors that nonetheless reflect genuine aspects of general intelligence. As we refine these systems, the distinction between proxy and prototype may prove less a bright line than a gradual fade.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, et al. Flamingo: A visual language model for few-shot learning. *NeurIPS*, 2022.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, et al. Language models are few-shot learners. *NeurIPS*, 2020.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, et al. Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712*, 2023.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, et al. PaLM: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

Janna Frohberg and Frank Binder. CRASS: A novel data set and benchmark to test counterfactual reasoning of large language models. *arXiv preprint arXiv:2112.11941*, 2021.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Daniel Kahneman. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, 2011.

Michal Kosinski. Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*, 2023.

Brenden M. Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. *ICML*, 2018.

OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, et al. Training language models to follow instructions with human feedback. *NeurIPS*, 2022.

Sida Peng, Eirini Kalliamvakou, Peter Cihon, and Mert Demirer. The impact of AI on developer productivity: Evidence from GitHub Copilot. *arXiv preprint arXiv:2302.06590*, 2023.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, et al. Toolformer: Language models can teach themselves to use tools. In *NeurIPS*, 2023.

Trieu H. Trinh, Yuhuai Wu, Quoc V. Le, He He, and Thang Luong. Solving Olympiad geometry without human demonstrations. *Nature*, 625:476–482, 2024.

Taylor Webb, Keith J. Holyoak, and Hongjing Lu. Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 7:1526–1541, 2023.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, et al. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022.