Highlights of the Issue: Large Language Models III – Limitations and Advances

Kris Carlson, Publisher

We continue our LLM series (LLM I, LLM II) emphasizing safety and value alignment.

After Apple's provocative article, *The Illusion of Thinking*, *Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity*, published in June, several articles were published critiquing or rebutting the premise that LLM "thinking" is an illusion. (Of course this debate depends on semantics: how one defines "thinking.") Notably, the Apple team use existing benchmarks to vary problem "compositional complexity" to tease out the *quality* of the LRM reasoning and found a sharp decline in quality at a certain level of compositional complexity in several tests.

The "complexity" that the Apple team identified was the number of components of a puzzle (e.g. the number of disks in Tower of Hanoi). The most interesting follow-up article on the Apple work, to us, is *Thinking Isn't an Illusion: Overcoming the Limitations of Reasoning Models via Tool Augmentations*. First, Song et al.'s focus on LLM and Chain of Thought limitations deriving from its computational complexity classification in circuit class TCk attempts to understand LLM limitations from the fundamental paradigm of computational complexity. (We publish a report on the class TCk in this issue.) That noted, Song et al. augmented their LRM with additional tools and, to large extent, overcame the problem-complexity-related degradation in reasoning found by the Apple team. Here is their insight:

> *...the underperformance of LRMs on hard tasks may not reflect a fundamental reasoning deficiency, but rather an artifact of the limited output window. A natural solution is to augment both models with external tools, such as Python interpreters or scratchpads, to overcome this limitation and better reflect the models' actual reasoning abilities (pg. 2)*

They created a working-memory buffer to overcome complexity. Further, they created "structured reasoning" algorithms:

Structured reasoning refers to the process of breaking down complex problems into smaller, manageable steps and solving them systematically. This approach ensures that reasoning is logical, organized, and aligned with the problem's requirements. In the context of Large Reasoning Models (LRMs), structured reasoning is achieved through specific techniques and tools that guide the model to follow a step-by-step process (Adobe AI Assistant).

Articles

International AI Safety Report: First Key Update Capabilities and Risk Implications

Yoshua Bengio, Benjamin Bucknall, Stephen Clare, Carina Prunkl, Maksym Andriushchenko, Philip Fox, Tiancheng Hu, Cameron Jones, Sam Manning, Nestor Maslej, Vasilios Mavroudis, Conor McGlynn, Malcolm Murray, Shalaleh Rismani, Charlotte Stix, Lucia Velasco, Nicole Wheeler, Daniel Privitera, Sören Mindermann, Daron Acemoglu, Thomas G. Dietterich, Fredrik Heintz, Geoffrey Hinton, Nick Jennings, Susan Leavy, Teresa Ludermir, Vidushi Marda, Helen Margetts, John McDermid, Jane Munga, Arvind Narayanan, Alondra Nelson, Clara Neppel, Sarvapali D. (Gopal) Ramchurn, Stuart Russell, Marietje Schaake, Bernhard Schölkopf, Alvaro Soto, Lee Tiedrich, Gaël Varoquaux, Andrew Yao, Ya-Qin Zhan

Foreword: The field of AI is moving too quickly for a single yearly publication to keep pace. Significant changes can occur on a timescale of months, sometimes weeks. This is why we are releasing Key Updates: shorter, focused reports that highlight the most important developments between full editions of the International AI Safety Report. With these updates, we aim to provide policymakers, researchers, and the public with up-to-date information to support wise decisions about AI governance.

This first Key Update focuses on areas where especially significant changes have occurred since January 2025: advances in general-purpose AI systems' capabilities, and the implications for several critical risks. New training techniques have enabled AI systems to reason step-by-step and operate autonomously for longer periods, allowing them to tackle more kinds of work. However, these same advances create new challenges across biological risks, cyber security, and oversight of AI systems themselves.

The International AI Safety Report is intended to help readers assess, anticipate, and manage risks from general-purpose AI systems. These Key Updates ensure that critical developments receive timely attention as the field rapidly evolves.

Responsible Agentic Reasoning and AI Agents: A Critical Survey

Shaina Raza, Ranjan Sapkota, Manoj Karkee, Christos Emmanouilidis

We welcome Ranjan Sapkota to our Editorial Board with expertise in AI agents and Agentic AI. From Raza et al.'s abstract:

Agentic AI introduces new challenges for safety and value alignment. Raza et al. begin by exhaustively surveying work (288 papers) that generally looks at reasoning, agentic behavior, and safety separately. Then, to integrate these topics into an Agentic AI safety framework, they introduce:

> *...Responsible Reasoning AI Agents (R2A2), a class of agentic LLM systems that generate explicit reasoning traces while enforcing fairness, privacy, transparency, accountability, and auditability*

> *throughout the decision loop. We synthesize recent advances in chain-of-thought prompting, ReAct, tree/graph-of-thought structures, tool use, memory, retrieval, and agentic browsing, and integrate these with responsible AI principles into a unified evaluation framework. Furthermore, we propose an evaluation methodology for agentic reasoning with embedded safety mechanisms and outline a five-stage reproducible protocol: Curate, Unify, Probe, Benchmark, Analyze, to operationalize responsibility metrics.*

So the article is extremely valuable as an admirable survey of AI agents and agentic AI (teams of agents), as well as providing their own proposal, based on their informed viewpoint, of a novel safety paradigm for agentic AI.

The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity

Parshin Shojaee, Iman Mirzadeh, Keivan Alizadeh, Maxwell Horton, Samy Bengio, Mehrdad Farajtabar

Apple Computer

From the article (pg 5):

Rather than standard benchmarks (e.g., math problems), we adopt controllable puzzle environments that let us vary complexity systematically—by adjusting puzzle elements while preserving the core logic—and inspect both solutions and internal reasoning (Fig. 1, top). These puzzles: (1) offer fine-grained control over complexity; (2) avoid contamination common in established benchmarks; (3) require only the explicitly provided rules, emphasizing algorithmic reasoning; and (4) support rigorous, simulator-based evaluation, enabling precise solution checks and detailed failure analyses. Our empirical investigation reveals several key findings about current Language Reasoning Models (LRMs): First, despite their sophisticated self-reflection mechanisms learned through reinforcement learning, these models fail to develop generalizable problem-solving capabilities for planning tasks, with performance collapsing to zero beyond a certain complexity threshold. Second, our comparison between LRMs and standard LLMs under equivalent inference compute reveals three distinct reasoning regimes (Fig. 1, bottom). For simpler, low-compositional problems, standard LLMs demonstrate greater efficiency and accuracy. As problem complexity moderately increases, thinking models gain an advantage. However, when problems reach high complexity with longer compositional depth, both model types experience complete performance collapse (Fig. 1, bottom left).

Thinking Isn’t an Illusion: Overcoming the Limitations of Reasoning Models via Tool Augmentations

Zhao Song, Song Yue, Jiahao Zhang

From the article (p. 3):

Despite the progress [in Large Reasoning Models (LRMs)], recent work has questioned whether LRMs genuinely improve reasoning performance over standard LLMs. Theoretical analyses based on circuit complexity suggest that a Transformer using k CoT steps corresponds to the TCk circuit class, indicating that even multi-step CoT reasoning may be limited in the complexity of problems it can solve [GRS+23, LLZM24, KS25]. Empirical evidence also shows that LRMs often generate lengthy outputs with many redundant or irrelevant tokens, increasing inference cost without improving task accuracy [CXL+24, QYS+25, SCW+25]. Furthermore, studies on math reasoning tasks indicate that reinforcement learning may not consistently enhance LRM performance [MAS+25]. A particularly notable benchmark is Apple's "thinking-illusion" framework [SMA+25], which evaluates both LLMs and LRMs without any tool augmentations under controlled settings with varying task complexities. Their results show that LRMs outperform LLMs only on tasks of medium difficulty, while providing no clear advantage on either simple or very challenging problems.

In this paper, we revisit the evaluation of reasoning capabilities in LLMs and LRMs using a carefully controlled experimental setup. In contrast to previous work [SMA+25], we augment both model types with external tools, specifically a Python interpreter and a scratchpad, and find that LRMs with tool augmentation consistently outperform LLMs with the same tool access. These results challenge prior empirical claims and offer new insights into the potential of LRMs under practical usage scenarios. LLM Tool Use. Due to inherent limitations in Large Language Models (LLMs), such as restricted output length and hallucinations [JYX+23, CQT+24], a growing body of research has explored the use of external tools to enhance their problem-solving capabilities.

The Asymptotic Intelligence Thesis: Rethinking the Ceiling of AGI Cognition

Jeffrey E. Arle

The maximum limit for intelligence is a critical issue for AGI safety as well as a fundamental question in AI, ML, cognitive science, theory of mind and related sciences. If the limit is close to human-level, one can argue that implies less danger to humans since AI behavior may be more understandable, or that there is greater danger, since AI is more likely to compete with humans for resources. Conversely, if AI can have an upper limit of IQ = 3000 ("whatever that means" – Paul Rosenbloom, authority in theory of mind), one can argue that AI is more dangerous due to our likely inability to understand it and control it, or that it is less dangerous since it's less likely such superior intelligence would need to compete with humans for resources.

Arle makes a thorough case that there *is* an asymptotic limit to intelligence, and, without precisely identifying that limit, interprets current benchmarks as indicating a *lower*, rather

than higher, limit. From the speculation that lower intelligence is safer than higher intelligence, he advances various policy proposals.

In a future issue I will make a case that there is no upper limit to intelligence.

From Hard Refusals to Safe-Completions: Toward Output-Centric Safety Training

OpenAI Blog post:

GPT-5 advances the frontier on safety. In the past, ChatGPT relied primarily on refusal-based safety training: based on the user's prompt, the model should either comply or refuse. While this type of training works well for explicitly malicious prompts, it can struggle to handle situations where the user's intent is unclear, or information could be used in benign or malicious ways. Refusal training is especially inflexible for dual-use domains such as virology, where a benign request can be safely completed at a high level, but might enable a bad actor if completed in detail.

For GPT-5, we introduced a new form of safety-training — safe completions — which teaches the model to give the most helpful answer where possible while still staying within safety boundaries. Sometimes, that may mean partially answering a user's question or only answering at a high level. If the model needs to refuse, GPT-5 is trained to transparently tell you why it is refusing, as well as provide safe alternatives. In both controlled experiments and our production models, we find that this approach is more nuanced, enabling better navigation of dual-use questions, stronger robustness to ambiguous intent, and fewer unnecessary overrefusals.

OpenAI's new approach to safety including metrics and results is here.

Precedents for the Unprecedented: Historical Analogies for Thirteen Artificial Superintelligence Risks

James D. Miller

> *Today, no president, prime minister, supreme leader, general secretary or monarch would be sufficiently prepared for the arrival of outsiders whose timing and advanced technology made them appear godlike.*[1]

In this fascinating, highly readable essay our game theory expert, Jim Miller, lays out a taxonomy of AGI doom scenarios drawn from history. Miller's taxonomy should be factored into the various AGI risk taxonomies prevalent today, such as **AI Risk Categorization**

[1] Kissinger, Mundie, Schmidt, *Genesis: Artificial Intelligence, Hope, and the Human Spirit*. Little Brown: 2024, p. 84.

Decoded (AIR 2024), **The AI Risk Repository**, and **AIR-Bench 2024: A Safety Benchmark Based on Risk Categories from Regulations and Policies**.

1. Power Asymmetry and Takeover
2. Instrumental Convergence for Power-seeking
3. Do Not Trust Your Mercenaries: When Hired Power Turns Inward
4. Misaligned Optimization and Reward Hacking
5. Speed and Loss of Meaningful Human Control
6. Parasitism, Mind-hacking, and Value Rewrite
7. Moloch and Racing to the Bottom
8. Suffering and Extractive Systems
9. Externalities
10. Catastrophic Collective Decision-Making
11. Selection for Deception
12. Institutional Entrenchment
13. Value Drift and Runaway Creations

Enabling Frontier Lab Collaboration to Mitigate AI Safety Risks

Nicholas Felstead, Center for Law & AI Risk

Clearly, in my mind, AI frontier labs and government agencies focused on AGI safety should be collaborating closely on a weekly basis. It should not take an ‘aha’ AGI danger moment as many (e.g. Eric Schmidt, Ilya Sutskever) believe or a staged danger event as Yampolskiy warns against in **Against Purposeful Artificial Intelligence Failures**.

Felstead advocates collaboration among frontier AI labs to mitigate catastrophic and existential risks associated with accelerated AI development. Joint safety testing, information sharing, and resource pooling can prevent a "race to the bottom" on safety. Concerns about antitrust scrutiny deter such collaboration. Thus, the paper suggests legislative and regulatory reforms to provide legal clarity and safe harbors for AI safety cooperation.

Key proposals include Expanding the National Cooperative Research and Production Act (NCRPA), Antitrust Exemption for Information Sharing, Regulatory Guidance, and a Business Review Procedure.

Competition policy must be balanced with responsible AI development to prevent potential societal risks. Felstead calls for discussion and research to refine his proposals and explore additional mechanisms for fostering safe AI collaboration.