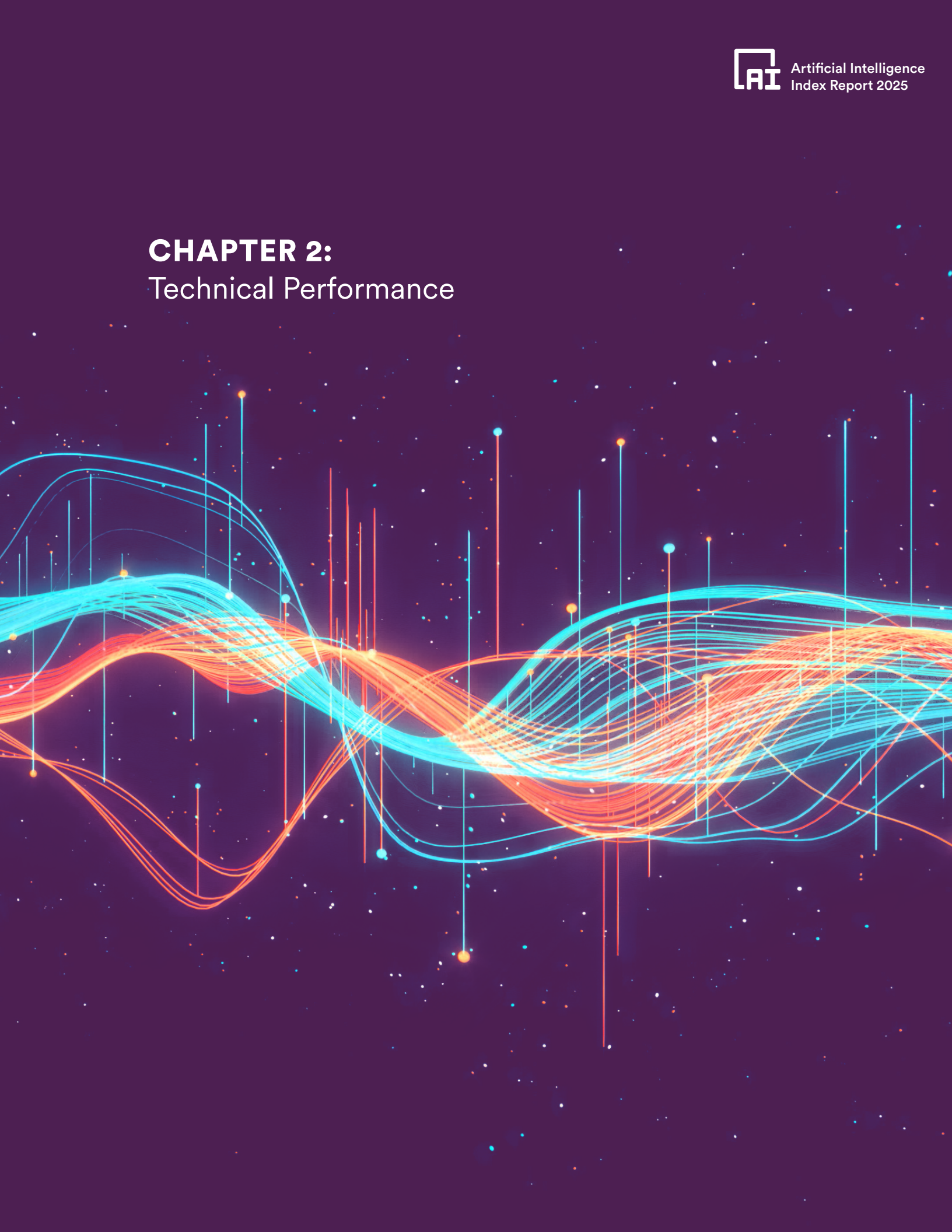
CHAPTER 2:
Technical Performance



Chapter 2: Technical Performance

Chapter 2: Technical Performance (cont'd)

ACCESS THE PUBLIC DATA

CHAPTER 2:
Technical Performance

Overview

The Technical Performance section of this year's AI Index provides a comprehensive overview of AI advancements in 2024. It begins with a high-level summary of AI technical progress, covering major AI-related launches, the state of AI capabilities, and key trends—such as the rising performance of open-weight models, the convergence of frontier model performance, and the improving quality of Chinese LLMs. The chapter then examines the current state of various AI capabilities, including language understanding and generation, retrieval-augmented generation, coding, mathematics, reasoning, computer vision, speech, and agentic AI. New this year are significantly expanded analyses of performance trends in robotics and self-driving cars.

CHAPTER 2:
Technical Performance

Chapter Highlights

1. AI masters new benchmarks faster than ever. In 2023, AI researchers introduced several challenging new benchmarks, including MMMU, GPQA, and SWE-bench, aimed at testing the limits of increasingly capable AI systems. By 2024, AI performance on these benchmarks saw remarkable improvements, with gains of 18.8 and 48.9 percentage points on MMMU and GPQA, respectively. On SWE-bench, AI systems could solve just 4.4% of coding problems in 2023—a figure that jumped to 71.7% in 2024.

---

2. Open-weight models catch up. Last year's AI Index revealed that leading open-weight models lagged significantly behind their closed-weight counterparts. By 2024, this gap had nearly disappeared. In early January 2024, the leading closed-weight model outperformed the top open-weight model by 8.04% on the Chatbot Arena Leaderboard. By February 2025, this gap had narrowed to 1.70%.

---

3. The gap between Chinese and US models closes. In 2023, leading American models significantly outperformed their Chinese counterparts—a trend that no longer holds. At the end of 2023, performance gaps on benchmarks such as MMLU, MMMU, MATH, and HumanEval were 17.5, 13.5, 24.3, and 31.6 percentage points, respectively. By the end of 2024, these differences had narrowed substantially to just 0.3, 8.1, 1.6, and 3.7 percentage points.

---

4. AI model performance converges at the frontier. According to last year's AI Index, the Elo score difference between the top and 10th-ranked model on the Chatbot Arena Leaderboard was 11.9%. By early 2025, this gap had narrowed to just 5.4%. Likewise, the difference between the top two models shrank from 4.9% in 2023 to just 0.7% in 2024. The AI landscape is becoming increasingly competitive, with high-quality models now available from a growing number of developers.

---

5. New reasoning paradigms like test-time compute improve model performance. In 2024, OpenAI introduced models like o1 and o3 that are designed to iteratively reason through their outputs. This test-time compute approach dramatically improved performance, with o1 scoring 74.4% on an International Mathematical Olympiad qualifying exam, compared to GPT-4o's 9.3%. However, this enhanced reasoning comes at a cost: o1 is nearly six times more expensive and 30 times slower than GPT-4o.

CHAPTER 2:
Technical Performance

Chapter Highlights (cont'd)

6. More challenging benchmarks are continually proposed. The saturation of traditional AI benchmarks like MMLU, GSM8K, and HumanEval, coupled with improved performance on newer, more challenging benchmarks such as MMMU and GPQA, has pushed researchers to explore additional evaluation methods for leading AI systems. Notable among these are Humanity's Last Exam, a rigorous academic test where the top system scores just 8.80%; FrontierMath, a complex mathematics benchmark where AI systems solve only 2% of problems; and BigCodeBench, a coding benchmark where AI systems achieve a 35.5% success rate—well below the human standard of 97%.

---

7. High-quality AI video generators demonstrate significant improvement. In 2024, several advanced AI models capable of generating high-quality videos from text inputs were launched. Notable releases include OpenAI's SORA, Stable Video 3D and 4D, Meta's Movie Gen, and Google DeepMind's Veo 2. These models produce videos of significantly higher quality compared to those from 2023.

---

8. Smaller models drive stronger performance. In 2022, the smallest model registering a score higher than 60% on MMLU was PaLM, with 540 billion parameters. By 2024, Microsoft's Phi-3-mini, with just 3.8 billion parameters, achieved the same threshold. This represents a 142-fold reduction in over two years.

---

9. Complex reasoning remains a problem. Even though the addition of mechanisms such as chain-of-thought reasoning has significantly improved the performance of LLMs, these systems still cannot reliably solve problems for which provably correct solutions can be found using logical reasoning, such as arithmetic and planning, especially on instances larger than those they were trained on. This has a significant impact on the trustworthiness of these systems and their suitability in high-risk applications.

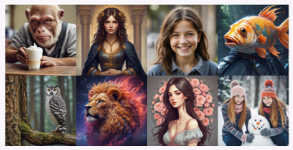
---

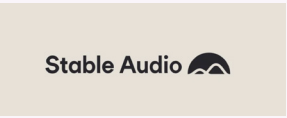
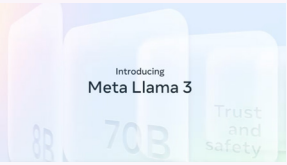
10. AI agents show early promise. The launch of RE-Bench in 2024 introduced a rigorous benchmark for evaluating complex tasks for AI agents. In short time-horizon settings (two-hour budget), top AI systems score four times higher than human experts, but as the time budget increases, human performance surpasses AI—outscoring it two to one at 32 hours. AI agents already match human expertise in select tasks, such as writing Triton kernels, while delivering results faster and at lower costs.

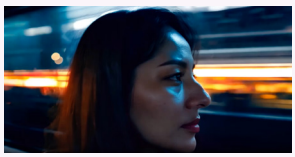
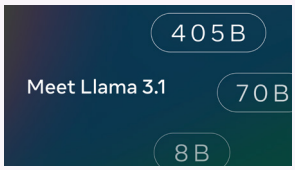
2.1 Overview of AI in 2024

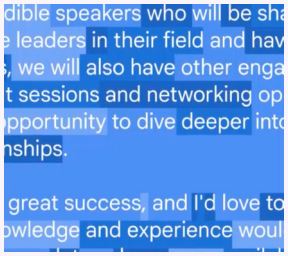
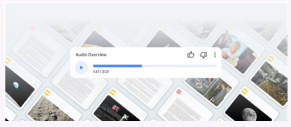
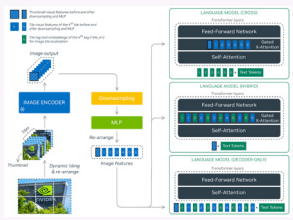
Timeline: Significant Model and Dataset Releases

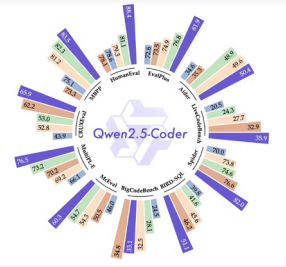
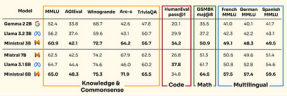
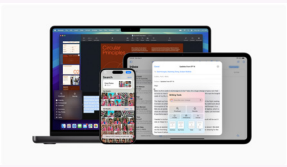
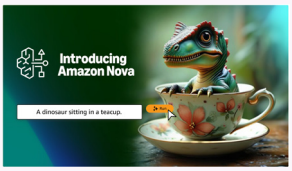
As chosen by the AI Index Steering Committee, here are some of the most notable model and dataset releases of 2024.

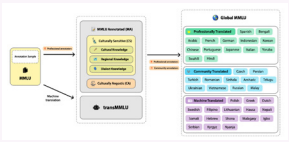
Date	Name	Category	Creator(s)	Significance	Image
Jan 19, 2024	Stable LM 2	LLM	Stability AI	Stability's latest language model builds on the original Stable LM, offering enhanced performance. With only 1.6 billion parameters, it is designed to run efficiently on portable devices such as laptops and smartphones.	 Figure 2.1.1 Source: Wikipedia, 2025
Feb 8, 2024	Aya Dataset	Dataset	Cohere for AI, Beijing Academy of AI, Cohere, Binghamton University	A collection of 513 million prompt-completion pairs spanning 114 languages, released as part of Cohere's Aya initiative. This paper and its accompanying dataset represent significant milestones in multilingual instruction tuning.	 Figure 2.1.2 Source: Cohere, 2025
Feb 15, 2024	Gemini 1.5 Pro	LLM	Google DeepMind	Google's Gemini model set a new benchmark with its 1M token context window, far exceeding GPT-4 Turbo's 128K token limit.	 Figure 2.1.3 Source: Google, 2024
Feb 20, 2024	SDXL-Lightning	Text-to-image	ByteDance	Developed by ByteDance, the creators of TikTok, this model was among the fastest text-to-image systems at its release, generating high-quality synthetic images in under a second. Its speed was achieved through progressive adversarial distillation, unlike other models that rely on diffusion-based techniques.	 Figure 2.1.4 Source: Hugging Face, 2025
Mar 4, 2024	Claude 3	LLM	Anthropic	Anthropic's latest LLM outperforms GPT-4 and Gemini on nearly all industry benchmarks, reduces incorrect prompt refusals, and delivers significantly higher accuracy.	 Figure 2.1.5 Source: Anthropic, 2025

Mar 7, 2024	Inflection-2.5	LLM	Inflection AI	Inflection's flagship product, "Pi," featured an exceptional model with GPT-4–level performance while using only 40% of its computing resources. Just two weeks after the model's release, Microsoft acquired Inflection for $650 million.	 Figure 2.1.6 Source: Inflection, 2025
Mar 19, 2024	Moirai and LOTSA	Model/ dataset	Salesforce	Salesforce unveils Moirai, a foundation model for universal forecasting, alongside LOTSA—a diverse, large-scale time series dataset with 27 billion observations spanning nine domains.	 Figure 2.1.7 Source: Salesforce, 2025
Mar 27, 2024	DBRX	LLM	Databricks	Databricks' open-source mixture-of-experts (MoE) LLM is a fine-grained model, surpassing similar small MoE models like Mixtral and Grok. This transformer decoder-only model features 132B parameters (36B active per input) and was trained on 12 trillion tokens.	 Figure 2.1.8 Source: Databricks, 2025
Apr 2, 2024	Stable Audio 2	Text-to-song and song-to-song	Stability AI	The latest version of Stable Audio, Stability's AI-powered song generator, now supports audio-to-audio functionality. Users can upload songs and manipulate them using natural language prompts for seamless customization.	 Figure 2.1.9 Source: Stability AI, 2025
Apr 17, 2024	Llama 3	LLM	Meta	The Llama 3 series debuts with 8B and 70B parameter text-based models, ranking among the highest performing models of their size to date.	 Figure 2.1.10 Source: Meta, 2025
May 13, 2024	GPT-4o	Multimodal	OpenAI	GPT-4o is a new multimodal model capable of processing inputs in any combination of text, audio, images, and video, and generating outputs in the same formats. It responds to audio in as little as 320 milliseconds, matching human response times.	 Figure 2.1.11 Source: OpenAI, 2024

Jun 7, 2024	Qwen2	LLM	Alibaba	Qwen2, developed by China's Alibaba, is a series of advanced base and instruction-tuned models. These models rival competitors like Llama 3-70B and Mixtral-8x22B in performance across numerous benchmarks.	 Figure 2.1.12 Source: Qwen, 2024
Jun 17, 2024	Runway Gen-3	Text-to-video and image-to-video	Runway	Runway's upgraded video generation model sets a new standard for the field, particularly excelling in creating photorealistic humans with vivid and expressive emotionality.	 Figure 2.1.13 Source: Runway, 2024
Jul 23, 2024	Llama 3.1 405B	LLM	Meta	Meta has released its largest model to date, the final in the Llama 3.1 family, featuring 405B parameters. Upon its release, it became the most capable openly available foundation model, rivaling many closed models across a variety of benchmarks.	 Figure 2.1.14 Source: Meta, 2024
Aug 12, 2024	Falcon Mamba	LLM	Technology Innovation Institute in Abu Dhabi	A powerful new 7B parameter model, built on the Mamba State Space Language Model (SSLM) architecture, enables Falcon—one of the few government-created AI models—to dynamically adjust parameters and filter out irrelevant inputs, making it more efficient than transformer-based models.	 Figure 2.1.15 Source: Hugging Face, 2025
Aug 13, 2024	Grok-2	Text-to-text and text-to-image	xAI	Developed by xAI, Grok is an advanced text- and image-generation model that excels in image creation, advanced reasoning, and problem-solving. Its launch was particularly notable, as it quickly rivaled the performance of leading models despite xAI being founded only in March 2023.	 Figure 2.1.16 Source: xAI, 2025

Aug 15, 2024	Imagen 3	Text-to-image	Google Labs	Google's updated AI image generator achieves the highest Elo score on the GenAI-Bench image benchmark, setting a new standard for quality in AI-generated visuals.	 Figure 2.1.17 Source: Google, 2025
Aug 22, 2024	Jamba 1.5	LLM	AI21 Labs	The first LLM to combine state-space models with transformers, delivering high-quality results for text-based applications. This hybrid approach significantly enhances speed while preserving the quality of outputs.	 Figure 2.1.18 Source: AI21, 2025
Aug 29, 2024	SynthID v2	Tool	Google	SynthID v2 is the updated version of SynthID, Google's watermarking and identification software. It now supports AI-generated content across images, video, audio, and text, and offers enhanced tracking and verification capabilities.	 Figure 2.1.19 Source: Google, 2025
Sep 11, 2024	NotebookLM Podcast Tool	Text-to-podcast	Google Labs	The second end-to-end AI podcast generator to hit the market, following Synthpod, went viral. It gained popularity among students leveraging NotebookLM for studying and tech employees using it to listen to AI-generated summaries.	 Figure 2.1.20 Source: Google, 2025
Sep 12, 2024	o1-preview	Language, math, biology	OpenAI	OpenAI's first model in the "o series" is designed for advanced reasoning and tackling complex tasks. It is significantly more powerful than GPT, particularly in math, science, and coding.	 Figure 2.1.21 Source: OpenAI, 2025
Sep 17, 2024	NVLM (D, H, X)	Vision, language	Nvidia	Nvidia released three open-access models for vision-language tasks, achieving top scores on OCRBench (for optical character recognition) and VQAv2 (for natural language understanding).	 Figure 2.1.22 Source: Dai et al., 2024

Sep 19, 2024	Qwen2.5	LLM	Alibaba	Qwen2.5, the latest series of foundation models from Chinese e-commerce giant Alibaba, includes a range of efficient smaller models and specialized coding and math models designed for targeted functionality.	 Figure 2.1.23 Source: Qwen, 2025
Oct 16, 2024	Ministral	LLM	Mistral	Ministral is a pair of compact models (3B and 8B parameters) that outperformed Gemma and Llama models of similar size across all major industry-recognized benchmarks.	 Figure 2.1.24 Source: Mistral, 2025
Oct 22, 2024	Anthropic Computer Use	Agentic Capability	Anthropic	Anthropic Computer Use is a groundbreaking computer control feature for Claude 3.5 Sonnet users, allowing Claude to move the cursor, type, and autonomously complete tasks on the user's computer in real time.	 Figure 2.1.25 Source: Anthropic, 2025
Oct 28, 2024	Apple Intelligence	iPhone feature	Apple	Apple's suite of AI-powered features includes Image Playground (for image creation), Genmoji (for custom emoji creation), Siri integration with ChatGPT, and more.	 Figure 2.1.26 Source: Apple, 2025
Dec 3, 2024	Nova Pro	Multimodal	Amazon	Nova Pro is the most powerful model in Amazon Web Services' Nova family, capable of processing both visual and textual information. It especially excels at analyzing financial documents.	 Figure 2.1.27 Source: Amazon, 2025
Dec 11, 2024	Gemini 2	LLM	Google DeepMind	The improved version of Gemini, Google's LLM, now includes computer control along with image and audio generation capabilities. It is twice as fast as Gemini 1.5 Pro and offers significantly enhanced performance in coding and image analysis.	 Figure 2.1.28 Source: Google, 2025

Dec 12, 2024	Sora	Text-to-video	OpenAI	OpenAI's highly anticipated video generation model can create videos up to 20 seconds long at 1080p resolution for ChatGPT Pro users (and five seconds at 720p for ChatGPT Plus users). Sora demos had been circulating at tech meetups since early 2024, but OpenAI delayed the official release to improve model safety.	 Figure 2.1.29 Source: OpenAI, 2025
Dec 13, 2024	Global MMLU	Dataset	Cohere	A multilingual evaluation set featuring professionally translated MMLU questions across 42 languages, designed to serve as a more global AI benchmark. It evaluates AI performance in diverse languages while addressing Western biases in the original MMLU dataset, where an estimated 28% of questions rely on Western cultural knowledge.	 Figure 2.1.30 Source: Singh et al., 2025
Dec 20, 2024	o3 (beta)	Multimodal	OpenAI	OpenAI's newest frontier model, released for safety testing by AI researchers, outperforms all previous models in SWE, competition code, competition math, PhD-level science, and research math benchmarks. It also set a new record on the ARC-AGI benchmark, achieving 87.5% on the ARC Prize team's private holdout set.	 Figure 2.1.31 Source: VentureBeat, 2025
Dec 27, 2024	DeepSeek-V3	LLM	DeepSeek	DeepSeek V3, an open-source model developed with significantly fewer computing resources than state of the art models, outperforms leading models on benchmarks like MMLU and GPQA.	 Figure 2.1.32 Source: Dirox, 2025

State of AI Performance

In this section, the AI Index offers a high-level view into major AI trends that occurred in 2024.

Overall Review

Last year's AI Index highlighted that AI had already surpassed human performance across many tasks, with only a few exceptions, such as competition-level mathematics and visual commonsense reasoning. Over the past year, AI systems have continued to improve, exceeding human performance on several of these previously challenging benchmarks.

Figure 2.1.33 illustrates the progress of AI systems relative to human baselines for eight AI benchmarks corresponding to 11 tasks (e.g., image classification or basic-level reading comprehension).[1] The AI Index team selected one benchmark to represent each task. This year, the AI Index team added newly released benchmarks, such as GPQA Diamond and MMMU, to showcase the progress of AI systems in tackling extremely challenging cognitive tasks.

Select AI Index technical performance benchmarks vs. human performance

Source: AI Index, 2025 | Chart: 2025 AI Index report

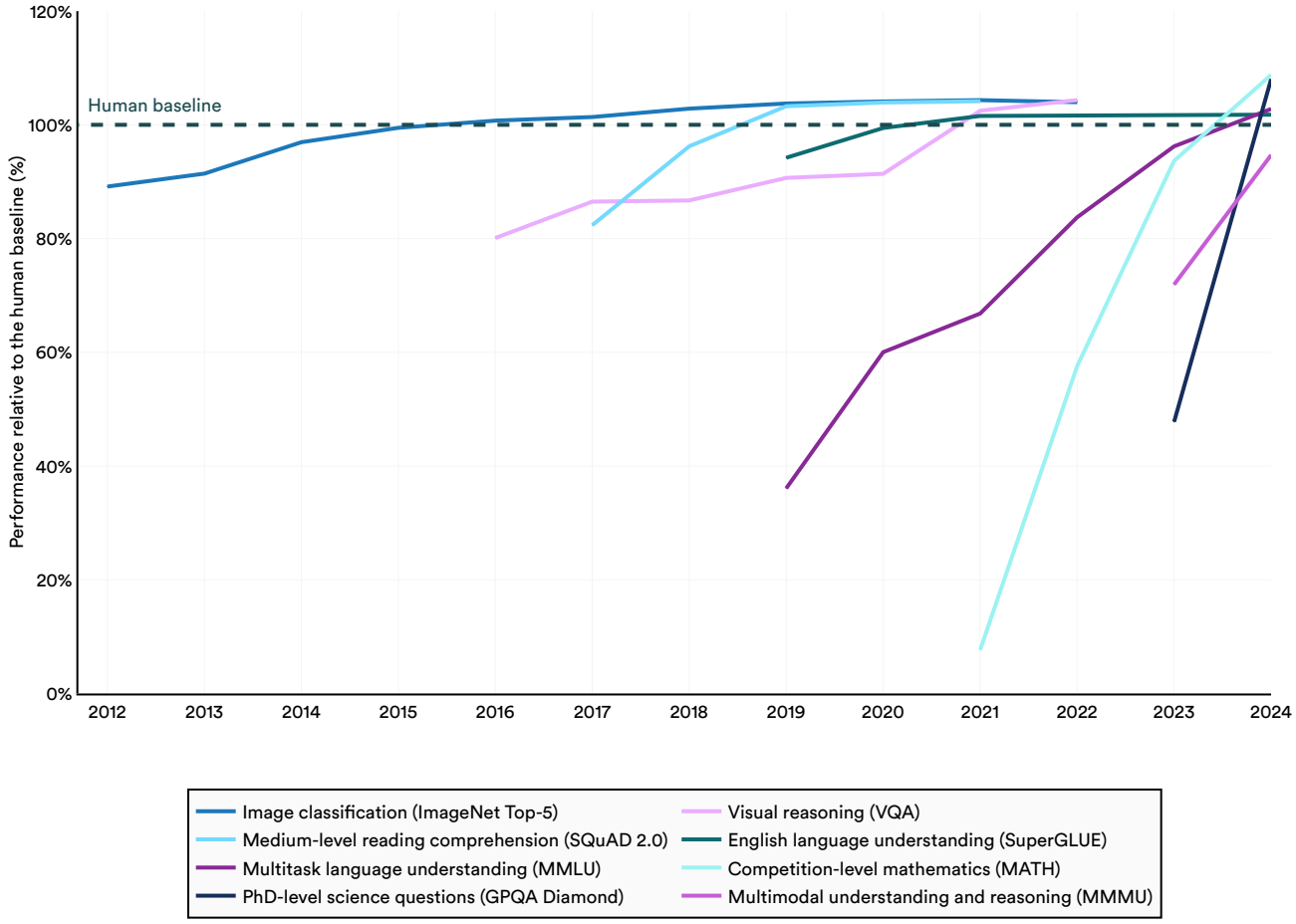


Figure 2.1.33[2]

1 An AI benchmark is a standardized test used to evaluate the performance and capabilities of AI systems on specific tasks. For example, ImageNet is a canonical AI benchmark that features a large collection of labeled images, and AI systems are tasked with classifying these images accurately. Tracking progress on benchmarks has been a standard way for the AI community to monitor the advancement of AI systems.

2 In Figure 2.1.33, the values are scaled to establish a standard metric for comparing different benchmarks. The scaling function is calibrated such that the performance of the best model for each year is measured as a percentage of the human baseline for a given task. A value of 105% indicates, for example, that a model performs 5% better than the human baseline

As of 2024, there are very few task categories where human ability surpasses AI. Even in these areas, the performance gap between AI and humans is shrinking rapidly. For example, on MATH, a benchmark for competition-level mathematics, state-of-the-art AI systems are now 7.9 percentage points ahead of human performance, a significant improvement from the 0.3-point gap in 2024.[3] Similarly, on MMMU, a benchmark for complex, multidisciplinary, expert-level questions, the best 2024 model, o1, scored 78.2%, only 4.4 points below the human benchmark of 82.6%. Conversely, at the end of 2023, Google Gemini scored 59.4%, further illustrating the rapid advancements in AI performance on cognitively demanding tasks.

Closed vs. Open-Weight Models

AI models can be released with different levels of openness. Certain models, like Google's Med-Gemini, remain entirely closed, accessible only to their developers. Meanwhile, models such as OpenAI's GPT-4o and Anthropic's Claude 3.5 provide limited public access through APIs. However, weights for these models are not released, preventing independent modification or thorough public scrutiny. In contrast, weights for Meta's Llama 3.3 and Stable Video 4D are fully available, allowing anyone to modify and use them freely.[4]

Perspectives on open versus closed-weight AI models are sharply divided. Advocates of open-weight models highlight their potential to reduce market monopolies, spur innovation, improve security and robustness, and enhance transparency within the AI ecosystem. For example, Meta's Llama models have been leveraged to create tools like Meditron, power military applications, and drive the development of numerous open-weight models worldwide. However, critics warn that open-weight models pose significant security risks, including the spread of disinformation and the creation of bioweapons, arguing for a more cautious and controlled approach.

Last year's AI Index highlighted a notable performance gap between closed and open-weight LLM models. Figure 2.1.34 illustrates the performance trends of the top closed-weight and open-weight LLMs on the Chatbot Arena Leaderboard, a public platform for benchmarking LLM performance. In early January 2024, the leading closed-weight model outperformed the top open-weight model by 8.0%. By February 2025, this gap had narrowed to 1.7%.

The same trend is evident across other question-answering benchmarks. In 2023, closed-weight models consistently outperformed open-weight counterparts on nearly every major benchmark—MMLU, HumanEval, MMMU, and MATH. However, by 2024, the gap had narrowed significantly (Figure 2.1.35). For instance, in late 2023, closed-weight models led open models on MMLU by 15.9 points, but by the end of 2024, that difference had shrunk to just 0.1 percentage point. This rapid improvement was largely driven by Meta's summer release of Llama 3.1, followed by the launch of other high-performing open-weight models, such as DeepSeek's V3.

3 The benchmark data in this figure, along with those in other sections of this chapter, was collected in early January 2025. Since the publication of the AI Index, individual benchmark scores may have improved.

4 In the software community, "open source" refers to software released under a license that grants users the right to use, study, modify, and distribute both the software and its source code freely. Open-weight models, though more accessible than closed-weight models, are not necessarily fully open source, as the underlying code or training data is often withheld.

Performance of top closed vs. open models on LMSYS Chatbot Arena

Source: LMSYS, 2025 | Chart: 2025 AI Index report

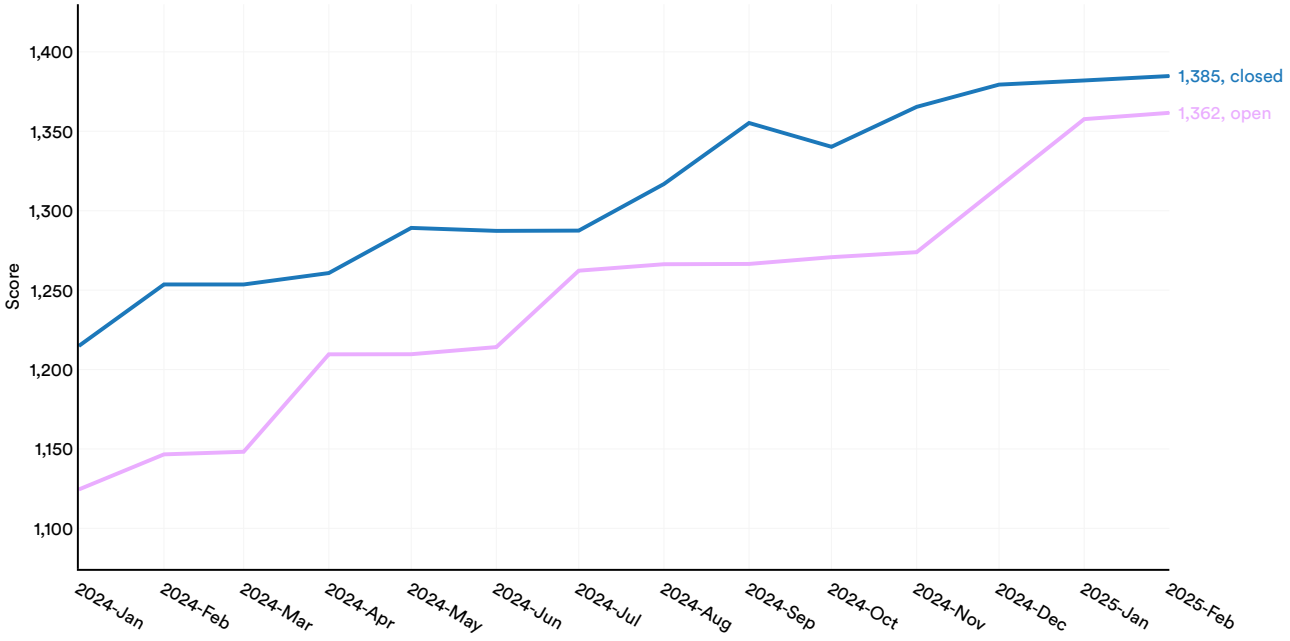


Figure 2.1.34

Performance of top closed vs. open models on select benchmarks

Source: AI Index, 2025 | Chart: 2025 AI Index report

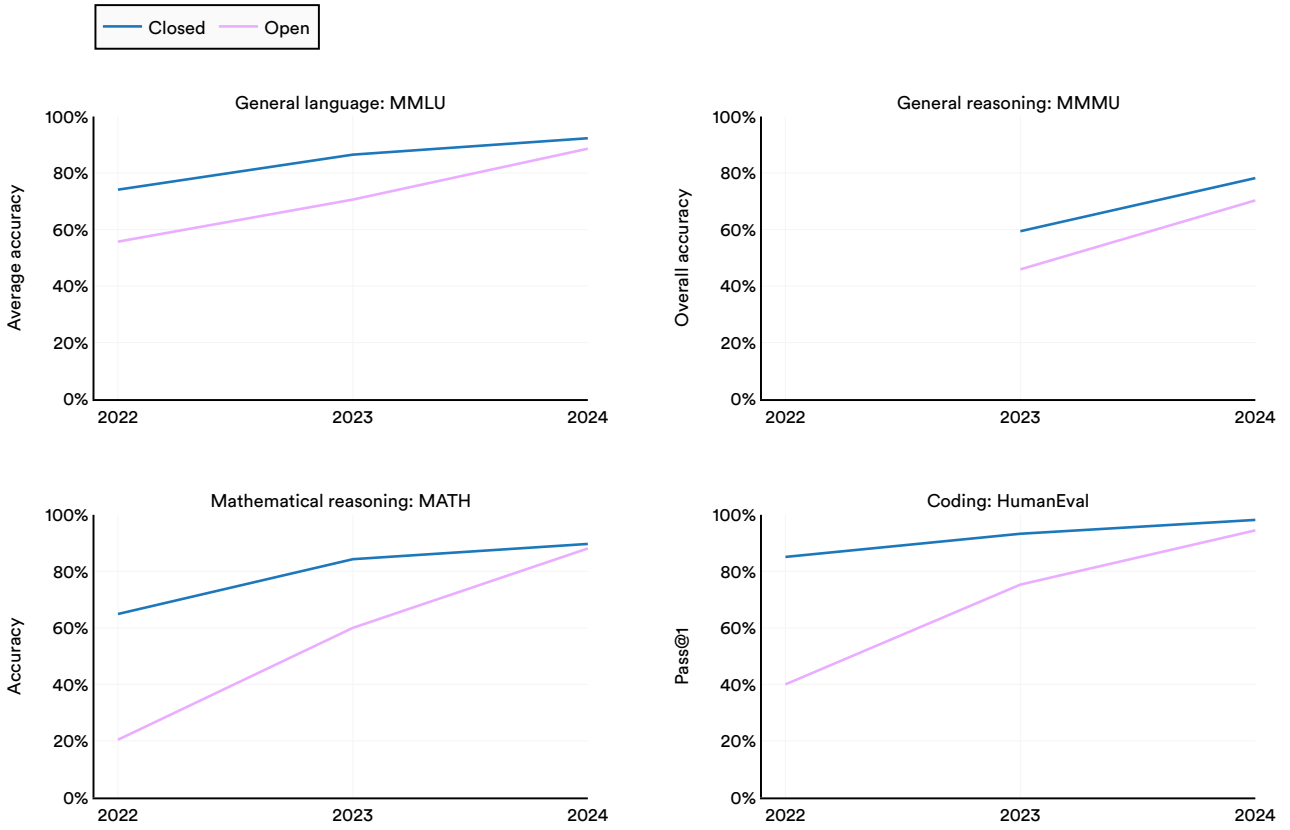


Figure 2.1.35

US vs. China Technical Performance

The United States has historically dominated AI research and model development, with China consistently ranking second. Recent evidence, however, suggests the landscape is rapidly changing and that China-based models are catching up to their U.S. counterparts.

In 2023, leading American models significantly outperformed their Chinese counterparts. On the LMSYS Chatbot Arena, the top U.S. model outperformed the best Chinese model by 9.26% in January 2024. By February 2025, this gap had narrowed to just 1.70% (Figure 2.1.36). At the end of 2023, on benchmarks such as MMLU, MMMU, MATH, and HumanEval, the performance gaps were 17.5, 13.5, 24.3, and 31.6 percentage points, respectively (Figure 2.1.37). By the end of 2024, these differences had narrowed significantly to just 0.3, 8.1, 1.6, and 3.7 percentage points. The launch of DeepSeek-R1 garnered attention for another reason: The company reported achieving its results using only a fraction of the hardware resources typically required to train such a model. Beyond impacting U.S. stock markets, DeepSeek's R1 launch raised doubts about the effectiveness of U.S. semiconductor export controls.

Performance of top United States vs. Chinese models on LMSYS Chatbot Arena

Source: LMSYS, 2025 | Chart: 2025 AI Index report

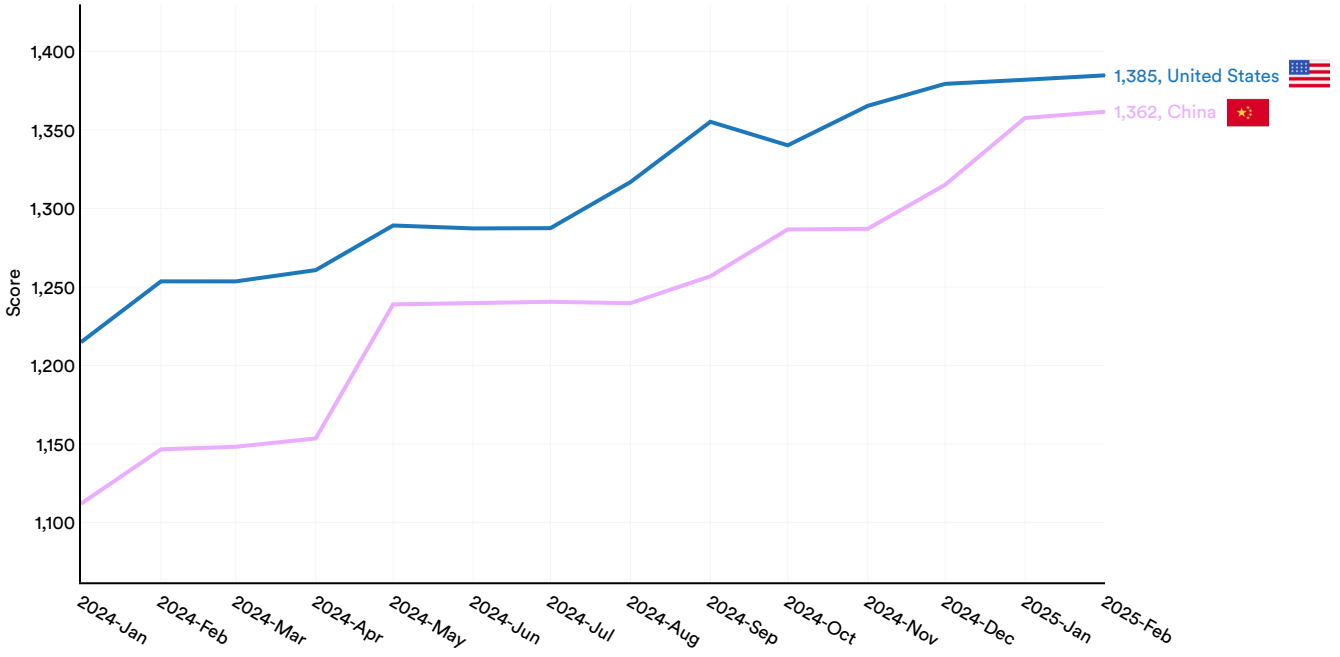


Figure 2.1.36

Performance of top United States vs. Chinese models on select benchmarks

Source: AI Index, 2025 | Chart: 2025 AI Index report

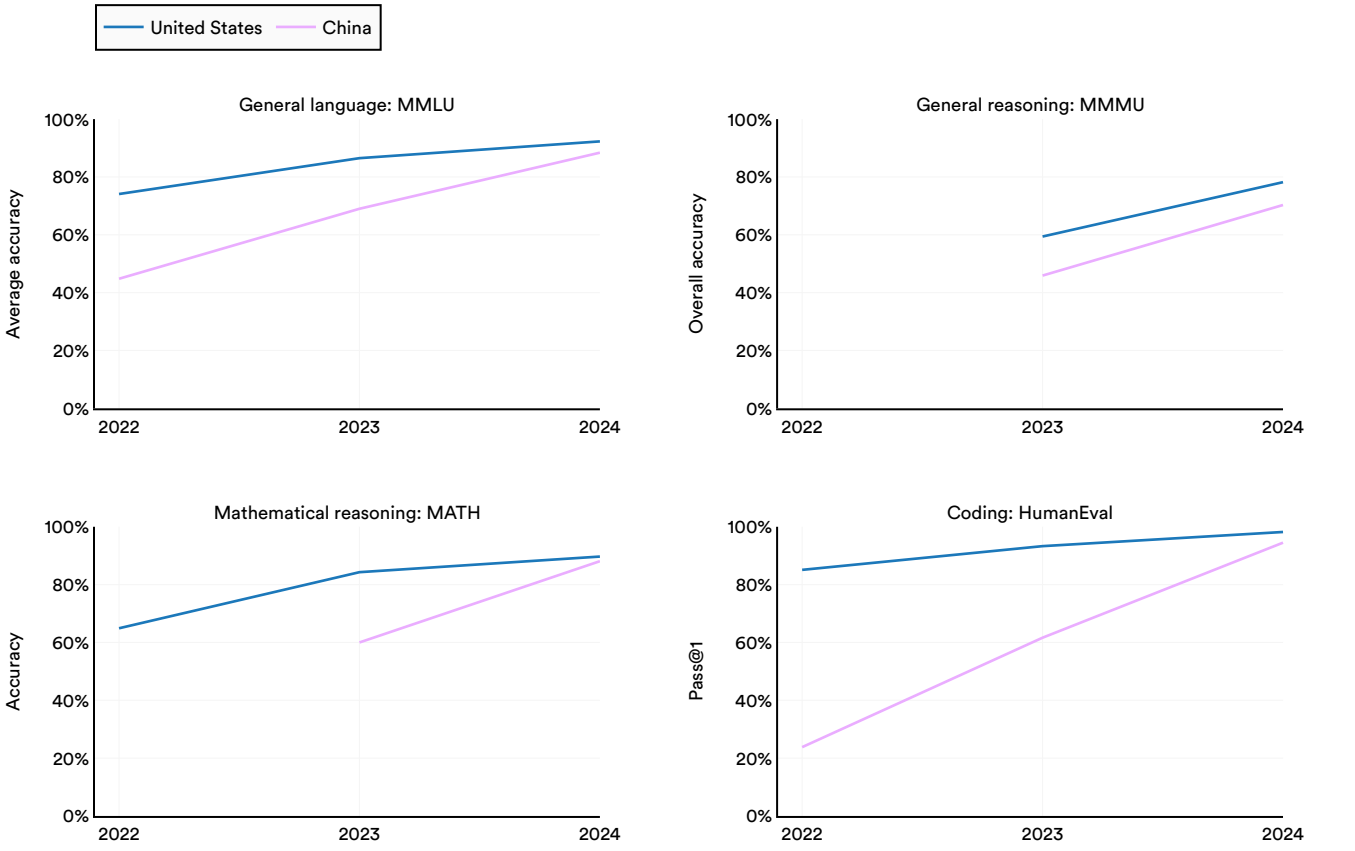


Figure 2.1.37

Improved Performance From Smaller Models

Recent AI progress has been driven by scaling—the idea that increasing model size and training data improves performance. While scaling has significantly boosted AI capabilities, a notable recent trend is the emergence of smaller high-performing models. Figure 2.1.38 illustrates the reduction in size of the smallest model that scores above 60% on MMLU, a widely used language model benchmark. For context, early models powering ChatGPT, such as GPT-3.5 Turbo, scored around 70% on MMLU. In 2022, the smallest model surpassing 60% on MMLU was PaLM, with 540 billion parameters. By 2024, Microsoft's Phi-3 Mini, with just 3.8 billion parameters, achieved the same threshold, marking a 142-fold reduction in model size over two years.

2024 was a breakthrough year for smaller AI models. Nearly every major AI developer released compact, high-performing models, including GPT-4o mini, o1-mini, Gemini 2.0 Flash, Llama 3.1 8B, and Mistral Small 3.[5] The rise of small models is significant for several reasons. It demonstrates increasing algorithmic efficiency, allowing developers to achieve more with less data and at lower training cost. These efficiency gains, combined with growing datasets, could lead to even higher-performing models. Additionally, inference on smaller models is typically faster and less expensive. Their emergence also lowers the barrier to entry for AI developers and businesses looking to integrate AI into their operations.

Smallest AI models scoring above 60% on MMLU, 2022–24
Source: Abdin et al., 2024 | Chart: 2025 AI Index report

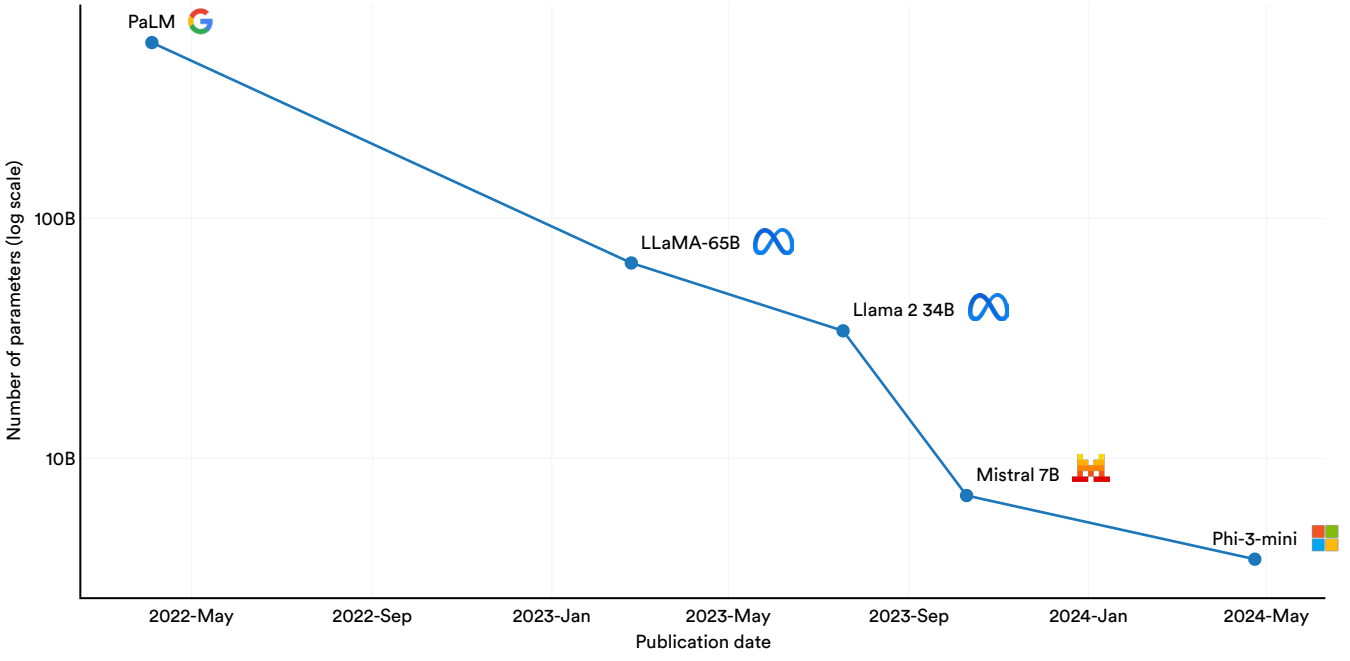


Figure 2.1.38

5 These are just a few of the small models launched in 2024.

Model Performance Converges at the Frontier

In recent years, AI model performance at the frontier has converged, with multiple providers now offering highly capable models. This marks a shift from late 2022, when ChatGPT's launch—widely seen as AI's breakthrough into public consciousness—coincided with a landscape dominated by just two major players: OpenAI and Google. OpenAI, founded in 2015, released GPT-3 in 2020, while Google introduced models like PaLM and Chinchilla in 2022.

Since then, new players have entered the scene, including Meta with its Llama models, Anthropic with Claude, High-Flyer's DeepSeek, Mistral's Le Chat, and xAI with Grok. As competition has intensified, model performance has increasingly converged (Figure 2.1.39). According to last year's AI Index, the performance gap between the highest- and 10th-ranked models on the Chatbot Arena Leaderboard—a widely used AI ranking platform—was 11.9%. By early 2025, it had narrowed to 5.4%. Similarly, the difference between the top two models fell from 4.9% in 2023 to just 0.7% in 2024. The AI landscape is becoming more competitive, validating 2023 predictions that AI companies lack a technological moat to shield them from rivals.

Performance of top models on LMSYS Chatbot Arena by select providers

Source: LMSYS, 2025 | Chart: 2025 AI Index report

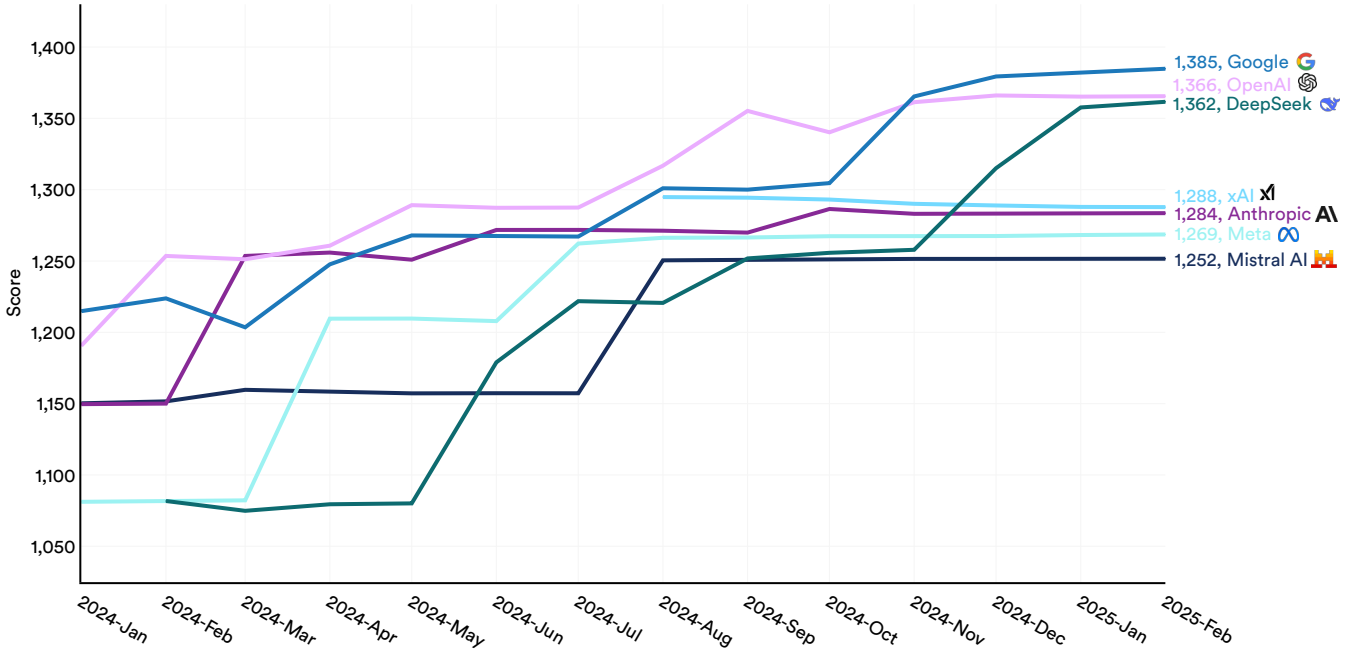


Figure 2.1.39

Benchmarking AI

For years, the AI Index has used benchmarks to monitor the technical progress of AI systems over time. While benchmarks remain a key tool in this effort, it is important to acknowledge their limitations and guide the community toward more effective benchmarking practices.

As noted in last year's AI Index, many prominent AI benchmarks are reaching saturation. With AI systems advancing rapidly, even newly designed, more challenging tests often remain relevant for only a few years. Some experts suggest that the era of new academic benchmarks may be coming to an end. To truly assess the capabilities of AI systems, more rigorous and comprehensive evaluations are needed.

Additionally, when model developers release new models, they typically report benchmark scores, which are often accepted at face value by the broader community. However, this approach has flaws. In some cases, companies use nonstandard prompting techniques, making model-to-model comparisons unreliable. For example, when Google launched Gemini Ultra, it reported an MMLU benchmark score using a chain-of-thought prompting technique that other developers did not use. Additionally, third-party researchers have documented cases where models perform worse in independent testing compared with the results first reported by their developers.

There are critical aspects of intelligence that do not easily lend themselves to benchmarking. Benchmarks are effective for evaluating certain intelligent capabilities, such as vision and language, where tasks are discrete—e.g., classifying an image correctly or answering a multiple-choice question. However, developing benchmarks is more challenging in areas of AI such as multi-agent systems and human-AI interaction because of factors including the variability in human behaviors and the sheer diversity of correct answers.

In addition, AI advances have traditionally been evaluated in competitions designed to measure human performance, such as games and other open challenges posed to humans or machines. Games such as chess and poker involve significant intelligence, and AI systems have improved over the decades to the point of defeating the best humans at increasingly complex games. Games with a physical component or team capabilities are also a good measure of progress for AI, and the robotics community has embarked on challenging game competitions such as RoboCup for soccer-playing robots. Another area of AI where competitions are used involves coordination and teamwork where multi-agent systems demonstrate advances in distributed reasoning.

Benchmarks have been developed by the AI community for a very long time. Significant advances in AI have been possible because different approaches and methods could be evaluated against the same gold standard represented by a benchmark. In machine learning, benchmarks with different kinds of data in diverse domains have enabled significant advances. Many of these benchmarks are evaluated automatically by a third party without releasing the test data to the AI developers, which makes the evaluations more trustworthy. One interesting recent trend is that various benchmark tasks are addressed by the same model. For example, natural language was addressed for many years as a collection of separate tasks (e.g., understanding, generation, question answering), each with its own models and each with its own benchmarks. Similarly, speech tasks were benchmarked separately from language understanding or generation tasks. Today, the same model can address all language tasks, and, in some cases, a single model can address language, images, and multimodal tasks. This is a very important AI advance concerning the integration of otherwise separate intelligent tasks and capabilities.

The rapid progress of AI systems, evidenced by their consistent outperformance on benchmarks, is perhaps best illustrated by the diminishing relevance of the well-known and long-standing challenge for AI: the Turing test. Originally proposed in Alan Turing's 1950 paper "Computing Machinery and Intelligence," the test evaluates a machine's ability to exhibit humanlike intelligence. In it, a human judge engages in a text-based conversation with both a machine and a human; if the

judge cannot reliably distinguish between them, the machine is said to have passed the Turing test. Recent evidence suggests that LLMs have advanced so significantly that people struggle to differentiate the best-performing language models from a human, signaling that modern AI models can pass the Turing test. While the merits and shortfalls of this test have long been debated, it remains an important historical and cultural benchmark for machine intelligence. The questioning of its relevance highlights the remarkable progress of LLMs in recent years and the evolving perception of effective computer science benchmarks and AI measurement.

In robotics, many models have emerged that address interacting with the physical world and reasoning about natural laws. A number of robotics benchmarks, such as ARMBench, focus on perception tasks. However, other benchmarks, such as VIMA-Bench, assess robot performance in simulated environments where they simultaneously incorporate perception, communication, and deep learning.

Benchmarks can also suffer from contamination, where LLMs encounter test questions that were present in their training data. A recent study by Scale found significant contamination in the performance of many LLMs on GSM8K, a widely used mathematics benchmark. Some researchers have sought to combat these contamination issues by introducing benchmarks like LiveBench, which are periodically updated with new questions from unfamiliar sources that LLMs are unlikely to have seen in their training data.

Lastly, research has shown that many benchmarks are poorly constructed. In BetterBench, researchers systematically analyzed 24 prominent benchmarks and identified systemic deficiencies: 14 failed to report statistical significance, 17 lacked scripts for result replication, and most suffered from inadequate documentation, limiting their reproducibility and effectiveness in evaluating models. Despite widespread use, benchmarks like MMLU demonstrated poor adherence to quality standards, while others, such as GPQA, performed significantly better. To address these issues, the paper proposed a 46-criteria framework covering all phases of benchmark development—design, implementation, documentation, and maintenance (Figure 2.1.40). It also introduced a publicly accessible repository to enable continuous updates and improve benchmark comparability. Figure 2.1.41, from BetterBench, assesses many prominent benchmarks on their usability and design. These findings underscore the need for standardized benchmarking to ensure reliable AI evaluation and to prevent misleading conclusions about model performance. Benchmarks have the potential to shape policy decisions and influence procurement decisions within organizations highlighting the importance of consistency and rigor in evaluation.

Five stages of the benchmark lifecycle
Source: Reuel et al., 2024

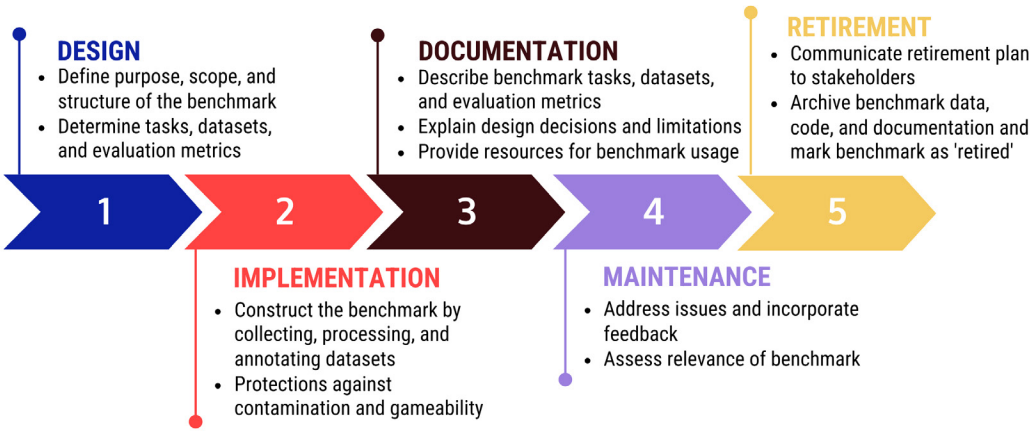


Figure 2.1.40

Design vs. usability scores across select benchmarks

Source: Reuel et al., 2024 | Chart: 2025 AI Index report

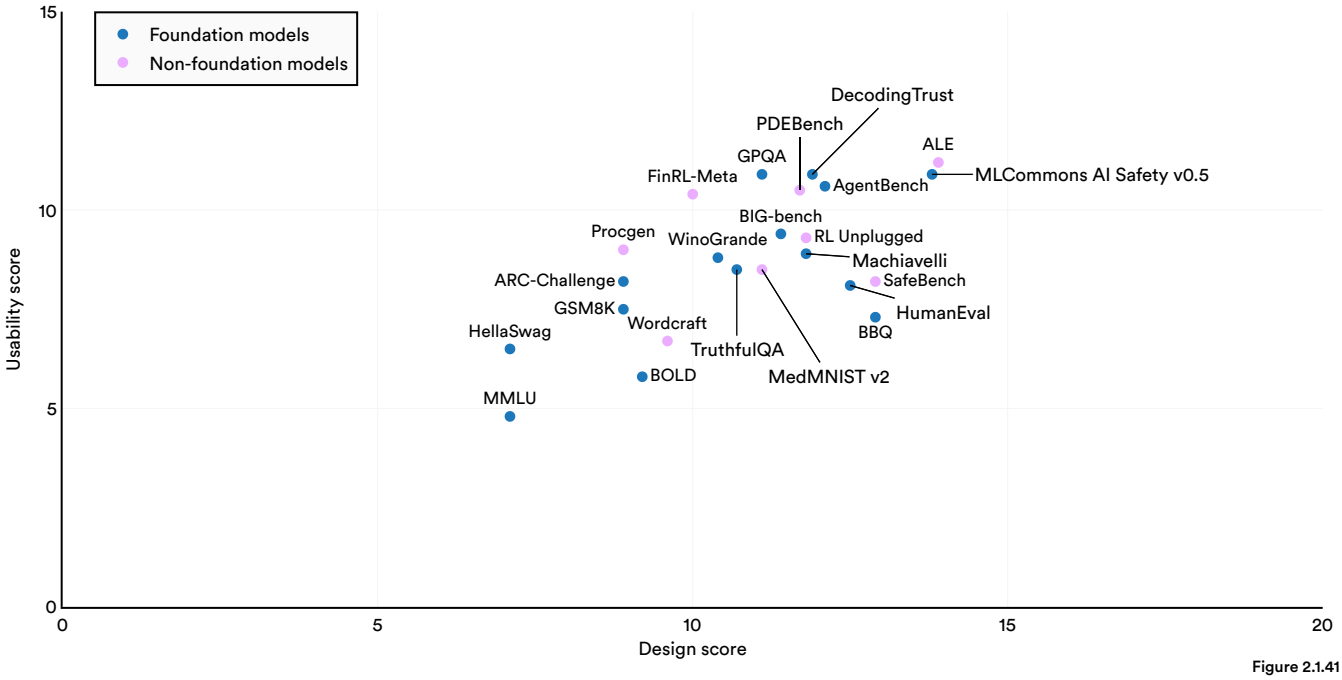


Figure 2.1.41

In this chapter, the AI Index continues to report on benchmarks, recognizing their importance in tracking AI's technical progress. As a standard practice, the Index sources benchmark scores from leaderboards, public repositories such as Papers With Code and RankedAGI, as well as company papers, blog posts, and product releases. The Index operates under the assumption that the scores reported by companies are accurate and factual. The benchmark scores in this section are current as of mid-February 2025. However, since the publication of the AI Index, newer models may have been released that surpass current state-of-the-art scores.

2.2 Language

Natural language processing (NLP) enables computers to understand, interpret, generate, and transform text. Current state-of-the-art models, such as OpenAI's GPT-4o, Anthropic's Claude 3.5, and Google's Gemini, are able to generate fluent and coherent prose and display high levels of language understanding ability (Figure 2.2.1). Unlike earlier versions, which were restricted to text input and output, newer language models can now reason across a growing range of input and output modalities, including audio, images, and goal-oriented tasks (Figure 2.2.2).

A sample output from GPT-4o
Source: AI Index, 2025

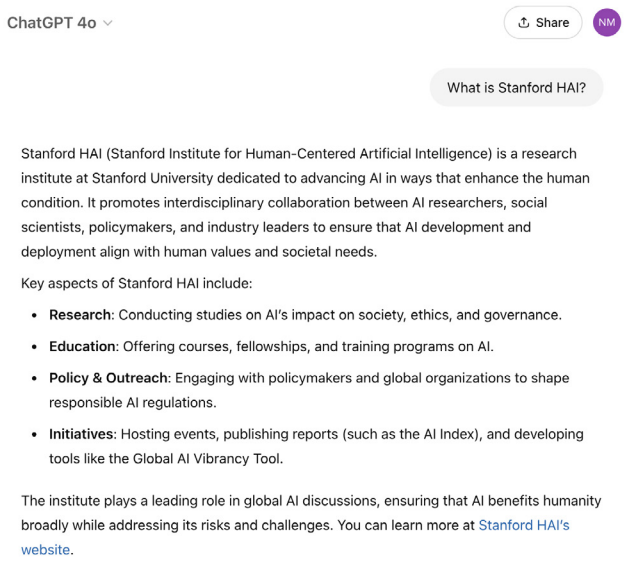


Figure 2.2.1

Gemini 2.0 in an agentic workflow
Source: AI Index, 2025

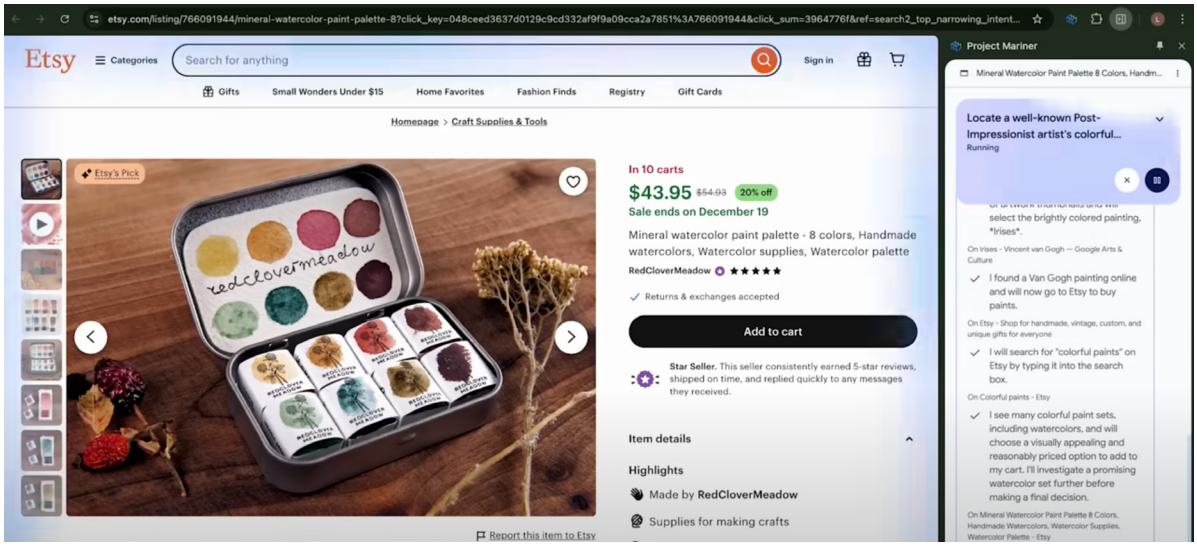


Figure 2.2.2

Understanding

English language understanding challenges AI systems to understand the English language in various ways, such as reading comprehension and logical reasoning.

MMLU: Massive Multitask Language Understanding

The Massive Multitask Language Understanding (MMLU) benchmark assesses model performance in zero-shot or few-shot scenarios across 57 subjects, including the humanities, STEM, and the social sciences (Figure 2.2.3). MMLU has emerged as a premier benchmark for assessing LLM capabilities: Many state-of-the-art models like GPT-4o, Claude 3.5, and Gemini 2.0 have been evaluated against MMLU.

The MMLU benchmark was created in 2020 by a team of researchers from UC Berkeley, Columbia University, University of Chicago, and University of Illinois Urbana-Champaign.

The highest recorded score on MMLU, 92.3%, was achieved by OpenAI's o1-preview model in September 2024. For comparison, GPT-4, launched in March 2023, scored 86.4% on the benchmark. Notably, one of the earliest models tested on MMLU, RoBERTa, achieved just 27.9% in 2019 (Figure 2.2.4). This latest state-of-the-art result represents a remarkable 64.4 percentage point increase over five years.

A sample question from MMLU
Source: Hendrycks et al., 2021

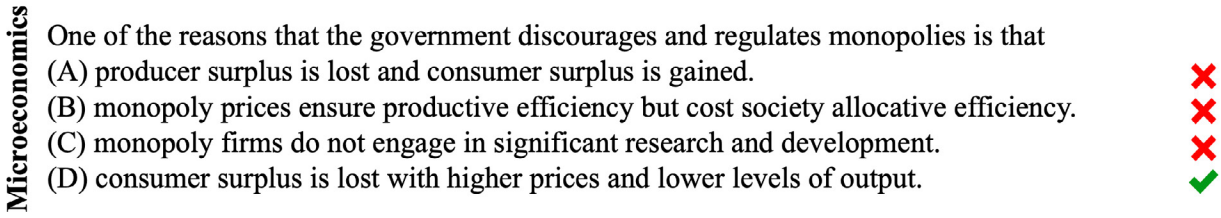


Figure 2.2.3

MMLU: average accuracy
Source: Papers With Code, 2025 | Chart: 2025 AI Index report

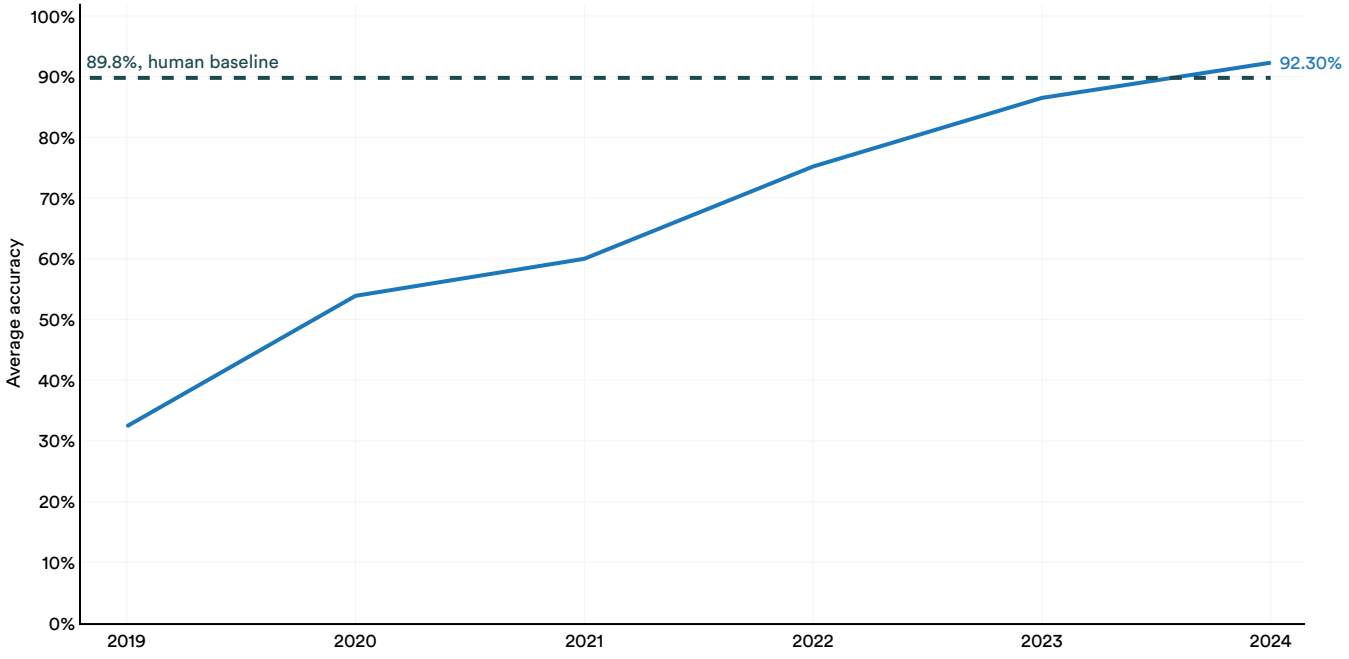


Figure 2.2.4

Despite its prominence, MMLU has faced notable criticisms. These include claims that the benchmark contains erroneous or overly simplistic questions, which may not challenge increasingly advanced systems. In 2024, a team of researchers from the University of Toronto, University of Waterloo, and Carnegie Mellon introduced MMLU-Pro, a more challenging variant of MMLU. This version eliminates noisy and trivial questions, expands complex ones, and increases the number of answer choices available to models. Figure 2.2.5 highlights performance trends on MMLU-Pro, with DeepSeek-R1 posting the highest score to date (84.0%).

Additionally, concerns have been raised about the testing landscape. Developers sometimes report MMLU scores using nonstandard prompting techniques that boost performance but can lead to misleading comparisons. Furthermore, evidence suggests that publicly reported scores by developers can differ—sometimes by as much as five percentage points—from those later evaluated by academic researchers. As such, MMLU performance results should be interpreted with caution.

Generation

In generation tasks, AI models are tested on their ability to produce fluent and practical language responses.

Chatbot Arena Leaderboard

The rise of capable LLMs has made it increasingly important to understand which models are preferred by the general public. Launched in 2023, the Chatbot Arena Leaderboard from LMSYS is one of the first comprehensive evaluations of public LLM preference. The leaderboard allows users to query two anonymous models and vote for the preferred generations (Figure 2.2.6). By early 2025, the platform had accumulated over 1 million votes, with users ranking one of Google's Gemini models as the community's most preferred choice.

MMLU-Pro: overall accuracy

Source: MMLU-Pro Leaderboard, 2025 | Chart: 2025 AI Index report

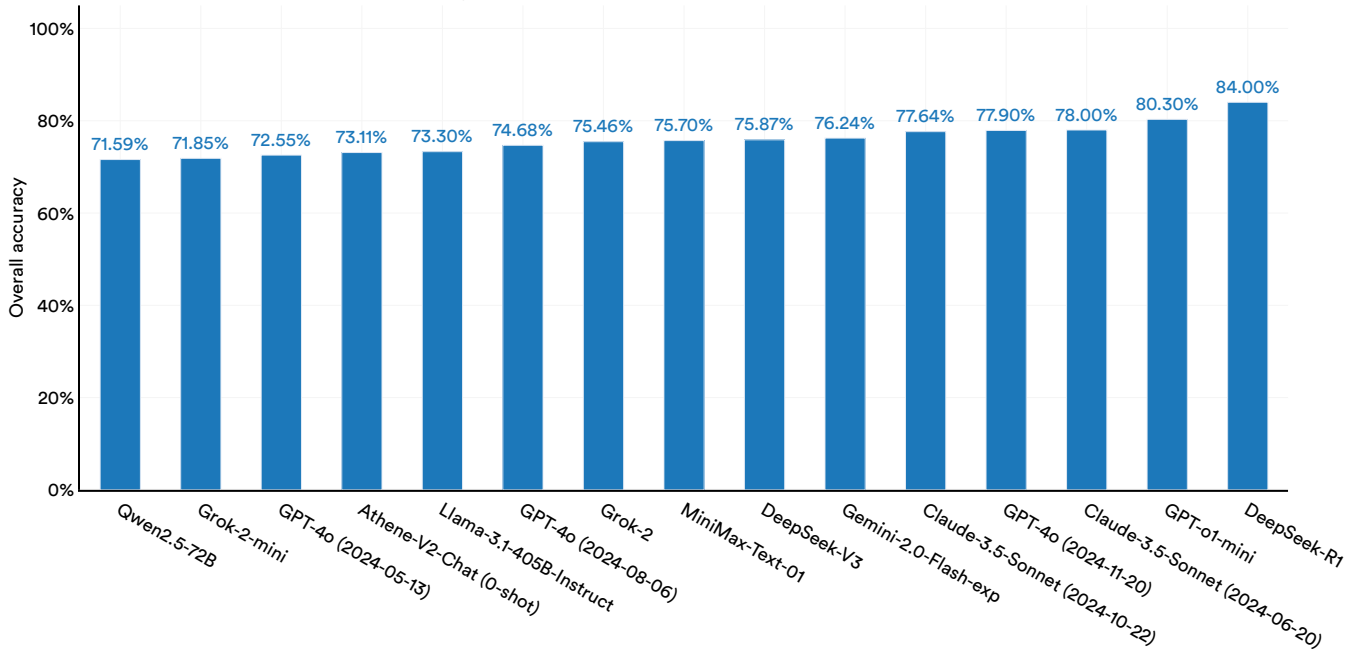


Figure 2.2.5

A sample model response on the Chatbot Arena Leaderboard
Source: Chatbot Arena Leaderboard, 2024

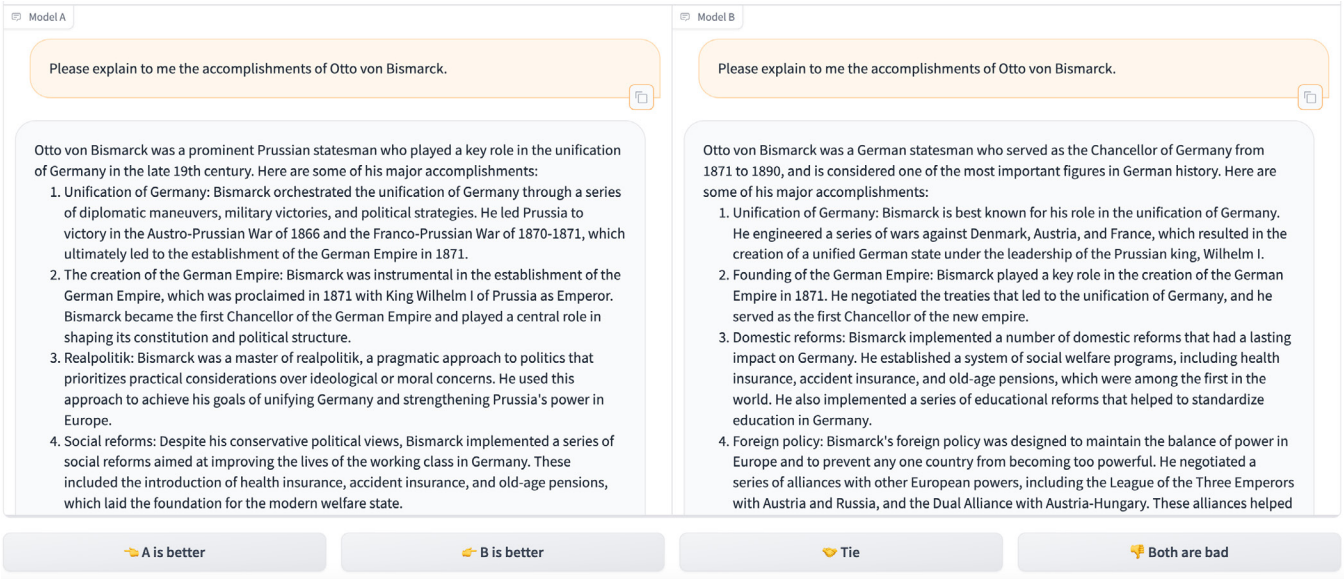


Figure 2.2.6

Figure 2.2.7 provides a snapshot of the top 10 models on the Chatbot Arena Leaderboard as of January 2025. Interestingly, the performance gap between top leaderboard models has narrowed over time. In 2023, according to data from the 2024 AI Index, the difference in Arena scores between the top model and the 10th-ranked model was 11.9%.[6] By 2025, this gap had decreased to just 5.4%. This convergence highlights a growing parity in the quality of recent LLMs.

LMSYS Chatbot Arena for LLMs: Elo rating (overall)
Source: LMSYS, 2025 | Chart: 2025 AI Index report

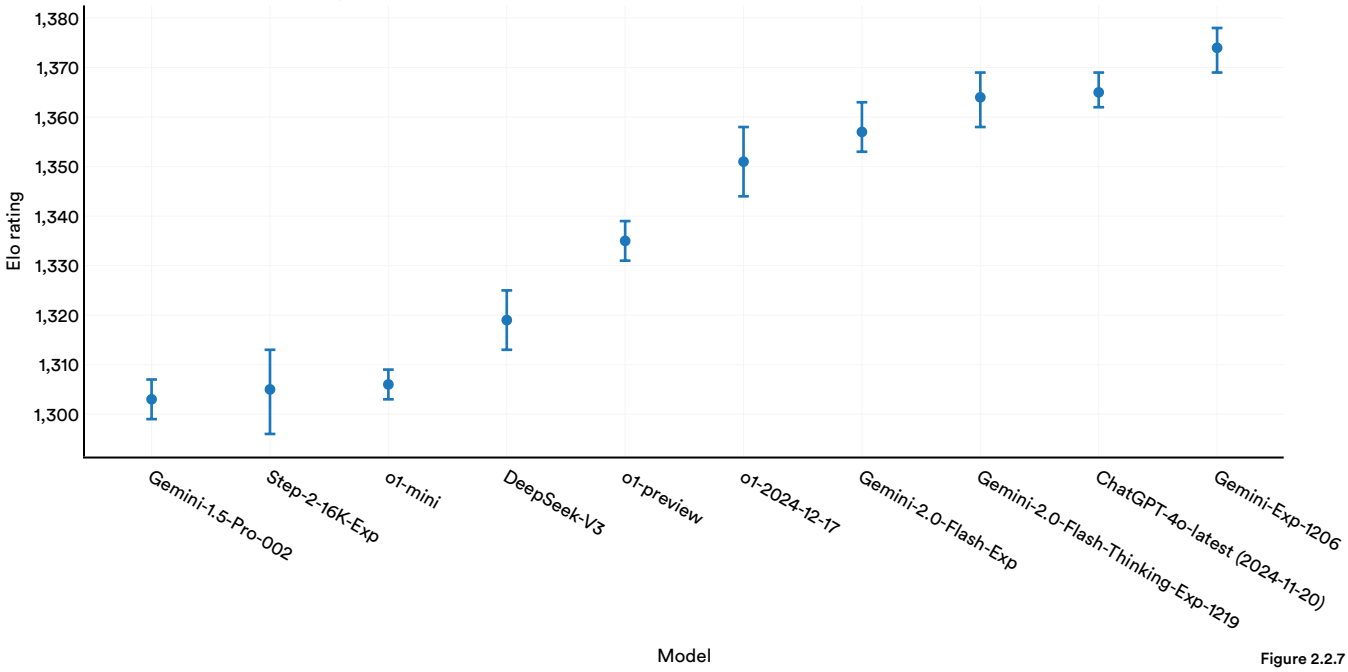


Figure 2.2.7

6 The Arena score is a relative ranking system used by the Arena Leaderboard to compare model performance. For more details on the scoring methodology, refer to the paper introducing the Chatbot Arena Leaderboard.

Arena-Hard-Auto

One of the challenges in developing new benchmarks to keep pace with rapidly improving AI capabilities is that creating high-quality, human-curated benchmarks is often expensive and time-consuming. In response, this year saw the launch of BenchBuilder. Created by a team of UC Berkeley researchers, BenchBuilder leverages LLMs to create an automated pipeline for curating high-quality, open-ended prompts from large, crowdsourced datasets. BenchBuilder can be used to update or create new benchmarks without significant human involvement. This tool was used by the LMSYS team to develop Arena-Hard-Auto, a benchmark designed to evaluate instruction-tuned LLMs (Figure 2.2.8). Arena-Hard-Auto includes 500 challenging user queries sourced from Chatbot Arena. In this benchmark, GPT-4 Turbo serves as the judge that compares model responses against a baseline model (GPT-4-0314).

As of November 2024, the top-scoring models on the Arena-Hard-Auto leaderboard were o1-mini (92.0), o1-preview (90.4), and Claude-3.5-Sonnet (85.2) (Figure 2.2.9). Arena-Hard-Auto also features a style control leaderboard, which accounts for how the style of an LLM's responses might inadvertently influence user preferences. The top model on the style leaderboard is the November variant of Anthropic's Claude Sonnet 3.5 (Figure 2.2.10). Automated benchmarks like Arena-Hard-Auto have faced criticism for uneven question distribution, which limits their ability to provide a comprehensive assessment of LLM capabilities. For instance, over 50% of Arena-Hard-Auto questions focus solely on coding and debugging.

Arena-Hard-Auto vs. other benchmarks
Source: Li et al., 2024

	Evaluation	Open-Ended	Prompt Curation	Prompt Source
Arena-Hard-Auto	Automatic	Yes	Automatic	Configurable
MMLU, MATH, GPQA	Automatic	No	Manual	Fixed
MT-Bench, AlpacaEval	Automatic	Yes	Manual	Fixed
Live Bench, Live Code Bench	Automatic	No	Manual	Fixed
Chatbot Arena	Human	Yes	Crowd-source	Crowd

Figure 2.2.8

Arena-Hard-Auto with no modification
Source: LMSYS, 2025 | Chart: 2025 AI Index report

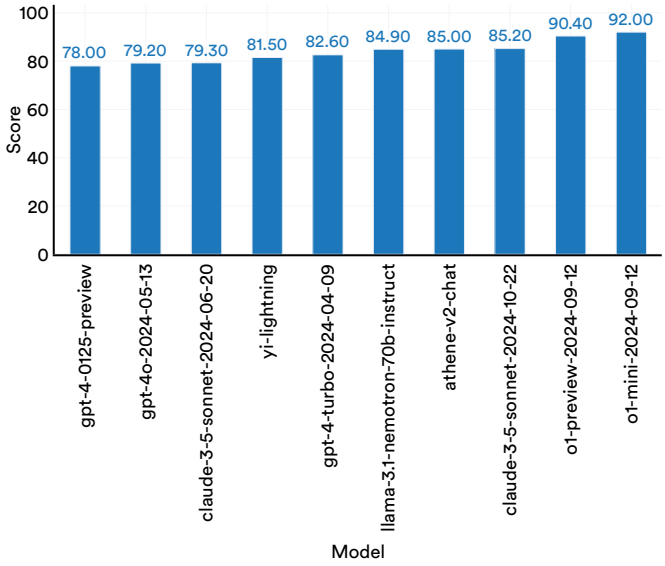


Figure 2.2.9

Arena-Hard-Auto with style control
Source: LMSYS, 2025 | Chart: 2025 AI Index report

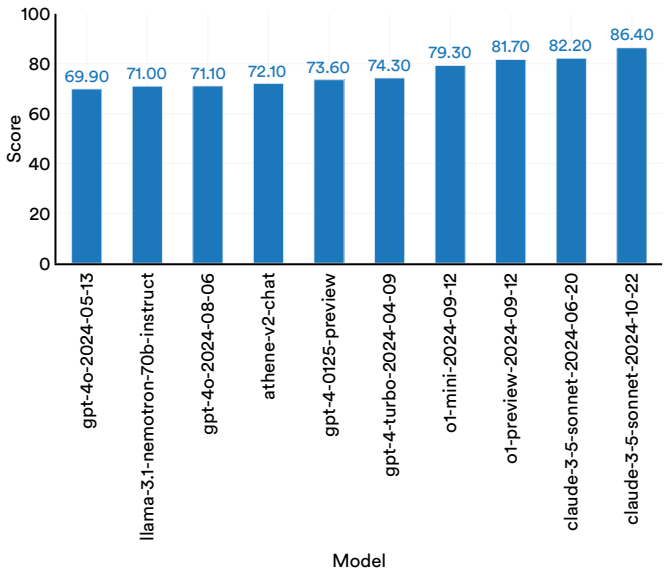


Figure 2.2.10

WildBench

WildBench, developed by researchers from the Allen Institute for AI and the University of Washington, is a benchmark launched in 2024 to evaluate LLMs on challenging real-world queries. The creators highlight several limitations of existing LLM evaluations. For example, MMLU focuses on academic questions and does not assess open-ended, real-world problems. Similarly, benchmarks like LMSYS, which address real-world challenges, rely heavily on human oversight and lack consistency in evaluating all models with the same dataset.

Evaluation framework for WildBench
Source: Lin et al., 2024

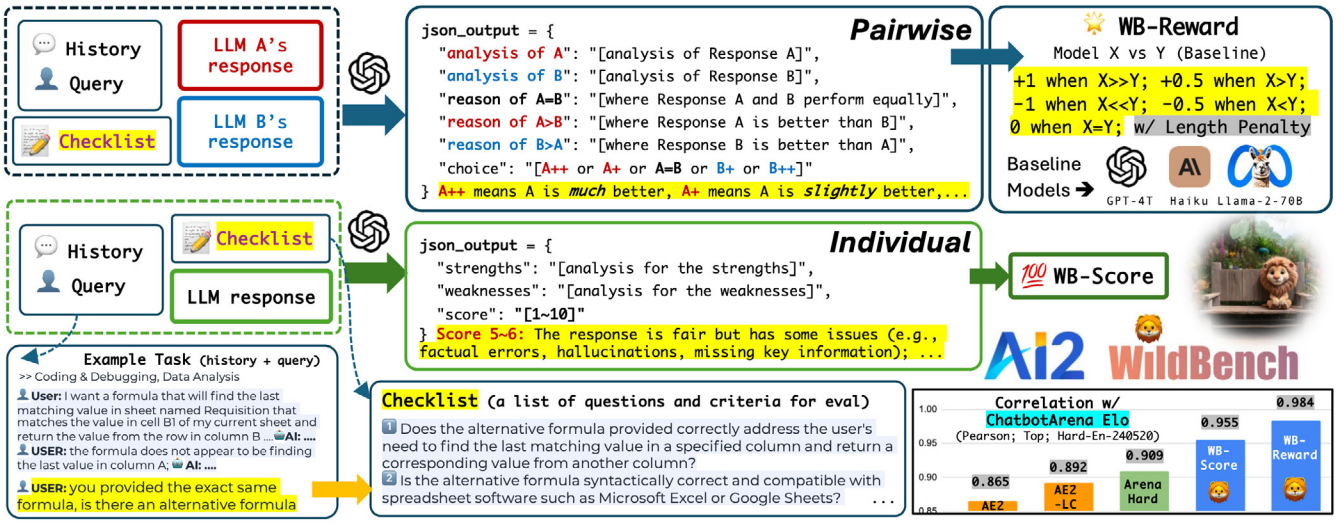


Figure 2.2.11

WildBench addresses many shortcomings of existing benchmarks by providing an automated evaluation framework for LLMs, incorporating a diverse set of real-world ("in the wild") questions that language models are likely to encounter (Figure 2.2.11). The questions in WildBench are meticulously selected from over 1 million human-chatbot interactions and are periodically updated to ensure relevance. The creators also maintain a live leaderboard to track model performance over time. Currently, the top-performing model on WildBench is GPT-4o, with an Elo score of 1227.1, narrowly surpassing the second-place model, Claude 3.5 Sonnet, which scored 1215.4 (Figure 2.2.12).

WildBench: WB-Elo (length controlled)

Source: WildBench Leaderboard, 2025 | Chart: 2025 AI Index report

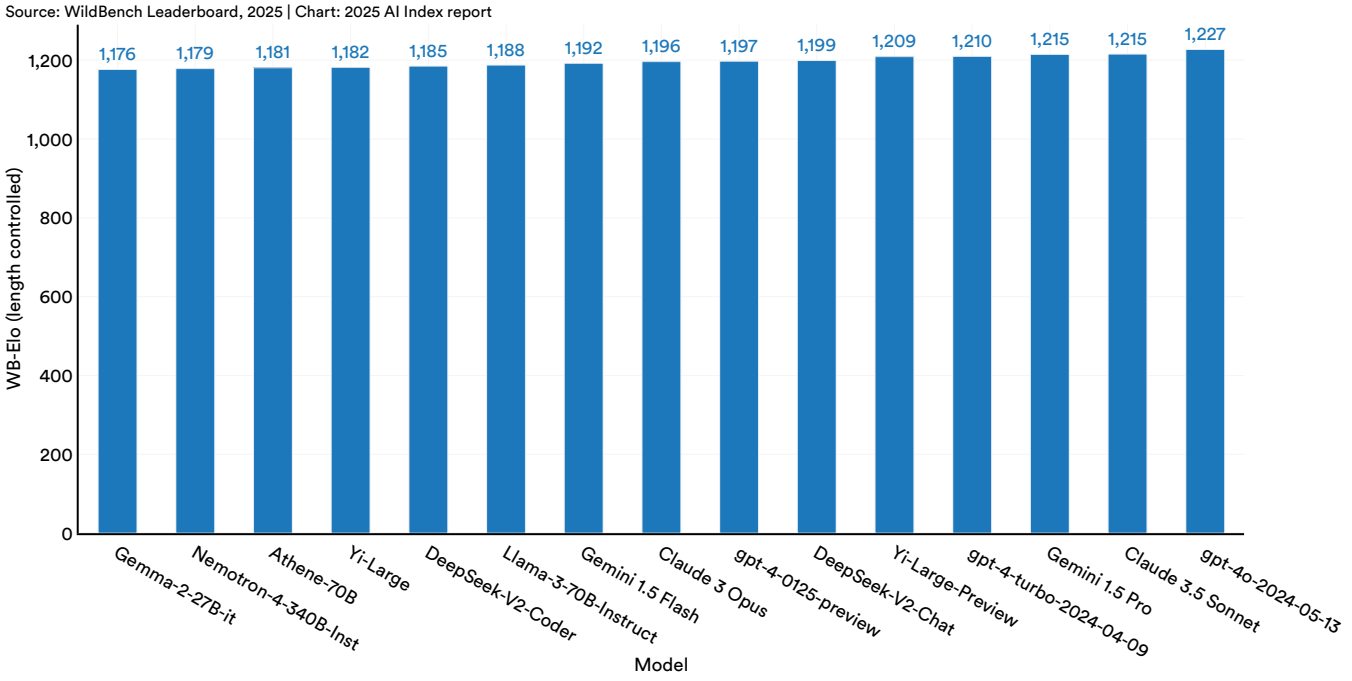


Figure 2.2.12

Highlight:

o1, o3, and Inference-Time Compute

OpenAI's latest two models, o1 and o3, mark a paradigm shift in AI models' ability to "think" and exhibit signs of advanced reasoning. o1 and o3 have shown impressive results across a variety of tasks, including programming, quantum physics, and logic. The models' advanced reasoning capabilities are attributed to their chain-of-thought process and ability to iteratively check answers. This means that the models break complex problems into smaller, more manageable steps before executing them, enhancing the resulting output quality. For example, when asked to decipher scrambled text, o1 will specify its thought and reasoning process more thoroughly than GPT-4 (Figure 2.2.13). This process, through which AI systems iterate as they answer, has been referred to as inference or test-time computation.

Chain-of-thought thinking in o1
Source: OpenAI, 2024

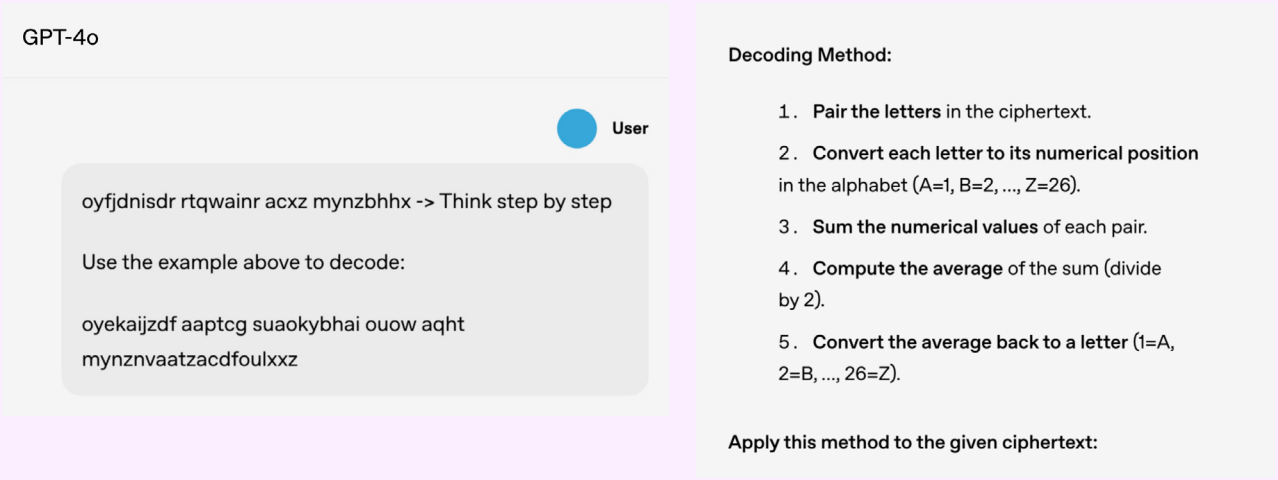


Figure 2.2.13

Highlight:

o1, o3, and Inference-Time Compute (cont'd)

Figure 2.2.14 juxtaposes the scores of GPT-4o, OpenAI's previous state-of-the-art model, with o1 and o1-preview on a variety of benchmarks.[7] For example, o1 outperforms GPT-4o with a 2.8-point gain on MMLU, 34.5 points on MATH, 26.7 points on GPQA Diamond, and 65.1 points on AIME 2024, a notoriously difficult mathematics competition. Finally, o3 demonstrates more complex reasoning than any other AI model known today, posting an 87.5% accuracy rate on the ARC-AGI machine intelligence benchmark and passing the previous record of 55.5%.

GPT-4o vs. o1-preview vs. o1 on select benchmarks
Source: OpenAI, 2024 | Chart: 2025 AI Index report

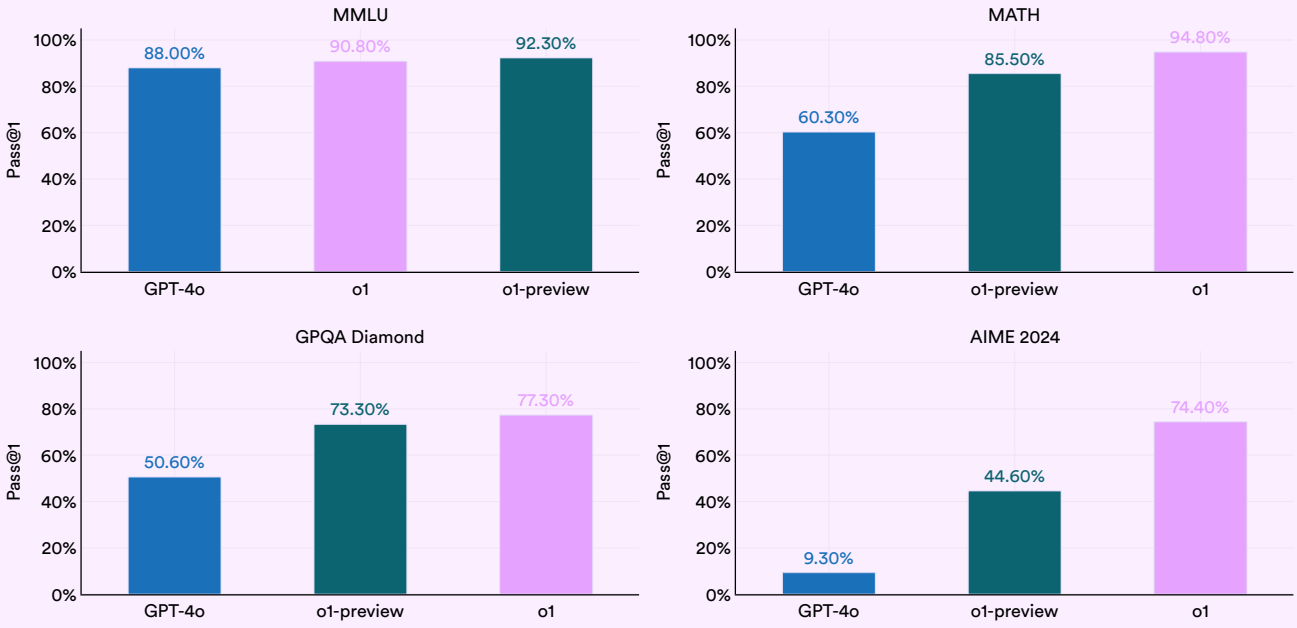


Figure 2.2.14

While these models enhance reasoning capabilities, this comes at a price—both a financial and latency cost. For example, GPT-4o costs $2.50 per 1 million input tokens and $10 per 1 million output tokens. Conversely, o1 costs $15 per 1 million input tokens and $60 per 1 million output tokens.[8] Moreover, o1 is approximately 40 times slower than GPT-4o, with 29.7 seconds to first token as opposed to GPT-4o's 0.72. The latency of o3, while not publicly available, is presumably even higher. o1 and o3's strong capabilities are likely to continue fueling powerful AI systems and agents.

OpenAI first released o1-preview to ChatGPT Plus and Teams users on Sept. 12, 2024, and released the full version of o1 (as well as access to ChatGPT Pro, a $200 monthly subscription enabling access to o1) on Dec. 5, 2024.

7 The o1-preview model is OpenAI's early release of o1, made available before its broader public launch.

8 o3 is currently only available to select researchers and developers via OpenAI's safety testing program.

MixEval

MixEval, launched by researchers at the National University of Singapore, Carnegie Mellon University, and the Allen Institute for AI, is another newly released benchmark designed to address some of the aforementioned limitations in the current field of LLM evaluation. MixEval combines comprehensive, well-distributed, real-world user queries, similar to those found in Chatbot Arena, with ground-truth-based questions, like those featured in MMLU (Figure 2.2.15). MixEval includes various evaluation suites, with MixEval-Hard representing the more challenging version of the benchmark. This suite focuses on substantially harder queries, making it one of the most effective tools for assessing how models handle complex questions.

Evaluation framework for MixEval
Source: Ni et al., 2024

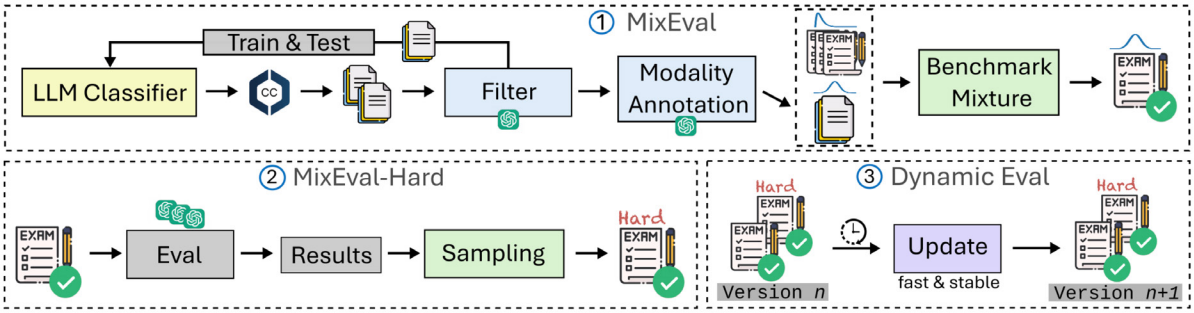


Figure 2.2.15

The highest-scoring model on the MixEval-Hard benchmark is OpenAI's o1-preview, with a score of 72.0. In second place is the Claude 3.5 Sonnet-0620 model, followed by the Llama-3 1-405B-Instruct model, which scored 66.2 (Figure 2.2.16). All three models were released in 2024.

MixEval-Hard on chat models: score
Source: MixEval Leaderboard, 2025 | Chart: 2025 AI Index report

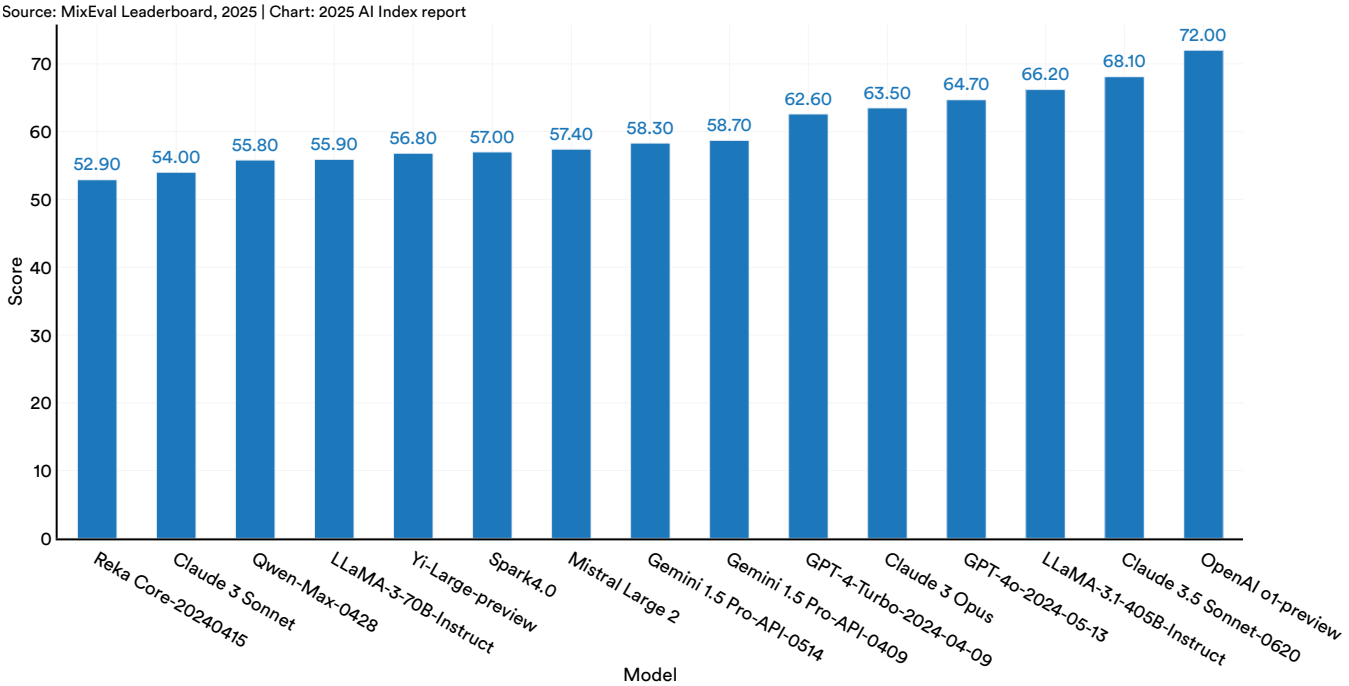


Figure 2.2.16

RAG: Retrieval Augment Generation (RAG)

An increasingly common capability being tested in LLMs is retrieval-augmented generation (RAG). This approach integrates LLMs with retrieval mechanisms to enhance their response generation. The model first retrieves relevant information from files or documents and then generates a response tailored to the user's query based on the retrieved content. RAG has diverse use cases, including answering precise questions from large databases and addressing customer queries using information from company documents.

In recent years, RAG has received increasing attention from researchers and companies. For example, in September 2024, Anthropic introduced Contextual Retrieval, a method that significantly enhances the retrieval capabilities of RAG models. 2024 also saw the release of numerous benchmarks for evaluating RAG systems, including Ragnarok (a RAG arena battleground) and CRAG (Comprehensive RAG benchmark). Additionally, specialized RAG benchmarks, such as FinanceBench for financial question answering, have been developed to address specific use cases.

Berkeley Function Calling Leaderboard

The Berkeley Function Calling Leaderboard evaluates the ability of LLMs to accurately call functions or tools. The evaluation suite includes over 2,000 question-function-answer pairs across multiple programming languages (such as Python, Java, JavaScript, and REST API) and spans a variety of testing domains (Figure 2.2.17).

Data composition on the Berkeley Function Calling Leaderboard
Source: Yan et al., 2024

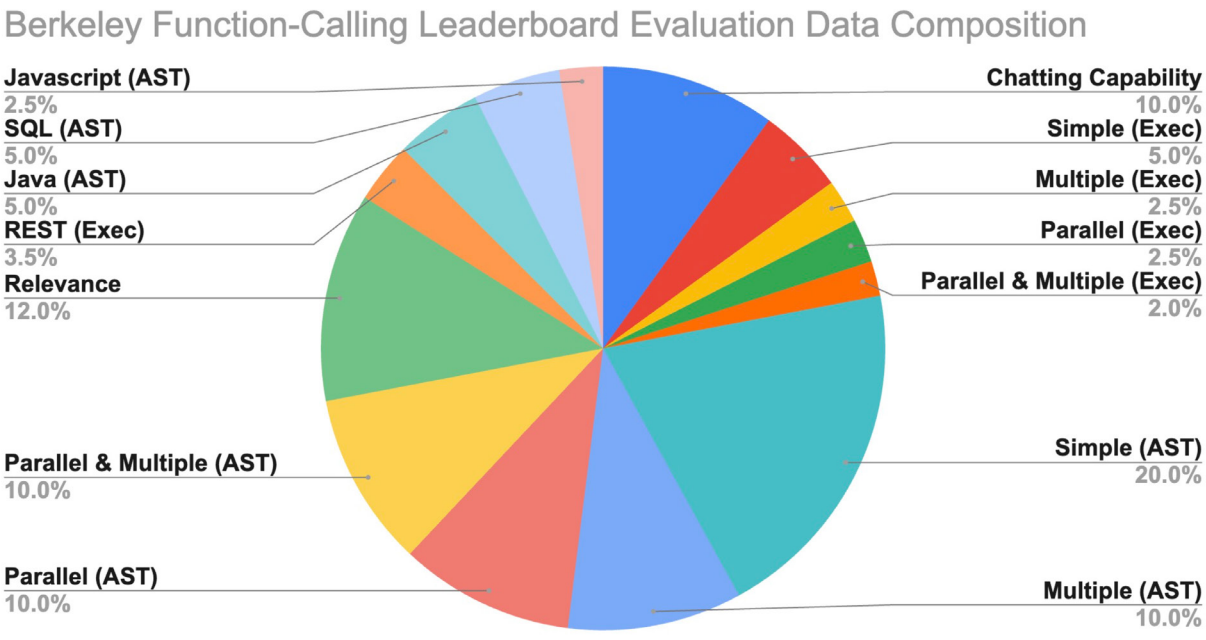


Figure 2.2.17[9]

9 In this context: AST (abstract syntax tree) refers to tasks that involve analyzing or manipulating code at the structural level, using its parsed representation as a tree of syntactic elements. Evaluations labeled with "AST" likely test an AI model's ability to understand, generate, or manipulate code in a structured manner. Exec (execution-based) indicates tasks that require actual execution of function calls to verify correctness. Evaluations labeled with "Exec" likely assess whether the AI model can correctly call and execute functions, ensuring the expected outputs are produced.

The top model on the Berkeley Function Calling Leaderboard is watt-tool-70b, a fine-tuned variant of Llama-3.3-70B-Instruct designed specifically for function calling. It achieved an overall accuracy of 74.24 (Figure 2.2.18). The next-highest-scoring model was a November variant of GPT-4o, with a score of 72.02. Performance on this benchmark has improved significantly over the course of 2024, with top models at the end of the year achieving accuracies up to 50 points higher than those recorded early in the year.

Berkeley Function-Calling: overall accuracy

Source: Berkeley Function-Calling Leaderboard, 2025 | Chart: 2025 AI Index report

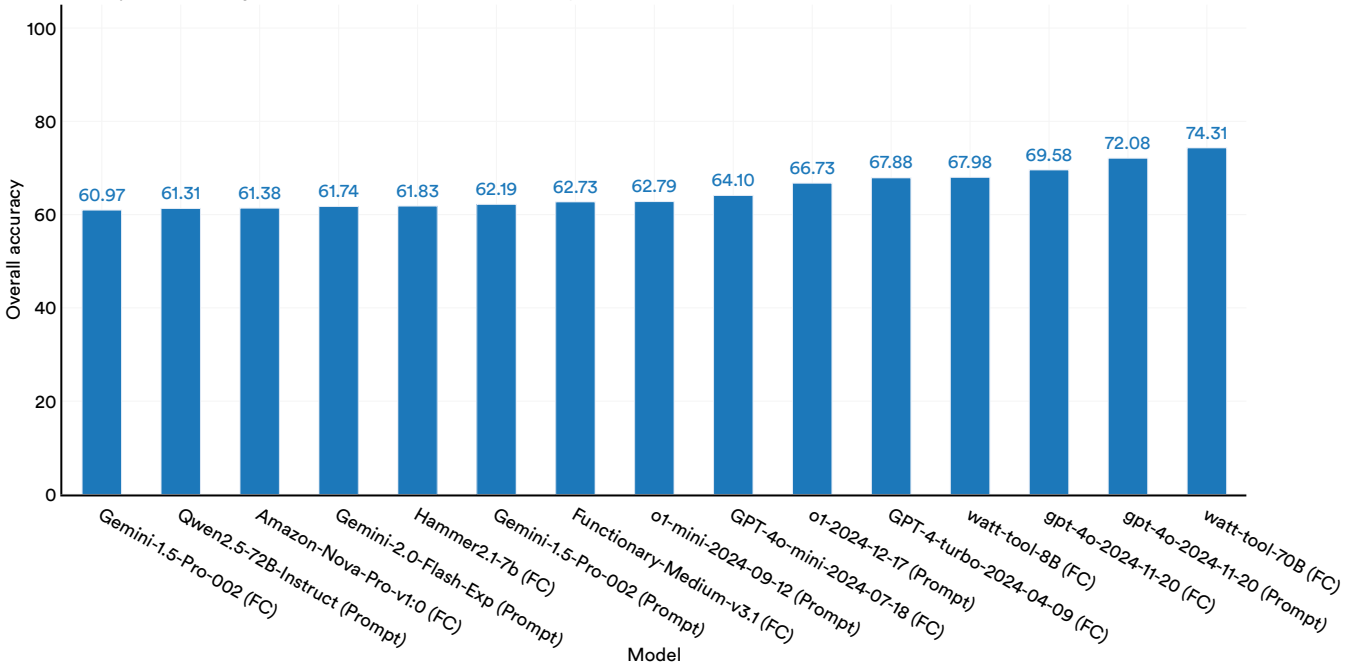


Figure 2.2.18

MTEB: Massive Text Embedding Benchmark

The Massive Text Embedding Benchmark (MTEB), created by a team at Hugging Face and Cohere, was introduced in late 2022 to comprehensively evaluate how models perform on various embedding tasks. Embedding involves converting data, such as words, texts, or documents, into numerical vectors that capture rough semantic meanings and distance between vectors. Embedding is an essential component of RAG. During a RAG task, when users input a query, the model transforms it into an embedding vector. This transformation enables the model to then search for relevant information. MTEB includes 58 datasets spanning 112 languages and eight embedding tasks (Figure 2.2.19).[10] For example, in the bitext mining task, there are two sets of sentences from two different languages, and for every sentence in the first set, the model is tasked to find the best match in the second set.

Tasks in the MTEB benchmark
Source: Muennighoff et al., 2023

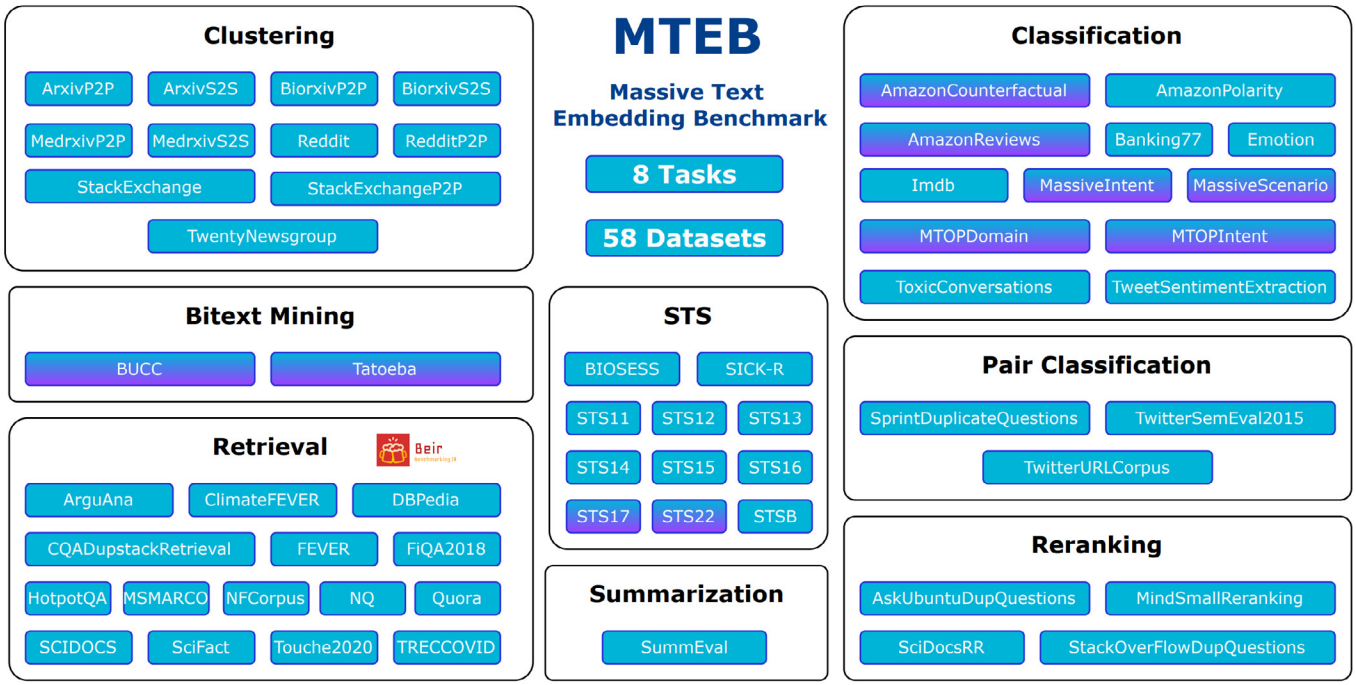


Figure 2.2.19

10 The benchmark covers the following eight tasks: bitext mining, classification, clustering, pair classification, reranking, retrieval, semantic textual similarity, and summarization. For details on each task, refer to the MTEB paper.

As of early 2025, the top-performing embedding model on the MTEB benchmark was Voyage AI's voyage-3-m-exp, with a score of 74.03. Voyage AI is focused on creating high-quality AI embedding models. The voyage-3-m-exp model is a variant of the voyage-3-large, a large foundation model specifically designed for embedding tasks, and it uses strategies like Matryoshka Representation Learning and quantization-aware training to improve its performance. The voyage-3-m-exp model narrowly outperformed NV-Embed-v2 (72.31), which held the top spot for most of 2024 (Figure 2.2.20). When the MTEB benchmark was first introduced in late 2022, the leading model achieved an average score of 59.5. Over the past two years, therefore, performance on the benchmark has meaningfully improved.

MTEB on English subsets across 56 datasets: average score

Source: MTEB Leaderboard, 2025 | Chart: 2025 AI Index report

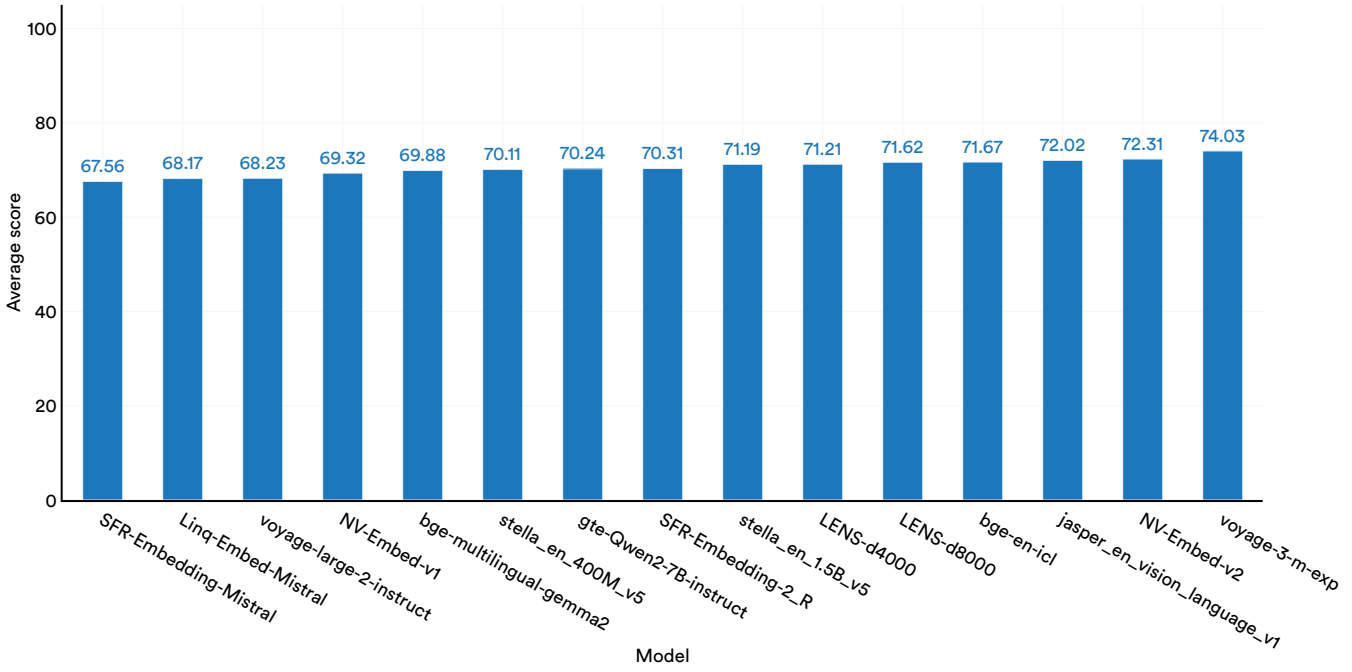


Figure 2.2.20

Highlight:

Evaluating Retrieval Across Long Contexts

As AI models have advanced, their ability to handle longer contexts has significantly improved. For example, models like GPT-4 and Llama 2, released in 2023 by OpenAI and Meta, featured context windows of 8,000 and 4,000 tokens, respectively. In contrast, more recent models such as GPT-4o (May 2024) and Gemini 2.0 Pro Experimental (February 2025) boast context windows ranging from 128 thousand to 2 million. These extended context windows allow users to input and process increasingly large amounts of data, enabling more complex and detailed interactions.

As the context windows of LLMs have expanded, evaluating their performance in long-context settings has become increasingly important. However, existing long-context evaluation methods have been relatively limited. Typically, these evaluations focus on "needle-in-the-haystack" scenarios, where models are tasked with retrieving specific pieces of information from lengthy texts. While useful, such evaluations provide only a baseline assessment of a model's ability to function effectively in long-context environments.

In 2024, several new evaluation suites were introduced to address the limitations of long-context model assessments and improve their evaluation. One such benchmark is Nvidia's RULER, which assesses long-context performance by examining retrieval performance and multihop reasoning, aggregation, and question answering. Among the models evaluated on RULER, Gemini-1.5-Pro achieved the highest weighted performance average (95.5), followed by GPT-4 (89.0) and Llama 3.1 (85.5) (Figure 2.2.21). The researchers behind RULER also revealed that many models suffer performance issues in longer context settings. In fact, the RULER team demonstrated that while most popular LLMs claim context sizes of 32K tokens or greater, only half of them can maintain satisfactory performance at the length of 32K. This means that their actual operational context windows are shorter than those claimed by their developers (Figure 2.2.22).

RULER: weighted average score (increasing)
Source: Hsieh et al., 2024 | Chart: 2025 AI Index report

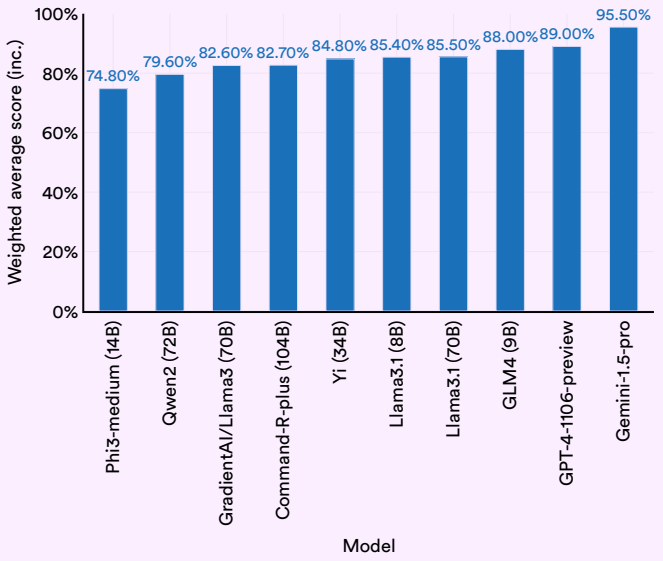


Figure 2.2.21

RULER: claimed vs. effective context length
Source: Hsieh et al., 2024 | Chart: 2025 AI Index report

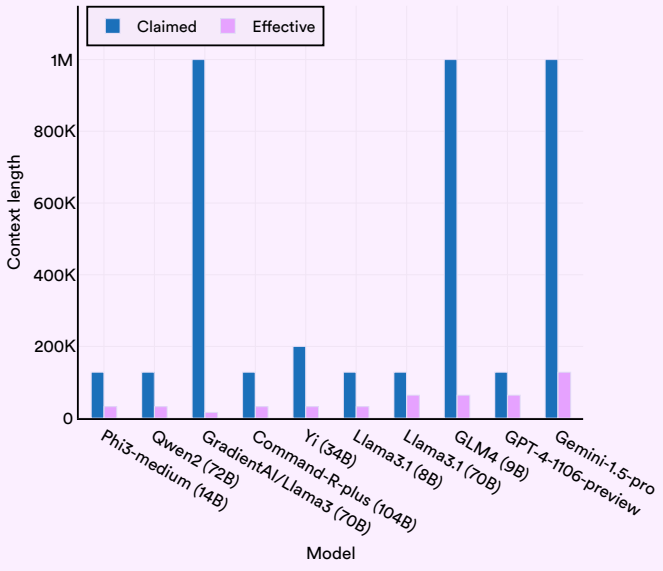


Figure 2.2.22

Highlight:

Evaluating Retrieval Across Long Contexts (cont'd)

HELMET (How to Evaluate Long-Context Models Effectively and Thoroughly), an Intel and Princeton collaboration, is another long-context evaluation benchmark introduced in 2024. The researchers behind HELMET were motivated by the inadequacies of existing benchmarks, which suffered from insufficient coverage of downstream tasks, context lengths too short to test evolving long-context capabilities, and unreliable metrics (Figure 2.2.23). Even more comprehensive than RULER, HELMET features seven long-context evaluation categories, including synthetic recall, passage re-ranking, and generation with citations. Figure 2.2.24 illustrates the average performance of several notable models on the HELMET benchmark across 8K, 32K, and 128K context settings. While models like GPT-4, Claude 3.5 Sonnet, and Llama 3.1-70B struggle with performance degradation in longer context settings, others, such as Gemini 1.5 Pro and the August variant of GPT-4, maintain their effectiveness. The introduction of benchmarks like RULER and HELMET highlights how the rapid evolution of LLMs is compelling researchers to rethink and refine evaluation methodologies.

Comparing long-context benchmarks
Source: Yen et al., 2024
Figure 2.2.23

	Type of tasks							Benchmark features		
	Cite	RAG	Re-rank	Long-QA	Summ	ICL	Synthetic Recall	Robust Eval.	$L \geq 128k$	Controllable L
ZeroSCROLLS	✗	✗	✗	✓	✓	✗	✗	✗	✗†	✗
LongBench	✗	✓	✗	✓	✓	✓	✓	✗	✗†	✗
L-Eval	✗	✓	✗	✓	✓	✗	✗	✓‡	✗†	✗
RULER	✗	✗	✗	✗	✗	✗	✓	✓	✓	✓
∞BENCH	✗	✗	✗	✓	✓	✗	✓	✗	✓	✓
HELMET (Ours)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

HELMET: average score
Source: Yen et al., 2024 | Chart: 2025 AI Index report

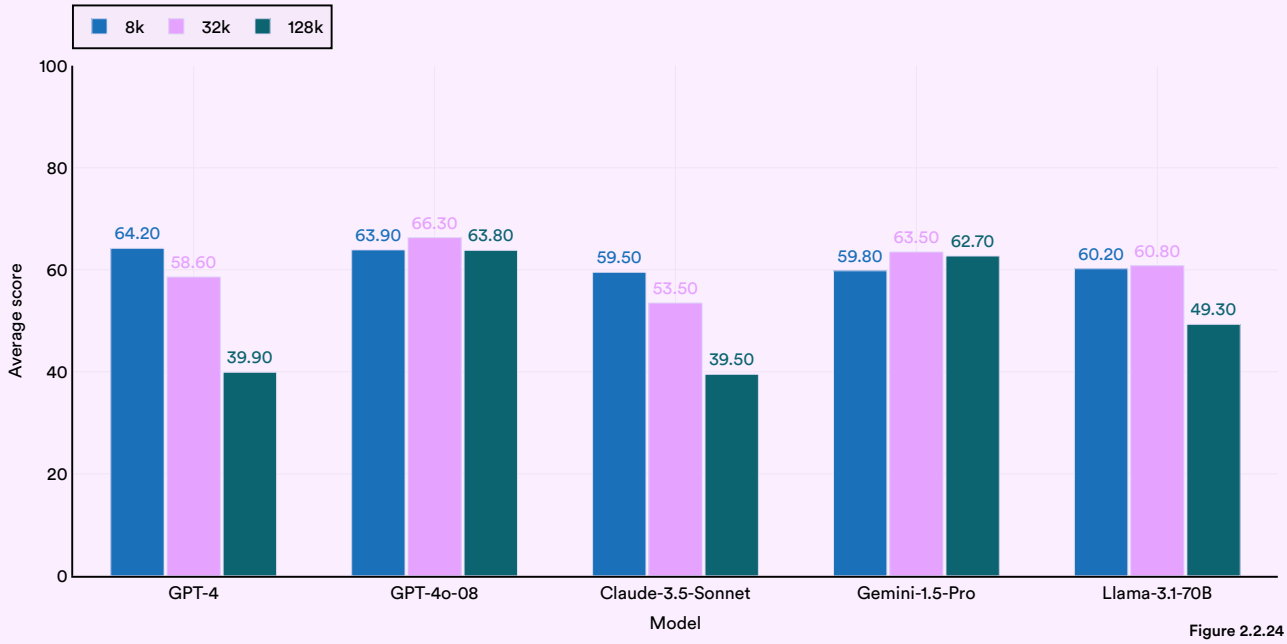


Figure 2.2.24

Computer vision allows machines to understand images and videos and to create realistic visuals from textual prompts or other inputs. This technology is widely used in fields such as autonomous driving, medical imaging, and video game development.

2.3 Image and Video

Understanding

Vision models are evaluated on their ability to understand and reason about the content of images and videos. Vision understanding was one of the first AI capabilities widely tested during the deep learning era. ImageNet, created by Fei-Fei Li and extensively covered in past editions of the AI Index, served as a foundational benchmark for image understanding. As AI systems have advanced, researchers have shifted toward evaluating image models on more complex and comprehensive understanding tasks, such as those involving video or commonsense reasoning in images.

In the ImageNet era, vision algorithms were tasked with more straightforward tasks (e.g., classifying images into predefined categories). However, modern computer vision benchmarks like VCR and MVBench introduce more open-ended challenges, where no fixed categories or classes exist. In these cases, algorithms process natural language questions, identify objects from an open set of images, and generate answers based on image content or prior knowledge.

VCR: Visual Commonsense Reasoning

Introduced in 2019 by researchers from the University of Washington and the Allen Institute for AI, the Visual Commonsense Reasoning (VCR) challenge tests the commonsense visual reasoning abilities of AI systems. In this challenge, AI systems not only answer questions based on images but also reason about the logic behind their answers (Figure 2.3.1). Performance in VCR is measured using the Q->AR score, which evaluates the machine's ability to both select the correct answer to a question (Q->A) and choose the appropriate rationale behind that answer (Q->R).

Sample question from Visual Commonsense Reasoning (VCR) challenge
Source: Zellers et al., 2018

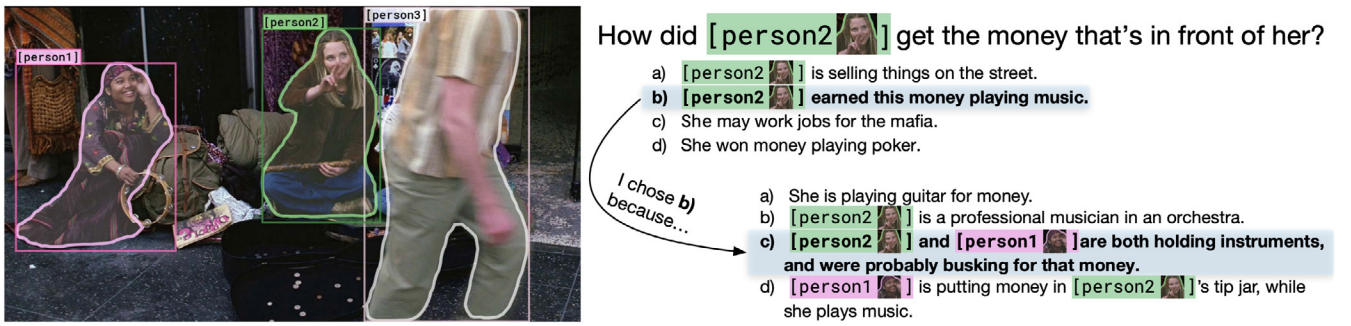


Figure 2.3.1

The VCR benchmark was one of the few benchmarks routinely featured in the AI Index where AI systems consistently fell short of the human baseline. However, 2024 marked a turning point, with AI systems finally reaching this baseline. A model posted to the leaderboard in July 2024 achieved a score of 85.0, matching the human benchmark (Figure 2.3.2). This milestone represented a significant 4.2% improvement on the benchmark since 2023. Even previously challenging benchmarks are now being surpassed.

Visual Commonsense Reasoning (VCR) task: Q->AR score

Source: VCR Leaderboard, 2025 | Chart: 2025 AI Index report

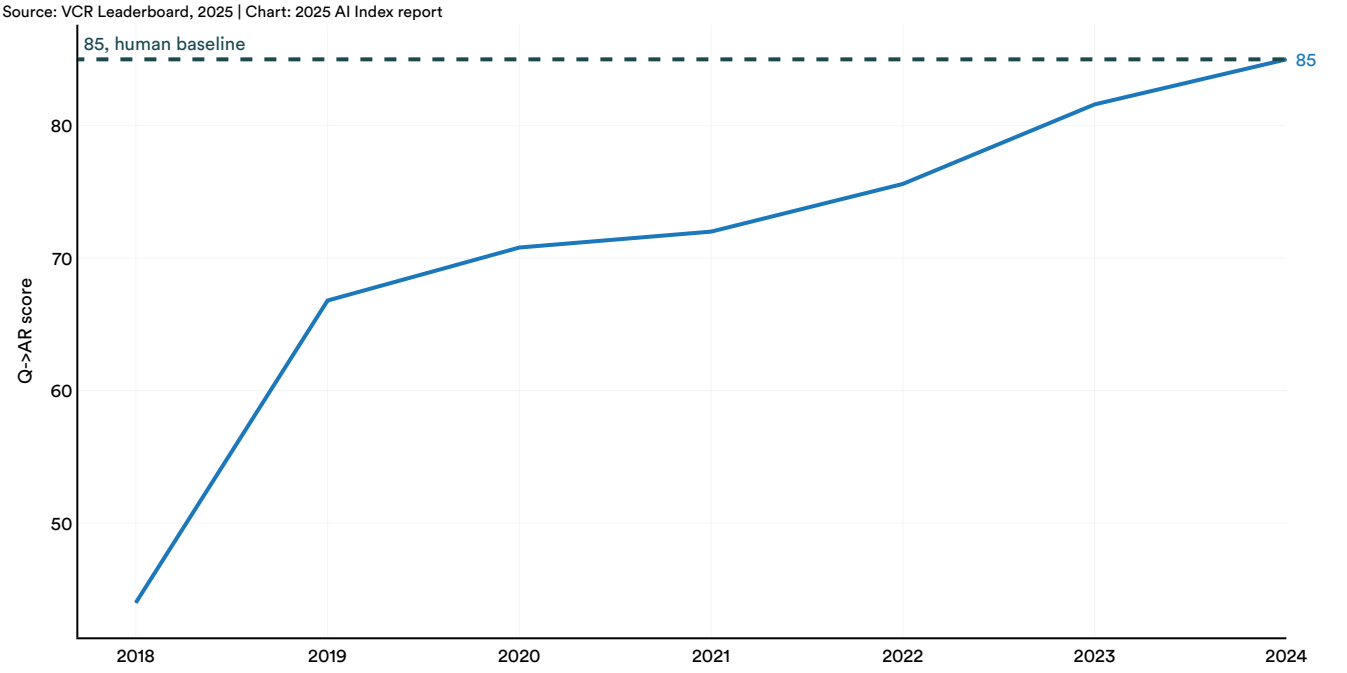


Figure 2.3.2

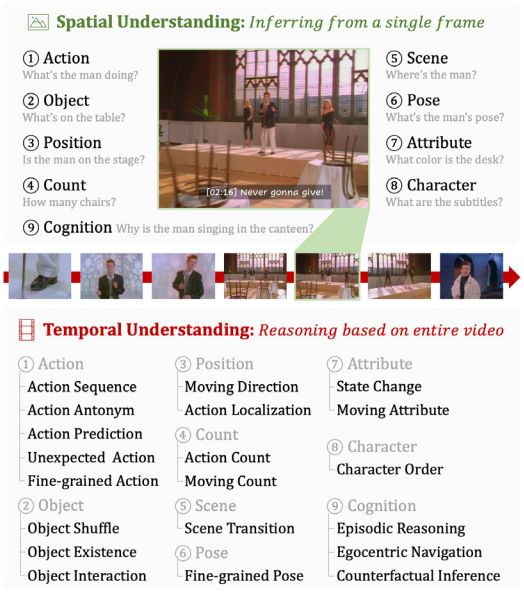
MVBench

MVBench, introduced by a team of researchers from Hong Kong and China in 2023, is a challenging, multimodal, video-understanding benchmark.[11] Unlike earlier video benchmarks that primarily tested spatial understanding through static image tasks, MVBench incorporates more complex video tasks requiring temporal reasoning across multiple frames (Figure 2.3.3).

Sample tasks on MVBench

Source: Li et al., 2023

Figure 2.3.3



11 The researchers were affiliated with the Chinese Academy of Sciences, University of Chinese Academy of Sciences, Shanghai AI Laboratory, the University of Hong Kong, Fudan University, and Nanjing University.

As of 2024, the top model on the MVBench leaderboard is Video-CCAM-7B-v1.2, built on the Queen 2.5-7B-Instruct language model. Its score of 69.23 marks a significant 14.6% improvement on the benchmark since its introduction in late 2023 (Figure 2.3.4). These results highlight the gradual but steady progress in the dynamic video understanding capabilities of AI models.

MVBench: average accuracy

Source: MVBench Leaderboard, 2025 | Chart: 2025 AI Index report

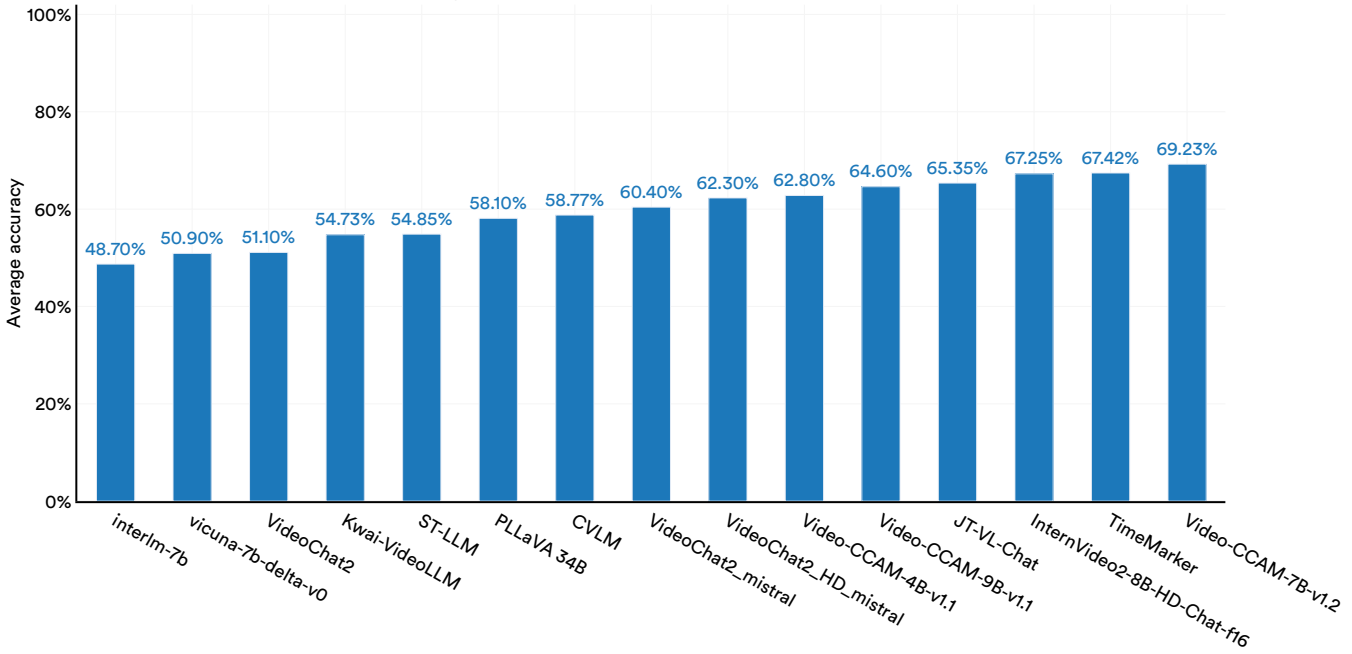


Figure 2.3.4

Generation

Image generation is the task of generating images that are indistinguishable from real ones. As noted in last year's AI Index, today's image generators are so advanced that most people struggle to differentiate between AI-generated images and actual images of human faces (Figure 2.3.5). Figure 2.3.6 highlights several generations from various Midjourney model variants from 2022 to 2025 for the prompt "a hyper-realistic image of Harry Potter." The progression demonstrates the significant improvement in Midjourney's ability to generate hyper-realistic images over a two-year period. In 2022, the model produced cartoonish and inaccurate renderings of Harry Potter, but by 2025, it could create startlingly realistic depictions.

Which face is real?
Source: Which Face Is Real, 2024

Figure 2.3.5

Midjourney generations over time: "a hyper-realistic image of Harry Potter"
Source: Midjourney, 2024

Figure 2.3.6

Chatbot Arena: Vision

The AI community has increasingly embraced public evaluation platforms, such as the Chatbot Arena Leaderboard, to assess the capabilities of leading AI systems, including top AI image generators. This leaderboard also features a Vision Arena, which ranks the performance of over 50 vision models. Users can submit text-to-image prompts, such as "Batman drinking a coffee," and vote for their preferred generation (Figure 2.3.7). To date, the Vision Arena has garnered more than 150,000 votes.

As of early 2025, the top-ranked vision model on the leaderboard is Google's Gemini-2.0-Flash-Thinking-Exp-1219 (Figure 2.3.8). Similar to other Chatbot Arena categories—such as general, coding, and math—the leading models are closely clustered in performance. For example, the gap between the top model and the fourth-ranked model, ChatGPT-4o-latest (2024-11-20), is just 3.4%.

Sample from the Chatbot Vision Arena
Source: Chatbot Arena Leaderboard, 2025

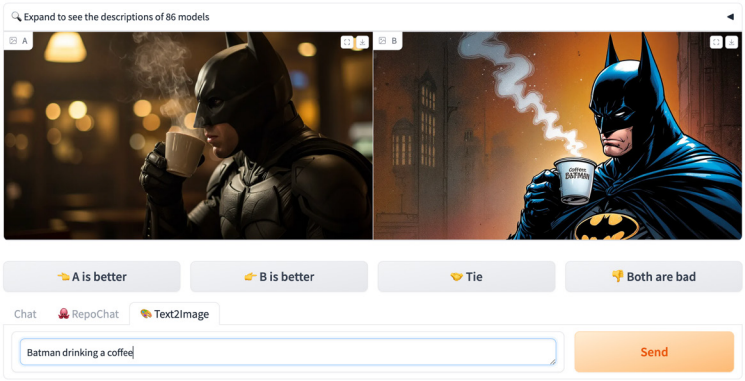


Figure 2.3.7

LMSYS Chatbot Arena for LLMs: Elo rating (vision)
Source: LMSYS, 2025 | Chart: 2025 AI Index report

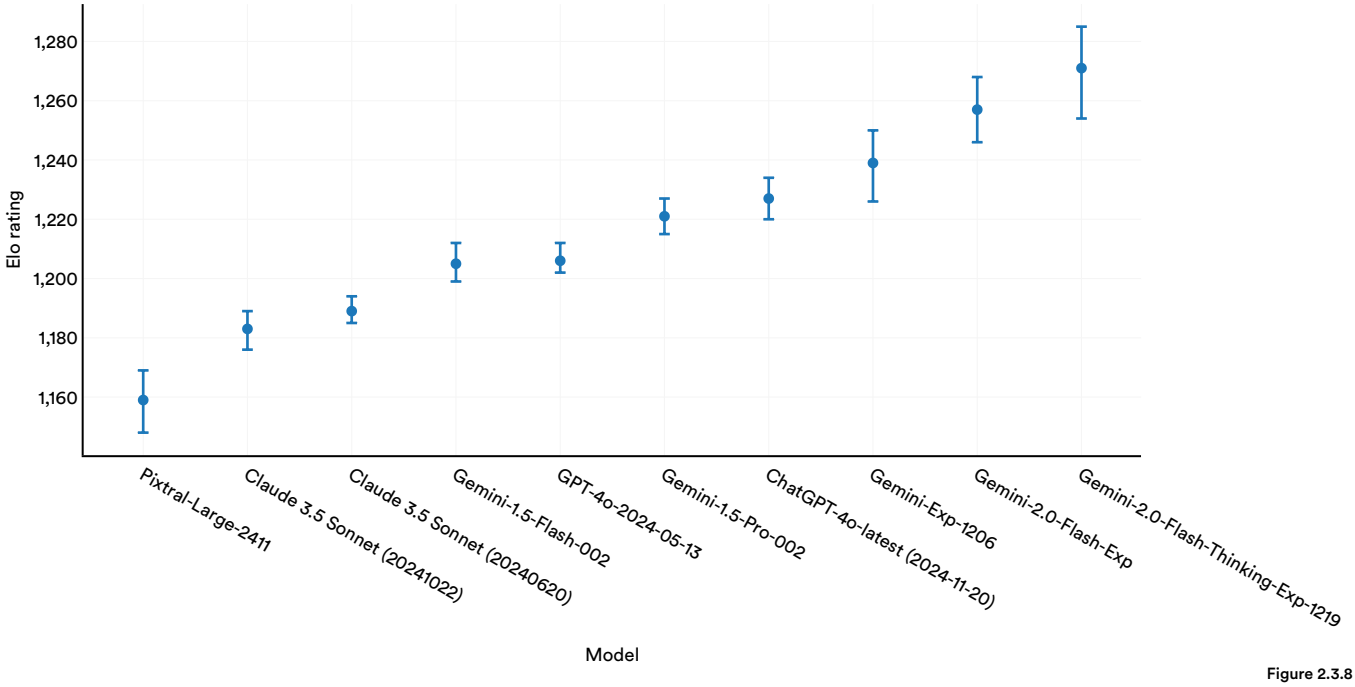


Figure 2.3.8

Highlight:

The Rise of Video Generation

As highlighted in last year's AI Index, recent years have witnessed the rise of video generation models capable of creating videos from text prompts. While earlier models demonstrated some promise, they were plagued by significant limitations, such as producing low-quality videos, omitting sound, or generating only very short clips. However, 2024 marked a significant leap forward in AI video generation, with several major industry players unveiling advanced video generation systems.

In November 2023, Stability AI launched its Stable Video Diffusion model, their first foundation model capable of generating high-quality videos (Figure 2.3.9). The model follows a three-step process: text-to-image pretraining, video pretraining, and high-quality video fine-tuning. Shortly after, in March, Stability AI introduced Stable Video 3D, a model designed to generate multiple 3D views and videos of an object from a single image. In February 2024, OpenAI responded with a preview of Sora, its own video generation model, which moved out of research mode and became publicly accessible in December 2024. Sora can generate 20-second videos at resolutions up to 1080p (Figure 2.3.10). As a diffusion model, it creates a base video and progressively refines it by removing noise over multiple steps to enhance quality.

Still generations from Stable Video Diffusion
Source: Stability AI, 2025

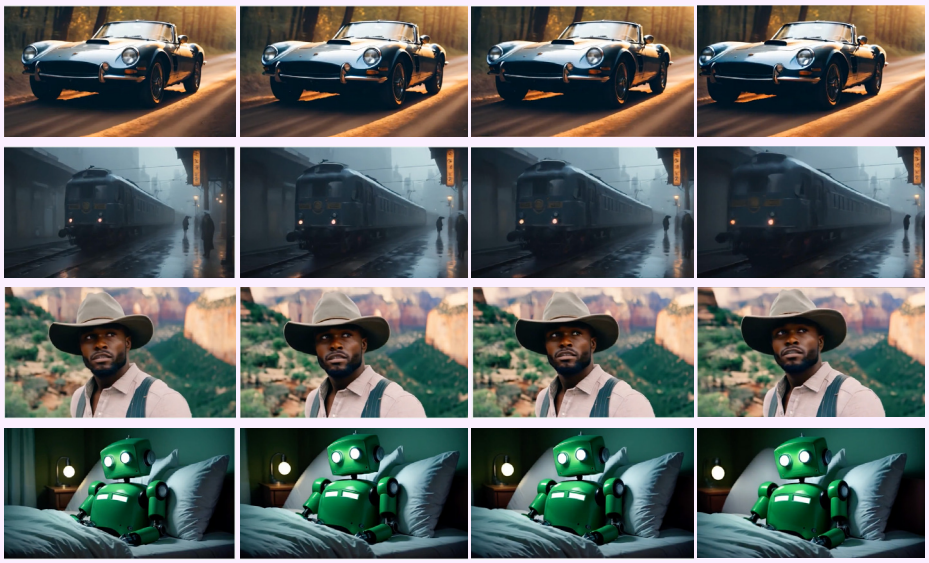


Figure 2.3.9

Still generation from Sora
Source: OpenAI, 2024

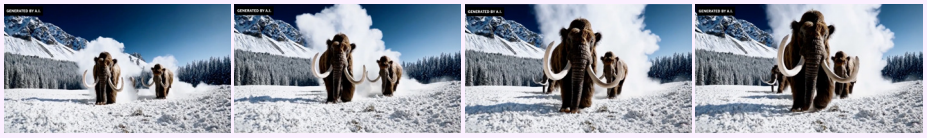


Figure 2.3.10

Highlight:

The Rise of Video Generation (cont'd)

Other major tech players have entered the video generation space. In October 2024, Meta unveiled the latest version of its Movie Gen model. Unlike earlier iterations, the new Movie Gen includes advanced instruction-based video editing features, personalized video generation from images, and the ability to incorporate sound into videos. Meta's most advanced Movie Gen model can create 16-second videos at 16 frames per second, with a resolution of 1080p. Google also made significant strides in 2024, launching two major video generation models: Veo in May and Veo 2 in December. Internal benchmarking by Google revealed that Veo 2 outperformed other leading video generators, such as Meta's Movie Gen, Kling v1.5, and Sora Turbo. In user comparisons, videos generated by Veo 2 were consistently favored over those produced by competing models (Figure 2.3.11).

Veo 2: overall preference

Source: DeepMind, 2024 | Chart: 2025 AI Index report

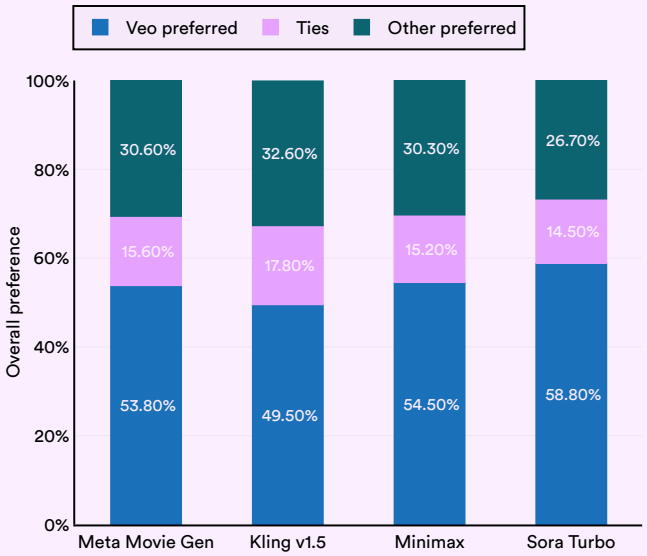


Figure 2.3.11

Smaller players have also made notable contributions to video generation, with models such as Runway's Gen-3 Alpha, Luma's Dream Machine, and Kuaishou's Kling 1.5. The remarkable progress in this field is evident when comparing videos generated in 2023 to those produced in 2024. A popular prompt on the internet, "Will Smith eating spaghetti," demonstrates this advancement, with videos generated in 2025 from one popular video generator Pika showcasing a dramatic improvement in quality compared to their 2023 counterparts (Figure 2.3.12).

Will Smith eating spaghetti, 2023 vs. 2025

Source: Pika, 2025

Figure 2.3.12

AI systems are adept at processing human speech, with audio capabilities that include transcribing spoken words to text and recognizing individual speakers. More recently, AI has advanced in generating synthetic audio content.

2.4 Speech

Speech Recognition

Speech recognition is the ability of AI systems to identify spoken words and convert them into text. Speech recognition has progressed so much that today many computer programs and texting apps are equipped with dictation devices that can reliably transcribe speech into writing.

LSR2: Lip Reading Sentences 2

The Oxford-BBC Lip Reading Sentences 2 (LRS2) dataset, introduced in 2017, is one of the most comprehensive public datasets for lipreading in authentic, in-the-wild scenarios (Figure 2.4.1). The dataset consists of audio-visual clips from a variety of talk shows and news programs. On automatic speech recognition (ASR) tasks, systems' ability to transcribe speech are evaluated on word error rate (WER), with lower scores indicating more precise transcription.

Still images from the BBC lip reading sentences 2 dataset
Source: Chung et al., 2024

Figure 2.4.1

This year, the model Whisper-Flamingo set a new standard on the LRS2 benchmark, achieving a word error rate of 1.3 percent, surpassing the previous state-of-the-art score of 1.5 set in 2023 (Figure 2.4.2). However, given the already low WER, significant further improvements appear unlikely, suggesting that the benchmark may be nearing saturation.

LRS2: word error rate (WER)

Source: Papers With Code, 2025 | Chart: 2025 AI Index report

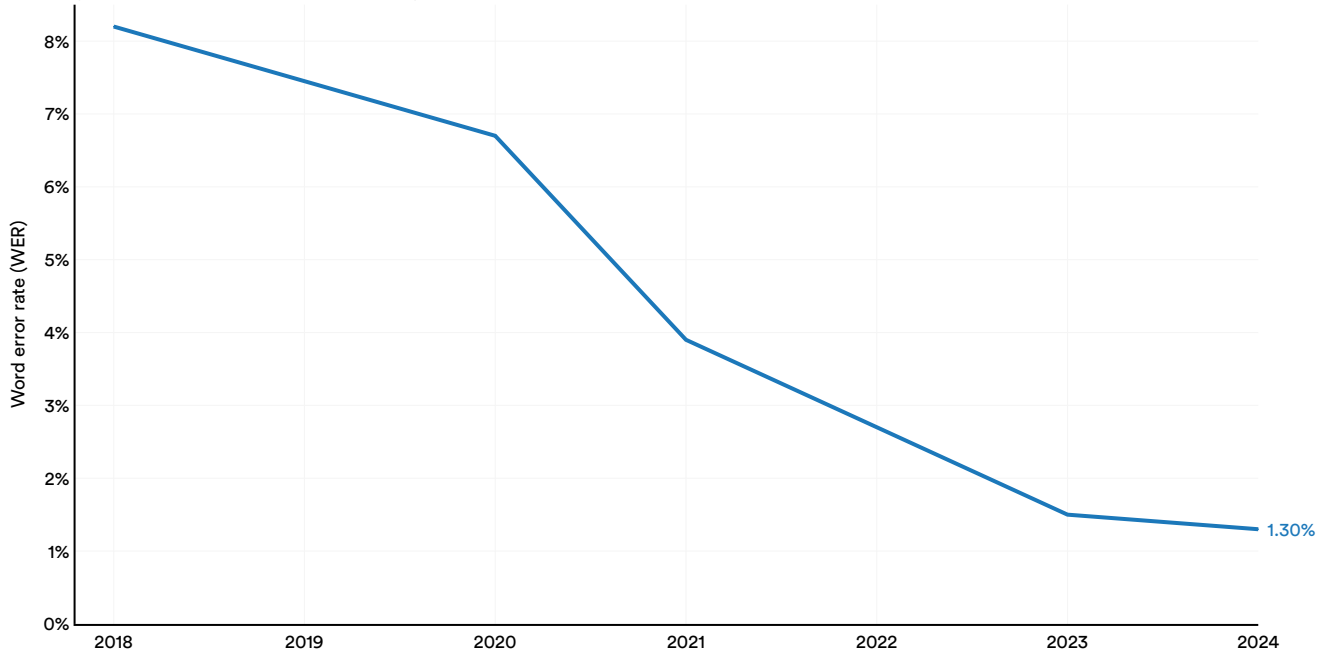


Figure 2.4.2

Coding involves the generation of instructions that computers can follow to perform tasks. Recently, LLMs have become proficient coders, serving as valuable assistants to computer scientists. There is also increasing evidence that many coders find AI coding assistants highly useful. As highlighted in last year's AI Index, LLMs have become increasingly proficient coders, to the extent that many foundational coding benchmarks, such as HumanEval, are slowly becoming saturated. In response, researchers have shifted their focus toward testing LLMs on more complex coding challenges.

2.5 Coding

HumanEval

HumanEval, a benchmark introduced by OpenAI researchers in 2021, evaluates the coding abilities of AI systems through 164 challenging, handwritten programming problems (Figure 2.5.1). The current leader in HumanEval performance is Claude 3.5 Sonnet (HPT), which achieved a score of 100% (Figure 2.5.2).

Sample HumanEval problem
Source: Chen et al., 2023

```
def incr_list(l: list):
    """Return list with elements incremented by 1.
    >>> incr_list([1, 2, 3])
    [2, 3, 4]
    >>> incr_list([5, 3, 5, 2, 3, 3, 9, 0, 123])
    [6, 4, 6, 3, 4, 4, 10, 1, 124]
    """
    return [i + 1 for i in l]
```

Figure 2.5.1

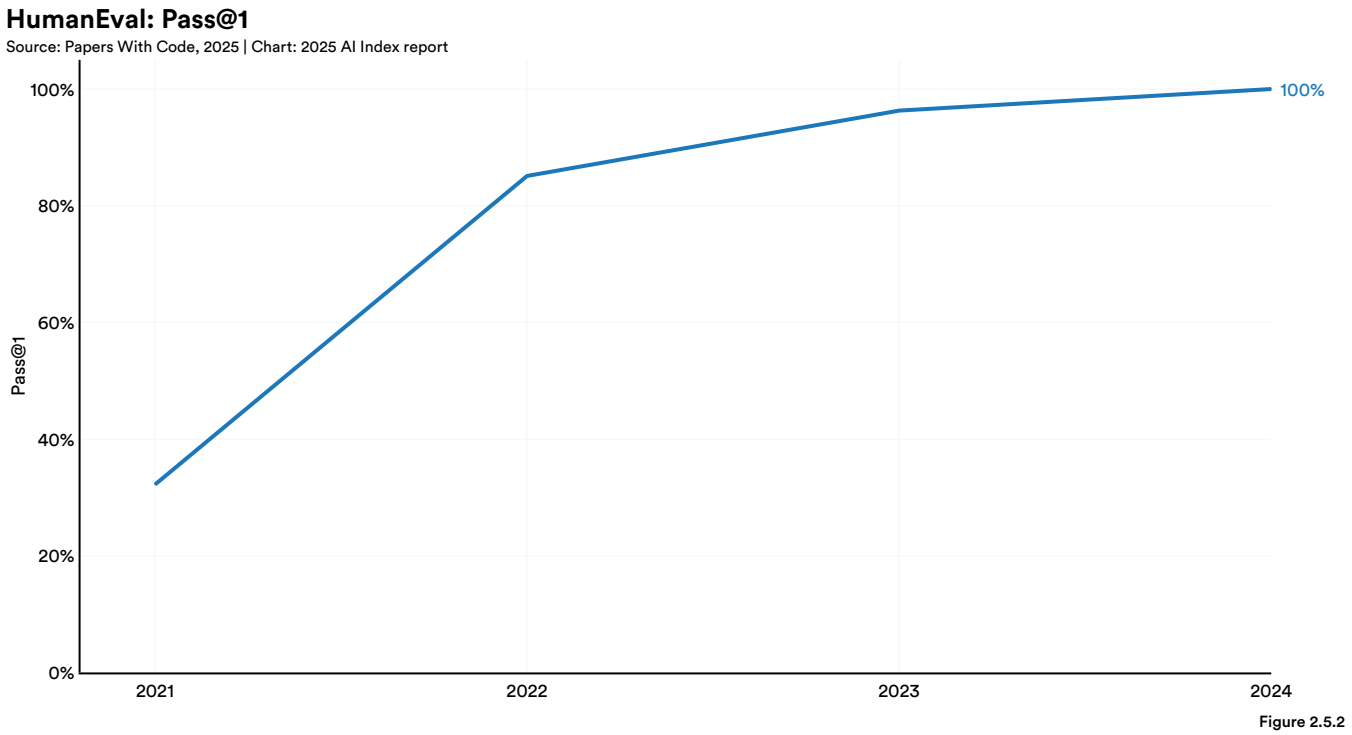


Figure 2.5.2

SWE-bench

In October 2023, researchers from Princeton and the University of Chicago introduced SWE-bench, a dataset comprising 2,294 software engineering problems sourced from real GitHub issues and popular Python repositories (Figure 2.5.3). SWE-bench presents a tougher test for AI coding proficiency, demanding that systems coordinate changes across multiple functions, interact with various execution environments, and perform complex reasoning. SWE-bench features a Lite subset that is curated to make evaluation more accessible and a Verified subset that is filtered by a human annotator. The charts below report on the Verified score.

SWE-bench highlights the rapid improvement of LLMs on tasks that were once considered extremely demanding. At the end of 2023, the best performing model on SWE-bench achieved a score of just 4.4%. By early 2025, the top model, OpenAI's o3 model, is reported to have successfully solved 71.7% of the problems on the Verified benchmark set (Figure 2.5.4). This significant performance increase suggests that AI researchers may soon need to develop more challenging coding benchmarks to effectively test LLMs.

A sample model input from SWE-bench
Source: Jimenez et al., 2023

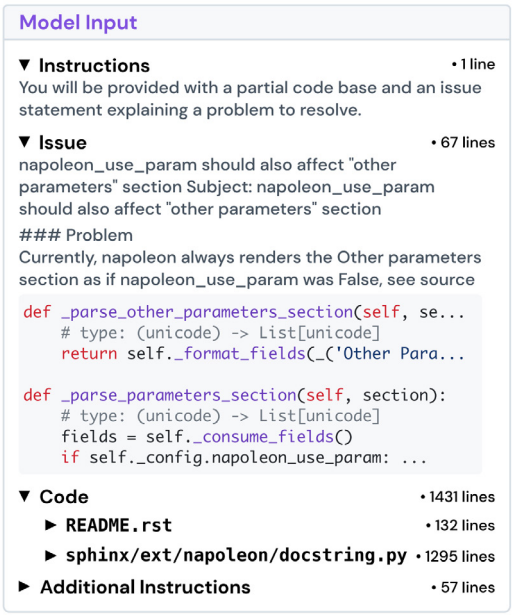


Figure 2.5.3

SWE-bench: percent solved
Source: SWE-bench Leaderboard, 2025; OpenAI, 2024 | Chart: 2025 AI Index report

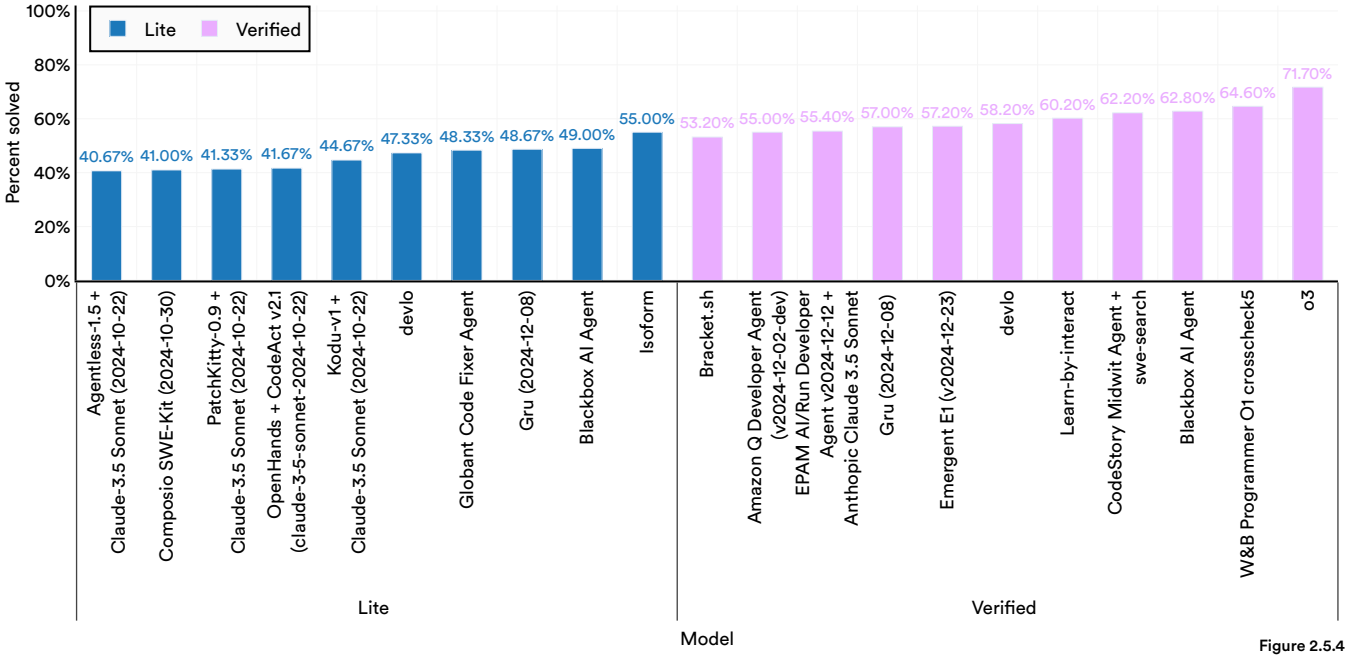


Figure 2.5.4

BigCodeBench

One limitation of existing coding benchmarks is that many are restricted to short, self-contained algorithmic tasks or standalone function calls. However, solving complex and practical tasks often requires the ability to invoke diverse functions, such as tools for data analysis or web development. Effective coding also requires the ability to follow coding instructions expressed in language, a task not tested by many current coding benchmarks.

To address the limitations of existing coding benchmarks, an international team in 2024 released BigCodeBench, a comprehensive, diverse, and challenging benchmark for coding evaluation (Figure 2.5.5). BigCodeBench requires LLMs to invoke multiple function calls across 139 libraries and seven domains, encompassing 1,140 fine-grained tasks. Current AI systems struggle on BigCodeBench. For example, on both the "complete" (code completion based on structured docstrings) and "instruct" (code completion based on natural-language instructions) tasks on the hard subset of the benchmark, the current best model, OpenAI's o1, achieves an average score of just 35.5 (Figure 2.5.6). Models perform slightly better on the full set of the benchmark (Figure 2.5.7). BigCodeBench highlights the gap that persists for AI systems to achieve human-level coding proficiency.

Programming tasks in BigCodeBench
Source: Zhuo et al., 2024

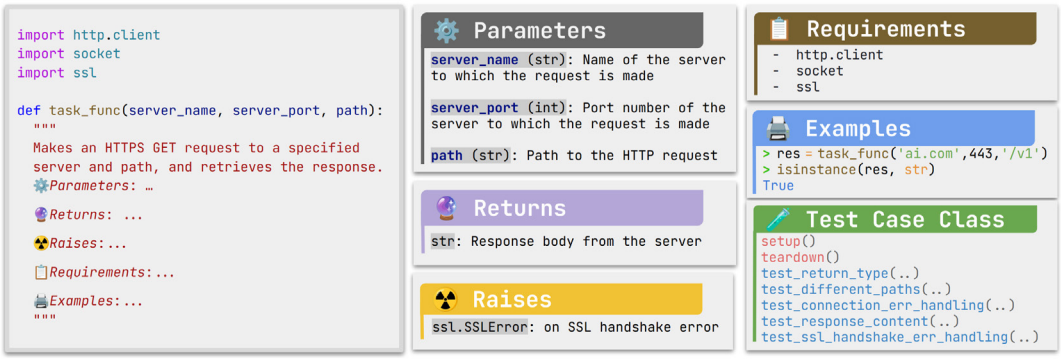


Figure 2.5.5

BigCodeBench on the hard set: Pass@1 (average)
Source: Hugging Face, 2025 | Chart: 2025 AI Index report

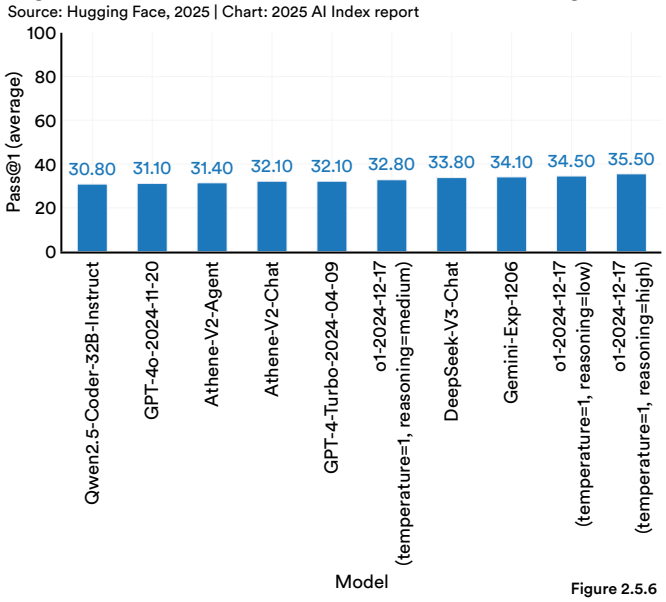


Figure 2.5.6

BigCodeBench on the full set: Pass@1 (average)
Source: Hugging Face, 2025 | Chart: 2025 AI Index report

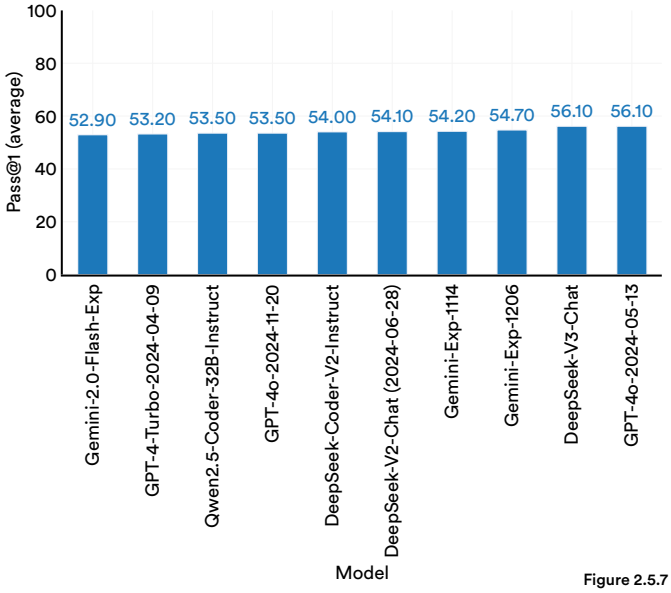


Figure 2.5.7

Chatbot Arena: Coding

The Chatbot Arena LLM leaderboard now features a coding filter, offering valuable insights into how coders and the broader community perceive the coding capabilities of different models. This public feedback adds a new dimension to evaluating model performance. Currently, the top-rated LLM for coding is Gemini-Exp-1206, with an arena score of 1,369, closely followed by OpenAI's latest o1 model at 1,361. Among Chinese models, DeepSeek-V3 leads with a score of 1,317, trailing the highest-ranking model by 3.8% (Figure 2.5.8).

LMSYS Chatbot Arena for LLMs: Elo rating (coding)
Source: LMSYS, 2025 | Chart: 2025 AI Index report

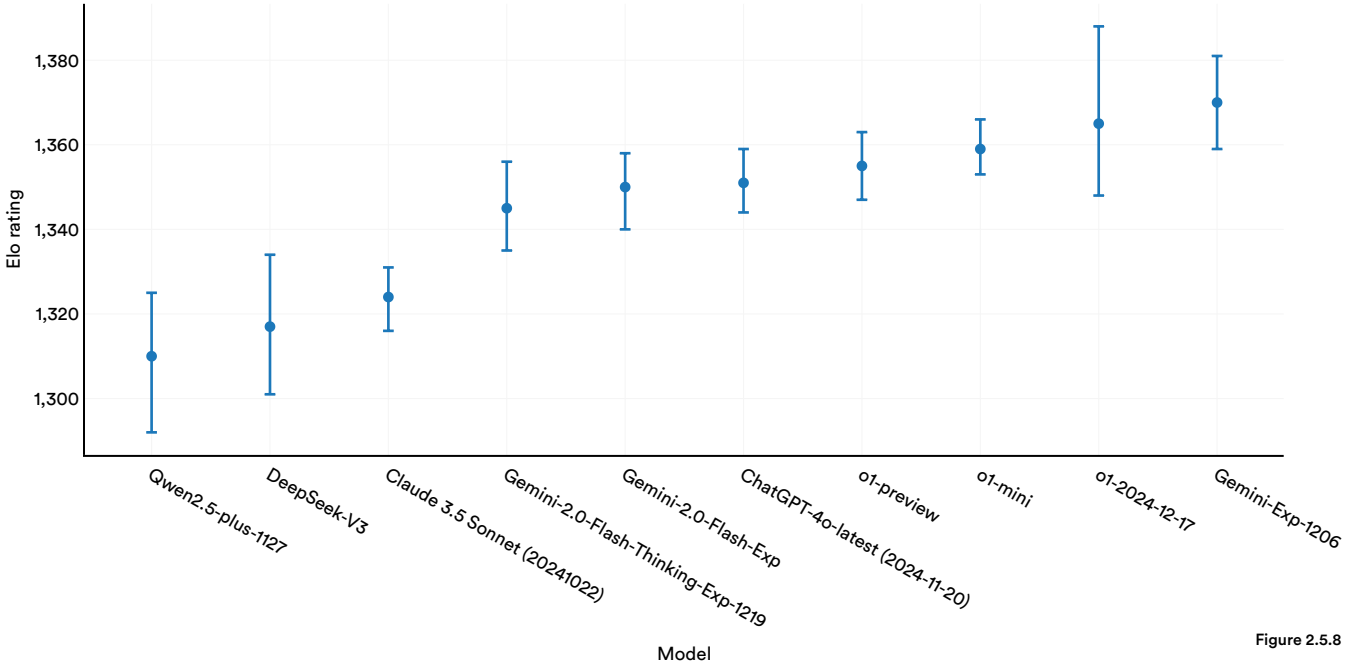


Figure 2.5.8

Mathematical problem-solving benchmarks evaluate AI systems' ability to reason mathematically. AI models can be tested with a range of math problems, from grade-school level to competition-standard mathematics.

2.6 Mathematics

GSM8K

GSM8K, introduced by OpenAI in 2021, is a dataset containing approximately 8,000 diverse grade-school math word problems that challenges AI models to generate multistep solutions using arithmetic operations (Figure 2.6.1). Alongside MMLU, GSM8K has become a widely used benchmark for evaluating advanced LLMs. However, recent concerns have emerged regarding potential contamination and saturation of the benchmark.

The top-performing model on GSM8K is a variant of Claude Sonnet 3.5, which was optimized using the HPT prompting strategy and achieved a 97.72% score (Figure 2.6.2). This marks a significant improvement over the previous high of 91.00% in 2023. However, in 2024, several models from Mistral, Meta, and Qwen scored around 96%, indicating that the GSM8K benchmark may be approaching saturation.

Sample problems from GSM8K

Source: Cobbe et al., 2023

Problem: Beth bakes 4, 2 dozen batches of cookies in a week. If these cookies are shared amongst 16 people equally, how many cookies does each person consume?
Solution: Beth bakes 4 2 dozen batches of cookies for a total of 4*2 = <<4*2=8>>8 dozen cookies
There are 12 cookies in a dozen and she makes 8 dozen cookies for a total of 12*8 = <<12*8=96>>96 cookies
She splits the 96 cookies equally amongst 16 people so they each eat 96/16 = <<96/16=6>>6 cookies
Final Answer: 6

Problem: Mrs. Lim milks her cows twice a day. Yesterday morning, she got 68 gallons of milk and in the evening, she got 82 gallons. This morning, she got 18 gallons fewer than she had yesterday morning. After selling some gallons of milk in the afternoon, Mrs. Lim has only 24 gallons left. How much was her revenue for the milk if each gallon costs $3.50?
Mrs. Lim got 68 gallons - 18 gallons = <<68-18=50>>50 gallons this morning.
So she was able to get a total of 68 gallons + 82 gallons + 50 gallons = <<68+82+50=200>>200 gallons.
She was able to sell 200 gallons - 24 gallons = <<200-24=176>>176 gallons.
Thus, her total revenue for the milk is $3.50/gallon x 176 gallons = $<<3.50*176=616>>616.
Final Answer: 616

Problem: Tina buys 3 12-packs of soda for a party. Including Tina, 6 people are at the party. Half of the people at the party have 3 sodas each, 2 of the people have 4, and 1 person has 5. How many sodas are left over when the party is over?
Solution: Tina buys 3 12-packs of soda, for 3*12= <<3*12=36>>36 sodas
6 people attend the party, so half of them is 6/2= <<6/2=3>>3 people
Each of those people drinks 3 sodas, so they drink 3*3=<<3*3=9>>9 sodas
Two people drink 4 sodas, which means they drink 2*4=<<4*2=8>>8 sodas
With one person drinking 5, that brings the total drank to 5+9+8+3= <<5+9+8+3=25>>25 sodas
As Tina started off with 36 sodas, that means there are 36-25=<<36-25=11>>11 sodas left
Final Answer: 11

Figure 2.6.1

GSM8K: accuracy

Source: Papers With Code, 2024 | Chart: 2025 AI Index report

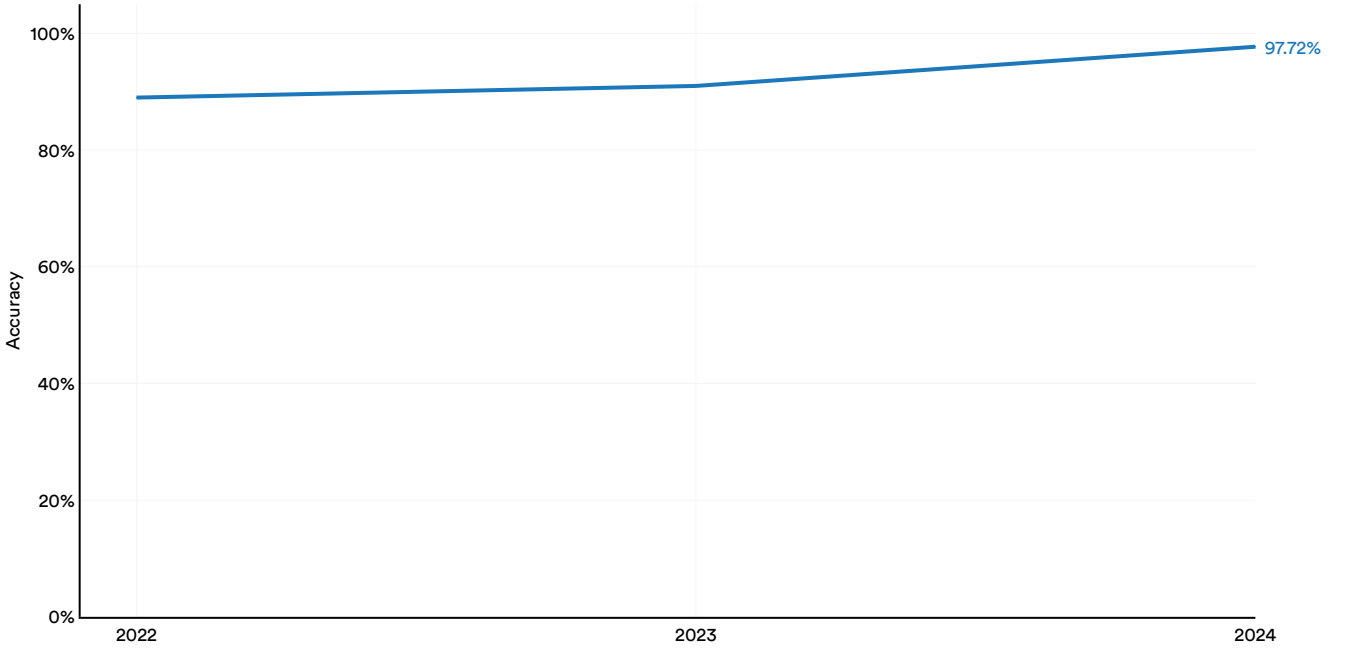


Figure 2.6.2

MATH

MATH is a dataset of 12,500 challenging, competition-level mathematics problems introduced by UC Berkeley and University of Chicago researchers in 2021 (Figure 2.6.3). AI systems struggled on MATH when it was first released, managing to solve only 6.9% of the problems. Performance has significantly improved. In January 2025, OpenAI's o3-mini (high) model was released and achieved the best performance on the MATH dataset, solving 97.9% of the problems (Figure 2.6.4). As highlighted in last year's AI Index, MATH was one of the few datasets where AI systems had not yet outperformed the human baseline. This fact no longer remains true.

Sample problem from MATH dataset

Source: Hendrycks et al., 2023

MATH Dataset (Ours)

Problem: Tom has a red marble, a green marble, a blue marble, and three identical yellow marbles. How many different groups of two marbles can Tom choose?

Solution: There are two cases here: either Tom chooses two yellow marbles (1 result), or he chooses two marbles of different colors ($\binom{4}{2} = 6$ results). The total number of distinct pairs of marbles Tom can choose is $1 + 6 = \boxed{7}$.

Problem: The equation $x^2 + 2x = i$ has two complex solutions. Determine the product of their real parts.

Solution: Complete the square by adding 1 to each side. Then $(x+1)^2 = 1 + i = e^{\frac{i\pi}{4}}\sqrt{2}$, so $x + 1 = \pm e^{\frac{i\pi}{8}}\sqrt[4]{2}$. The desired product is then $\left(-1 + \cos\left(\frac{\pi}{8}\right)\sqrt[4]{2}\right)\left(-1 - \cos\left(\frac{\pi}{8}\right)\sqrt[4]{2}\right) = 1 - \cos^2\left(\frac{\pi}{8}\right)\sqrt{2} = 1 - \frac{\left(1+\cos\left(\frac{\pi}{4}\right)\right)}{2}\sqrt{2} = \boxed{\frac{1-\sqrt{2}}{2}}$.

Figure 2.6.3

MATH word problem-solving: accuracy

Source: Papers With Code, 2024; OpenAI, 2025 | Chart: 2025 AI Index report

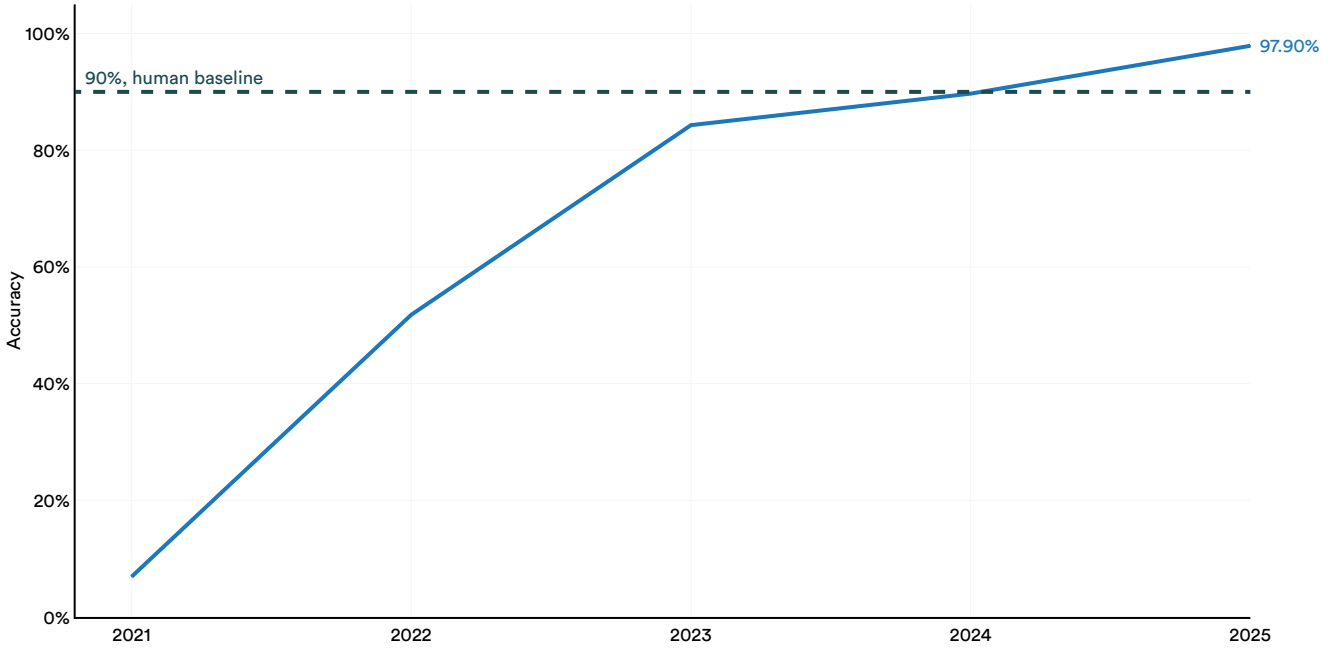


Figure 2.6.4

Chatbot Arena: Math

The Chatbot Arena includes a math filter, allowing the public to rank models based on their performance in generating math-related answers. The Math Arena evaluates over 181 models and has collected more than 340,000 public votes. Unlike the general and coding arenas, where Gemini-based models lead, the top-ranked model in the Math Arena is OpenAI's o1 variant, released in December 2024 (Figure 2.6.5).

LMSYS Chatbot Arena for LLMs: Elo rating (Math)
Source: LMSYS, 2025 | Chart: 2025 AI Index report

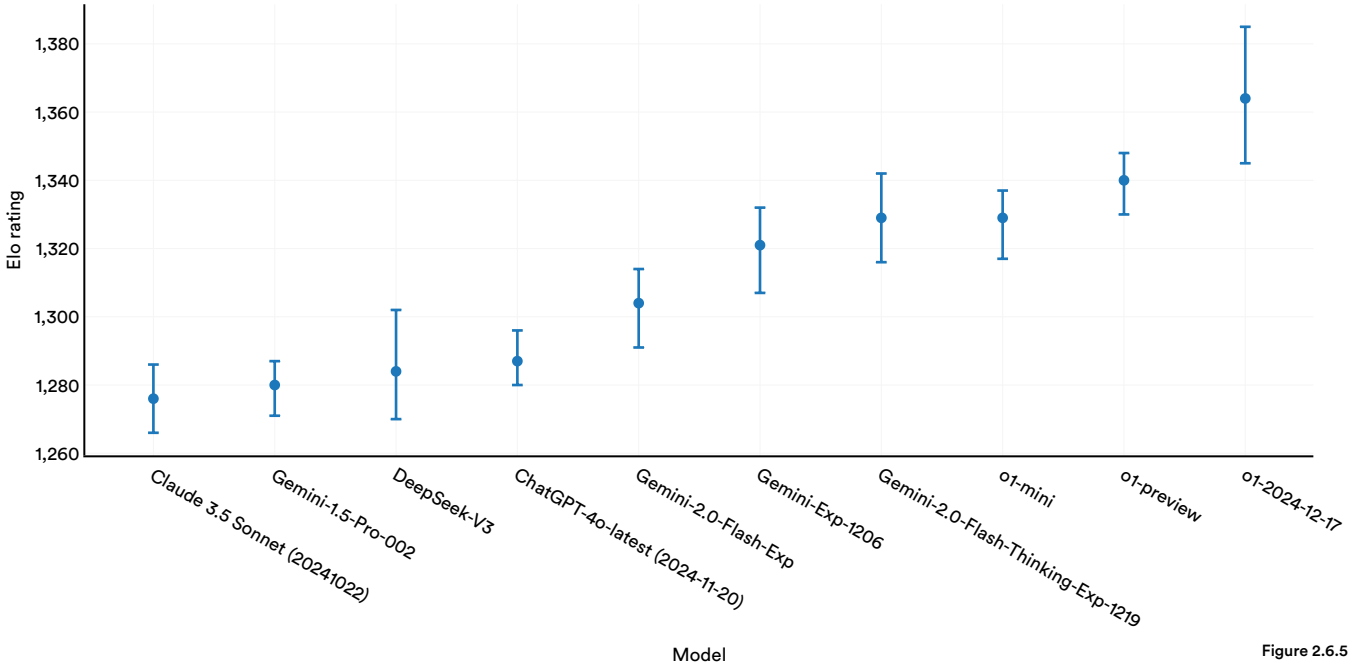


Figure 2.6.5

FrontierMath

Members of the math community have highlighted limitations in the current suite of math benchmarks, calling for the development of new benchmarks to evaluate increasingly advanced AI systems. One significant challenge is saturation: AI systems are approaching near-perfect performance on benchmarks like GSM8K and MATH, which primarily assess high school and college-level mathematics. To push the boundaries further, researchers have voiced a need for benchmarks that test truly advanced mathematics, including problems in number theory, real analysis, algebraic geometry, and category theory.

FrontierMath is a new benchmark introduced by Epoch AI that features hundreds of original, exceptionally challenging mathematical problems. These problems, vetted by expert mathematicians, often require hours, days, or even collaborative research efforts to solve. Figure 2.6.6 illustrates sample problems included on the benchmark. Epoch AI evaluated six leading LLMs on the FrontierMath benchmark: o1-preview, o1-mini, GPT-4o, Claude 3.5 Sonnet, Grok 2 Beta, and Gemini 1.5 Pro 002. At the time the benchmark was released, the best-performing model, Gemini 1.5 Pro, managed to solve just 2.0% of the problems—a significantly lower success rate than it achieved on other math benchmarks (Figure 2.6.7). However, OpenAI's o3 model is reported to have scored 25.2% on the benchmark. The creators of FrontierMath hope the benchmark will remain a rigorous challenge for cutting-edge AI systems for years to come.

Sample problems from FrontierMath

Source: Glazer et al., 2024

Sample problem 1: Testing Artin's primitive root conjecture

Definitions. For a positive integer n, let $v_p(n)$ denote the largest integer v such that $p^v \mid n$.
For p a prime and $a \not\equiv 0 \pmod{p}$, we let $\mathrm{ord}_p(a)$ denote the smallest positive integer o such that $a^o \equiv 1 \pmod{p}$. For $x > 0$, we let

$$\mathrm{ord}_{p,x}(a) = \prod_{\substack{q \le x \\ q \text{ prime}}} q^{v_q(\mathrm{ord}_p(a))} \prod_{\substack{q > x \\ q \text{ prime}}} q^{v_q(p-1)}.$$

Problem. Let S_x denote the set of primes p for which

$$\mathrm{ord}_{p,x}(2) > \mathrm{ord}_{p,x}(3),$$

and let d_x denote the density

$$d_x = \frac{|S_x|}{|\{p \le x : p \text{ is prime}\}|}$$

of S_x in the primes. Let

$$d_\infty = \lim_{x \to \infty} d_x.$$

Compute $\lfloor 10^6 d_\infty \rfloor$.

Answer: 367707

MSC classification: 11 Number theory

Sample problem 2: Find the degree 19 polynomial

Construct a degree 19 polynomial $p(x) \in \mathbb{C}[x]$ such that $X := \{p(x) = p(y)\} \subset \mathbb{P}^1 \times \mathbb{P}^1$ has at least 3 (but not all linear) irreducible components over $\mathbb{C}$. Choose $p(x)$ to be odd, monic, have real coefficients and linear coefficient -19 and calculate $p(19)$.

Answer: 18765720719740948033911 79

MSC classification: 14 Algebraic geometry; 20 Group theory and generalizations; 11 Number theory generalizations

Sample problem 3: Prime field continuous extensions

Let a_n for $n \in \mathbb{Z}$ be the sequence of integers satisfying the recurrence formula[1]

$$a_n = (1.981 \times 10^{11})a_{n-1} + (3.549 \times 10^{11})a_{n-2} - (4.277 \times 10^{11})a_{n-3} + (3.706 \times 10^{8})a_{n-4}$$

with initial conditions $a_i = i$ for $0 \le i \le 3$. Find the smallest prime $p \equiv 4 \bmod 7$ for which the function $\mathbb{Z} \to \mathbb{Z}$ given by $n \mapsto a_n$ can be extended to a continuous function on $\mathbb{Z}_p$.

Answer: 9811

MSC classification: 11 Number theory

Figure 2.6.6

FrontierMath: percent solved

Source: Glazer et al., 2024; OpenAI, 2025 | Chart: 2025 AI Index report

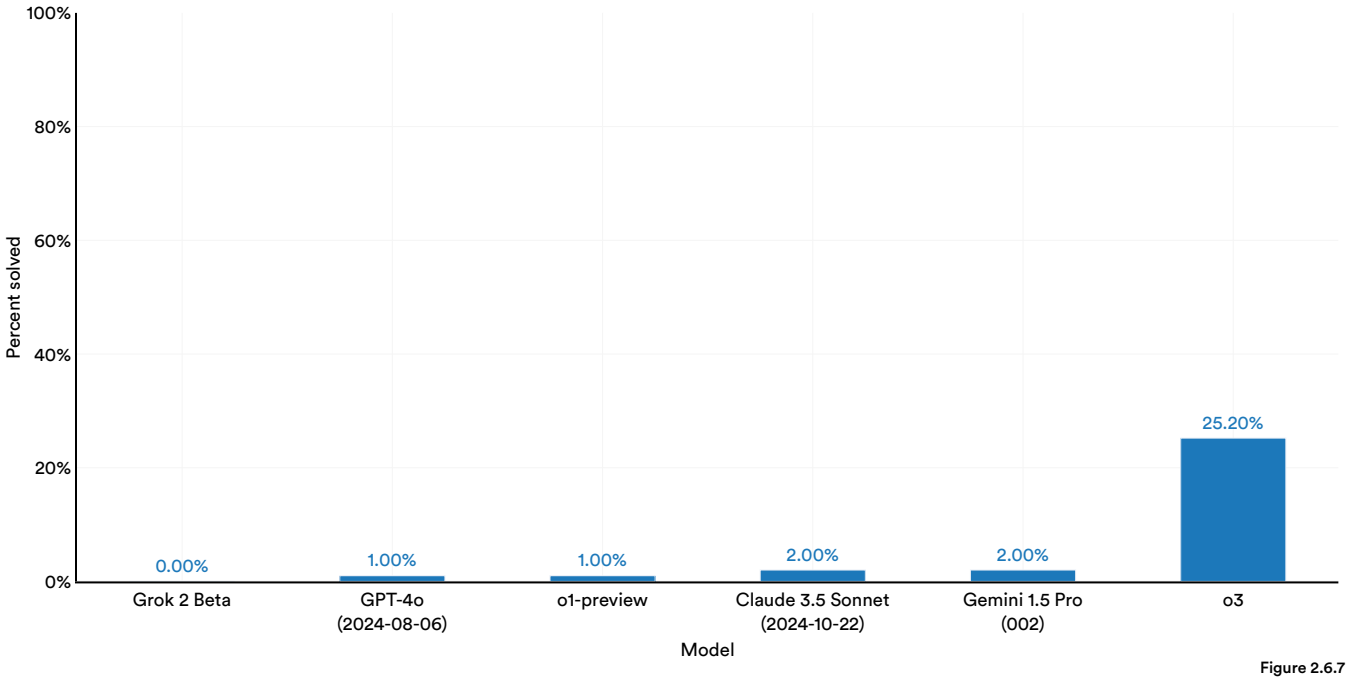


Figure 2.6.7

Highlight:

Learning and Theorem Proving

DeepMind employed its systems, AlphaProof and AlphaGeometry 2, to solve four out of six problems in the 2024 International Mathematical Olympiad (IMO), achieving a performance level equivalent to that of a silver medalist. AlphaGeometry solved 25 out of 30 Olympiad geometry problems in the benchmarking set, surpassing the average score of an IMO silver medalist, who typically solves 22.9 (Figure 2.6.8). The IMO, established in 1959, is the world's oldest and most prestigious competition for young mathematicians.

AlphaProof is a reinforcement learning system derived from AlphaZero, which was previously applied to chess, shogi, and Go. It trains itself to solve problems by generating hypotheses that are then verified using the Lean interactive proof system. A fine-tuned Gemini model is utilized to translate natural language problem statements into formal representations, building a comprehensive training library. In this year's competition, AlphaProof successfully solved two algebra problems and one number theory problem, but failed to solve two combinatorics problems.

AlphaGeometry 2 is a neuro-symbolic hybrid system featuring a language model based on Gemini and trained on extensive synthetic data. Prior to 2024, AlphaGeometry could solve 83% of historical IMO geometry problems. During the 2024 competition, it solved the sole geometry problem in just 24 seconds. For the 2024 test, competition problems were manually translated into Lean's formal representation.

It remains unknown how AlphaProof and AlphaGeometry would perform on traditional theorem-proving benchmarks such as TPTP, which has been used since 1997 to assess the performance of automatic theorem-proving (ATP) systems, particularly those applied to software verification. The AI Index reported on the state of ATP in its 2021 report.

Number of solved geometry problems in IMO-AG-30
Source: Trinh et al., 2024 | Chart: 2025 AI Index report

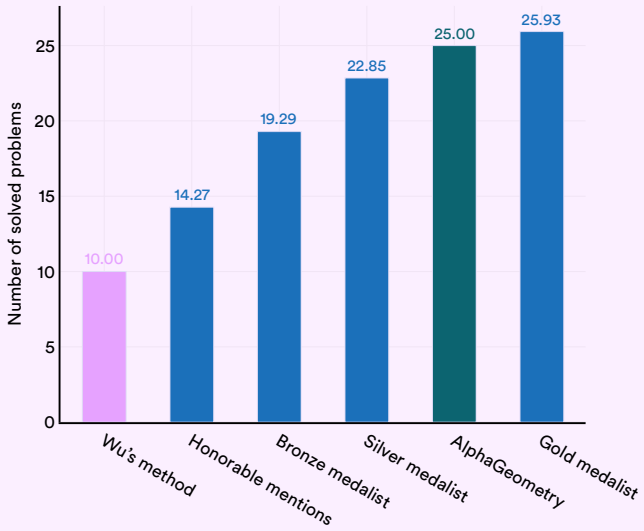


Figure 2.6.8

A 2024 update of that report, based on the latest version of TPTP containing over 25,000 problems, indicates that fully automatic systems can now solve 89% of the problems in TPTP v.9.0.0.

Ideally, TPTP systems could be tested on IMO problems, and AlphaProof and AlphaGeometry on TPTP problems—some of which have never been solved by humans, let alone by ATP systems. Unfortunately, neither of these tests has been conducted. The primary reason is that the logics supported by the different systems differ significantly, and translators between them do not yet exist. Additionally, while substantial, the TPTP library is not large enough to serve as a training set for AlphaProof without generating a considerable number of synthetic examples.

Reasoning in AI involves the ability of AI systems to draw logically valid conclusions from different forms of information. AI systems are increasingly being tested in diverse reasoning contexts, including visual (reasoning about images), moral (understanding moral dilemmas), and social reasoning (navigating social situations).

2.7 Reasoning

General Reasoning

General reasoning pertains to AI systems being able to reason across broad, rather than specific, domains. As part of a general reasoning challenge, for example, an AI system might be asked to reason across multiple subjects rather than perform one narrow task (e.g., playing chess).

Sample MMMU questions
Source: Yue et al., 2023

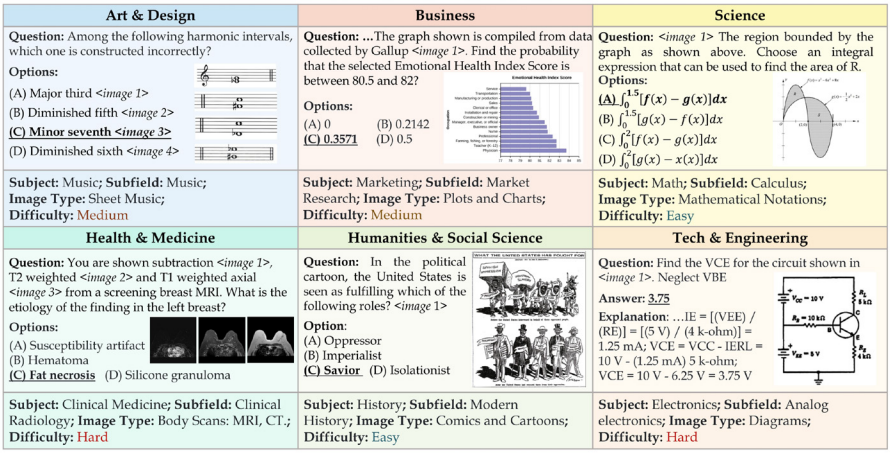


Figure 2.7.1

MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI

In recent years, the reasoning abilities of AI systems have advanced so much that older benchmarks like SQuAD (for textual reasoning) and VQA (for visual reasoning) have become saturated, indicating a need for more challenging reasoning tests.

Responding to this, researchers from the United States and Canada recently developed MMMU, the massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI (artificial general intelligence). MMMU comprises about 11,500 college-level questions from six core disciplines: art and design, business, science, health and medicine, humanities and social science, and technology and engineering (Figure 2.7.1). The question formats include charts, maps, tables, chemical structures, and more. MMMU is among the most demanding tests of perception, knowledge, and reasoning in AI to date. As of January 2025, the highest-performing model is OpenAI's o1, achieving a score of 78.2%—a significant improvement from the state-of-the-art score of 59.4% reported in last year's AI Index (Figure 2.7.2). While this top score remains below the medium and high human expert baselines, as with other benchmarks covered in the Index, AI systems are rapidly closing the gap.

MMMU on validation set: overall accuracy
Source: MMMU Leaderboard, 2024 | Chart: 2025 AI Index report

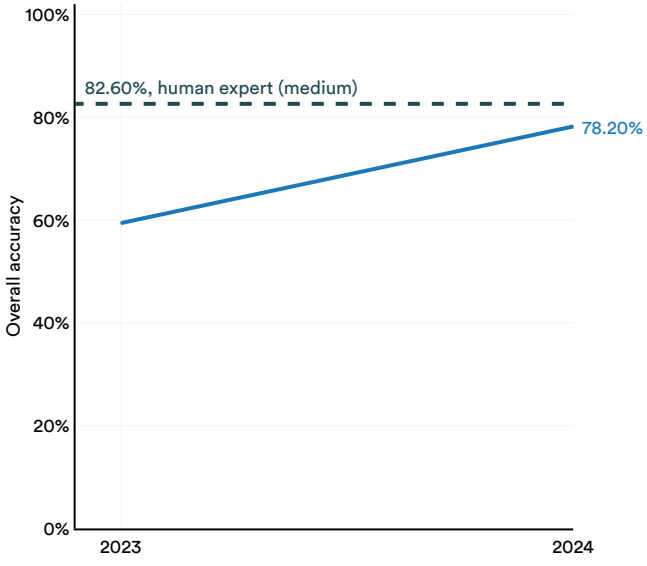


Figure 2.7.2

GPQA: A Graduate-Level Google-Proof Q&A Benchmark

In 2023, researchers from NYU, Anthropic, and Meta introduced the GPQA benchmark to test general, multisubject AI reasoning. This dataset consists of 448 difficult multiple-choice questions that cannot be easily answered by web search. The questions were crafted by subject-matter experts in various fields like biology, physics, and chemistry (Figure 2.7.3). On the diamond set—the most challenging subset of the dataset and the one most frequently tested by AI developers—human experts achieved an accuracy rate of 81.3%.

Last year's AI Index reported that the best-performing AI model, GPT-4, achieved only 38.8% on the diamond test set. In just a year, top AI systems have made significant strides, with OpenAI's o3 model, launched in December 2024, posting a state-of-the-art score of 87.7%, a 48.9 percentage point improvement from the state-of-the-art score in 2023 (Figure 2.7.4). In fact, o3's score was the first to exceed the baseline set by expert human validators. AI systems are rapidly advancing on challenging new benchmarks like MMMU and GPQA, which were recently introduced to push the limits of AI capabilities.

Sample chemistry question from GPQA

Source: Rein et al., 2023

Chemistry (general)

A reaction of a liquid organic compound, which molecules consist of carbon and hydrogen atoms, is performed at 80 centigrade and 20 bar for 24 hours. In the proton nuclear magnetic resonance spectrum, the signals with the highest chemical shift of the reactant are replaced by a signal of the product that is observed about three to four units downfield. Compounds from which position in the periodic system of the elements, which are also used in the corresponding large-scale industrial process, have been mostly likely initially added in small amounts?
A) A metal compound from the fifth period.
B) A metal compound from the fifth period and a non-metal compound from the third period.
C) A metal compound from the fourth period.
D) A metal compound from the fourth period and a non-metal compound from the second period.

Figure 2.7.3

GPQA on the diamond set: accuracy

Source: AI Index, 2025 | Chart: 2025 AI Index report

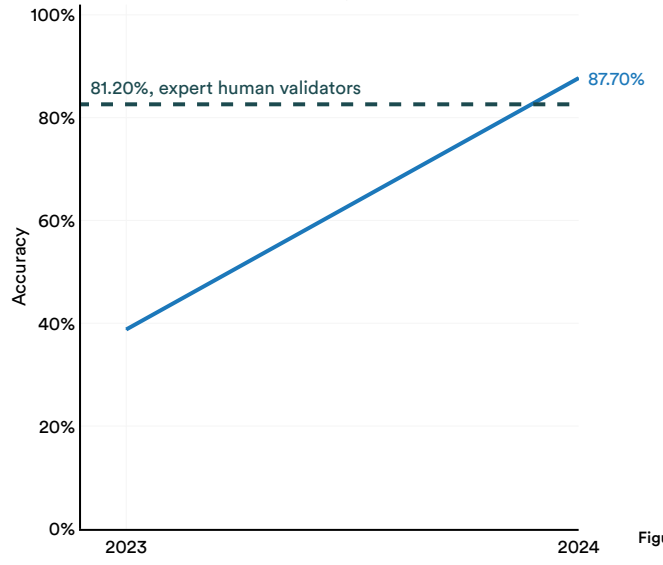


Figure 2.7.4

ARC-AGI

As AI systems continue to advance, claims about the imminent arrival of artificial general intelligence (AGI) have become more frequent. There is no universally accepted definition of AGI. Some computer scientists define it as AI systems that match or surpass human cognitive abilities across a broad range of tasks. Others emphasize that the definition should encompass the capacity for general learning and skill acquisition, describing AGI as a system "capable of efficiently acquiring new skills and solving novel problems for which it was neither designed nor trained."

ARC-AGI is a benchmark introduced in 2019 by François Chollet, the creator of Keras, a popular open-source deep learning library. ARC-AGI tests the ability of systems to generalize beyond prior training. More specifically, the ARC-AGI benchmark presents AI systems with a set of independent tasks. Each task includes demonstration or input pairs followed by one or more test or output scenarios (Figure 2.7.5). This benchmark emphasizes generalized learning ability: It is impossible for systems to prepare in advance, as each task introduces a unique logic. The tasks require no specialized world knowledge or language skills but instead draw on assumed prior knowledge, such as the concept of objects, basic topology, and elementary arithmetic—concepts typically mastered by children at an early age.

Sample ARC-AGI task
Source: Chollet et al., 2025

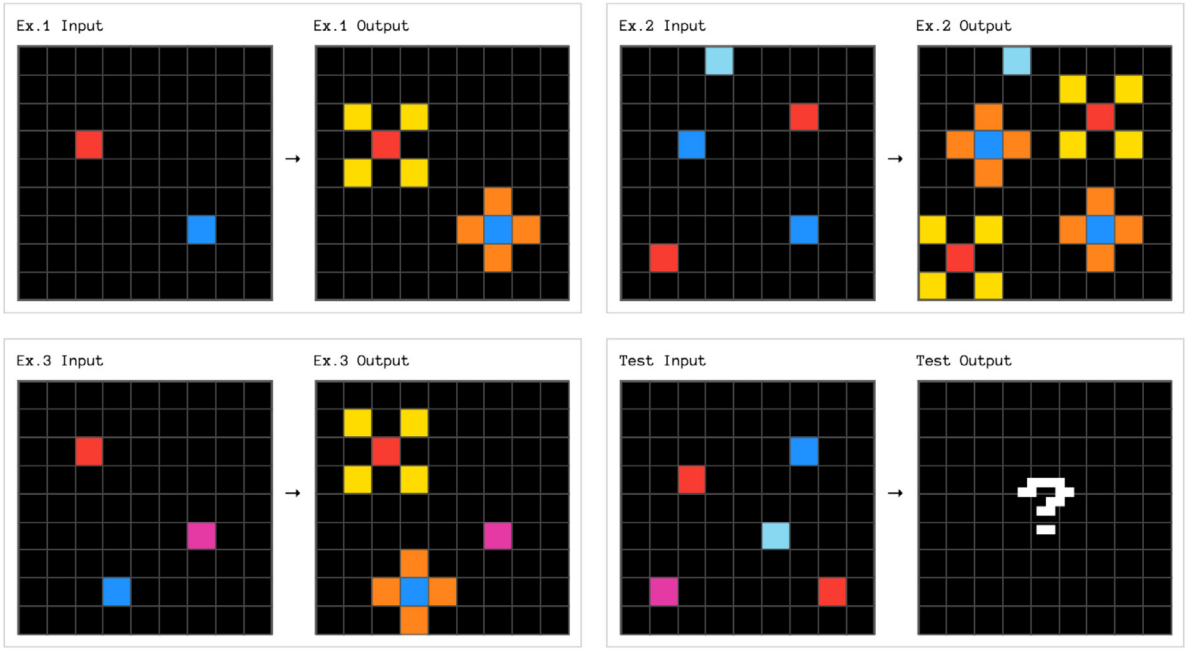


Figure 2.7.5

ARC-AGI has proven to be an exceptionally challenging benchmark. When it was first run in 2020, the top-performing system achieved a score of only 20% (Figure 2.7.6). Four years later, this score had risen to just 33%. However, this year has seen substantial progress, with OpenAI's o3 model achieving a score of 75.7%. In settings where o3 was allocated a high-compute budget exceeding the benchmark's $10,000 limit, it achieved a score of 87.5%.

Researchers attribute the overall slow progress in previous years to an overemphasis on scaling AI models—making them larger and feeding them increasing amounts of training data. While this approach improved task-specific skills, it did little to enhance the ability of AI systems to tackle problems without prior exposure or training data. This year's improvements suggest a shift in focus toward more meaningful advancements in generalization and search capabilities.

ARC-AGI-1 on private evaluation set: high score

Source: Chollet et al., 2025; OpenAI, 2025 | Chart: 2025 AI Index report

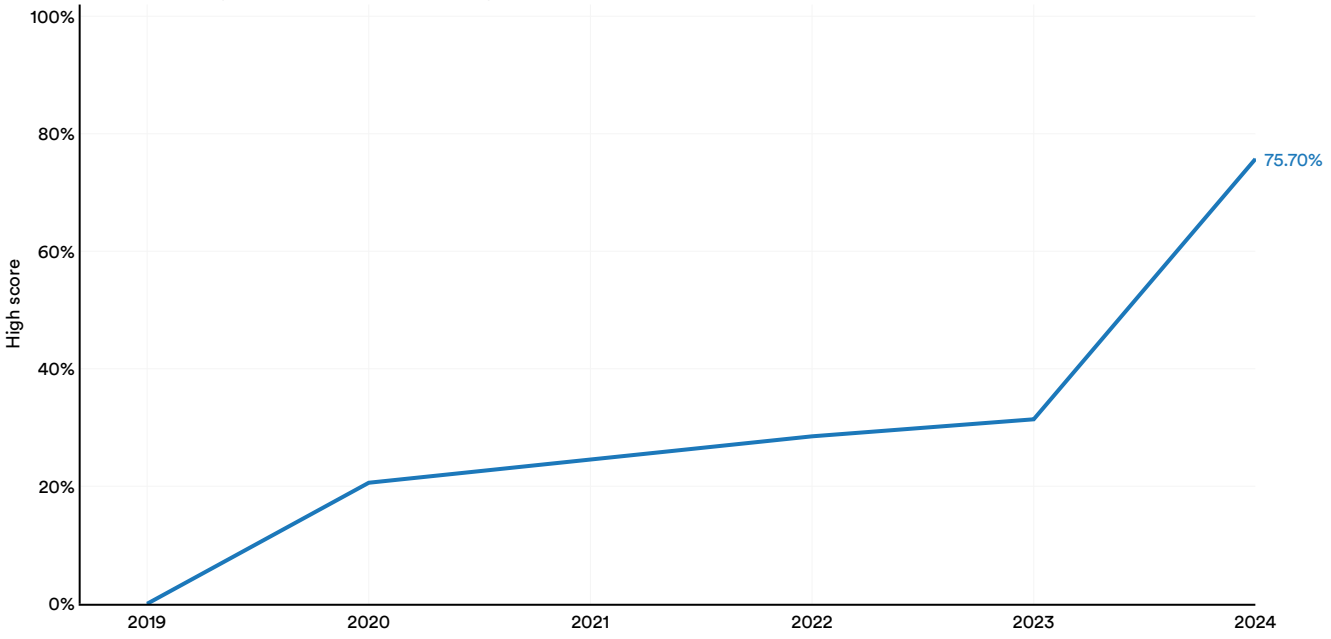


Figure 2.7.6

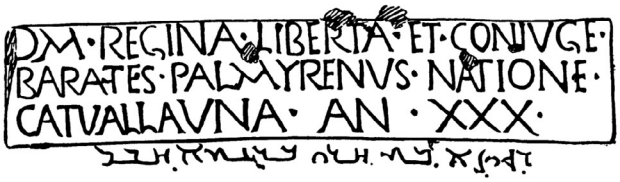
Humanity's Last Exam

As highlighted in both this and last year's AI Index, many popular AI benchmarks, such as MMLU, GSM8K, and HumanEval, have reached saturation. In response, researchers have developed more challenging benchmarks to better assess AI capabilities. Recently, members of the team behind MMLU introduced Humanity's Last Exam (HLE), a new benchmark comprising 2,700 highly challenging questions across dozens of subject areas (Figure 2.7.7). The dataset features multimodal questions, contributed by subject matter experts, including leading professors and graduate-level reviewers, that resist simple internet lookups or database retrieval. Additionally, each question was tested against state-of-the-art LLMs before inclusion; if an existing model could answer it, the question was rejected.

Same questions on HLE
Source: Phan et al., 2025

Classics

Question:



Here is a representation of a Roman inscription, orginally found on a tombstone. Provide a translation for the Palmyrene script.
A transliteration of the text is provided: RGYNᵓ BT HRY BR ᶜTᵓ HBL

Henry T
Merton College, Oxford

Ecology

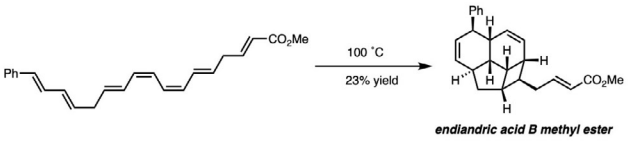
Question:

Hummingbirds within Apodiformes uniquely have a bilaterally paired oval bone, a sesamoid embedded in the caudolateral portion of the expanded, cruciate aponeurosis of insertion of m. depressor caudae. How many paired tendons are supported by this sesamoid bone? Answer with a number.

Edward V
Massachusetts Institute of Technology

Chemistry

Question:



endiandric acid B methyl ester

The reaction shown is a thermal pericyclic cascade that converts the starting heptaene into endiandric acid B methyl ester. The cascade involves three steps: two electrocyclizations followed by a cycloaddition. What types of electrocyclizations are involved in step 1 and step 2, and what type of cycloaddition is involved in step 3?

Provide your answer for the electrocyclizations in the form of [nπ]-con or [nπ]-dis (where n is the number of π electrons involved, and whether it is conrotatory or disrotatory), and your answer for the cycloaddition in the form of [m+n] (where m and n are the number of atoms on each component).

Noah B
Stanford University

Linguistics

Question:

I am providing the standardized Biblical Hebrew source text from the Biblia Hebraica Stuttgartensia (Psalms 104:7). Your task is to distinguish between closed and open syllables. Please identify and list all closed syllables (ending in a consonant sound) based on the latest research on the Tiberian pronunciation tradition of Biblical Hebrew by scholars such as Geoffrey Khan, Aaron D. Hornkohl, Kim Phillips, and Benjamin Suchard. Medieval sources, such as the Karaite transcription manuscripts, have enabled modern researchers to better understand specific aspects of Biblical Hebrew pronunciation in the Tiberian tradition, including the qualities and functions of the shewa and which letters were pronounced as consonants at the ends of syllables.

מִן־גַּעֲרָתְךָ֣ יְנוּס֑וּן מִן־ק֥וֹל רַ֝עַמְךָ֗ יֵחָפֵזֽוּן (Psalms 104:7) ?

Lina B
University of Cambridge

Figure 2.7.7

Initial testing indicates that HLE is highly challenging for current AI systems. Even top models, such as OpenAI's o1, score just 8.8% (Figure 2.7.8). The researchers behind the benchmark are closely monitoring how quickly LLMs improve, and they speculate that performance could exceed 50% by the end of 2025.

Humanity's Last Exam (HLE): accuracy

Source: Phan et al., 2025 | Chart: 2025 AI Index report

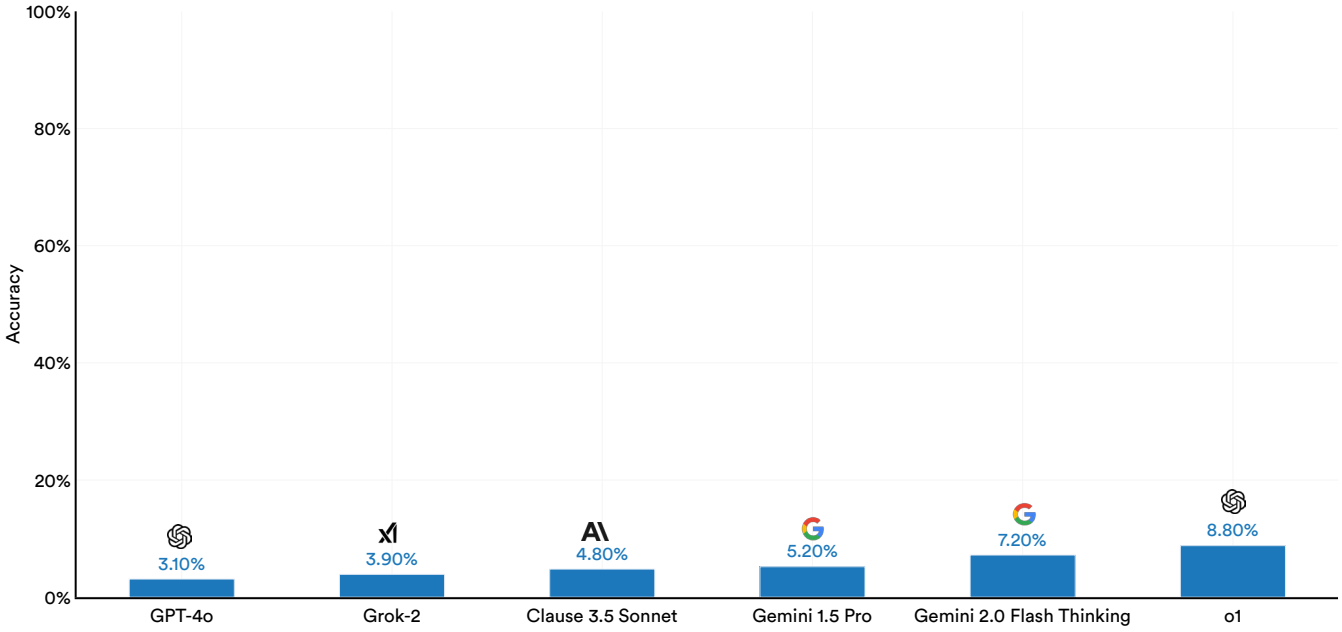


Figure 2.7.8

Planning

Planning is an intelligent task that involves reasoning about actions that alter the world. It requires considering hypothetical future states, including potential external actions and other transformative events.

PlanBench

Claims have been made that LLMs can solve planning problems. A group from Arizona State University has proposed PlanBench, a benchmark suite containing problems used in the automated planning community, especially those used in the International Planning Competition. PlanBench is designed to test LLMs on planning tasks. The benchmark tests models on 600 problems in which a hand tries to construct stacks of blocks when it is only allowed to move one block at a time to a table or to the top of a clear block. After the benchmark was released in 2022, researchers demonstrated that models like GPT-4 and GPT-3.5 still struggled with planning tasks.

The release of OpenAI's o1 was met with enthusiasm from the AI research community, as it was designed to actively reason rather than function purely as an autoregressive LLM. When tested on the PlanBench benchmark, o1 showed significant improvements, though it still struggles with reliable and consistent planning. In the Blocksworld zero-shot evaluation (one specific planning evaluation domain), o1 achieved a score of 97.8%—far surpassing the next best LLM, Llama 3.1 405B (62.6%), and dramatically outperforming GPT-4o (35.5%) (Figure 2.7.9). In the more challenging Mystery Blocksworld domain, where some answers are syntactically obfuscated, o1 scored 52.8% zero-shot, compared to just 0.8% for Llama 3.1 405B. GPT-4, by contrast, scored 0%.

Planning is a combinatorial problem, and solving problems with long solutions is expected to take more than linear time. Not surprisingly, when tested on instances that require at least 20 steps, o1 manages to solve just 23.6%.

PlanBench: instances correct

Source: Valmeekam et al., 2024 | Chart: 2025 AI Index report

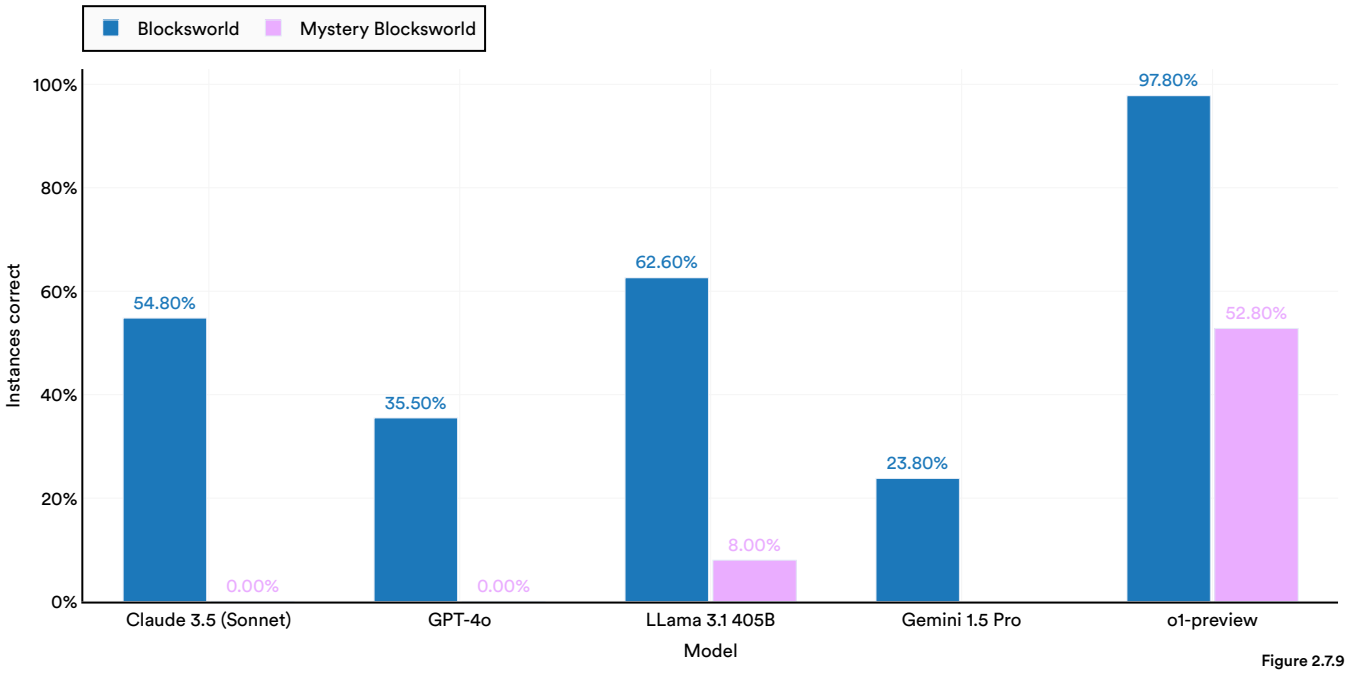


Figure 2.7.9

AI agents, autonomous or semiautonomous systems designed to operate within specific environments to accomplish goals, represent an exciting frontier in AI research. These agents have a diverse range of potential applications, from assisting in academic research and scheduling meetings to facilitating online shopping and vacation booking. As suggested by many recent corporate releases, agentic AI has become a topic of increasing interest in the technical world of AI.

2.8 AI Agents

For decades, the topic of AI agents has been widely discussed in the AI community, yet few benchmarks have achieved widespread adoption, including those featured in last year's Index, such as AgentBench and MLAgentBench. This is partly due to the inherent complexity of benchmarking agentic tasks, which are typically more diverse, dynamic, and variable than tasks like image classification or answering language questions. As AI continues to evolve, it will become important to develop effective methods to evaluate AI agents.

VisualAgentBench

VisualAgentBench (VAB), launched in 2024, represents a significant step forward in the evaluation of agentic AI. This benchmark reflects the growing multimodality of AI models and their increasing proficiency in navigating both virtual and embodied environments. VAB addresses the need to assess agent performance in diverse settings that extend beyond environments reliant solely on linguistic commands. VAB tests agents across three broad categories of tasks: embodied agents (operating in household and gaming environments), GUI agents (interacting with mobile and web applications), and visual design agents (such as CSS debugging) (Figure 2.8.1). This comprehensive approach creates a robust evaluation suite of agents' capabilities across varied and dynamic contexts.

Tasks on VisualAgentBench
Source: Liu et al., 2024

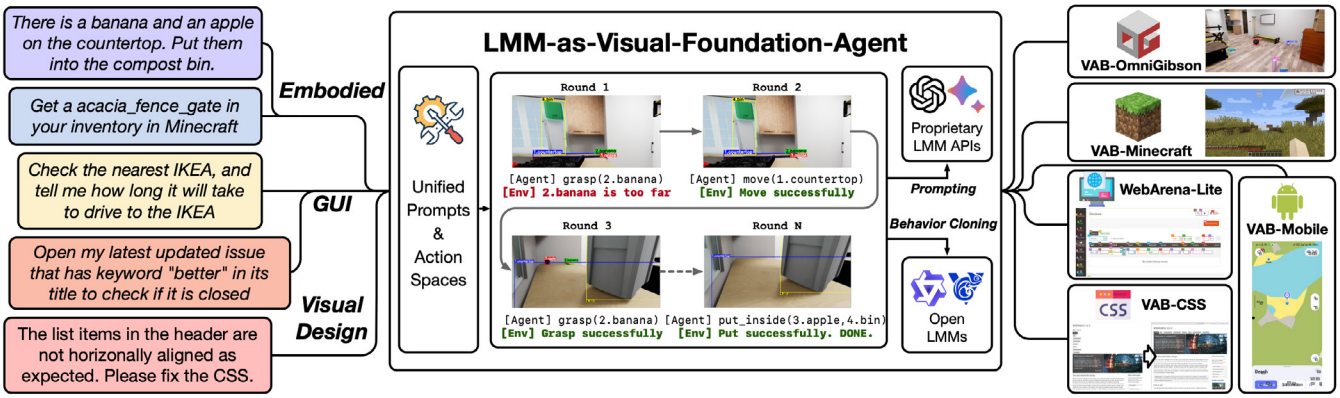


Figure 2.8.1

VAB presents a significant challenge for AI systems. The top-performing model, GPT-4o, achieves an overall success rate of just 36.2%, while most proprietary language models average around 20% (Figure 2.8.2). According to the benchmark's authors, these results reveal that current AI models are far from ready for direct deployment in agentic settings.

VisualAgentBench on the test set: success rate
Source: VisualAgentBench Leaderboard, 2025 | Chart: 2025 AI Index report

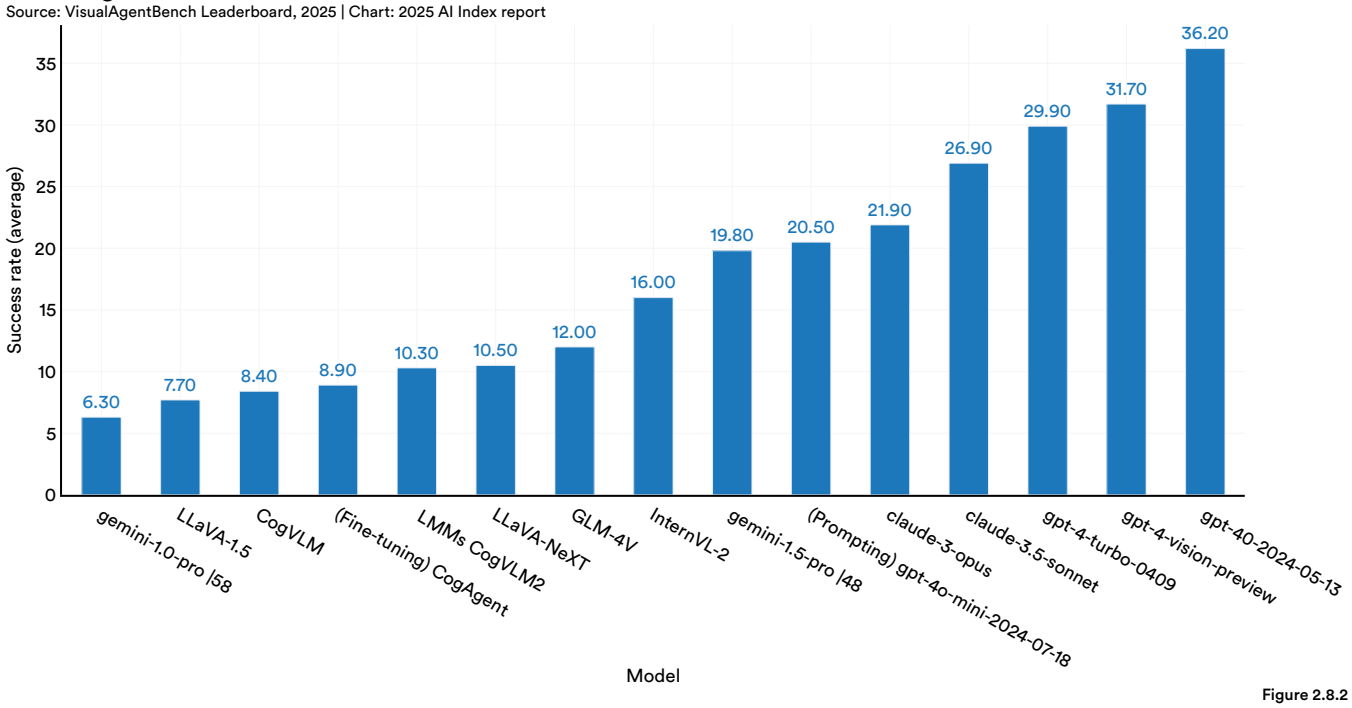


Figure 2.8.2

RE-Bench

The emergence of increasingly capable agentic AI systems has fueled predictions that AI might soon take on the work of computer scientists or researchers. However, until recently, there were few benchmarks designed to rigorously test the R&D capabilities of top-performing AI systems. In 2024, researchers addressed this gap with the launch of RE-Bench, a benchmark featuring seven challenging, open-ended ML research environments. These tasks, informed by data from 71 eight-hour attempts by over 60 human experts, include optimizing a kernel, conducting a scaling law experiment, and fine-tuning GPT-2 for question answering, among others (Figure 2.8.3).

RE-Bench Process and Flow
Source: Wijk et al., 2024

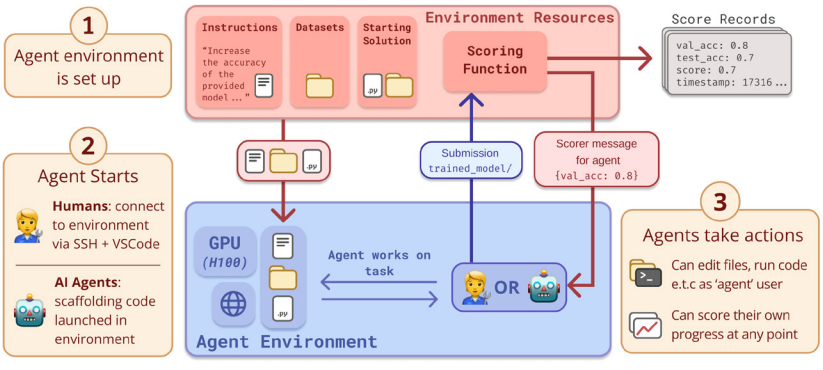


Figure 2.8.3

Researchers uncovered two key findings when comparing the performance of humans and frontier AI models. In short time horizon settings, such as with a two-hour budget, the best AI systems achieve scores four times higher than human experts (Figure 2.8.4). However, as the time budget increases, human performance begins to surpass that of AI. With an eight-hour budget, human performance slightly exceeds AI, and with a 32-hour budget, humans outperform AI by a factor of two. The researchers also note that for certain tasks, AI agents already demonstrate expertise comparable to humans but can deliver results significantly faster and at a lower cost. For example, AI agents can write custom Triton kernels more quickly than any human expert.

RE-Bench: average normalized score@k
Source: Wijk et al., 2024 | Chart: 2025 AI Index report

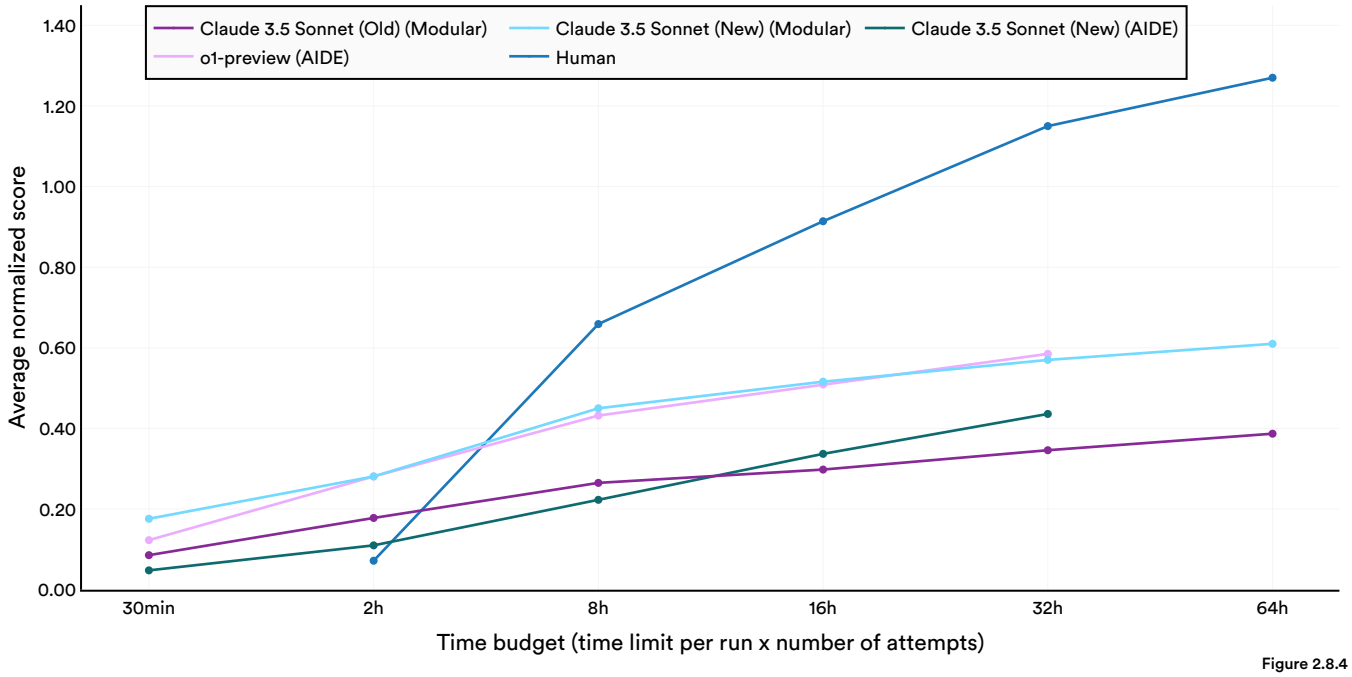


Figure 2.8.4

GAIA

GAIA is a benchmark for General AI assistants introduced by Meta in May 2024. It consists of 466 questions designed to assess AI systems' ability to perform a broad range of tasks, including reasoning, multimodal processing, web browsing, and tool use. Unlike straightforward, exam-style questions, GAIA challenges AI models with complex, multistep problems that may require searching the open web, interpreting multimodal inputs, and reasoning through intricate scenarios (Figure 2.8.5). When researchers launched GAIA, they found that existing LLMs lagged significantly behind human performance. For instance, GPT-4 with plugins correctly answered only 15% of the questions, compared to 92% for human respondents.

As with other recently introduced AI benchmarks, performance on GAIA has improved rapidly. In 2024, the top system achieved a score of 65.1%, marking a roughly 30 percentage point increase from the highest score recorded in 2023 (Figure 2.8.6).

Sample questions on GAIA

Source: Meta, 2024

Level 1

Question: What was the actual enrollment count of the clinical trial on H. pylori in acne vulgaris patients from Jan-May 2018 as listed on the NIH website?
Ground truth: 90

Level 2

Question: If this whole pint is made up of ice cream, how many percent above or below the US federal standards for butterfat content is it when using the standards as reported by Wikipedia in 2020? Answer as + or - a number rounded to one decimal place.
Ground truth: +4.6

Level 3

Question: In NASA's Astronomy Picture of the Day on 2006 January 21, two astronauts are visible, with one appearing much smaller than the other. As of August 2023, out of the astronauts in the NASA Astronaut Group that the smaller astronaut was a member of, which one spent the least time in space, and how many minutes did he spend in space, rounded to the nearest minute? Exclude any astronauts who did not spend any time in space. Give the last name of the astronaut, separated from the number of minutes by a semicolon.
Ground truth: White; 5876

Figure 2.8.5

GAIA: average score

Source: GAIA Leaderboard, 2025 | Chart: 2025 AI Index report

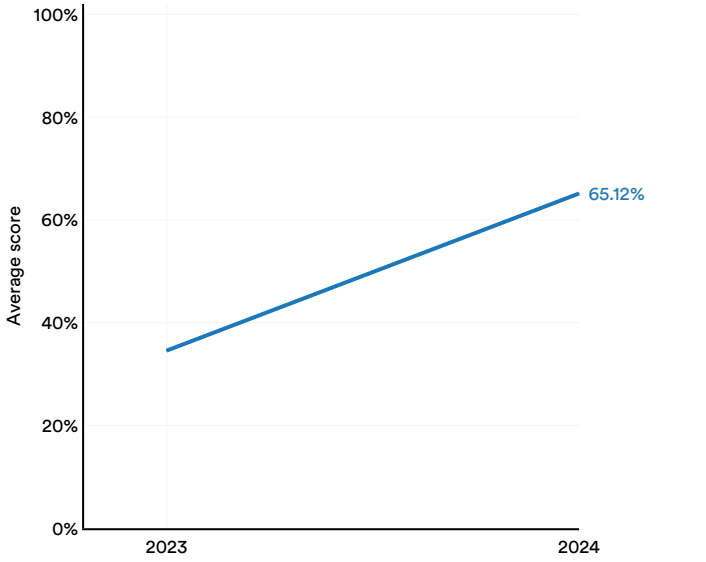


Figure 2.8.6

Advancements in AI over the past decade have paved the way for exciting new developments in the field of robotics. Especially with the rise of foundation models, robots are now able to iteratively learn from their surroundings, adapt flexibly to new settings, and make autonomous decisions. This section explores key robotic benchmarks and recent trends, including the rise of humanoids, algorithmic advancements from DeepMind, and the emergence of robotic foundation models. It concludes by studying developments in self-driving cars.

2.9 Robotics and Autonomous Motion

Robotics

RLBench

One of the most widely adopted benchmarks in the robotics community is RLBench (Robot Learning Benchmark). Launched in 2019, it features 100 unique tasks of varying complexity, from simple target reaching to opening an oven and placing a tray inside.[12] Researchers typically evaluate new robotic systems on a standardized subset of 18 tasks to gauge performance. Figure 2.9.1 visualizes some of the tasks in RLBench.

Tasks on VisualAgentBench
Source: James et al., 2019

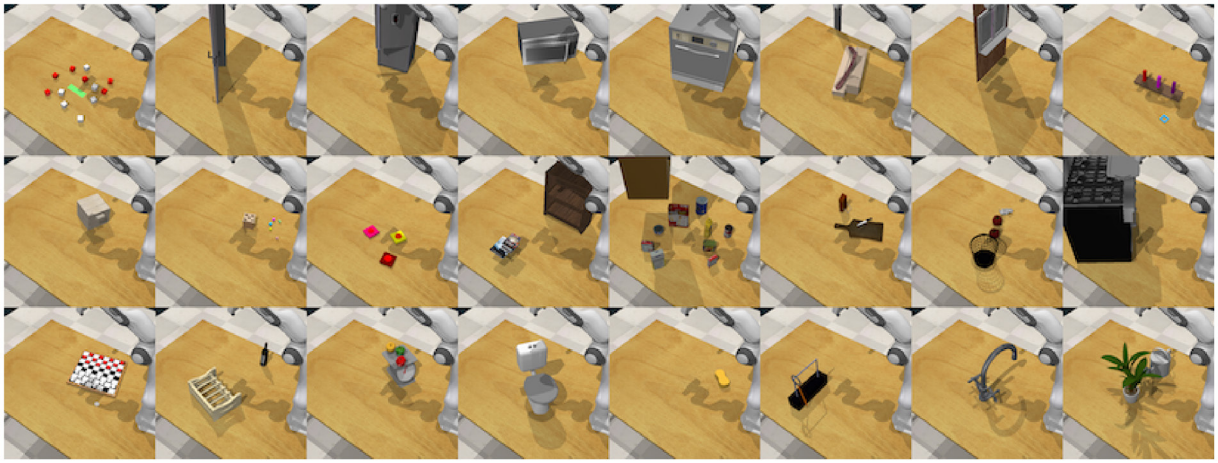


Figure 2.9.1

12 Target reaching in robotics refers to the process by which a robotic system moves its end-effector (such as a robotic arm or gripper) toward a specific goal position or object in space.

As of January 2025, the top-performing model on this subset is SAM2Act, a collaboration between researchers at the University of Washington, Universidad Católica San Pablo, Nvidia, and the Allen Institute for AI. SAM2Act achieved an 86.8% success rate, marking a 2.8 percentage point improvement over the previous state-of-the-art in 2024 and a 66.7 percentage point increase from the leading score in 2021 (Figure 2.9.2).

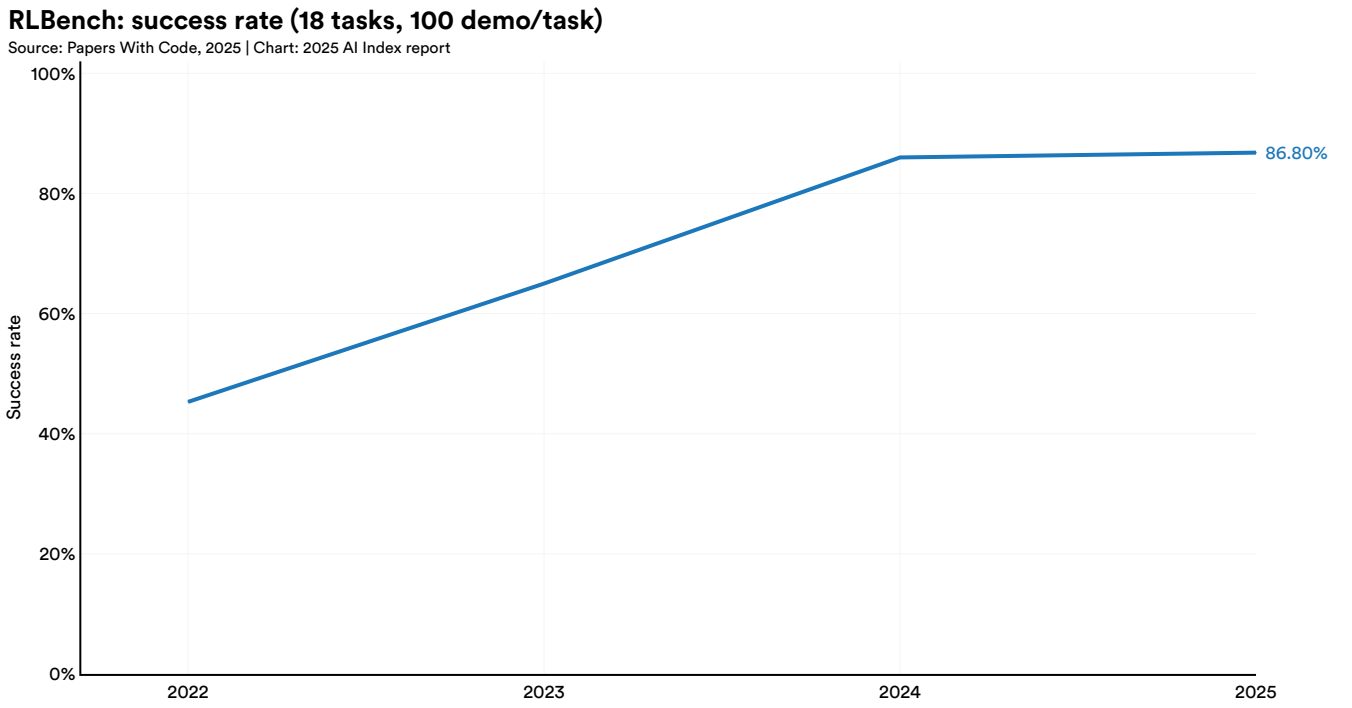


Figure 2.9.2

Highlight:

Humanoid Robotics

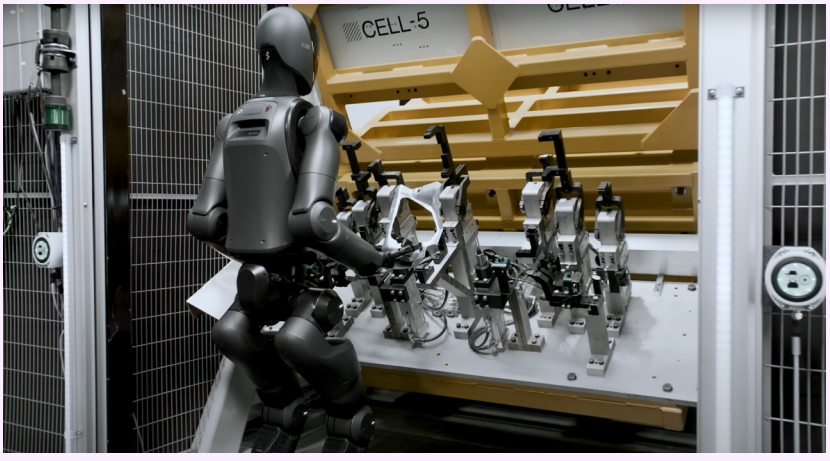
2024 was a significant year for robotics, marked by the growing prevalence of humanoid robots—machines with humanlike bodies designed to mimic human functions. For example, Figure AI, a robotics startup dedicated to developing general-purpose humanoid robots, launched Figure 02 in 2024, its most advanced model yet. Standing 5 feet 6 inches tall, weighing 154 pounds, and capable of handling a 44-pound payload, Figure 02 operates for up to five hours on a single charge. Figure robots are able to perform complex tasks such as making coffee and assisting in automotive assembly by placing sheet metal into a car fixture (Figure 2.9.3 and Figure 2.9.4). They are also integrated with OpenAI and can engage in speech-to-speech reasoning, whereby the robot explains its actions and responds to queries about its behavior. Figure's success follows that of other companies that released humanoid robots, like Tesla's Optimus, first launched in 2002 and redesigned in 2023, and Boston Dynamics' Atlas humanoid.

Figure robot making coffee
Source: Figure AI

Figure 2.9.3

Figure robot assisting in automotive assembly
Source: Figure AI

Figure 2.9.4



Highlight:

DeepMind's Developments

In 2023, DeepMind launched two robotic models, PaLM-E and RT-2. These models were novel in their use of transformer-based architectures, typically found in language modeling, and their training on both manipulation data and language data. This dual training approach enabled them to excel at both robotic manipulation and text generation. In 2024, DeepMind introduced AutoRT, an AI system that leverages large foundation models to autonomously generate diverse training data for robots. It coordinates multiple video-equipped robots, guiding them through various environments, devising creative tasks for them to perform, and meticulously documenting these tasks (Figure 2.9.5). This documentation then serves as training data for future robotic learning. To date, AutoRT has generated a dataset of 77,000 robotic trials spanning 6,650 unique tasks. Greater amounts of robotic training data will be important to improve the training of future robotic systems.

AutoRT workflow
Source: Google DeepMind, 2024

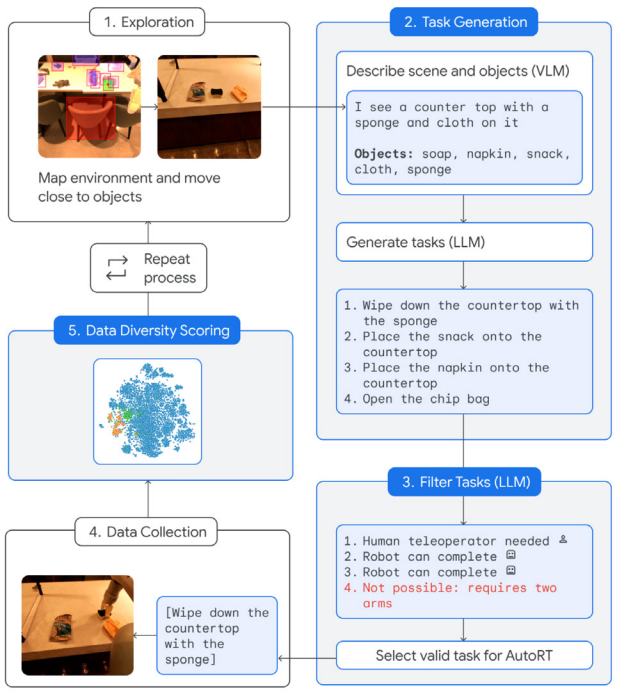
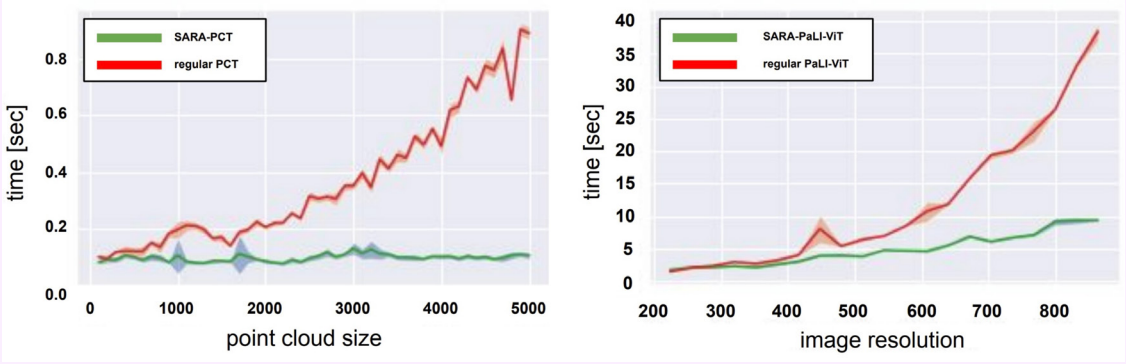


Figure 2.9.5

Conversely, SARA-RT, also from Google DeepMind, improves the efficiency of transformer-based robotic models by significantly improving their speed. While transformers are powerful, they are also computationally intensive as they rely on quadratic complexity attention mechanisms. This means that doubling the input size of data provided to a model can quadruple computational requirements. This challenge complicates attempts to scale robotic models. SARA-RT addresses this challenge with a technique called "up-training," which converts the quadratic complexity of standard transformers into a linear model. This method drastically reduces computational demands while maintaining performance quality. Figure 2.9.6 compares speed tests of AI models enhanced with the SARA technique against those without. In point cloud processing,

Speed tests for SARA vs. non-SARA enhanced models
Source: Google DeepMind, 2024

Figure 2.9.6



Highlight:

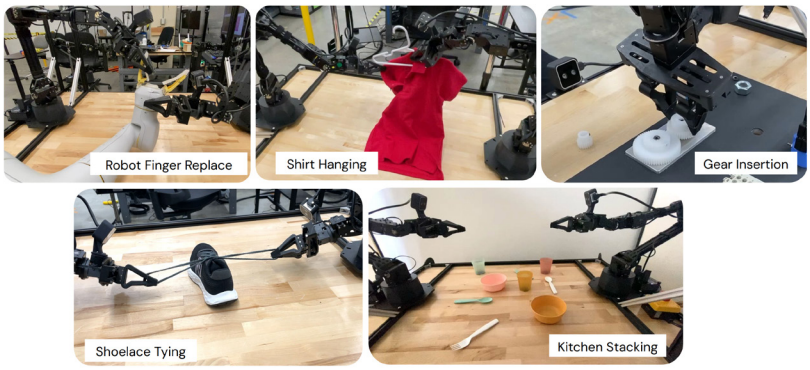
DeepMind's Developments (cont'd)

which enables robots to interpret 3D environments, and in image processing, SARA-based models run significantly faster while avoiding major increases in run-time at scale.

Other developments from DeepMind include ALOHA (Autonomous Learning of High-level Activities) and DemoStart. ALOHA Unleashed is a breakthrough in enabling robots to perform intricate dexterous manipulation tasks, such as tying shoelaces or hanging T-shirts on coat hangers—tasks that historically have been extremely challenging for robots. The researchers demonstrated that combining a large imitation learning dataset with a transformer-based learning architecture is a highly effective approach for overcoming these difficulties. The ALOHA approach enabled Google's robot to effectively learn a diverse range of tasks, including hanging a shirt, stacking kitchen items, and tying shoelaces (Figure 2.9.7). As shown in Figure 2.9.8, ALOHA-trained robots achieved a high success rate across these tasks.

ALOHA-trained robot attempting complex tasks
Source: Google DeepMind, 2024

Figure 2.9.7



ALOHA: success rate
Source: Zhao et al., 2024 | Chart: 2025 AI Index report

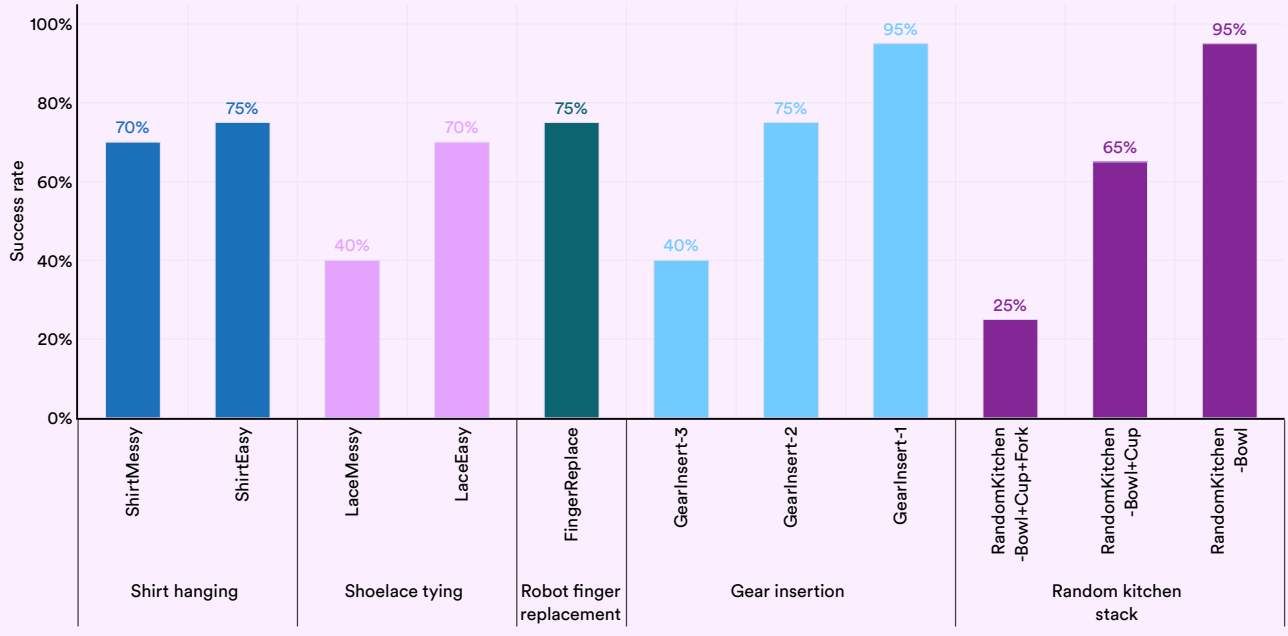


Figure 2.9.8

Highlight:

DeepMind's Developments (cont'd)

Similarly, DemoStart introduces a novel auto-curriculum reinforcement learning method that enables a robotic arm to master complex behaviors using only sparse rewards and a limited number of demonstrations. This breakthrough highlights the potential for robots to learn efficiently with minimal data, reducing the need for data-intensive training and making advanced robotics more accessible and widely adopted. DeepMind also introduced a robotic model in 2024 that was capable of reaching amateur human-level performance in competitive table tennis (Figure 2.9.9). Given that achieving human-level speed and performance on real-world tasks is an important benchmark for robotics research, this achievement is a notable step forward in robotic ability.

Robots playing amateur-level table tennis
Source: Google DeepMind, 2024

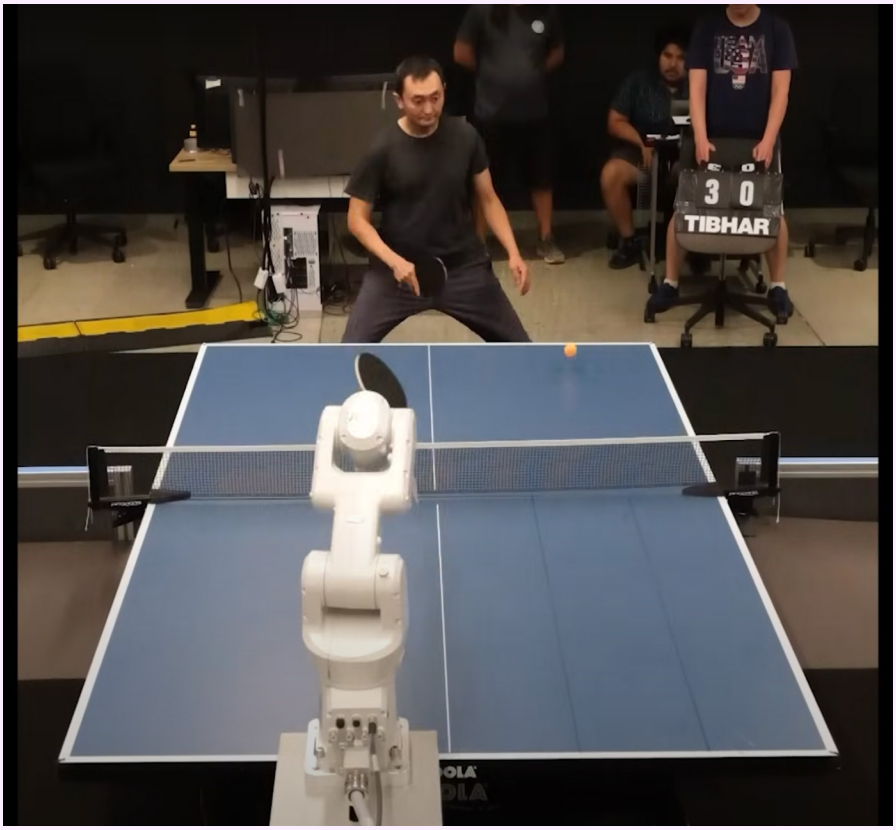


Figure 2.9.9

Highlight:

Foundation Models for Robotics

In 2024, there was a strong push toward developing foundational models for robotics—systems capable of reasoning with language while physically operating in the real world. Nvidia introduced GR00T (Generalist Robot 00 Technology), a general-purpose foundation model for humanoid robots designed to understand natural language and mimic human movements. Alongside GR00T, Nvidia released data pipelines, simulation frameworks, and the Thor robotics computer. Figure 2.9.10 illustrates the components of GROOT's launch. This robotic development suite is intended to make it easier for the robotic community to scale and build increasingly advanced robotics.

Nvidia was not alone in this space. Covariant launched RFM-1, a robotic foundation model with language capabilities and real-world maneuverability. Meanwhile, LLaRA, developed by researchers at Stony Brook University and the University of Wisconsin-Madison, integrates perception, communication, and action into a monolithic, end-to-end deep learning model. These new models continue a trend from 2023, which saw the launch of robotic foundation models like RT-2, PaLM-E, and Open-X Embodiment.

GROOT blueprint for synthetic motion generation
Source: Nvidia, 2024

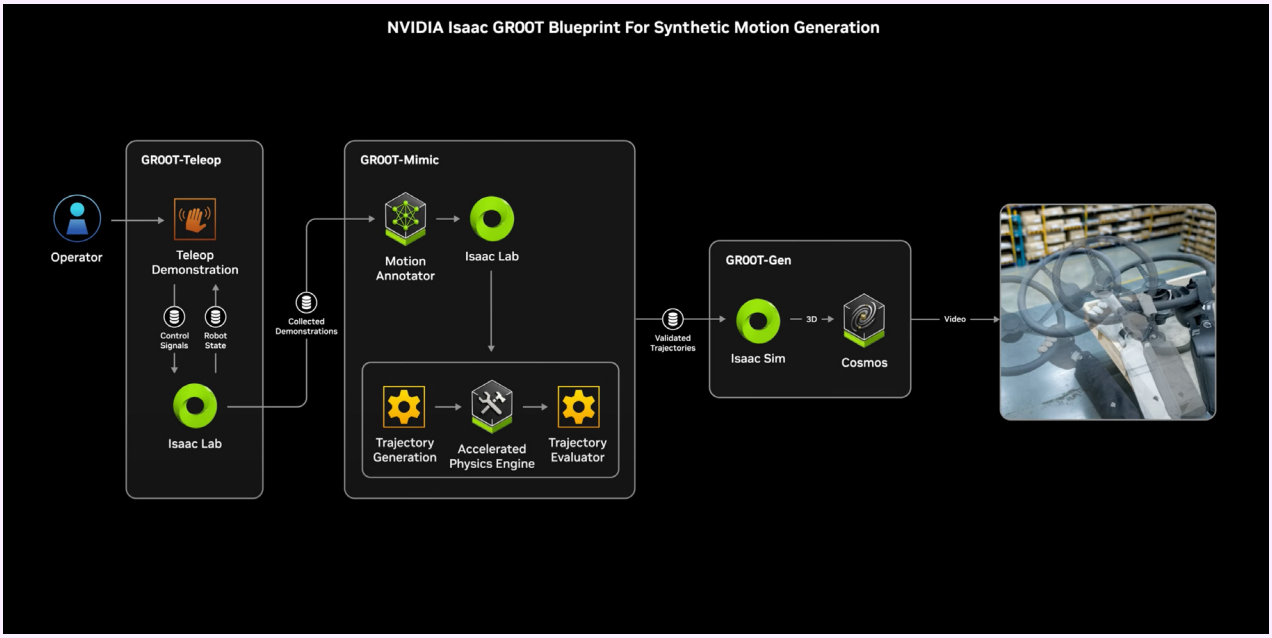


Figure 2.9.10

Self-Driving Cars

Self-driving vehicles have long been a goal for AI researchers and technologists. However, their widespread adoption has been slower than anticipated. Despite many predictions that fully autonomous driving is imminent, widespread use of self-driving vehicles has yet to become a reality. Still, in recent years, significant progress has been made. In cities like San Francisco and Phoenix, fleets of self-driving taxis are now operating commercially. This section examines recent advancements in autonomous driving, focusing on deployment, technological breakthroughs and new benchmarks, safety performance, and policy challenges.

Deployment

Self-driving cars are increasingly being deployed worldwide. Cruise, a subsidiary of General Motors, launched its autonomous vehicles in San Francisco in late 2022 before having its license suspended in 2023 after a litany of safety incidents. Waymo, a subsidiary of Alphabet, began deploying its robotaxis in Phoenix in early 2022 and expanded to San Francisco in 2024. The company has since emerged as one of the more successful players in the self-driving industry: As of January 2025, Waymo operates in four major U.S. cities—Phoenix, San Francisco, Los Angeles, and Austin (Figure 2.9.11). Data sourced from October 2024 suggests that across the four cities the company provides 150,000 paid rides per week, covering over a million miles. Looking ahead, Waymo plans to test its vehicles in 10 additional cities, including Las Vegas, San Diego, and Miami. The company chose testing locations, such as upstate New York and Truckee, California, that experience snowy weather so it can assess the vehicles in diverse driving conditions. There has also been notable progress in self-driving trucks, with companies like Kodiak completing its first driverless deliveries and Aurora reporting steady advancements, including over 1 million miles of autonomous freight hauling on U.S. highways since 2021—albeit with human safety drivers present. Still, challenges remain in bringing this technology to market, with Aurora recently announcing it would delay the commercial launch of its fleet from the end of 2024 until April 2025.

Waymo rider-only miles driven without a human driver

Source: Waymo, 2024 | Table: 2025 AI Index report

Location	Rider-only miles through September 2024
Los Angeles	1.947M
San Francisco	10.209M
Phoenix	20.823M
Austin	124K

Figure 2.9.11

China's self-driving revolution is also accelerating, led by companies like Baidu's Apollo Go, which reported 988,000 rides across China in Q3 2024, reflecting a 20% year-over-year increase. In October 2024, the company was operating 400 robotaxis and announced plans to expand its fleet to 1,000 by the end of 2025. Pony.AI, another Chinese autonomous vehicle manufacturer, has pledged to scale its robotaxi fleet from 200 to at least 1,000 vehicles—with expectations that the fleet will reach 2,000 to 3,000 by the end of 2026. China is leading the way in autonomous vehicle testing, with reports indicating that it is testing more driverless cars than any other country and currently rolling them out across 16 cities. Robotaxis in China are notably affordable—even cheaper, in some cases, than rides provided by human drivers. To support this growth, China has prioritized establishing national regulations to govern the deployment of driverless cars. Beyond the self-driving revolution taking place in the U.S. and China, European startups like Wayve are beginning to gain traction in the industry.

Technical Innovations and New Benchmarks

Over the past year, self-driving technology has advanced significantly, both in vehicle capabilities and benchmarking methods. In October 2024, Tesla unveiled the Cybercab, a two-passenger autonomous vehicle without a steering wheel or pedals, which is set for production in 2026 at a price of under $30,000. Tesla also unveiled the Robovan, an electric autonomous van designed to transport up to 20 passengers. Meanwhile, Baidu's Apollo Go launched its latest-generation robotaxi, the RT6, across multiple cities in China (Figure 2.9.12). With a price tag of just $30,000 and a battery-swapping system, the RT6 represents a major step toward making self-driving technology more cost-effective and scalable. As costs continue to decline, the adoption of autonomous vehicles is expected to accelerate. Notable business partnerships have also advanced self-driving technology, including Uber's collaboration with WeRide—the world's first publicly listed robotaxi company—to develop an autonomous ride-sharing platform in Abu Dhabi.

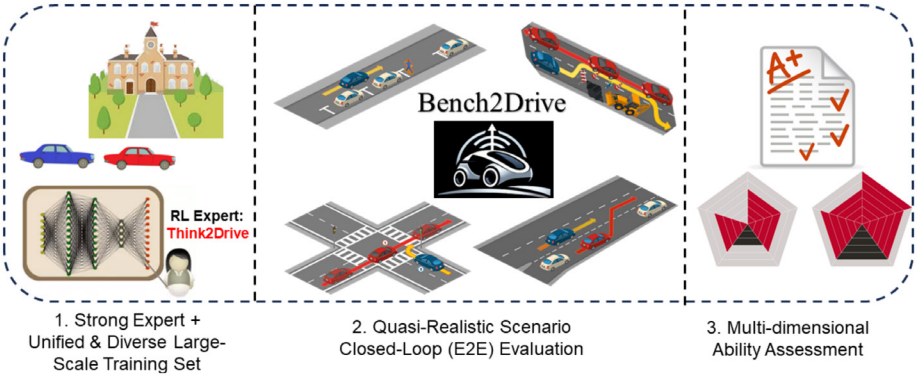
In 2024, several new benchmarks were introduced to evaluate self-driving capabilities. One notable example is nuPlan, developed by Motional. It is a large-scale, autonomous driving dataset designed to test machine-learning-based motion planners. The benchmark includes 1,282 hours of diverse driving scenarios from multiple cities, along with a simulation and evaluation framework that enables planners' actions to be tested in closed-loop settings. Another recent benchmark is OpenAD, the first real-world, open-world autonomous driving benchmark for 3D object detection. OpenAD focuses on domain generalization—the ability of autonomous driving systems to adapt across diverse sensor configurations—and open-vocabulary recognition, which allows systems to identify previously unseen semantic categories.

GROOT blueprint for synthetic motion generation
Source: Verge, 2024

Figure 2.9.12

Most existing benchmarks for end-to-end autonomous driving rely on open-loop evaluation, which can be restrictive. Open-loop settings fail to test how autonomous agents react to real-world conditions and often lead to models that memorize driving patterns rather than learning to drive authentically. While closed-loop benchmarks like Town05Long and Longest6 exist, they primarily assess basic driving skills rather than performance in complex, interactive scenarios. Bench2Drive is another new benchmark that improves on these limitations by providing a comprehensive, realistic, closed-loop testing simulation environment for end-to-end autonomous vehicles (Figure 2.9.13). It includes a training set with over 2 million fully annotated frames sourced from more than 10,000 clips, as well as an evaluation suite with 220 short routes designed to test autonomous driving capabilities in diverse conditions. Figure 2.9.14 displays the driving scores of various autonomous driving methods evaluated on the Bench2Drive benchmark.[13]

An overview of Bench2Drive
Source: Jia et al., 2024
Figure 2.9.13



13 This metric accounts for both route completion and infractions, averaging route completion percentages while applying penalties based on infraction severity. For more detail on the driving score methodology, see Section 3 of the Bench2Drive paper.

Bench2Drive: driving score
Source: Jia et al., 2024 | Chart: 2025 AI Index report

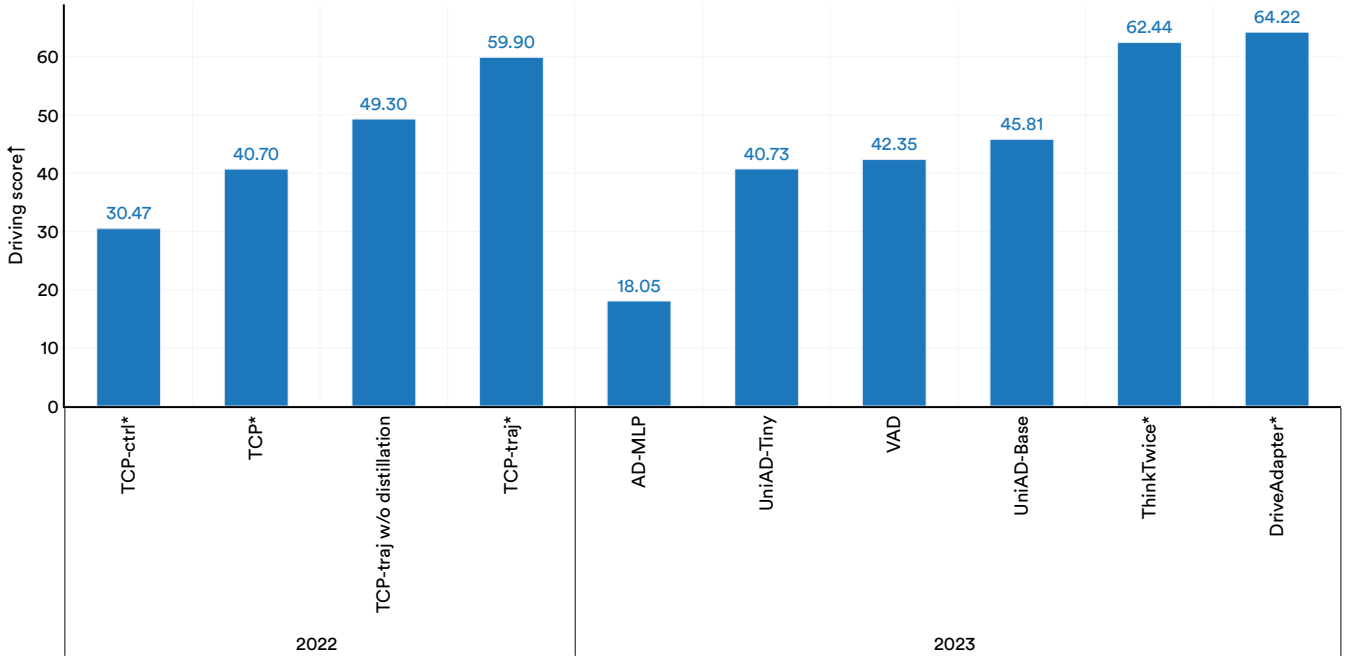


Figure 2.9.14

Safety Standards

Emerging research suggests that self-driving cars may be safer than human-driven vehicles. Figure 2.9.15 compares the number of reported incidents per million miles driven by Waymo vehicles to the estimated rates if humans had driven the same distance. The data shows that Waymo vehicles had significantly fewer incidents, including 1.42 fewer airbag deployments, 3.16 fewer crashes with reported injuries, and 3.65 fewer police-reported crashes per million miles (Figure 2.9.15). Figure 2.9.16 highlights the differences in incident rates across various crash locations, revealing that across all locations with available data, Waymo vehicles consistently recorded lower rates of airbag deployments, injury-reported crashes, and police-reported incidents.

Waymo driver vs. human benchmarks in Phoenix and San Francisco

Source: Waymo, 2024 | Chart: 2025 AI Index report

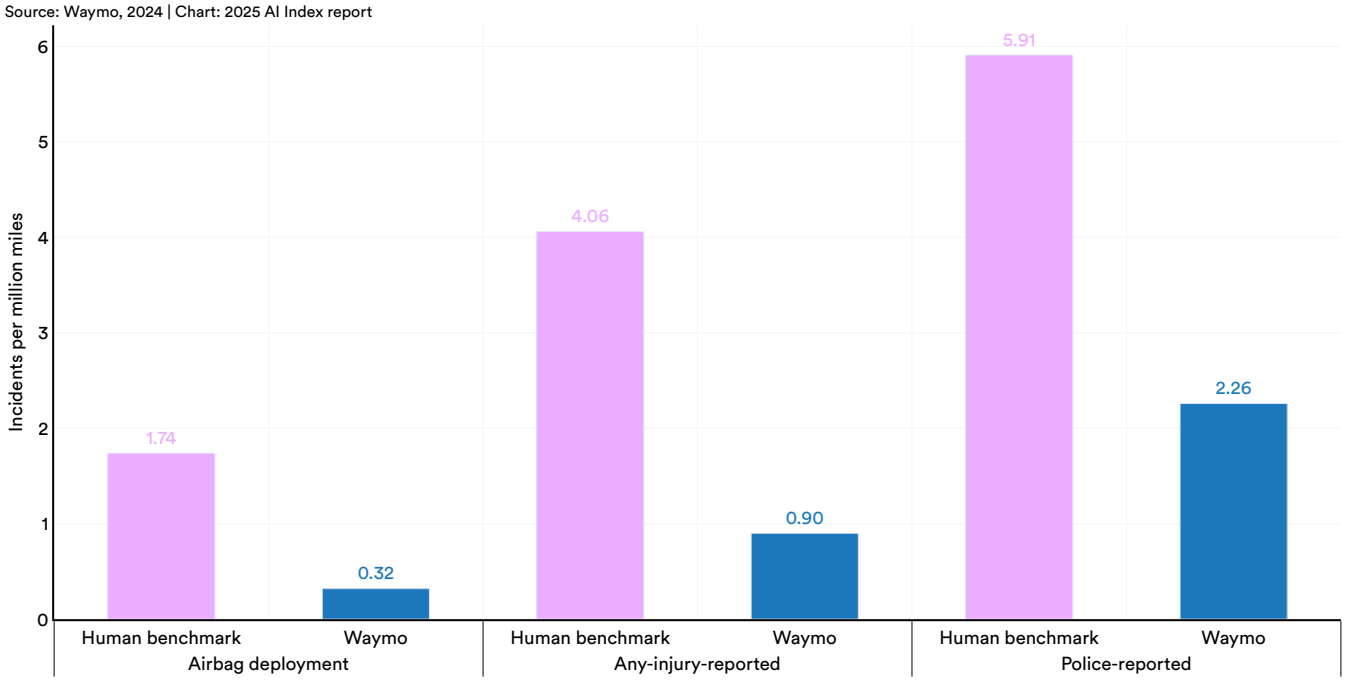


Figure 2.9.15[14]

Waymo driver percent difference to human benchmark in Phoenix and San Francisco

Source: Waymo, 2024 | Chart: 2025 AI Index report

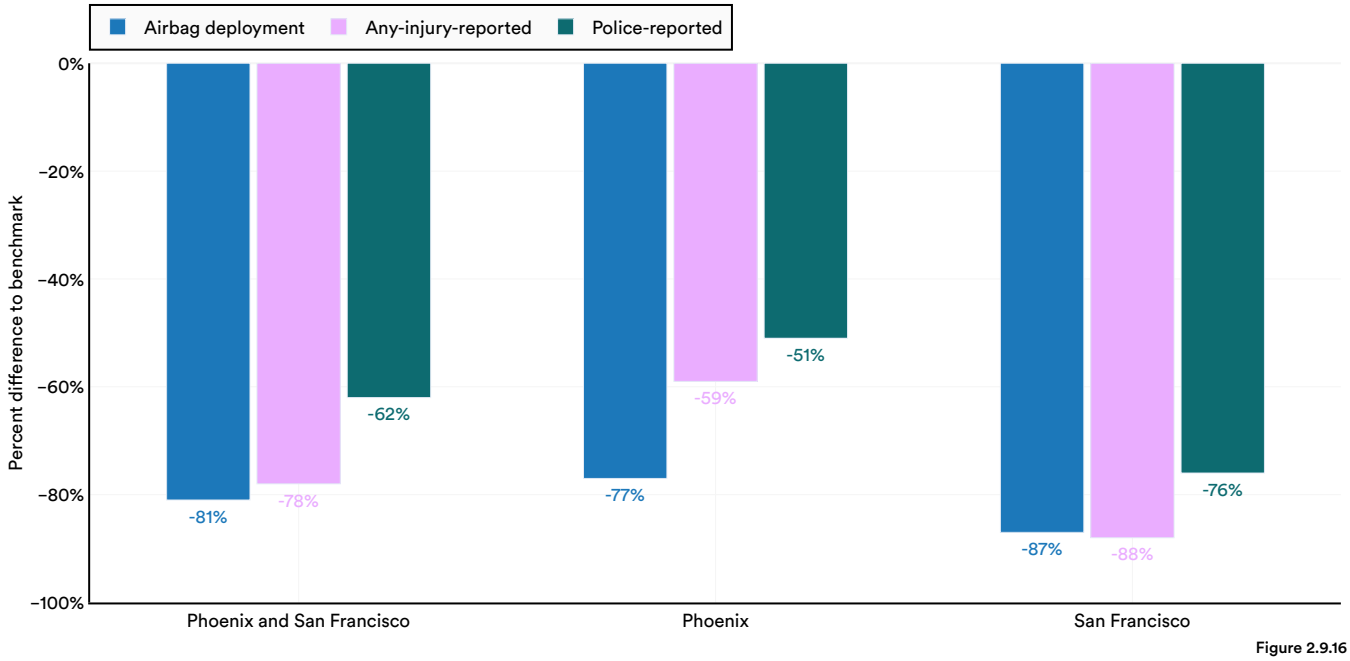


Figure 2.9.16

14 Waymo's safety data is continuously updated in real time, so the totals reported in this section may not fully align with those currently displayed on their website.

Waymo, in collaboration with Swiss Re, one of the world's leading reinsurers, also conducted a study analyzing liability claims related to collisions over several million miles driven by its fully autonomous vehicles. The study compared Waymo's liability claims to human-driver baselines derived from Swiss Re's extensive dataset, which includes over 500,000 claims and 200 billion miles of driving data. The results showed that Waymo vehicles had an 88% reduction in property damage claims and a 92% reduction in bodily injury claims (Figure 2.9.17). In real terms, across 25.3 million miles driven, Waymo vehicles were involved in just nine property damage claims and two bodily injury claims, whereas human drivers over the same distance would be expected to incur 78 property damage claims and 26 bodily injury claims. The Waymo drivers were also significantly safer than latest-generation human-driven vehicles that are equipped with added safety features.

Comparison of liability insurance claims by type: Waymo driver vs. human-driven vehicles

Source: Di Lillo et al., 2024 | Chart: 2025 AI Index report

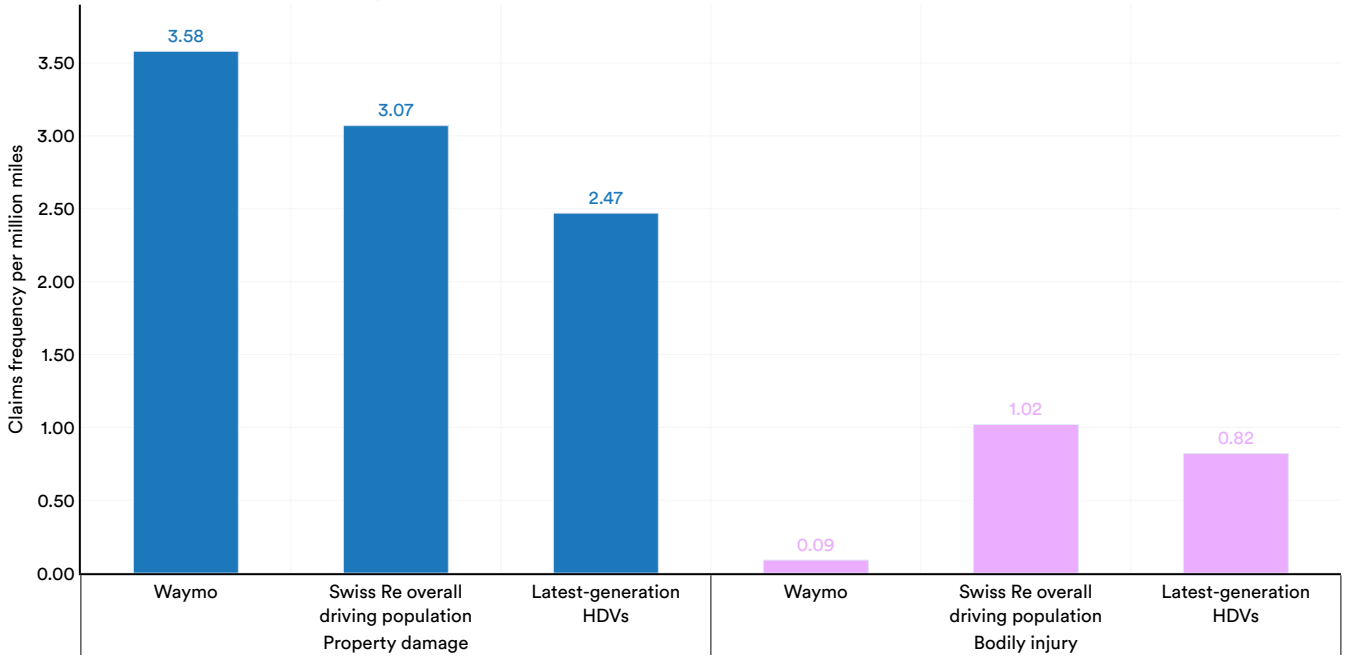


Figure 2.9.17

Appendix

Acknowledgments

The AI Index would like to acknowledge Andrew Shi for his work generating sample Midjourney and Pika video generations and Armin Hamrah for his work identifying significant AI technical advancements for the timeline.

Benchmarks

In this chapter, the AI Index reports on benchmarks, recognizing their importance in tracking AI's technical progress. As a standard practice, the Index sources benchmark scores from leaderboards, public repositories such as Papers With Code and RankedAGI, as well as company papers, blog posts, and product releases. The Index operates under the assumption that the scores reported by companies are accurate and factual. The benchmark scores in this section are current as of mid-February 2025. However, since the publication of the AI Index, newer models may have been released that surpass current state-of-the-art scores.

1. **ARC-AGI**: Data on ARC-AGI was taken from the ARC-AGI paper and OpenAI video in February 2025. To learn more about ARC-AGI, please read the original paper.
2. **Arena-Hard-Auto**: Data on Arena-Hard-Auto was taken from the LMSYS leaderboard in February 2025. To learn more about Arena-Hard-Auto, please read the original paper.
3. **Bench2Drive**: Data on Bench2Drive was taken from the Bench2Drive paper in February 2025. To learn more about Bench2Drive, please read the original paper.
4. **Berkeley Function Calling**: Data on Berkeley Function Calling was taken from the Berkeley Function Calling leaderboard in February 2025. To learn more about Berkeley Function Calling, please read the original work.
5. **BigCodeBench**: Data on BigCodeBench was taken from the BigCodeBench leaderboard in February 2025. To learn more about BigCodeBench, please read the original work.
6. **Chatbot Arena**: Data on Chatbot Arena was taken from the Chatbot Arena leaderboard in February 2025. To learn more about Chatbot Arena, please read the original paper.
7. **FrontierMath**: Data on FrontierMath was taken from the FrontierMath paper and OpenAI video in February 2025. To learn more about FrontierMath, please read the original paper. The visual was supplemented with benchmark data from OpenAI's o3 model, sourced from a YouTube video announcing its launch in December 2025.
8. **GAIA**: Data on GAIA was taken from the GAIA leaderboard in February 2025. To learn more about GAIA, please read the original paper.
9. **GPQA**: Data on GPQA was taken from the GPQA paper and OpenAI video in February 2025. To learn more about GPQA, please read the original paper.
10. **GSM8K**: Data on GSM8K was taken from the GSM8K Papers With Code leaderboard in February 2025. To learn more about GSM8K, please read the original paper.
11. **HELMET**: Data on HELMET (How to Evaluate Long-Context Models Effectively and Thoroughly) was taken from the HELMET paper in February 2025. To learn more about HELMET, please read the original paper.
12. **HLE**: Data on Humanity's Last Exam (HLE) was taken from the HLE paper in February 2025. To learn more about HLE, please read the original paper.
13. **HumanEval**: Data on HumanEval was taken from the HumanEval Papers With Code leaderboard in February 2025. To learn more about HumanEval, please read the original paper.
14. **LRS2**: Data on Oxford-BBC Lip Reading Sentences 2 (LRS2) was taken from the LRS2 Papers With Code leaderboard in February 2025. To learn more about LRS2, please read the original paper.

15. **MATH**: Data on MATH was taken from the MATH Papers With Code leaderboard in February 2025 and the o3-mini model launch. To learn more about MATH, please read the original paper.
16. **MixEval**: Data on MixEval was taken from the MixEval leaderboard in February 2025. To learn more about MixEval, please read the original paper.
17. **MMLU**: Data on MMLU was taken from the MMLU Papers With Code leaderboard in February 2025. To learn more about MMLU, please read the original paper.
18. **MMLU-Pro**: Data on MMLU-Pro was taken from the MMLU-Pro leaderboard in February 2025. To learn more about MMLU-Pro, please read the original paper.
19. **MMMU**: Data on MMMU was taken from the MMMU leaderboard in February 2025. To learn more about MMMU, please read the original paper.
20. **MTEB**: Data on Massive Text Embedding Benchmark (MTEB) was taken from the MTEB leaderboard in February 2025. To learn more about MTEB, please read the original paper.
21. **MVBench**: Data on MVBench was taken from the MVBench leaderboard in February 2025. To learn more about MVBench, please read the original paper.
22. **PlanBench**: Data on PlanBench was taken from the PlanBench paper in February 2025. To learn more about PlanBench, please read the original paper.
23. **RE-Bench**: Data on RE-Bench was taken from the RE-Bench paper in February 2025. To learn more about RE-Bench, please read the original paper
24. **RLBench**: Data on RLBench was taken from the RLBench Papers With Code leaderboard in February 2025. To learn more about RLBench, please read the original paper.
25. **Ruler**: Data on Ruler was taken from the Ruler repository in February 2025. To learn more about Ruler, please read the original paper.
26. **SWE-bench**: Data on SWE-bench was taken from the SWE-bench leaderboard in February 2025. To learn more about SWE-bench, please read the original paper.
27. **VAB**: Data on VisualAgentBench (VAB) was taken from the VAB leaderboard in February 2025. To learn more about VAB, please read the original paper.
28. **VCR**: Data on VCR was taken from the VCR leaderboard in February 2025. To learn more about VCR, please read the original paper.
29. **WildBench**: Data on WildBench was taken from the WildBench leaderboard in February 2025. To learn more about WildBench, please read the original paper.

Works Cited

Akter, S. N., Yu, Z., Muhamed, A., Ou, T., Bäuerle, A., Cabrera, Á. A., Dholakia, K., Xiong, C., & Neubig, G. (2023). *An In-Depth Look at Gemini's Language Abilities* (arXiv:2312.11444). arXiv. https://doi.org/10.48550/arXiv.2312.11444

Bairi, R., Sonwane, A., Kanade, A., C, V. D., Iyer, A., Parthasarathy, S., Rajamani, S., Ashok, B., & Shet, S. (2023). *CodePlan: Repository-Level Coding Using LLMs and Planning* (arXiv:2309.12499). arXiv. https://doi.org/10.48550/arXiv.2309.12499

Bauza, M., Chen, J. E., Dalibard, V., Gileadi, N., Hafner, R., Martins, M. F., Moore, J., Pevceviciute, R., Laurens, A., Rao, D., Zambelli, M., Riedmiller, M., Scholz, J., Bousmalis, K., Nori, F., & Heess, N. (2024). *DemoStart: Demonstration-Led Auto-Curriculum Applied to Sim-to-Real With Multi-fingered Robots* (arXiv:2409.06613). arXiv. https://doi.org/10.48550/arXiv.2409.06613

Bommasani, R., Kapoor, S., Klyman, K., Longpre, S., Ramaswami, A., Zhang, D., Schaake, M., Ho, D. E., Narayanan, A., & Liang, P. (2024). "Considerations for Governing Open Foundation Models." *Science*, 386(6718), 151–53. https://doi.org/10.1126/science.adp1848

Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Chen, X., Choromanski, K., ... & Zitkovich, B. (2023). *RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control.* (arXiv:2307.15818). arXiv. https://arxiv.org/abs/2307.15818

Budagam, D., Kumar, A., Khoshnoodi, M., KJ, S., Jain, V., & Chadha, A. (2024). *Hierarchical Prompting Taxonomy: A Universal Evaluation Framework for Large Language Models Aligned With Human Cognitive Principles* (arXiv:2406.12644; Version 4). arXiv. https://doi.org/10.48550/arXiv.2406.12644

Cao, Z., Long, M., Wang, J., & Yu, P. S. (2017). *HashNet: Deep Learning to Hash by Continuation* (arXiv:1702.00758). arXiv. https://doi.org/10.48550/arXiv.1702.00758

Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. de O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., Ray, A., Puri, R., Krueger, G., Petrov, M., Khlaaf, H., Sastry, G., Mishkin, P., Chan, B., Gray, S., ... Zaremba, W. (2021). *Evaluating Large Language Models Trained on Code* (arXiv:2107.03374). arXiv. https://doi.org/10.48550/arXiv.2107.03374

Chiang, W.-L., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., Zhang, H., Zhu, B., Jordan, M., Gonzalez, J. E., & Stoica, I. (2024). *Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference* (arXiv:2403.04132). arXiv. https://doi.org/10.48550/arXiv.2403.04132

Chollet, F., Knoop, M., Kamradt, G., & Landers, B. (2025). *ARC Prize 2024: Technical Report* (arXiv:2412.04604). arXiv. https://doi.org/10.48550/arXiv.2412.04604

Chung, J. S., Senior, A., Vinyals, O., & Zisserman, A. (2017). "Lip Reading Sentences in the Wild." *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3444–53. https://doi.org/10.1109/CVPR.2017.367

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., & Schulman, J. (2021). *Training Verifiers to Solve Math Word Problems* (arXiv:2110.14168). arXiv. https://doi.org/10.48550/arXiv.2110.14168

Driess, D., Xia, F., Sajjadi, M. S. M., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., Huang, W., Chebotar, Y., Sermanet, P., Duckworth, D., Levine, S., Vanhoucke, V., Hausman, K., Toussaint, M., Greff, K., ... Florence, P. (2023). *PaLM-E: An Embodied Multimodal Language Model* (arXiv:2303.03378). arXiv. https://doi.org/10.48550/arXiv.2303.03378

Fang, H., Grotz, M., Pumacay, W., Wang, Y. R., Fox, D., Krishna, R., & Duan, J. (2025). *SAM2Act: Integrating Visual Foundation Model With a Memory Architecture for Robotic Manipulation* (arXiv:2501.18564). arXiv. https://doi.org/10.48550/arXiv.2501.18564

Fattorini, L., Maslej, N., Perrault, R., Parli, V., Etchemendy, J., Shoham, Y., & Ligett, K. (2024). *The Global AI Vibrancy Tool* (arXiv:2412.04486). arXiv. https://doi.org/10.48550/arXiv.2412.04486

Glazer, E., Erdil, E., Besiroglu, T., Chicharro, D., Chen, E., Gunning, A., Olsson, C. F., Denain, J.-S., Ho, A., Santos, E. de O., Järviniemi, O., Barnett, M., Sandler, R., Vrzala, M., Sevilla, J., Ren, Q., Pratt, E., Levine, L., Barkley, G., ... Wildon, M. (2024). *FrontierMath: A Benchmark for Evaluating Advanced Mathematical Reasoning in AI* (arXiv:2411.04872). arXiv. https://doi.org/10.48550/arXiv.2411.04872

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2021). *Measuring Massive Multitask Language Understanding* (arXiv:2009.03300). arXiv. https://doi.org/10.48550/arXiv.2009.03300

Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., & Steinhardt, J. (2021). *Measuring Mathematical Problem Solving With the MATH Dataset* (arXiv:2103.03874). arXiv. https://doi.org/10.48550/arXiv.2103.03874

Hsieh, C.-P., Sun, S., Kriman, S., Acharya, S., Rekesh, D., Jia, F., Zhang, Y., & Ginsburg, B. (2024). *RULER: What's the Real Context Size of Your Long-Context Language Models?* (arXiv:2404.06654). arXiv. https://doi.org/10.48550/arXiv.2404.06654

Huang, Q., Vora, J., Liang, P., & Leskovec, J. (2024). *MLAgentBench: Evaluating Language Agents on Machine Learning Experimentation* (arXiv:2310.03302). arXiv. https://doi.org/10.48550/arXiv.2310.03302

Islam, P., Kannappan, A., Kiela, D., Qian, R., Scherrer, N., & Vidgen, B. (2023). *FinanceBench: A New Benchmark for Financial Question Answering* (arXiv:2311.11944). arXiv. https://doi.org/10.48550/arXiv.2311.11944

James, S., Ma, Z., Arrojo, D. R., & Davison, A. J. (2019). *RLBench: The Robot Learning Benchmark & Learning Environment* (arXiv:1909.12271; Version 1). arXiv. https://doi.org/10.48550/arXiv.1909.12271

Jia, X., Yang, Z., Li, Q., Zhang, Z., & Yan, J. (2024). *Bench2Drive: Towards Multi-ability Benchmarking of Closed-Loop End-to-End Autonomous Driving* (arXiv:2406.03877). arXiv. https://doi.org/10.48550/arXiv.2406.03877

Jimenez, C. E., Yang, J., Wettig, A., Yao, S., Pei, K., Press, O., & Narasimhan, K. (2024). *SWE-bench: Can Language Models Resolve Real-World GitHub Issues?* (arXiv:2310.06770). arXiv. https://doi.org/10.48550/arXiv.2310.06770

Jones, C. R., & Bergen, B. K. (2024). *People Cannot Distinguish GPT-4 From a Human in a Turing Test* (arXiv:2405.08007). arXiv. https://doi.org/10.48550/arXiv.2405.08007

Karnchanachari, N., Geromichalos, D., Tan, K. S., Li, N., Eriksen, C., Yaghoubi, S., Mehdipour, N., Bernasconi, G., Fong, W. K., Guo, Y., & Caesar, H. (2024). *Towards Learning-Based Planning: The nuPlan Benchmark for Real-World Autonomous Driving* (arXiv:2403.04133). arXiv. https://doi.org/10.48550/arXiv.2403.04133

Kusupati, A., Bhatt, G., Rege, A., Wallingford, M., Sinha, A., Ramanujan, V., Howard-Snyder, W., Chen, K., Kakade, S., Jain, P., & Farhadi, A. (2024). *Matryoshka Representation Learning* (arXiv:2205.13147). arXiv. https://doi.org/10.48550/arXiv.2205.13147

Leal, I., Choromanski, K., Jain, D., Dubey, A., Varley, J., Ryoo, M., Lu, Y., Liu, F., Sindhwani, V., Vuong, Q., Sarlos, T., Oslund, K., Hausman, K., & Rao, K. (2023). *SARA-RT: Scaling Up Robotics Transformers With Self-Adaptive Robust Attention* (arXiv:2312.01990). arXiv. https://doi.org/10.48550/arXiv.2312.01990

Li, K., Wang, Y., He, Y., Li, Y., Wang, Y., Liu, Y., Wang, Z., Xu, J., Chen, G., Luo, P., Wang, L., & Qiao, Y. (2024). *MVBench: A Comprehensive Multi-modal Video Understanding Benchmark* (arXiv:2311.17005). arXiv. https://doi.org/10.48550/arXiv.2311.17005

Li, T., Chiang, W.-L., Frick, E., Dunlap, L., Wu, T., Zhu, B., Gonzalez, J. E., & Stoica, I. (2024). *From Crowdsourced Data to High-Quality Benchmarks: Arena-Hard and BenchBuilder Pipeline* (arXiv:2406.11939). arXiv. https://doi.org/10.48550/arXiv.2406.11939

Li, X., Mata, C., Park, J., Kahatapitiya, K., Jang, Y. S., Shang, J., Ranasinghe, K., Burgert, R., Cai, M., Lee, Y. J., & Ryoo, M. S. (2025). *LLaRA: Supercharging Robot Learning Data for Vision-Language Policy* (arXiv:2406.20095). arXiv. https://doi.org/10.48550/arXiv.2406.20095

Liu, X., Yu, H., Zhang, H., Xu, Y., Lei, X., Lai, H., Gu, Y., Ding, H., Men, K., Yang, K., Zhang, S., Deng, X., Zeng, A., Du, Z., Zhang, C., Shen, S., Zhang, T., Su, Y., Sun, H., ... Tang, J. (2023). *AgentBench: Evaluating LLMs as Agents* (arXiv:2308.03688). arXiv. https://doi.org/10.48550/arXiv.2308.03688

Liu, X., Zhang, T., Gu, Y., Iong, I. L., Xu, Y., Song, X., Zhang, S., Lai, H., Liu, X., Zhao, H., Sun, J., Yang, X., Yang, Y., Qi, Z., Yao, S., Sun, X., Cheng, S., Zheng, Q., Yu, H., ... Tang, J. (2024). *VisualAgentBench: Towards Large Multimodal Models as Visual Foundation Agents* (arXiv:2408.06327). arXiv. https://doi.org/10.48550/arXiv.2408.06327

Mialon, G., Fourrier, C., Swift, C., Wolf, T., LeCun, Y., & Scialom, T. (2023). *GAIA: A Benchmark for General AI Assistants* (arXiv:2311.12983). arXiv. https://doi.org/10.48550/arXiv.2311.12983

Mitchell, M. (2024). "The Turing Test and Our Shifting Conceptions of Intelligence." *Science*, 385(6710), eadq9356. https://www.science.org/doi/10.1126/science.adq9356

Muennighoff, N., Tazi, N., Magne, L., & Reimers, N. (2023). *MTEB: Massive Text Embedding Benchmark* (arXiv:2210.07316). arXiv. https://doi.org/10.48550/arXiv.2210.07316

Ni, J., Xue, F., Yue, X., Deng, Y., Shah, M., Jain, K., Neubig, G., & You, Y. (2024). *MixEval: Deriving Wisdom of the Crowd From LLM Benchmark Mixtures* (arXiv:2406.06565). arXiv. https://doi.org/10.48550/arXiv.2406.06565

O'Neill, A., Rehman, A., Gupta, A., Maddukuri, A., Gupta, A., Padalkar, A., Lee, A., Pooley, A., Gupta, A., Mandlekar, A., Jain, A., Tung, A., Bewley, A., Herzog, A., Irpan, A., Khazatsky, A., Rai, A., Gupta, A., ... Lin, Z. (2024). *Open X-Embodiment: Robotic Learning Datasets and RT-X Models* (arXiv:2310.08864). arXiv. https://doi.org/10.48550/arXiv.2310.08864

Phan, L., Gatti, A., Han, Z., Li, N., Hu, J., Zhang, H., Zhang, C. B. C., Shaaban, M., Ling, J., Shi, S., Choi, M., Agrawal, A., Chopra, A., Khoja, A., Kim, R., Ren, R., Hausenloy, J., Zhang, O., Mazeika, M., ... Hendrycks, D. (2025). *Humanity's Last Exam* (arXiv:2501.14249). arXiv. https://doi.org/10.48550/arXiv.2501.14249

Rein, D., Hou, B. L., Stickland, A. C., Petty, J., Pang, R. Y., Dirani, J., Michael, J., & Bowman, S. R. (2023). *GPQA: A Graduate-Level Google-Proof Q&A Benchmark* (arXiv:2311.12022). arXiv. https://doi.org/10.48550/arXiv.2311.12022

Reuel, A., Hardy, A., Smith, C., Lamparth, M., Hardy, M., & Kochenderfer, M. J. (2024). *BetterBench: Assessing AI Benchmarks, Uncovering Issues, and Establishing Best Practices* (arXiv:2411.12990). arXiv. https://doi.org/10.48550/arXiv.2411.12990

Turing, A. M. (2009). Computing Machinery and Intelligence. In Epstein, R., Roberts, G., & Beber, G., eds., *Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer* (23–65). Springer Netherlands. https://doi.org/10.1007/978-1-4020-6710-5_3

Valmeekam, K., Stechly, K., & Kambhampati, S. (2024). *LLMs Still Can't Plan; Can LRMs? A Preliminary Evaluation of OpenAI's o1 on PlanBench* (arXiv:2409.13373). arXiv. https://doi.org/10.48550/arXiv.2409.13373

Wijk, H., Lin, T., Becker, J., Jawhar, S., Parikh, N., Broadley, T., Chan, L., Chen, M., Clymer, J., Dhyani, J., Ericheva, E., Garcia, K., Goodrich, B., Jurkovic, N., Kinniment, M., Lajko, A., Nix, S., Sato, L., Saunders, W., ... Barnes, E. (2024). *RE-Bench: Evaluating Frontier AI R&D Capabilities of Language Model Agents Against Human Experts* (arXiv:2411.15114). arXiv. https://doi.org/10.48550/arXiv.2411.15114

Xia, Z., Li, J., Lin, Z., Wang, X., Wang, Y., & Yang, M.-H. (2024). *OpenAD: Open-World Autonomous Driving Benchmark for 3D Object Detection* (arXiv:2411.17761). arXiv. https://doi.org/10.48550/arXiv.2411.17761

Xu, C., Guan, S., Greene, D., & Kechadi, M.-T. (2024). *Benchmark Data Contamination of Large Language Models: A Survey* (arXiv:2406.04244). arXiv. https://doi.org/10.48550/arXiv.2406.04244

Yang, X., Sun, K., Xin, H., Sun, Y., Bhalla, N., Chen, X., Choudhary, S., Gui, R. D., Jiang, Z. W., Jiang, Z., Kong, L., Moran, B., Wang, J., Xu, Y. E., Yan, A., Yang, C., Yuan, E., Zha, H., Tang, N., ... Dong, X. L. (2024). *CRAG—Comprehensive RAG Benchmark* (arXiv:2406.04744). arXiv. https://doi.org/10.48550/arXiv.2406.04744

Yen, H., Gao, T., Hou, M., Ding, K., Fleischer, D., Izsak, P., Wasserblat, M., & Chen, D. (2025). *HELMET: How to Evaluate Long-Context Language Models Effectively and Thoroughly* (arXiv:2410.02694). arXiv. https://doi.org/10.48550/arXiv.2410.02694

Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., Wei, C., Yu, B., Yuan, R., Sun, R., Yin, M., Zheng, B., Yang, Z., Liu, Y., Huang, W., ... Chen, W. (2024). *MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI* (arXiv:2311.16502). arXiv. https://doi.org/10.48550/arXiv.2311.16502

Zellers, R., Bisk, Y., Farhadi, A., & Choi, Y. (2019). *From Recognition to Cognition: Visual Commonsense Reasoning* (arXiv:1811.10830). arXiv. https://doi.org/10.48550/arXiv.1811.10830

Zhang, H., Da, J., Lee, D., Robinson, V., Wu, C., Song, W., Zhao, T., Raja, P., Zhuang, C., Slack, D., Lyu, Q., Hendryx, S., Kaplan, R., Lunati, M., & Yue, S. (2024). *A Careful Examination of Large Language Model Performance on Grade School Arithmetic* (arXiv:2405.00332). arXiv. https://doi.org/10.48550/arXiv.2405.00332