Advancing AI: Reasoning

A vision forward for networks

White paper

Authors: Anne Lee, Alonso Silva, Chun-Nam Yu, David Rinaldi, Elie De Panafieu,
Gerard Burnside, Johann Daigremont, Karun Nithi, Liubov Tupikina, Sean Kennedy

Since the introduction of large language models in late 2022, the subfield of AI reasoning has evolved significantly. While showing promise in basic reasoning capabilities, early models exhibited limitations such as hallucinations, mathematical challenges, brittleness and a lack of explainability. Recent work on AI reasoning has attempted to create more accurate and efficient reasoning models that require less computing, training time and costs. In this paper we review the continuing innovations in model architecture, training approaches and evaluation methods and assess them with respect to creating AI systems that can reason effectively across multiple dimensions and demonstrate true AI wisdom. Addressing these challenges, while adhering to principles of universality, flexibility, explainability and scalability, will help us to realize the full potential of AI reasoning capabilities across various domains from telecommunications to broader real-world applications.

Contents

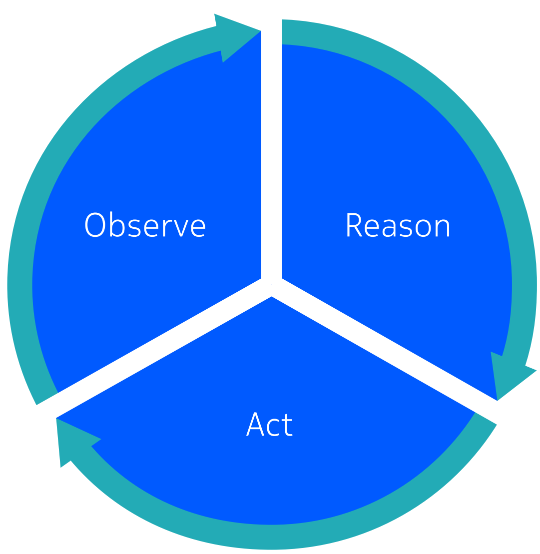
Introduction

The AI landscape shifted dramatically in late 2022 with the public debut of large language models (LLMs) and AI chatbots, the highest profile being ChatGPT, powered by GPT3.5 LLM. This marked the dawn of generative AI systems capable of producing human-like text, audio, images and code. These LLMs could simulate human dialogue, perform semi-complex math and summarize content, suggesting an emerging capacity for logic. They could, if prompted in the right way, reason through a problem and provide a correct answer. These systems exhibited not only fluent language capabilities but also the illusion of understanding and basic reasoning. However, the limitations of these early models soon became apparent when subsequent usage exposed critical flaws:

- **Hallucinations**—fabrication of erroneous statements that cannot be verified
- **Mathematical challenges**—struggles with advanced, abstract, intuitive mathematical reasoning
- **Brittleness**—inability to generalize for out-of-distribution scenarios
- **Opacity**—black box "logic" with limited explainability.

These limitations would also restrict their utility in real-world applications. But the launch of LLMs was just the beginning. Their early abilities, however flawed, sparked the imagination of researchers and investors. And soon, researchers pivoted to working on AI. Investors poured money into AI applications and solutions. Meanwhile, governments rushed and struggled to keep up with needed regulations to ensure that AI would responsibly improve the way people work and live.

Since the introduction of ChatGPT, new offerings have rapidly evolved from simple AI chatbots that provided a user-friendly interface for one primary AI model to AI agents. A general AI agent is an autonomous or semi-autonomous system that uses AI models as its core to observe its environment, reason about the information it processes—including making decisions—and takes actions to achieve specific goals through function calls and tools. From general AI agents emerged agentic AI agents. An agentic AI agent is a special type of AI agent that exhibits autonomous decision-making and self-directed behavior with the capacity to set its own goals, often with a high degree of independence. Agentic AI agents are more advanced with the ability to modify their objectives or strategies without direct human input.



Currently, there is a significant effort and emphasis on developing and enhancing the reasoning capabilities of AI models to advance artificial intelligence [1][2]. The definition of "reasoning" in the context of AI is the ability of an AI system to study a given situation, use existing knowledge to draw conclusions, make decisions and solve problems. There are multiple types of reasoning, we have listed seven that are common:

1. **Deductive reasoning**—drawing a conclusion based on premises that are generally assumed to be true, e.g., "all mammals are warm-blooded; whales are mammals; therefore, whales are warm-blooded"
2. **Inductive reasoning**—drawing a general conclusion from observations of specific examples, e.g., "you observe the sun setting in the west every day that you check, and you conclude that the sun always sets in the west"
3. **Abductive reasoning**—drawing a probabilistic conclusion to explain an observation based on incomplete knowledge, e.g., "the lawn is wet; therefore, it must have rained, however, it is possible that the sprinkler turned on"
4. **Causal reasoning**—drawing a conclusion on cause-and-effect relationships based on observations, evidence, and/or prior knowledge, e.g., "evidence shows that smoking (cause) increases the risk of developing lung cancer (effect)"
5. **Analogical reasoning**—drawing a conclusion for a new situation by comparing it to a known situation that shares similarities, e.g., "electricity flowing through a wire is like water flowing through a pipe; just as narrowing the pipe reduces the water flow, increasing resistance in the path of a wire reduces the electric current"
6. **Search and planning reasoning**—drawing a conclusion on the best sequence of steps to take to go from a starting point to the goal, e.g., "to take the optimal route from your home to your friend's house, you first search all possible paths and steps given travel options such as walking, driving and taking a bus and constraints such as traffic, time and costs, then you plan your route, putting together the best combination of steps to reach your friend
7. **Mathematical reasoning**—drawing a conclusion to a mathematical problem or proof utilizing deductive and inductive reasoning, e.g., "since any number for which the sum of its digits is divisible by 3 is also divisible by 3, then the number 51 is divisible by 3 because 5+1=6 is divisible by 3."

To achieve more accurate outcomes and solve more complex problems, AI models need to go beyond simple pattern recognition. They must be able to collect initial relevant information, make hypotheses, reduce the search space, determine missing information if any, collect more information, maintain memory over a relevant period, and draw conclusions to make decisions. In short, AI models must be able to reason.

Advanced AI models should NOT be simply memorizing information or just correlatively recognizing patterns.

Current AI models are still lacking capabilities to address reasoning fully. For example, most LLMs rely on an approach that is token-level only. Lack of transparency must be addressed for better explainability. And hallucinations are still a major issue that limits the applicability of AI to solving critical problems. There are also inefficiencies in how reasoning is performed in today's models that must be improved for the sake of cost, latency and sustainability.

Resolving these limitations will lead to more effective tools for root cause analysis, troubleshooting, operation management, network planning and more. Addressing these limitations will also go beyond addressing the telecom problem space. These are fundamental improvements that would apply to many domains and ultimately change the way we currently solve problems.

Goals

Strong reasoning capabilities will overcome the limitations of current models and accelerate adoption of AI. In the long run, strong reasoning enables and is enabled by:

- Explainability and transparency
- Hallucination mitigation
- Self-checking and self-correction
- Problem solving
- Planning
- Autonomy

Reasoning is foundational to achieving artificial general intelligence (AGI) or superintelligence. But, even if AGI is not a goal, the ability to reason well is useful and many times vital to task-specific AI solutions. Reasoning well is important for creating agentic agents and their specialized large action model (LAM) components—function-calling agents that can ultimately act autonomously.

AI can be utilized for root cause analysis, troubleshooting, trouble ticket assignment, operations management, resource management, energy management, self-healing, network security, network planning, intelligent consumer services, and much more. All applications mentioned can benefit from AI that demonstrates strong reasoning capabilities.

It should be noted that there are basic research efforts ongoing in the AI field to better mimic the way human brains work using minimal energy, requiring minimal training data, refining learning approaches, etc., that may produce new architectures that go beyond the original transformer [3]. This area is dynamic and will continue to be monitored.

Tenets

Tenets are guiding principles that inform the decision-making of a company. They are often inspirational and define north stars.

The long-term vision of AI is that it will become ubiquitous and native to all software systems. AI will be the future of software. Tactically, there is still a lot of work to get there. To achieve this vision, developers must focus on creating more intuitive interfaces and enhancing the integration of AI capabilities across various platforms. Additionally, addressing ethical considerations and ensuring data privacy will be crucial as we move toward this AI-driven future.

While classical software today is not without flaws, AI has unique impediments as well as areas where improvements would widen its scope of applicability. Enhancing and expanding the reasoning capabilities of AI models will guide us to this long-term vision of nativeness and ubiquity.

The ability to reason in AI models is important for practical purposes. This capability will improve the performance and accuracy of the outputs of AI models while broadening the scope of actions that an AI system can perform.

The guiding principles for AI reasoning are:

1. **Universality**—generalizability of a reasoning solution to any use case is vital; avoid bespoke approaches dependent on the problem, when possible
2. **Flexibility**—services that can adapt flexibly to changing environments will offer the most value
3. **Explainability**—ensuring the transparency and explainability of AI decision-making processes is crucial, especially in safety-critical applications
4. **Scalability**—necessary to ensure that AI solutions can effectively address the needs of large-scale environments.

Reasoning in AI models can be improved in multiple dimensions, which include the seven types of reasoning mentioned earlier along with abstract reasoning, common sense reasoning, and ethical reasoning. Ultimately, reasoning is a fundamental component of AI wisdom. AI wisdom is the ability of an AI system to make sound decisions and provide insightful recommendations based on a deep understanding of complex situations. Such capability doesn't just involve processing vast amounts of data and recognizing patterns but also ensuring consideration of ethics, contextual awareness, and the long-term implications of its decision-making processes. Further research is necessary to address foundational areas and ensure the feasibility of reasoning AI systems.

State of the business: today's reasoning models and approaches

When ChatGPT launched, the strategy for enhancing the intelligence and accuracy of AI foundational models involved increasing the model size (i.e., scaling up to trillions of parameters), augmenting the training data, and boosting computational resources. This is called pre-training scaling. Since then, two more scaling laws have emerged: post-training scaling and test-time scaling [4].

In post-training scaling, a pretrained foundational model is used to create a derivative model that is adapted to a specific set of use cases or domain. This is done with techniques such as supervised fine-tuning (SFT), pruning, quantization, distillation, reinforcement learning (RL), reinforcement learning with human feedback (RLHF), and synthetic data augmentation.

Test-time scaling, also known as long thinking, takes place during inference. Models use this technique to reason through multiple potential responses before arriving at the best answer. This approach requires more time and computing but provides higher quality responses. Implementation approaches include chain-of-thought prompting, sampling with majority voting, and search (exploring multiple paths of a tree-like structure of responses).[1]

In addition to the scaling laws, work is now underway to abstract up above the token level. Models are being researched to operate at a "concept" level using higher-level semantic representations that could be in the form of words, phrases, sentences, or more. These are called large concept models (LCMs) [5]. New models may also go beyond linguistics to be multi-modal, or they may be non-linguistic, modeling concepts in a latent space [6] that find parallels in mathematics or are analogous to human reasoning; the results eventually get translated back to linguistics as output.

State of the art (SOTA) reasoning models

At the time of this writing (August 2025), the following reasoning models were introduced, and additional new models are expected to follow soon [7]. You can find a listing of reasoning models on the Hugging Face leaderboard [8].

Reasoning models

- OpenAI o1/mini/pro, o3/mini+high, etc.
- DeepSeek R1 and distilled models
- Hugging Face Open-R1 initiative
- Anthropic Claude 3.7 Sonnet
- Google Gemini 2.0 Flash & Pro Thinking
- Alibaba Qwen QwQ/QvQ models
- ByteDance Doubao-1.5-pro
- Moonshot AI Kimi k1.5
- Microsoft Research rStar-math
- UC Berkeley TinyZero

1 Note that the data used during training may also have a large influence on the inference-time capability. See https://arxiv.org/abs/2503.01307

- Cognitive Computations Dolphin R1 tuned models (e.g., latest Mistral).

Reasoning datasets

- Dolphin-r1, OpenR1-Math-220k, etc.
- Curated list on Hugging Face [9]

To measure the performance of these models, there are many evolving and emerging benchmarks to gauge and compare the models against each other. Performance metrics and goals may cover the accuracy of the knowledge about facts in specific domains, the skill to derive correct answers from test questions, or "intelligence."

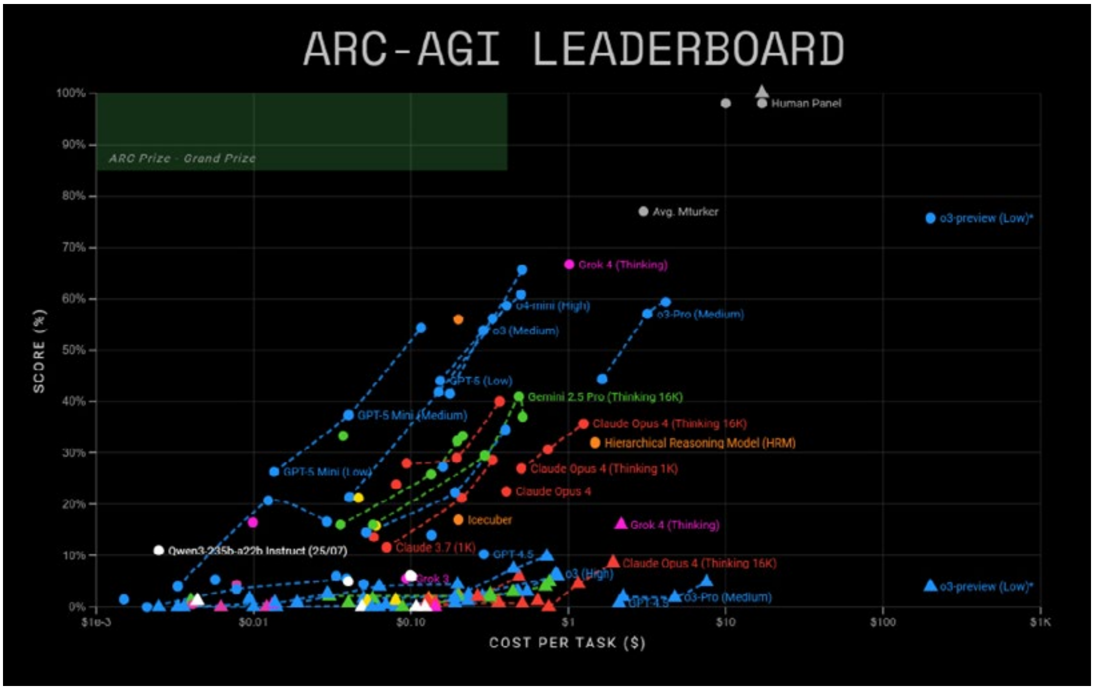
There is no standard definition of "intelligence" for AI. Therefore, for this paper, we will adopt the definition provided by Stanford University's Human-Centered Artificial Intelligence [10].

Intelligence might be defined as the ability to learn and perform suitable techniques to solve problems and achieve goals appropriate to the context in an uncertain, ever-varying world. A fully pre-programmed factory robot is flexible, accurate and consistent but not intelligent.

Reasoning ability in an AI model is therefore an aspect of intelligence. The Abstraction and Reasoning Corpus (ARC) benchmark, in figure 1, is one that claims to measure intelligence while also claiming that other benchmarks only measure skill or memorization. Regardless, the following are currently popular benchmarks for reasoning AI models:

- **ARC-AGI**—the Abstract and Reasoning Corpus can be considered as a general artificial intelligence benchmark, as a program synthesis benchmark, or as a psychometric intelligence test for both humans and artificially intelligent systems
- **HLE**—Humanity's Last Exam is a multi-modal benchmark at the frontier of human knowledge, designed to be the final closed-ended academic benchmark of its kind with broad subject coverage
- **MMLU**—Massive Multitask Language Understanding assesses knowledge across 57 academic subjects with 16,000 multiple-choice questions
- **GPQA**—Graduate-level Google-Proof QA covers 448 questions in biology, physics, and chemistry
- **GAIA**—Meta's General AI Assistants is a set of 466 real-world questions, which, although conceptually simple for humans, may be difficult for current AIs and requires fundamental abilities such as reasoning, multi-modality handling, web browsing, and general tool-use proficiency
- **Math 500**—a comprehensive mathematics benchmark that includes 500 problems spanning various topics such as algebra, calculus, probability, and more, which assess both computational ability and mathematical reasoning
- **AIME 2024**—the American Invitational Mathematics Examination is a selective and prestigious 15-question, 3-hour test (started in 1983) available to those who rank in the top 5% on the AMC 12 high school mathematics examination (formerly known as the AHSME), and starting in 2010, those who rank in the top 2.5% on the AMC 10 (two different versions of the test are administered, the AIME I and AIME II).

Figure 1. Abstraction and Reasoning Corpus (ARC) benchmark



These are just a sample of benchmarks. One should be aware that benchmarks are a dynamically evolving area of research. Models continue to master existing benchmarks rapidly, leading to the introduction of new ones.

Approaches to improving AI reasoning architectures

In the 1980s, symbolic AI architectures undergirded expert system solutions. Due to the nature of the rules-based design, they inherently provided good reasoning abilities, however, there were practical scaling issues with such approaches.

Today, the most powerful AI systems utilize transformer architectures driven by the self-attention mechanism utilizing neural networks, which function as learning algorithms rather than rules-based systems. While these have good scaling properties, early learning models are inherently black boxes, unlike expert systems.

However, rapid innovations, such as the introduction of Chain-of-Thought (CoT) mechanisms [2], are creating powerful reasoning capabilities, making explainability and transparency increasingly possible while also improving accuracy consistently. Because of this, reasoning AI models are now a dominant focus of the field. The aim is to make these models more accurate while requiring less compute, less training and inferencing time, and lower costs. Minimizing training data and reducing the size of the model are also goals.

As an example of how the field is evolving, just as ChatGPT rocked the world in November 2022, so did DeepSeek-R1, in January 2025. DeepSeek-R1 is a reasoning model that was able to perform on par with some of the best models to date using drastically less compute, time and costs [11-14].

Furthermore, researchers at the Shanghai AI Laboratory have shown that with the right tools and test-time scaling techniques, improvements can be realized that make it possible to create a small language model (SLM) with one billion parameters that is able to outperform a 405B LLM on complicated math benchmarks [15].

2 Derivatives of CoT are Chain of Continuous Thought (Coconut), Chain-of-Associated-Thoughts (CoAT), Tree of Thought (TOT), Chain of Draft (CoD), etc.

Lessons learned: limitations and future directions

Until now, most AI models have been able to mimic human behavior convincingly enough to create the illusion of understanding and reasoning. However, we have yet to fully comprehend a model's environment and the contextual meaning of its knowledge. This phenomenon becomes apparent when models are faced with scenarios outside their training dataset. For instance, the benchmark on the new Math Olympiad, where all models performed poorly within hours of release, illustrates how fragile they can be [16].

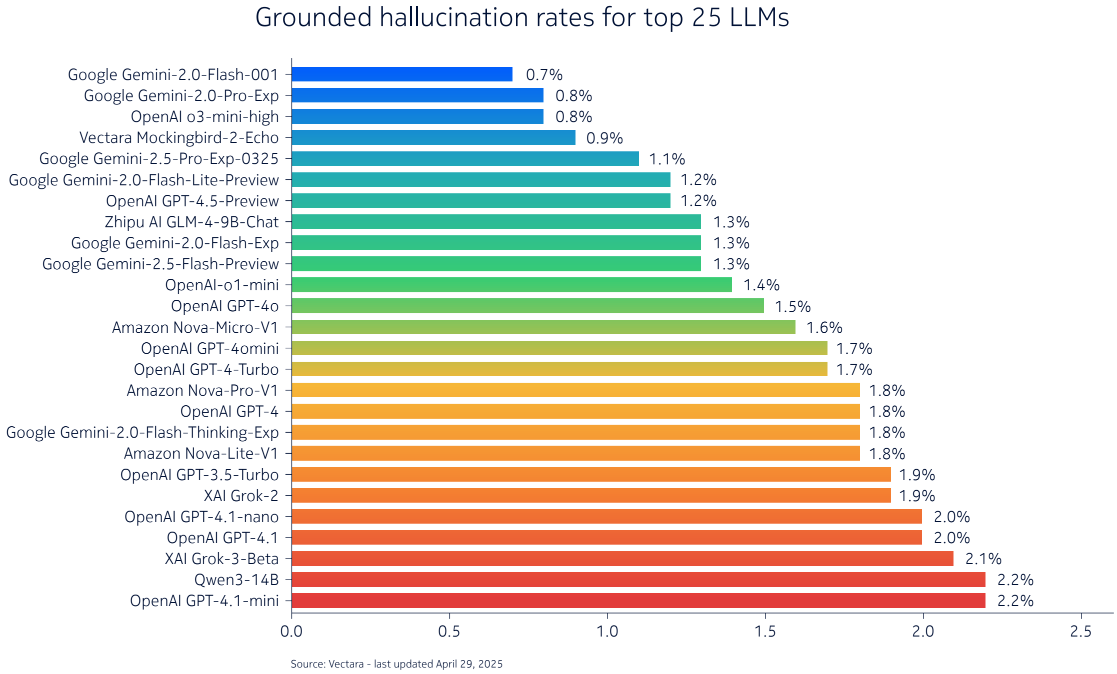
Current reasoning models, which use test-time scaling, nonetheless, demonstrate progress towards full reasoning capabilities in AI models, although there are still limitations. Test-time scaling during inferencing, for instance, requires more computation, which results in higher costs. Also, to reach true intelligence, models need an understanding of causality rather than simply recognizing correlation patterns. Being able to perceive nuances, intent, contradictions or ambiguities are required capabilities that are necessary to make an LLM a full reasoning model.

From an architectural perspective, models will need better long-term memory components. In addition, models will need to be able to source external, multi-modal data that is beyond their training data—in some instances, real-time data. Retrieval augmented generation (RAG) techniques and tools sprang up immediately after the launch of ChatGPT to provide rudimentary solutions to incorporate external data. This area continues to evolve rapidly. Models also need improvements in how they engage with external tools. As an example, Stanford University recently released a framework called OctoTools that aims to provide an efficient, structured approach to external tool selection [17].

In a similar vein, Anthropic open-sourced their model context protocol (MCP) at the end of 2024 [18], introducing a new standard for connecting AI assistants to the systems where data lives, including content repositories, business tools and development environments. MCP has been gaining significant numbers of adoptees since. Finally, hallucinations are a phenomenon of generative AI that impacts the performance of reasoning models. These must be minimized greatly to increase trust in AI model results. This topic is another area of active AI research.

Just as there are AI performance benchmarks, there are multiple hallucination detector models available. For example, in August 2024, Vectara announced HHEM-2.1, an improved version of HHEM-2.0, along with an open-source version [19] and a new [20] leaderboard powered by HHEM-2.1. They claim that their detector outperforms both GPT-3.5-Turbo and GPT-4 for hallucination detection. The following (figure 2) shows Vectara' s leaderboard in August for the top 25 LLMs [21].

Figure 2. Vectara hallucination leaderboard (April 29, 2025)



Conclusion

The subfield of AI reasoning has evolved significantly since the introduction of large language models in late 2022. While early models showed promise in basic reasoning capabilities, they were hampered by limitations such as hallucinations, mathematical challenges, brittleness, and a lack of explainability. Today's focus has shifted toward creating more accurate and efficient reasoning models that require less compute, training time, and costs.

Key developments in the field include:

- Evolution from simple AI chatbots to more sophisticated AI agents and agentic AI systems
- Emergence of new scaling approaches such as pre-training, post-training, and test-time scaling
- Development of Large Concept Models (LCMs) that operate at higher semantic levels and research into non-linguistic modeling concepts in a latent space
- Introduction of promising models, such as DeepSeek-R1, that demonstrate strong performance while using significantly reduced resources.

However, significant challenges remain to be addressed:

- True comprehension versus pattern recognition
- Better causality understanding
- Long-term memory capabilities
- Enhanced ability to handle external, multi-modal data
- Reduced hallucinations
- More efficient reasoning processes to optimize costs and latency.

As AI becomes ubiquitous in software systems, addressing these challenges while adhering to principles of universality, flexibility, explainability, and scalability will be crucial. The primary goal remains creating AI systems that can reason effectively across multiple dimensions and demonstrate true AI wisdom—making sound decisions based on deep understanding rather than mere pattern matching.

The path forward will require continued innovation in model architectures, training approaches, and evaluation methods to realize the full potential of AI reasoning capabilities across various domains, from telecommunications to broader real-world applications.

Acknowledgements

We would like to acknowledge work done by others at Nokia and Nokia Bell Labs that serves as input to this paper. This includes those who have researched reasoning AI models, particularly investigating DeepSeek. Our appreciation goes to Frederik Vandeputte, Laith Zumot, Antonio Massaro, Atefeh Mohajeri, Kedar Namjoshi, Lou Salaün, Mudita Shakya, and Zixiang Nie.

References

1. Woodie, A., "What are reasoning models and why you should care," Big Data Wire, 4 Feb 2025. Online: https://www.bigdatawire.com/2025/02/04/what-are-reasoning-models-and-why-you-should-care/
2. Thibaud, A.-L., "Where would reasoning AI leave human intelligence," World Economic Forum Annual Meeting, 22 Jan 2025. Online: https://www.weforum.org/stories/2025/01/in-a-world-of-reasoning-ai-where-does-that-leave-human-intelligence
3. Dickson, B., "Researchers find you don't need a ton of data to train LLMs for reasoning tasks," Venture Beat, 14 Feb 2025. Online: https://venturebeat.com/ai/researchers-find-you-dont-need-a-ton-of-data-to-train-llms-for-reasoning-tasks
4. Briski, K., "How scaling laws drive smarter, more powerful AI," Nvidia web site, 12 Feb 2025. Online: https://blogs.nvidia.com/blog/ai-scaling-laws
5. Barrault, L., et al., "Large Concept Models: Language Modeling in a Sentence Representation Space," 15 Dec 2024. Online: https://arxiv.org/abs/2412.08821
6. Hao, S., et al., "Training Large Language Models to Reason in a Continuous Latent Space," Meta, 11 Dec 2024. Online: https://arxiv.org/abs/2412.06769
7. Thompson, A.D., "Models Table," LifeArchitect.ai, viewed 5 Aug 2025. Online: https://lifearchitect.ai/models-table
8. Open LLM Leaderboard, "Open LLM Leaderboard," Hugging Face website, viewed Aug 2025. Online: https://huggingface.co/open-llm-leaderboard
9. Marten, R., "Reasoning Datasets," Hugging Face, 7 June 2025. Online: https://huggingface.co/collections/ryanmarten/reasoning-datasets-67aa88352e5d6cc89c1dd938
10. Manning, C., "Artificial intelligence definitions," Stanford University HAIF, Sep 2020. Online: https://hai-production.s3.amazonaws.com/files/2020-09/AI-Definitions-HAI.pdf
11. Raje, P., "Deepdive into Deepseek," LinkedIn post, 29 Jan 2025. Online: https://www.linkedin.com/pulse/deepdive-deepseek-prasad-raje-jakqc
12. Buchholz, K., "DeepSeek-R1 upsets AI market with low prices," Statista, 28 Jan 2025. Online: https://www.statista.com/chart/33839/prices-for-processing-one-million-input-output-tokens-on-different-ai-models
13. McConnon, A., "DeepSeek's reasoning AI shows power of small models, efficiently trained," IBM web site, Feb 2025. Online: https://www.ibm.com/think/news/deepseek-r1-ai
14. Li, S., "DeepSeek explained 4: Multi-Token Prediction," Medium, 20 Feb 2025. Online: https://medium.com/data-science-collective/deepseek-explained-4-multi-token-prediction-33f11fe2b868
15. Dickson, B., "How test-time scaling unlocks hidden reasoning abilities in small language models (and allows them to outperform LLMs)," World Economic Forum web site, 20 Feb 2025. Online: https://www.weforum.org/stories/2025/01/in-a-world-of-reasoning-ai-where-does-that-leave-human-intelligence/How test-time scaling unlocks hidden reasoning abilities in small language models (and allows them to outperform LLMs) | VentureBeat

16. Petrov, I., et al., Proof or Bluff? Evaluating LLMs on 2025 USA Math Olympiad, 27 Mar 2025. Online: https://arxiv.org/abs/2503.21934v1

17. Razzaq, A., “Stanford researchers introduce OctoTools: a training-free open-source agentic AI framework designed to tackle complex reasoning across diverse domains,” Marktechpost, 22 Feb 2025. Online: https://www.marktechpost.com/2025/02/22/stanford-researchers-introduce-octotools-a-training-free-open-source-agentic-ai-framework-designed-to-tackle-complex-reasoning-across-diverse-domains

18. Model Context Protocol, “Connect your AI applications to the world,” Model Context Protocol website, viewed Aug 2025. Online: https://modelcontextprotocol.io

19. Vectara, “Hallucination evaluation model,” Hugging Face website, viewed Aug 2025. Online: https://huggingface.co/vectara/hallucination_evaluation_model

20. Vectara, “Hughes hallucination evaluation model (HHEM) leaderboard,” Hugging Face web site, viewed Aug 2025. Online: https://huggingface.co/spaces/vectara/leaderboard

21. Mendelevitch, O., et al., “HHEM 2.1: A better hallucination detection model and a new leaderboard,” Vectara web site, 5 Aug 2024. Online: https://www.vectara.com/blog/hhem-2-1-a-better-hallucination-detection-model

Nokia OYJ
Karakaari 7
02610 Espoo
Finland
Tel. +358 (0) 10 44 88 000

Document code: 1666500 (October) CID214996