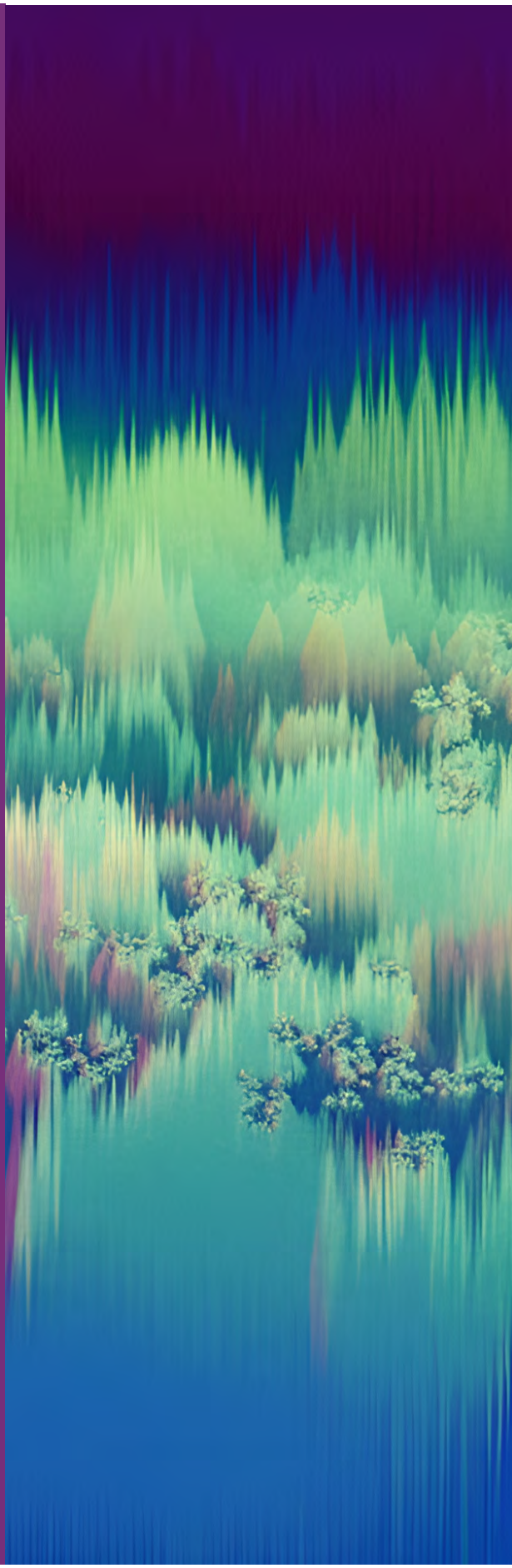
2

Technical Performance

Overview

AI models improved rapidly in 2025, with benchmark scores rising across language, reasoning, coding, and math. However, evaluations are being outpaced by the progress they were built to measure, and benchmarks face growing questions about their reliability. Even with those limitations, a clear pattern emerges: the gap between top models is shrinking. This narrowing extends geographically, as the distance between top U.S. and Chinese models has closed almost completely. With capability no longer a clear differentiator, competitive pressure is shifting toward cost, reliability, and real-world usefulness. In professional domains, evaluations in tax, legal reasoning, and corporate finance show stronger performance in some areas than others. The range of what AI systems can do is also expanding. AI agents are improving, but still fail roughly one in three attempts. Video generation models are no longer just producing realistic-looking content; some are beginning to learn how the physical world actually works, progress that could help bring AI into physical spaces. That transition is still early, as robots struggle in unstructured environments, though autonomous vehicles are a notable exception, having reached mass-scale deployment with promising early safety records. Overall, AI's technical advancement is a story of wonder and speed, faster than many of the evaluation, governance, and adoption frameworks discussed in later chapters.

Contents



Chapter Highlights

1. **AI capability is outpacing the benchmarks designed to measure it, and surpassing human-level performance.** Frontier models gained 30 percentage points in a single year on Humanity's Last Exam, a benchmark built to be hard for AI and favorable to human experts. Evaluations intended to be challenging for years are saturated in months, compressing the window in which benchmarks remain useful for tracking progress.

2. **Top model performance is converging, with 4 companies now clustered within 25 Elo points (inspired by chess ratings) when rated against one another by human voting in the Arena Leaderboard and benchmark.** As of March 2026, Anthropic (1,503), xAI (1,495), Google (1,494), OpenAI (1,481), Alibaba (1,449), and DeepSeek (1,424) all occupy the top tier of the Arena Elo ratings, shifting competitive pressure toward cost, reliability, and domain-specific performance.

3. **The open-weight performance gap reopened in 2025 after briefly closing the year before.** As of March 2026, the top closed-weight model leads the top open-weight model by 3.3%, up from 0.5% in August 2024. Six of the top 10 models on the Arena Leaderboard are now closed-weight.

4. **The U.S.-China AI model performance gap has effectively closed.** U.S. and Chinese models have traded places at the top of performance rankings multiple times since early 2025. In February 2025, DeepSeek-R1 briefly matched the top U.S. model. As of March 2026, the top U.S. model leads by 2.7%, with a gap that fluctuated over the past year while remaining in the single digits.

5. **The benchmarks used to measure AI progress face growing reliability and gaming concerns, with error rates up to 42% on widely used evaluations.** A review found invalid question rates ranging from 2% on MMLU Math to 42% on GSM8K. Separate research suggests that Arena leaderboard standing may partly reflect adaptation to the platform rather than general capability.

6. **Video generation models are starting to capture how objects behave.** Google DeepMind's Veo 3, tested across more than 18,000 generated videos, demonstrated abilities like simulating buoyancy and solving mazes without being trained on those tasks.

7. **AI models can win a gold medal at the International Mathematical Olympiad but still can't reliably tell time, illustrating what researchers call jagged intelligence.** Gemini Deep Think scored 35 points (gold) at the 2025 IMO, working end to end in natural language within the 4.5-hour time limit, up from the 28-point silver achieved in 2024. On ClockBench, the top model read analog clocks correctly 50.1% of the time, compared with 90.1% for humans.

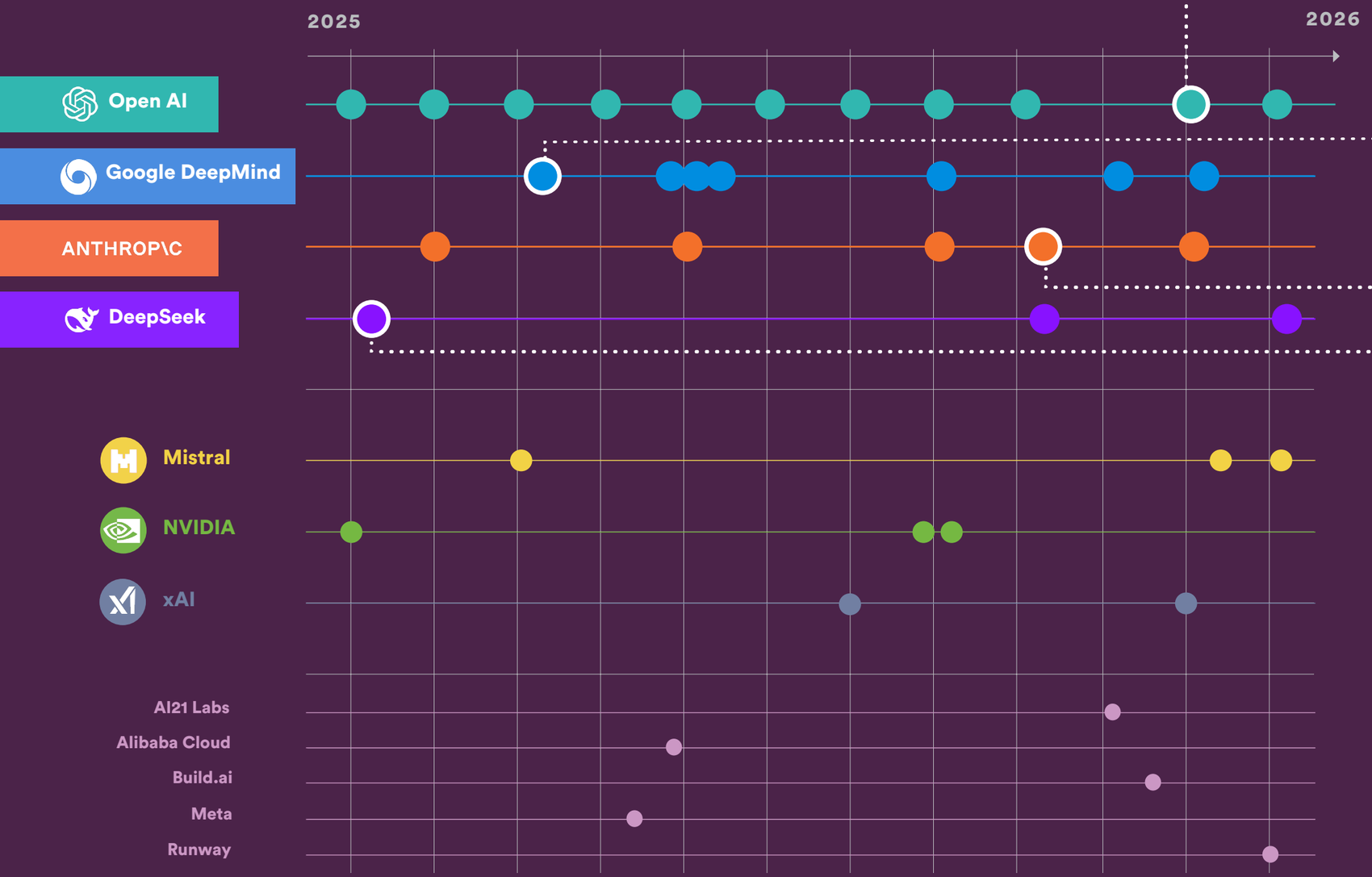
8. **AI models are expanding into professional domains, showing performance ranging from 60 to 90% in evaluations in tax, mortgage processing, corporate finance, and legal reasoning.** The performance of the top 15 models is separated by as little as 3 percentage points in each benchmark. These kinds of domains where high competency and reliability are required remain a great challenge for AI models.

9 **AI agents advanced from answering questions to completing tasks in 2025, though they still fail roughly one in three attempts on structured benchmarks.** On OSWorld, which tests agents on real computer tasks across operating systems, accuracy rose from roughly 12% to 66.3%, within 6 percentage points of human performance.

10 **Robots still fail at most household tasks, even as they excel in controlled environments.** Robots succeed in only 12% of real household tasks, highlighting how far AI is from mastering the physical world. On RLBench, robotic manipulation in software-based simulations has reached 89.4% success, but the gap between predictable lab settings and unpredictable household environments is wide.

11 **Autonomous vehicles reached mass-scale deployment in 2025.** Waymo reached approximately 450,000 weekly trips across five U.S. cities. In China, Apollo Go completed 11 million fully driverless rides, a 175% year-over-year increase. European operators are active but comparable deployment data is not publicly available, limiting the global picture. Deployments so far are in areas with generally favorable weather and humans are available off-site to take over when necessary.

Timeline: Significant Model Releases



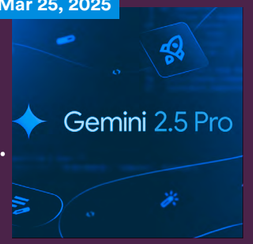
Nov 12, 2025

GPT-5.1

Multimodal

GPT-5.1 delivered significant improvements in both capability and efficiency. It runs faster than GPT-5, scores higher on coding and reasoning benchmarks (e.g., ~76.3% on SWE-bench Verified vs. ~72.8%), and dynamically adjusts reasoning effort based on task complexity.

Mar 25, 2025



Gemini 2.5 Pro

Multimodal

Gemini 2.5 was a major update from 2.0 that expanded context to 1M tokens, delivered strong reasoning and coding results (e.g., ~63.8% on SWE-Bench Verified), and reached #1 on LMArena.

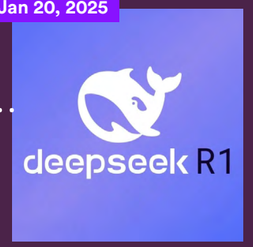
Sep 29, 2025

Claude Sonnet 4.5

LLM

Claude Sonnet 4.5 marked a major jump in real-world capability—hitting 61.4% on OSWorld computer-use tasks and 77.2%+ on SWE-bench Verified. It also shipped new tooling, including checkpoints, a VS Code extension, memory editing, and the Claude Agent SDK, enabling developers to build long-running autonomous workflows.

Jan 20, 2025



DeepSeek-R1

LLM

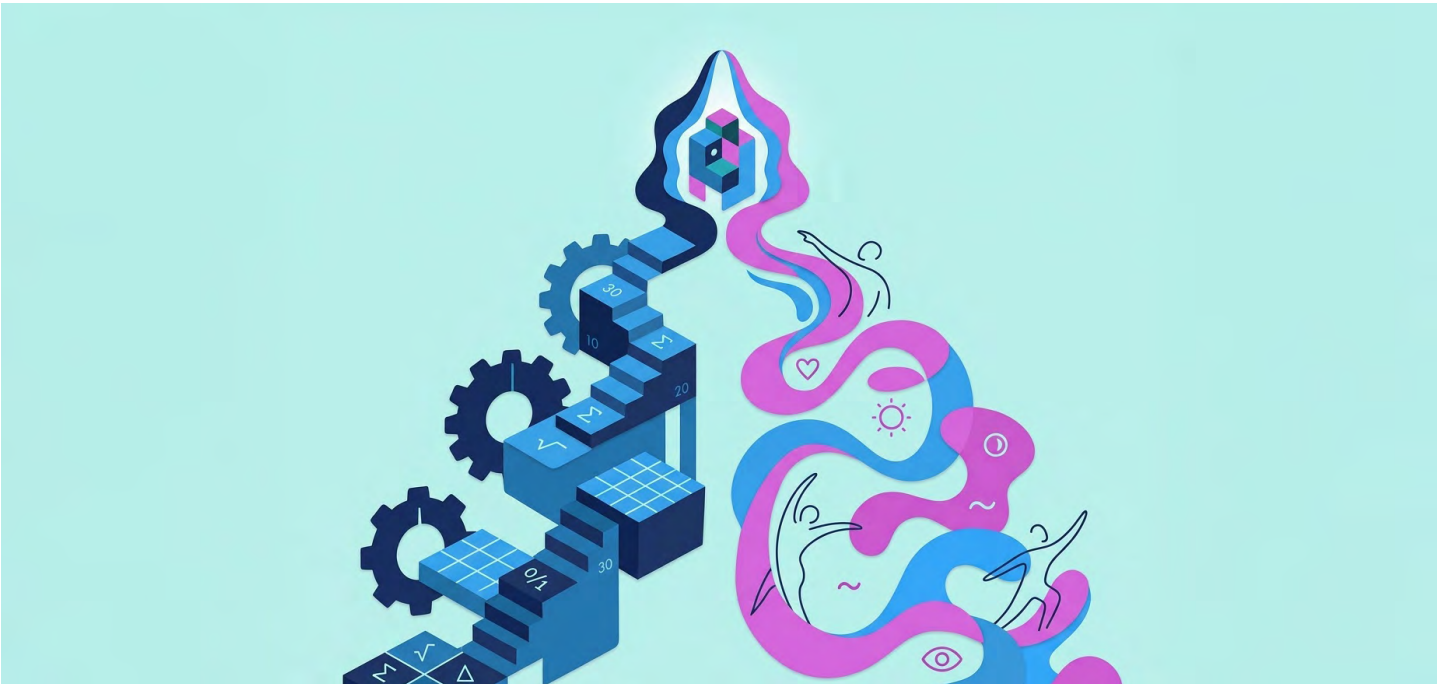
DeepSeek R1 introduced a reinforcement-learning approach called GRPO, which trains reasoning ability without relying on labeled data or a separate critic model. By comparing groups of generated outputs against predefined rules, the method reduces training complexity. The model's strong performance relative to higher-cost systems led some investors to reassess the competitive dynamics of the AI sector. Following its release, major U.S. technology stocks experienced a temporary decline of over one trillion dollars in market value amid concerns that more efficient training methods could affect existing business models.

2.1 Overall Performance Trends

This section examines patterns in AI performance, from the pace at which models are reaching human-level baselines to how competition among leading models and countries has narrowed. It also assesses where the tools used to measure this progress are themselves falling short. To enable comparison across diverse valuation tasks, performance metrics are scaled to a common reference point. The scaling methodology, developed by the AI Index team, calibrates each benchmark so that the best-performing model in a given year is measured as a percentage of the established human baseline for that task. For example, using this approach, a value of 105% indicates that a model performs 5% better than the human baseline. The benchmarks included in this analysis represent tasks that can be structurally evaluated. It may not fully capture the breadth of capabilities required for real-world AI deployment. The Benchmarking AI subsection later in this section explores these limitations in detail.

Technical Performance Benchmarks vs. Human Performance

AI performance continued to improve across a broad set of benchmark categories in 2025, with some of the largest gains appearing on tasks that were well below human baseline performance just a few years ago (Figure 2.1.1). Frontier systems now meet or exceed established human performance levels on long-running benchmarks, including ImageNet, SuperGLUE, and MMLU. Since last year's report, several benchmarks designed to test more advanced reasoning have reached or approached the human benchmark, including PhD-level science questions (GPQA Diamond), multimodal reasoning (MMMU), and mathematical reasoning (AIME). Models are still performing below the baseline in the areas of autonomous software engineering (SWE-bench Verified) and agent-based multimodal computer use (OSWorld), but the pace of improvement is rapidly accelerating. On SWE-bench Verified, for example, performance rose from approximately 60% in 2024 to close to 100% in 2025.



Select AI Index technical performance benchmarks vs. human performance
Source: AI Index, 2026 | Chart: 2026 AI Index report

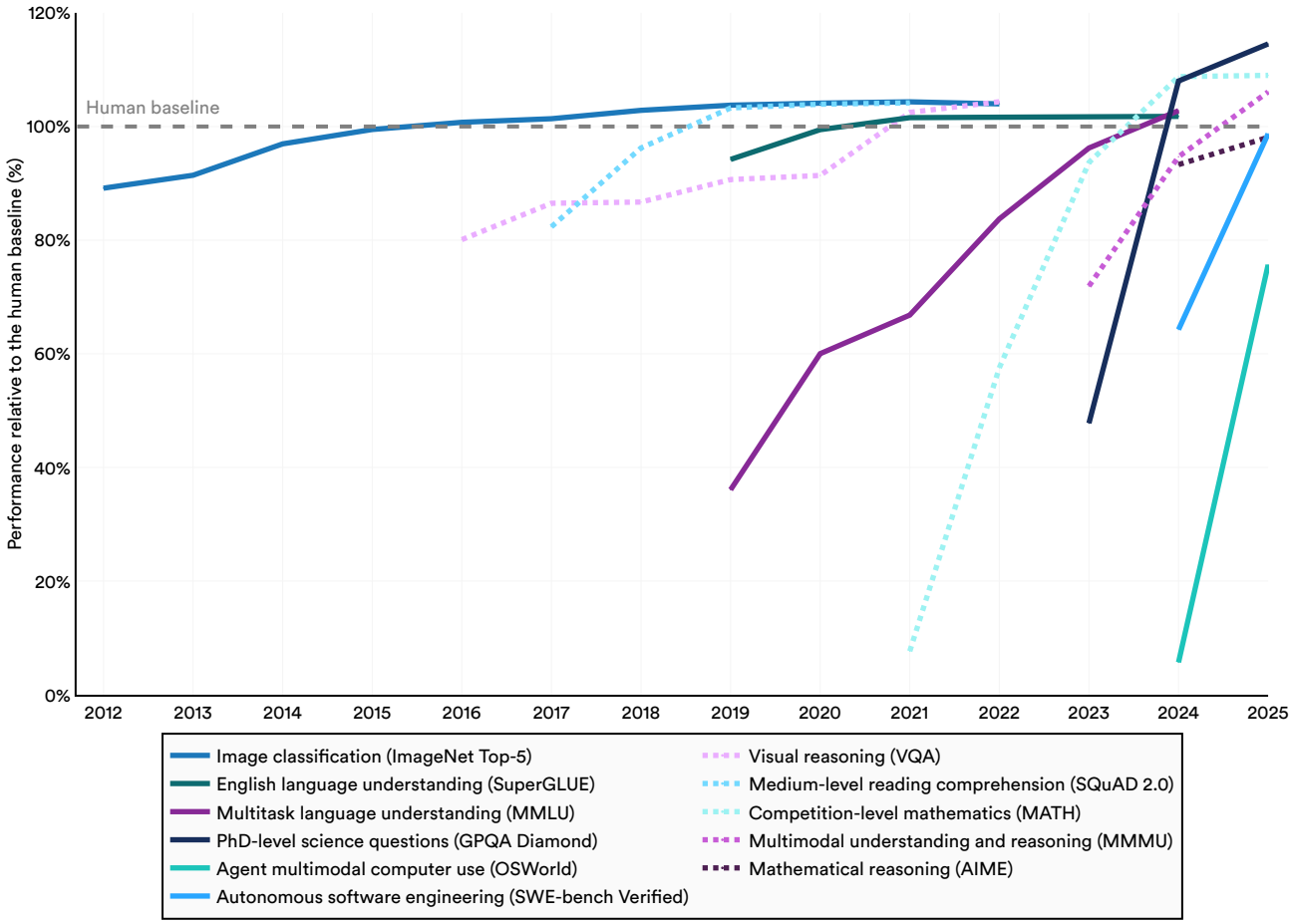


Figure 2.1.1[1]

Closed- vs. Open-Weight Models

The performance gap between leading closed-weight and open-weight models has fluctuated over the past three years, with open-weight systems closing in and then falling behind as new proprietary models are released (Figure 2.1.2). In May 2023, the leading closed-weight model (GPT-4-0314) outperformed the top open-weight model (Vicuna-13B) by 174 points (15.2%) on the Arena Leaderboard. Stronger open-weight releases, including Mixtral, WizardLM, and Llama-3.1-405B, narrowed the gap to just 7 points (0.5%) by August 2024. Over the past year, that trend reversed with the arrival of new closed-weight frontier systems such as o1-preview and Gemini 2.5 Pro. As of March 2026, the top closed-weight model, Claude Opus 4.6 (1,503), led the top open-weight model GLM-5 (1,454) by 49 points (3.4%). While closed-weight models still lead, open-weight models are far more competitive than they were a few years ago.

1 In Figure 2.1.1, 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.

Performance of top closed vs. open models on the Arena
Source: Arena, 2026 | Chart: 2026 AI Index report

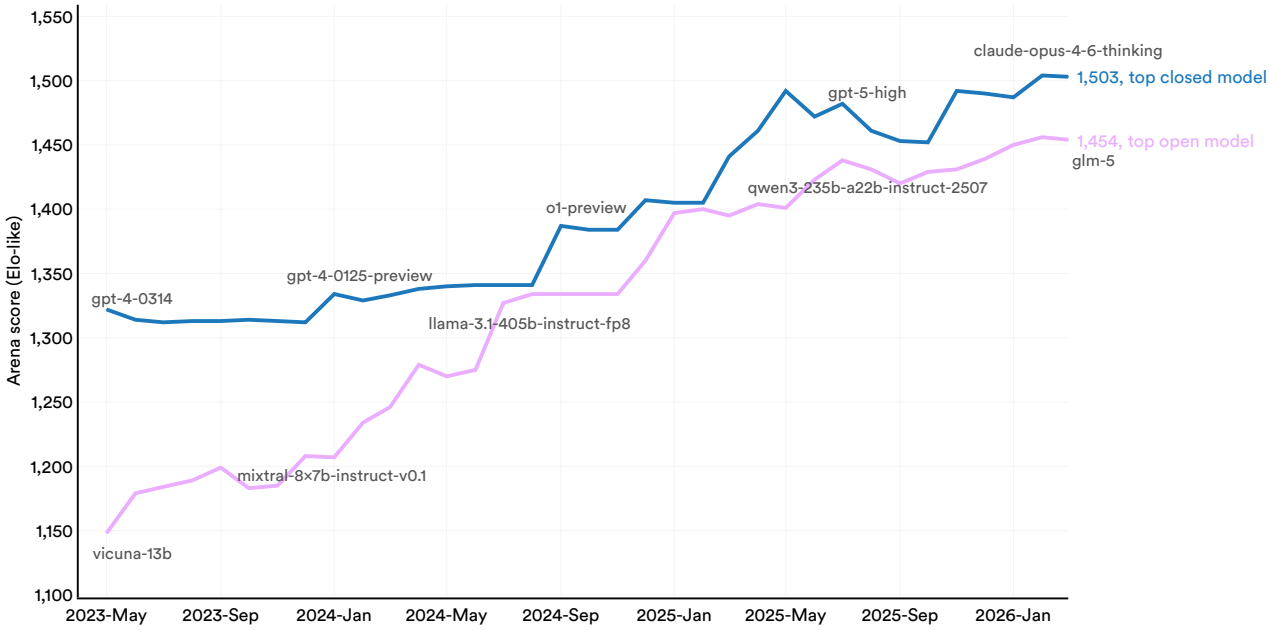


Figure 2.1.2[2]

US vs. China Technical Performance

The United States' substantial lead in 2023 shrank considerably by early 2025, and the performance gap has remained narrow since then (Figure 2.1.3). In February 2025, DeepSeek-R1 (1,400) trailed the leading U.S. model (o1-2024-12-17, 1,405) by just 5 Arena points (0.4%). As of March 2026, the top U.S. model (Claude Opus 4.6, 1,503) led the top Chinese model (Dola-Seed-2.0 Preview, 1,464) by 39 points (2.7%). Over the past year, the gap has fluctuated between near parity and low single digits. This convergence is particularly notable because it has emerged from two distinct development environments and institutional contexts, including the research dynamics examined in Chapter 1 and the investment patterns discussed in Chapter 4.

Performance of top United States vs. Chinese models on the Arena
Source: Arena, 2026 | Chart: 2026 AI Index report

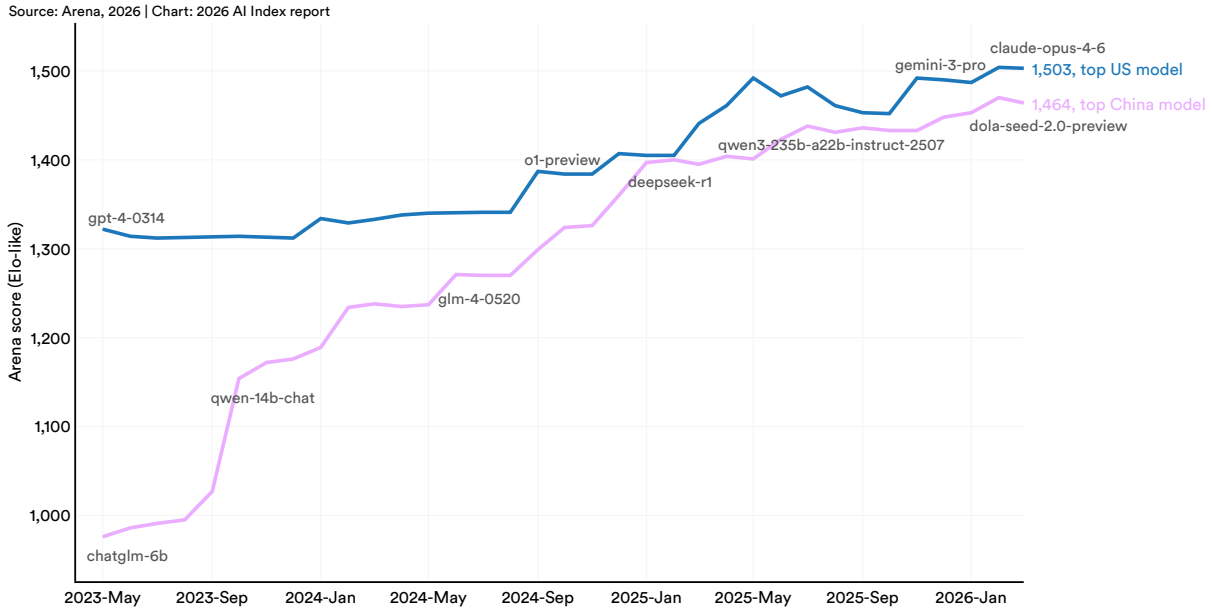


Figure 2.1.3[3]

2 Source: the Arena historical leaderboard (Public, Style Control On), exported in March 2026.

3 Source: the Arena historical leaderboard (Public, Style Control On), exported in March 2026.

Model Performance Converges at the Frontier

Frontier models became even more tightly clustered over the past year, as several companies moved into a very narrow performance band at the top of the Arena Leaderboard (Figure 2.1.4). In early 2023, OpenAI had a clear lead with its top model scoring 1,322 compared to Google's 1,117. This gap narrowed steadily through 2024 as Google, Anthropic, and others released stronger models. By February 2025, DeepSeek had briefly matched and surpassed the top U.S. systems on Arena. In last year's report, the top four models spanned roughly 97 points and, as of March 2026, the top four models are separated by fewer than 25 points. Anthropic leads at 1,503, followed closely by xAI (1,495), Google (1,494), and OpenAI (1,481). DeepSeek (1,424) and Alibaba (1,449) trail only modestly. Meta's Arena performance has flattened since early 2025, reflecting a slowdown in competitive releases, though newer models could be in the pipeline for 2026. As leading models become harder to distinguish on benchmark performance, factors such as cost, latency, reliability, and domain-specific optimization may play a greater role in user adoption.

Performance of top models on the Arena by select providers

Source: Arena, 2026 | Chart: 2026 AI Index report

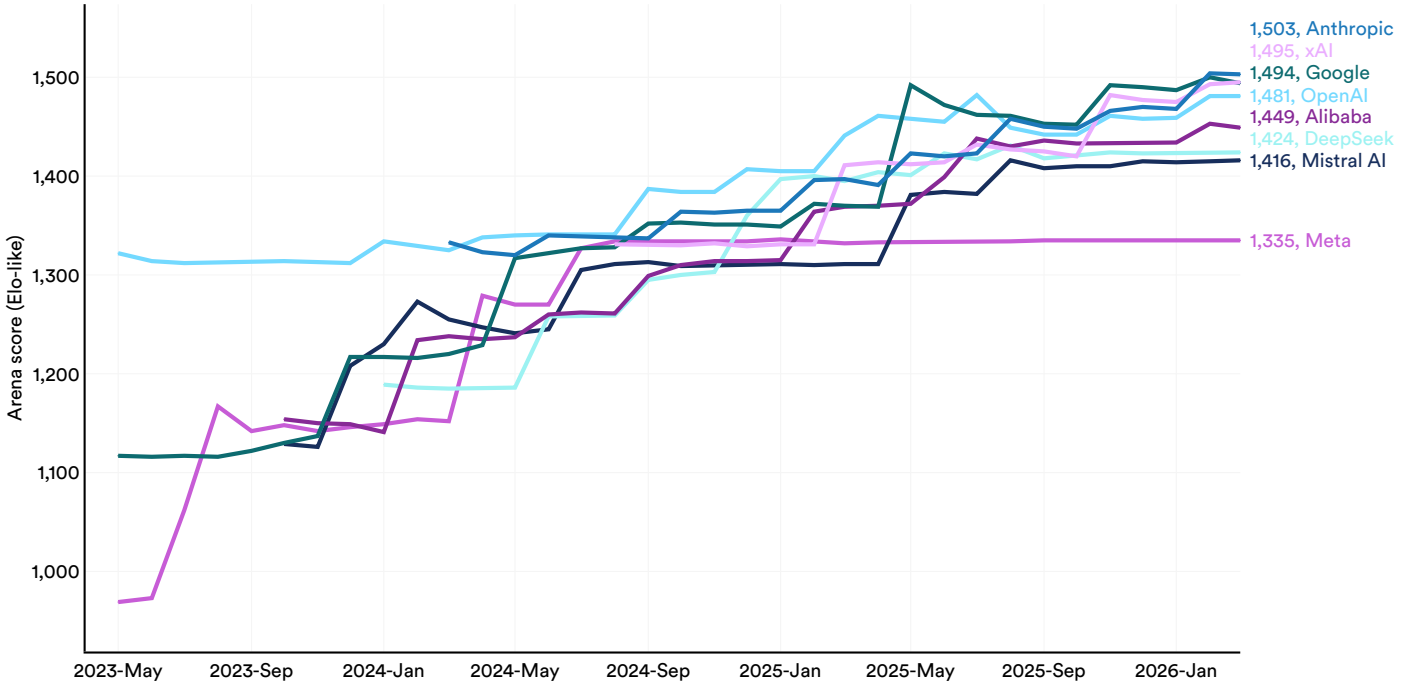


Figure 2.1.4[4]

Benchmarking AI

Benchmarks still anchor much of how AI's technical progress is measured, but their limitations are more visible. Since last year's report, the AI Index has expanded its analysis to examine where benchmarks remain useful, and where they fall short.

Several challenges highlighted in previous editions of this report persist. Benchmark saturation, where models reach scores so high that a test can no longer distinguish between them, remains a concern. Tests designed to be harder often remain useful for only a few years before systems surpass them. As Chapter 1 documents, reporting discrepancies continue, and the most capable modern models are now among

4 Source: the Arena historical leaderboard (Public, Style Control On), exported in March 2026.

the least transparent. The growing opacity and nonstandard prompting techniques make model-to-model comparisons unreliable, and third-party evaluations have documented cases where models perform more poorly in independent testing compared to developer-reported results. In addition, contamination—when models are exposed to test set data during training—can lead to falsely inflated scores. In 2025, Meta faced criticism that its Llama 4 model was optimized using specialized variants to improve leaderboard rankings and may have trained on benchmark test data, though the company disputed these claims. Additionally, audits of widely used benchmarks revealed that many remain poorly constructed, with inadequate documentation, no reporting of statistical significance, and a lack of replication scripts. Even when benchmark scores are technically valid, strong benchmark performance does not always translate to real-world utility.

Last year's report also highlighted how difficult it is to benchmark more complex, interactive forms of intelligence, which matter even more for current AI systems. Even though many benchmarks for multiagent coordination, human–AI interaction, tool-using agents, and physical-world robotics have been proposed (e.g., for robotic manipulation, embodied reasoning, and agentic tasks), they remain underdeveloped. These domains are inherently harder to standardize as physical tasks involve unpredictable environments, diverse hardware, and a range of valid approaches that resist repeatable scoring. Later sections of this chapter report on several of these benchmarks in detail.

The benchmarking landscape has seen several developments that extend beyond these recurring concerns. First, there is a growing case for evaluations that measure human–AI collaboration rather than AI performance in isolation. Most widely used benchmarks test systems without human involvement, even though many real deployments involve people supervising, steering, and integrating AI outputs. Recent work argues that the field should adopt centaur evaluations, assessments in which humans and AI jointly solve tasks, because these better reflect actual use and allow measurement of human-centered qualities like interpretability and helpfulness that conventional benchmarks ignore.

Second, new methods have emerged to address the invalid benchmark questions. A review by Stanford researchers identified the proportion of invalid questions across nine widely used benchmarks, with error rates ranging from 2% on MMLU Math to 42% on GSM8K (Figure 2.1.5). Truong et al., 2025, introduced a framework that uses statistical analysis of response patterns to flag problematic items for expert review, achieving up to 84% precision. Separately, Cheng et al., 2025, have proposed shifting toward "certificate-grade," peer-based evaluation frameworks that are community-governed, proctored systems with secure environments, continuously refreshed test items, and delayed result disclosure.

Invalid question detection across nine benchmarks

Source: Truong et al., 2025 | Chart: 2026 AI Index report

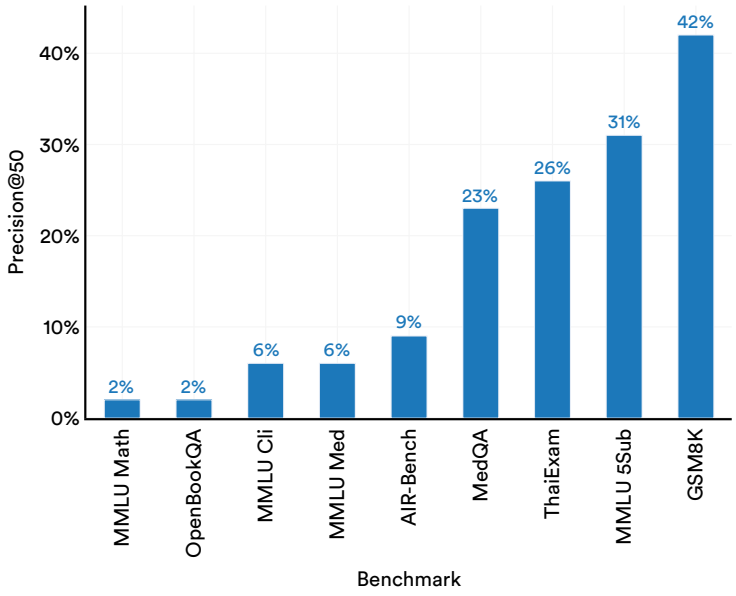
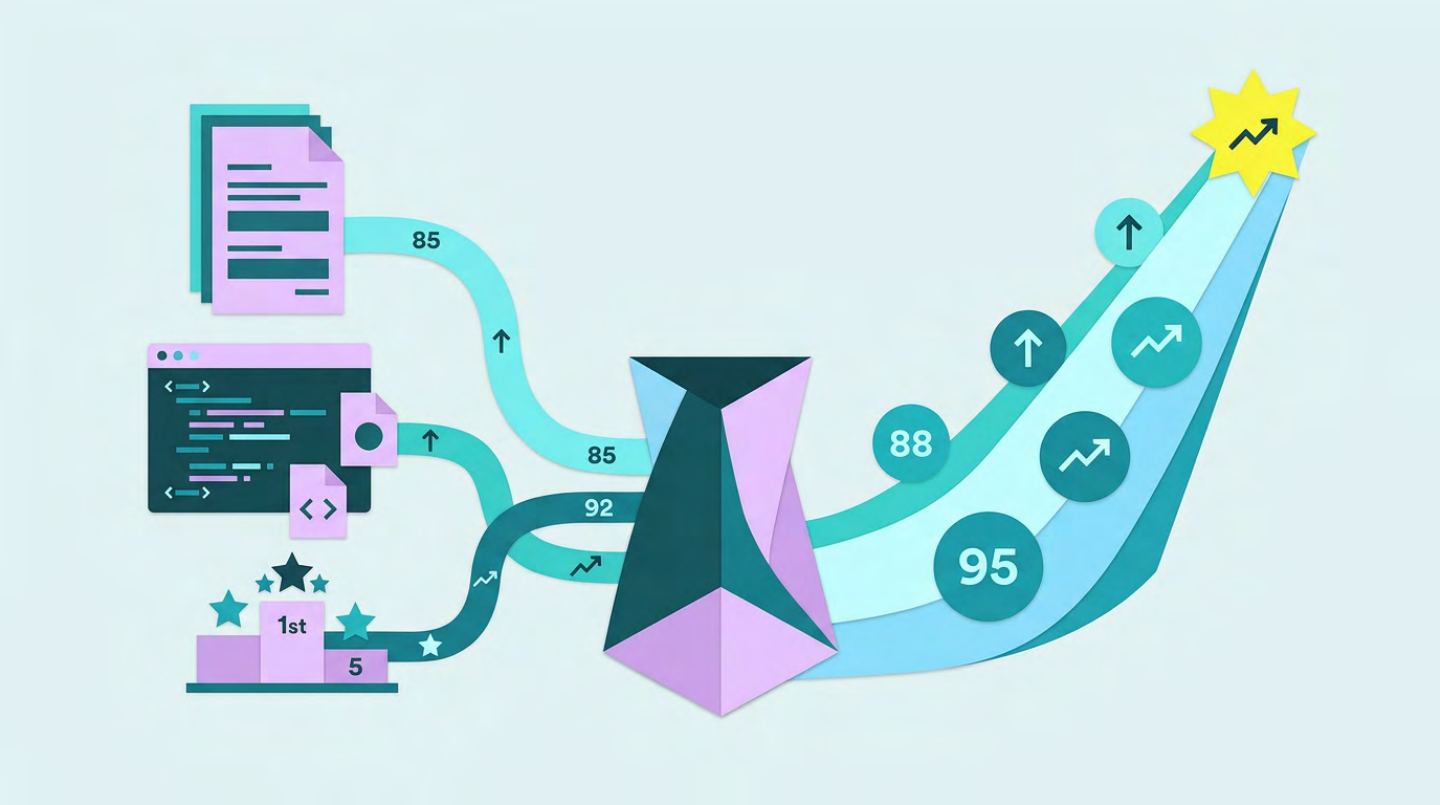


Figure 2.1.5

Third, questions have been raised about the reliability of popular public benchmarking platforms such as the Arena. A recent analysis (Singh et al., 2025) argues that platform dynamics could affect ranking accuracy. If providers are able to iterate on or swap model variants outside the public record, it introduces selection effects that make comparisons less straightforward. The study also points out data-access asymmetries and shows that additional Arena-style interaction data can improve performance on Arena-derived evaluations, suggesting that leaderboard standing may partly reflect adaptation to the platform rather than general capability alone.

Finally, while capability evaluations are widespread, assessments of social impacts remain fragmented and incomplete. Reul et al., 2025, found that developers' reporting of bias and environmental impact is often sparse and declining, while third-party researchers more rigorously assess harms such as harmful content and performance disparities. Because only developers can disclose key information about data, labor practices, and training infrastructure, current evaluation practices provide a strong picture of what models can do but a far weaker account of their societal consequences. Chapter 3 examines responsible AI evaluation in further detail.

In this chapter, the AI Index continues to report on benchmarks as key indicators of technical progress. Scores are sourced from leaderboards, public repositories, and company disclosures, including papers, blog posts, and product releases. The AI Index assumes that company-reported results are accurate. All scores reflect the state of the field as of early 2026; subsequent model releases may have surpassed these benchmarks.

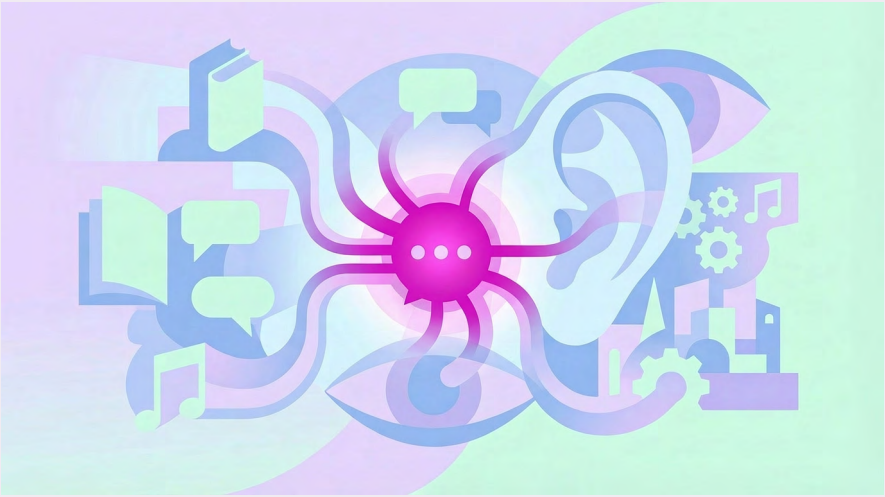


2.2 Language

Language understanding and generation continue to serve as foundational capabilities for modern AI systems. This section examines how models perform on tasks requiring comprehension of complex text, production of coherent responses, and execution of specialized language-based operations. The benchmarks also span general-purpose question answering to specific technical capabilities like function calling and text embedding.

Understanding

Language understanding benchmarks measure how well models can comprehend and reason over text across a broad range of subjects, from the humanities to highly technical materials. As performance has improved, evaluation has shifted toward harder test sets that are less susceptible to familiarity or memorization. The goal is to track where models are improving rather than reaching the upper limits of current benchmarking tools.



MMLU: Massive Multitask Language Understanding

MMLU remains a widely cited measure of broad knowledge across disciplines. Introduced in 2024, the MMLU-Pro benchmark assesses performance with over 12,000 questions and a 10-option, multiple-choice format designed to better test reasoning. This expanded answer base has a measurable impact on model performance evaluation. Compared to the original MMLU benchmark, model accuracy on MMLU-Pro typically drops by 16%–33%, which provides better differentiation between top models. For example, GPT-4o and GPT-4-Turbo appeared to have a 1% gap on standard MMLU, but on MMLU-Pro the spread widens to 9%. The newer benchmark's design reduces prompt sensitivity and strengthens reasoning evaluation. Previously, MMLU showed around 4%–5% sensitivity to prompt variations versus an estimated 2% with MMLU-Pro. In addition, reasoning methods such as chain-of-thought tend to yield much better performance on MMLU-Pro than direct answer strategies.

As of early 2026, top model performance on MMLU-Pro is tightly clustered, with the leading 15 models all scoring above 87% (Figure 2.2.1). Google's Gemini-3.1-Pro leads at 91.2%, followed by Gemini-3-Pro (Thinking) at 90.1% and GPT-o1 at 89.3%. Models that employ thinking strategies tend to appear higher in the rankings, outperforming their standard counterparts, which are grouped in the 87%–88% range. The overall spread between the top-ranked and 15th-ranked model is just over 4 percentage points, illustrating how competitive the frontier has become on broad knowledge tasks. This tight clustering is also consistent with the convergence pattern described in Section 2.1.

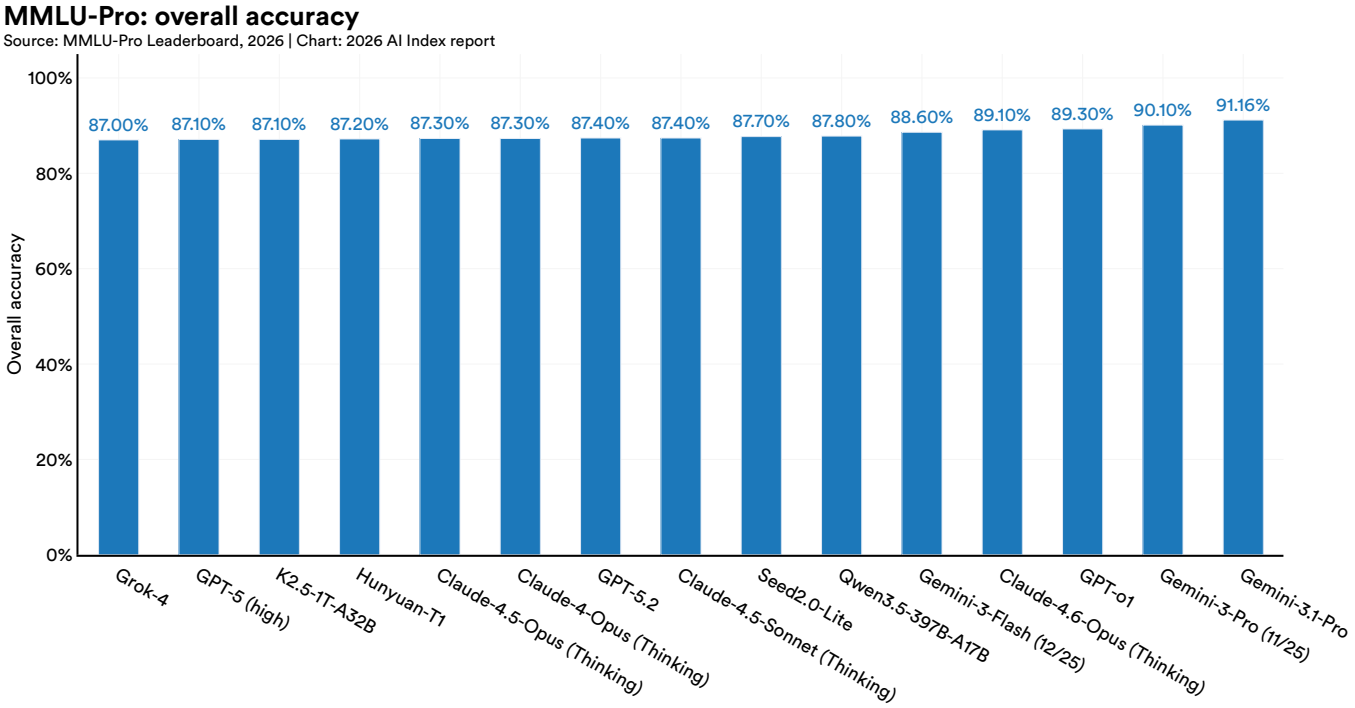


Figure 2.2.1[5]

Generation

Generation benchmarks focus on the quality of model outputs, looking at clarity, helpfulness, instruction-following, and style. Unlike knowledge-style tests, these evaluations often depend on human judgment since some dimensions are subjective and dependent on both the prompt and the user. Preference-based tests help measure that subjectivity and are a useful complement to traditional benchmarks for tracking how models perform in real-world settings.

Arena Leaderboard

The Arena (formerly LMArena) is an interactive platform with a community-driven ranking system that allows users to directly compare outputs of large language models (LLMs) on identical prompts and then vote on which they favor. Evaluations are blind to minimize bias toward particular model providers or architectures. By aggregating thousands of comparisons, the platform generates Elo ratings, a ranking system borrowed from chess. This approach emphasizes user experience and practical utility, capturing aspects of model quality that structured benchmarks cannot, including human judgment on real-world tasks.

The user-centered approach does have limitations, as preferences may not align with correctness and may not be fully representative of model use cases or contexts. Singh et al. (2025) highlight potential sources of bias, such as order bias, length bias, or style preferences, that are not correlated with output accuracy. As mentioned earlier, evaluations such as Arena can be a complementary view rather than an absolute score on model quality.

Elo ratings on the Text Arena are tightly clustered as of early 2026, with the top 15 models spanning roughly

5 Source: https://huggingface.co/spaces/TIGER-Lab/MMLU-Pro.

46 points (Figure 2.2.2). Claude-Opus-4-6-Thinking leads at approximately 1,510, followed closely by Gemini-3.1-Pro-Preview. The gap narrows further down the rankings, and confidence intervals overlap for many models. So, while Anthropic and Google models appear throughout the top ranks, no single model dominates the leaderboard.

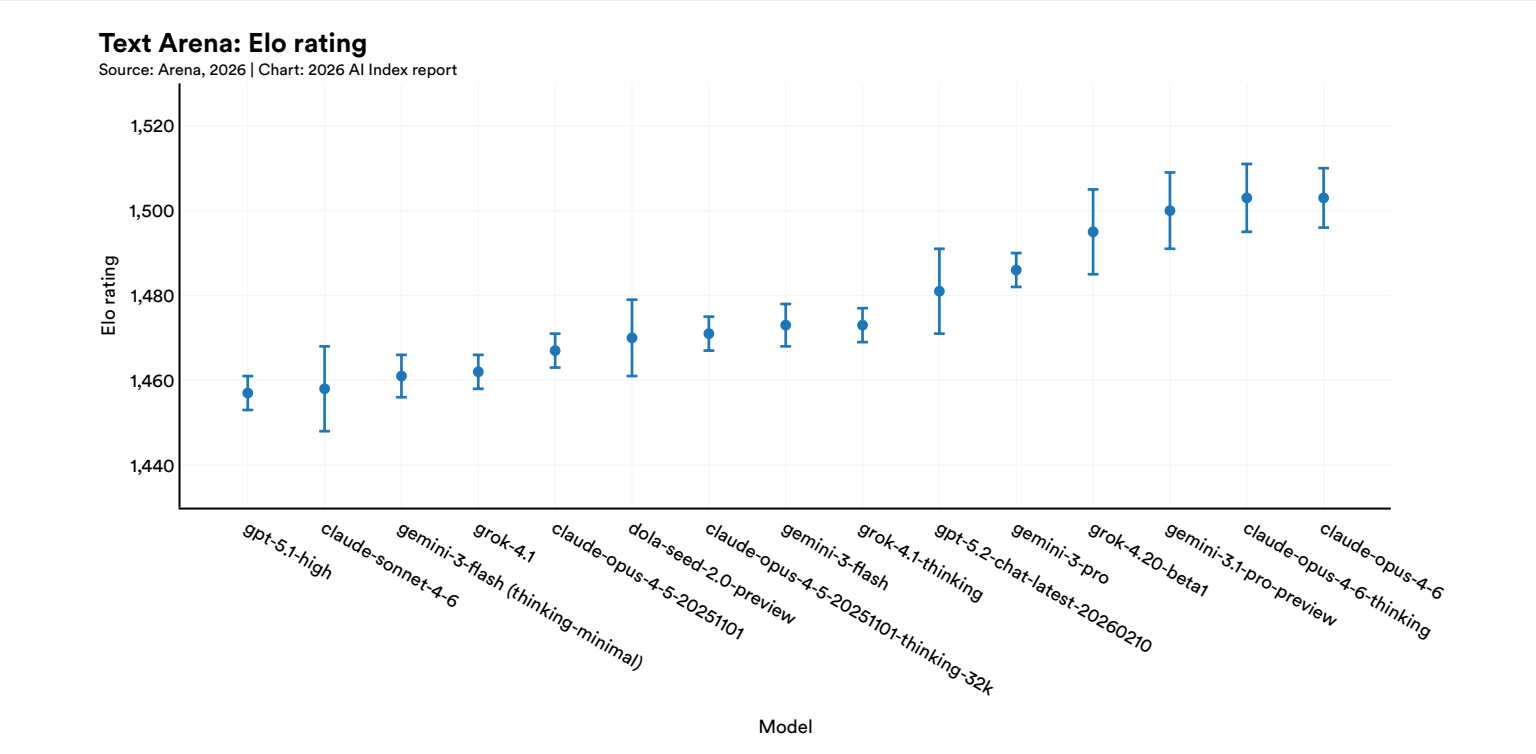


Figure 2.2.2[6]

Specialized Language Tasks

Beyond general understanding and generation, language models need to handle tasks that make them usable for practical deployment. Three key capabilities in deployed applications are retrieval-augmented generation (RAG), function calling, and text embedding. Benchmarks used to track these capabilities are particularly useful because they test fluency and whether models can operate as part of a larger system. It also makes it easier to compare models in settings where performance depends not just on the base model, but on issues such as retrieval quality or how outputs are parsed and executed.

RAG: Retrieval-Augmented Generation

Retrieval-augmented generation (RAG) provides a way for models to deliver accurate, up-to-date information beyond the knowledge encoded during training in model parameters. At inference time, RAG systems augment model responses with information retrieved from external sources.

Standard RAG pipelines retrieve individual text chunks based on query similarity, which can struggle when answering questions that require synthesizing information across documents. To address the problem, in 2024, Microsoft Research introduced Graph RAG, which enables more effective responses to queries by structuring source material into a knowledge graph and generating community summaries that capture high-

6 Source: https://arena.ai/leaderboard/text.

level themes. Other variants focus on improving multistep retrieval or reranking passages before generation. As expected, these choices in architecture involve trade-offs between answer quality, latency, and cost.

Context windows, discussed later in this section, have important implications for RAG systems. Extended context windows can support retrieval of more material, though that does not guarantee better performance since models have to parse through the information with reliable attention across the entire window.

Berkeley Function Calling Leaderboard

Function calling allows a model to use external tools and APIs by generating structured requests that another system can run, then folding the results back into its response. It is a foundational capability for agent frameworks, where models need to take actions or retrieve information beyond their training data.

The Berkeley Function Calling Leaderboard (BFCL) evaluates models on their function-calling ability and has evolved considerably since its initial release. The current iteration, BFCL V4, shifts the focus toward holistic agent evaluations. Agentic tasks account for 40% of the overall score, multiturn interactions are 30%, and the remainder is split across live, nonlive, and hallucination categories. The agentic component tests web search and memory while the multiturn component evaluates multistep dialogues. Earlier versions focused more narrowly on single-turn function calling.

The overall accuracy on the BFCL varies widely as of early 2026. The top 15 models span a roughly 21 percentage point range (Figure 2.2.3). Claude models occupy three of the top six positions, with Claude-Opus-4.5 leading at 77.5%. There is also a performance distinction with evaluation modes, showing the trade-offs between general capability and task-specific optimization. For example, Grok-4-0709 scores 63% in prompt mode but drops to 61.4% when using function-calling mode, while Grok-4-1 scores higher in its fast-reasoning variant (69.6%) than its nonreasoning counterpart (58.3%).

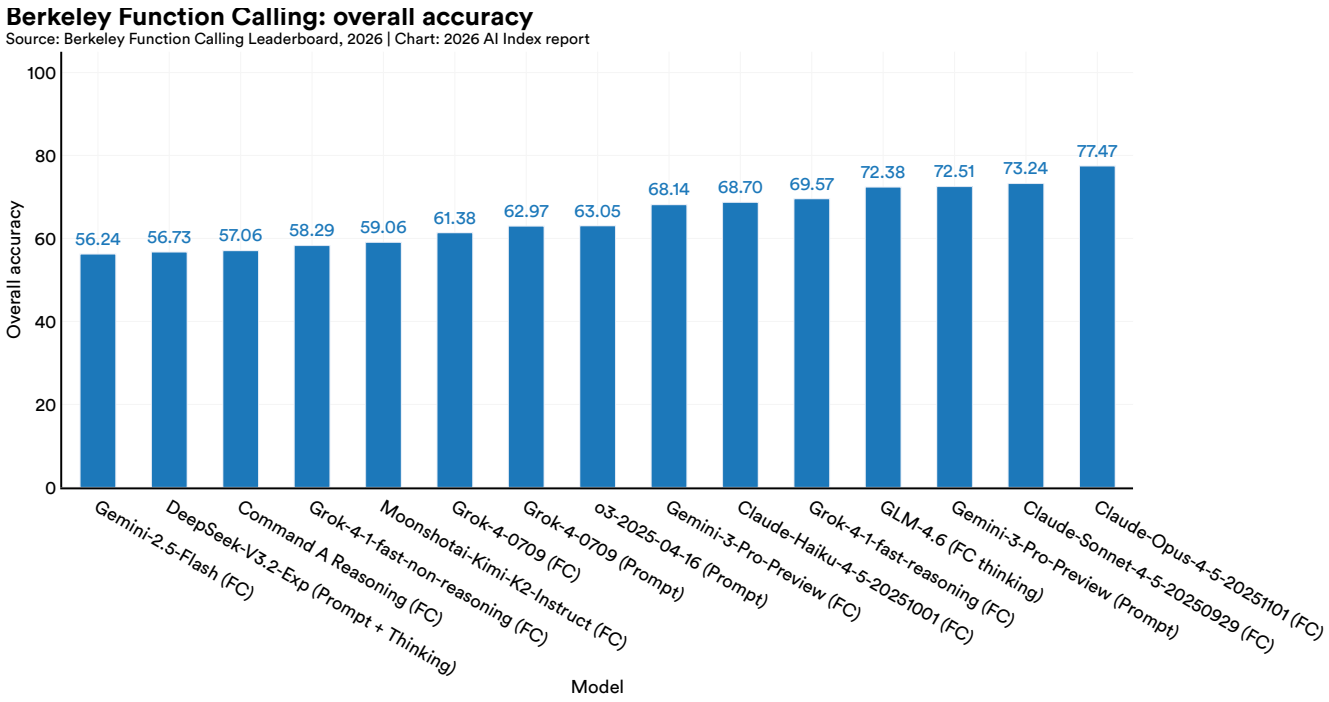


Figure 2.2.3[7]

7 Source: https://gorilla.cs.berkeley.edu/leaderboard.html.

MTEB: Massive Text Embedding Benchmark

The Massive Text Embedding Benchmark (MTEB) evaluates different embedding models across a set of tasks that require semantic understanding. It includes over 50 datasets, spanning eight task categories, which makes it harder for models to look strong by optimizing for a single use case rather than performing well across different settings.

The top average task score on MTEB (English v2) has risen steadily since 2022, coinciding with the broader adoption of large-scale pretraining techniques for embedding models. In 2025, the top score reached 76, rising approximately 11 points since 2023 (Figure 2.2.4). However, the best models still fall short of a perfect score.

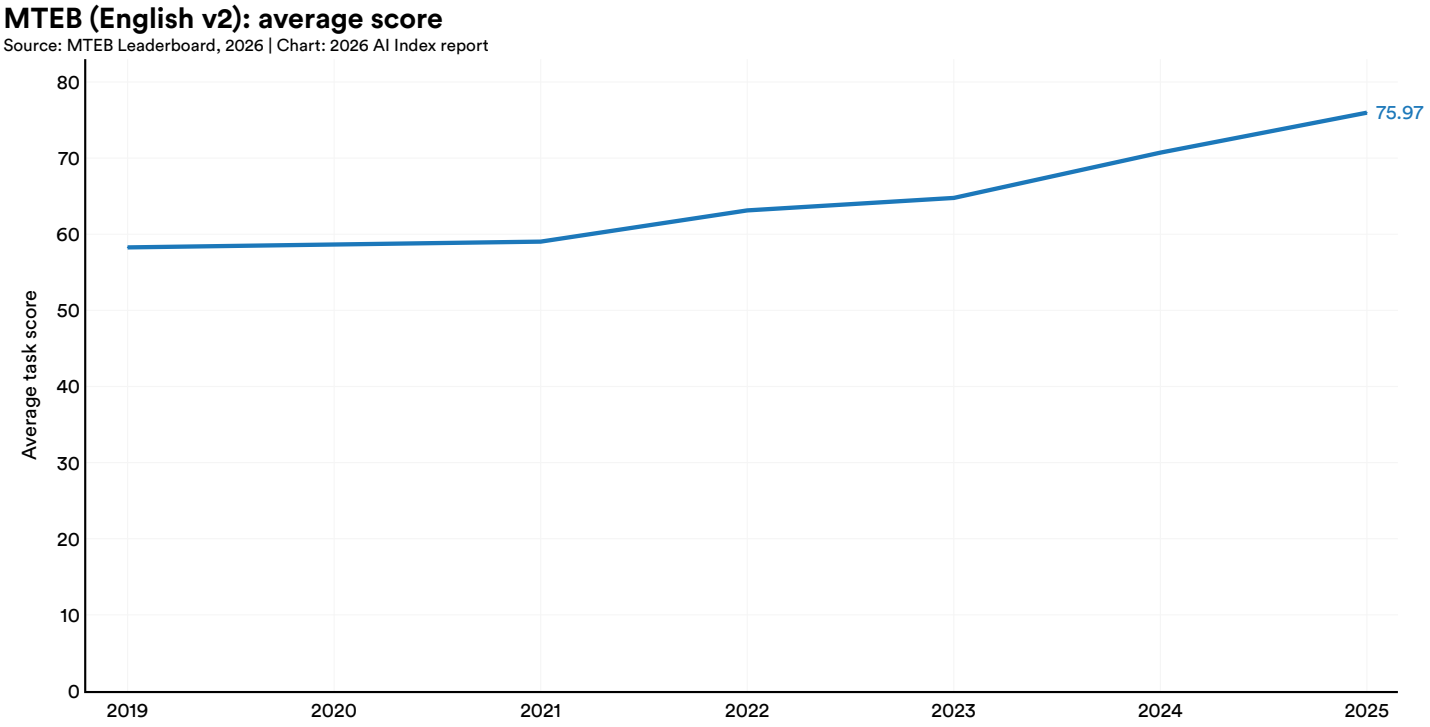


Figure 2.2.4[8]

8 Source: https://mteb-leaderboard.hf.space/?benchmark_name=MTEB%28eng%2C+v1%29 https://arxiv.org/abs/2502.13595.

HIGHLIGHT:

The Gap Between Long Context Windows and Deep Understanding

Context windows, the amount of text a model can process in a single input, have grown by almost 30x per year since mid-2023 (Figure 2.2.5). Models that once accepted a few thousand tokens can now process 1 million or more. At the upper end, this is equivalent to multiple books or an entire codebase in a single pass. On two long-context benchmarks, Fiction.liveBench, which measures narrative comprehension, and MRCR, which measures multi-needle retrieval, the input length at which leading models achieve 80% accuracy has increased even faster, at roughly 250x over a nine month period (Burnham and Adamczewski, 2025). However, bigger context windows do not translate into deeper understanding, as the gap between accepted and usable context length is wide.

Recent research points to different reasons for this gap. On one expert-level, long-context benchmark (LongBench v2), human experts scored just 53.7% accuracy under a 15-minute time limit, and the best model scored 57.7% (Bai et al., 2025). This is a narrow margin in contrast to the structured benchmarks where models have surpassed human baselines, and reflects the difficulty of deep comprehension over long inputs. Models that were prompted to reason through the material step by step did perform better than those asked to answer immediately, suggesting that how a model works through long text matters as much as the amount of text it can accept. Other research has found that models handle simple lookups well but struggle when asked to find multiple pieces of matching information or to apply conditions across a very long document—tasks that would be straightforward for a human scanning the same text (Yu et al., 2025). Models can complete these tasks if guided to check each one by one, but this approach is slow and expensive. Longer inputs come with practical costs of slower response times, higher operating expenses, and reduced accuracy for information that appears later in the input.

Measuring long context ability also remains difficult. When a model scores well on a long context test, it is not always clear whether it genuinely processed the full input or simply relied on knowledge it already had. Yang et al. (2025) introduced a metric designed to separate these two factors and found that model rankings shifted a lot. For example, a model that ranked seventh on raw scores ranked first when only long-context ability was measured, further underscoring why it is important to distinguish between a model specifically being able to better handle long inputs, rather than having overall better capabilities. If the gap between context window size and effective utilization becomes more precise, models may improve their ability to work on tasks that unfold over hours or days and sustain longer chains of reasoning (Denain and Ho, 2025). Developing evaluations that reliably distinguish true long-context ability from general model capability will be important for tracking that progress and ensuring that benchmark gains reflect real improvements.

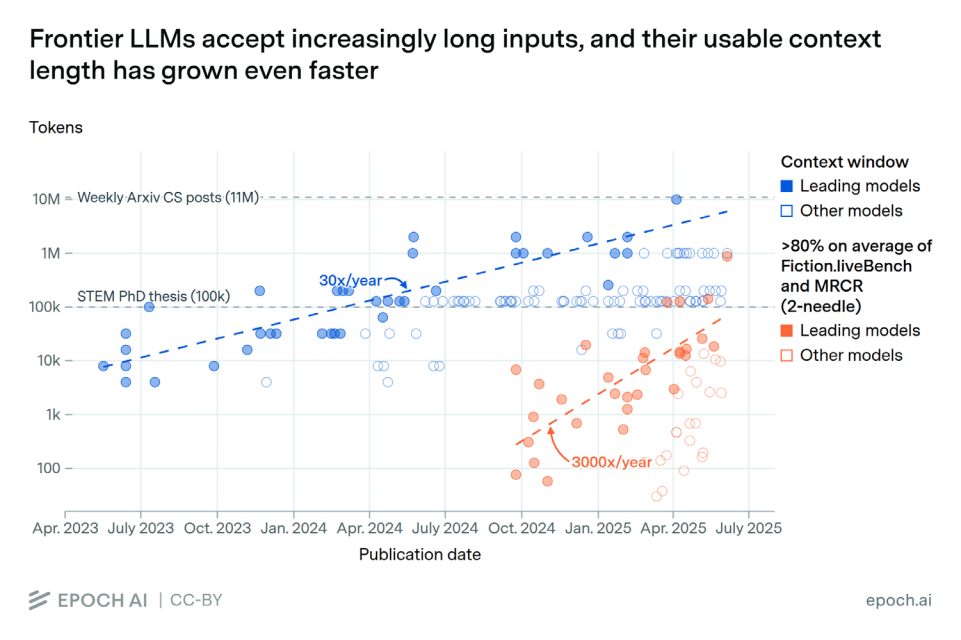


Figure 2.2.5

2.3 Image and Video

Beyond language, many models process visual inputs, and their video and image capabilities have advanced significantly. This section examines model performance across the dimensions of understanding—how well they comprehend and reason over video content—and generation, which evaluates the quality of AI-produced images and videos.

Understanding

Video understanding benchmarks measure how well models can track actions, objects, and events across frames rather than reasoning over a single image. As performance on earlier benchmarks has improved, evaluation has shifted toward tasks that demand multistep temporal reasoning and domain-specific knowledge applied to video.

MVBench

MVBench evaluates whether multimodal models can move beyond static image understanding to handle the complexities of video. This includes interpreting motion, temporal sequences, and shifting context across frames. Its focus on temporal reasoning makes it a useful benchmark for tracking performance in more dynamic visual environments.

The top-performing model on MVBench reaches 74.1% average accuracy, with JT-VL-Chat and JT3.5 tied at that score (Figure 2.3.1). In early 2026, across the top 15 models, performance spans a range of roughly 23 percentage points. VideoChat 2 has the lowest average accuracy (51.1%), while several VideoChat2 variants are grouped in the middle tier (60%–65%).

MVBench: average accuracy

Source: MVBench Leaderboard, 2026 | Chart: 2026 AI Index report

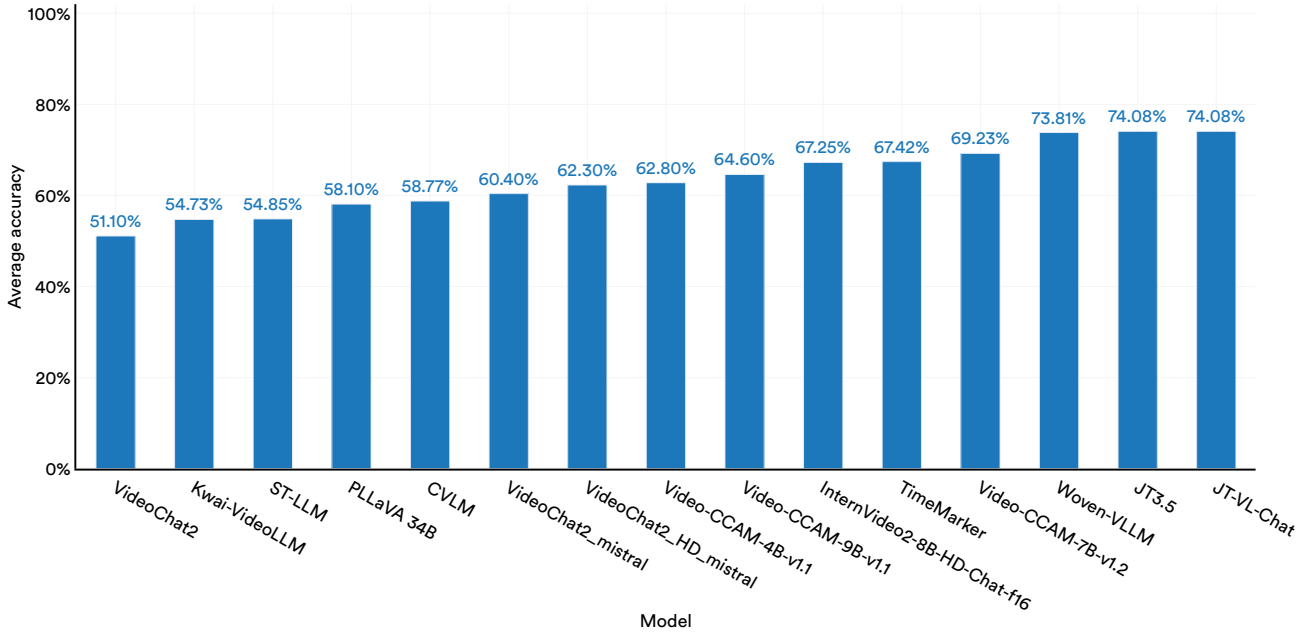


Figure 2.3.1[9]

9 Source: https://huggingface.co/spaces/OpenGVLab/MVBench_Leaderboard.

Video-MMMU

Video-MMMU is a large, multimodal, multidisciplinary benchmark for learning from educational videos, comprising 300 expert-level videos averaging roughly 506 seconds across six disciplines and 30 subjects. Each video is paired with three sets of questions that test progressively deeper understanding. Perception questions test whether a model can pull key details from text/audio; comprehension questions test whether it grasps the concept or solution strategy; and adaptation questions require applying that knowledge to a new scenario. Adaptation questions reuse MMMU/MMMU-Pro items for STEM fields and custom case studies for art/humanities, so models have to go beyond the specific video. The benchmarks also introduce a Δknowledge metric to track how much a model's performance improves after processing the video.

As of 2025, no model has reached the human baseline of 74.4% on Video-MMMU overall accuracy (Figure 2.3.2). The best performing model, Keye-VL-1.5-8B, scores 66%, followed closely by Claude -3.5-Sonnet (65.8%). The lowest score is VILA1.5-8B at 20.9%, leaving a 45 percentage point range across the leaderboard.

The Δknowledge metric results reveal a further gap between human and model learning (Figure 2.3.3). Human experts gain 33.1 percentage points after watching the video, while the best model on this metric, GPT-4o, gains only about half of that (15.6 points). About a third of models even show negative Δknowledge, as their performance actually declines after processing the video.

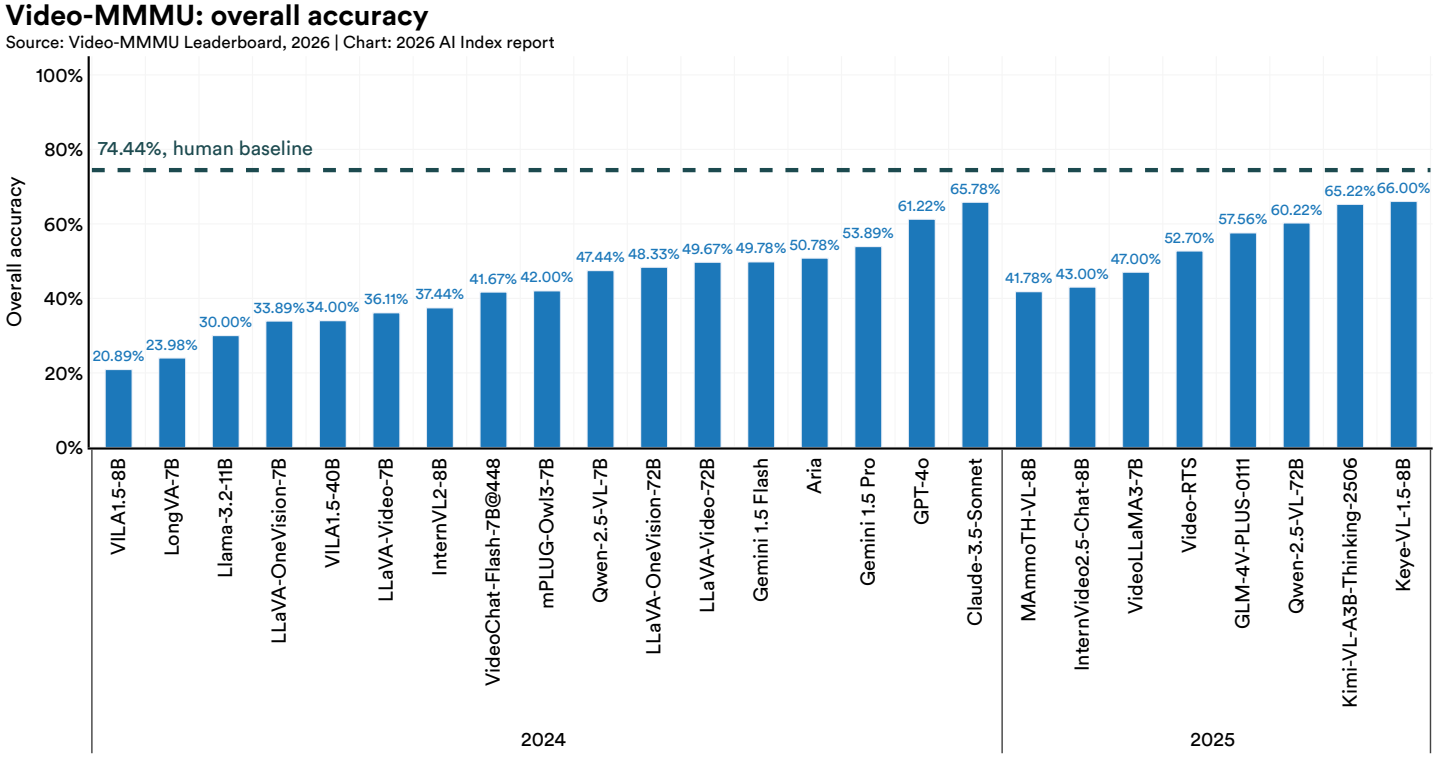


Figure 2.3.2[10]

10 Source: https://videommmu.github.io/#Leaderboard.

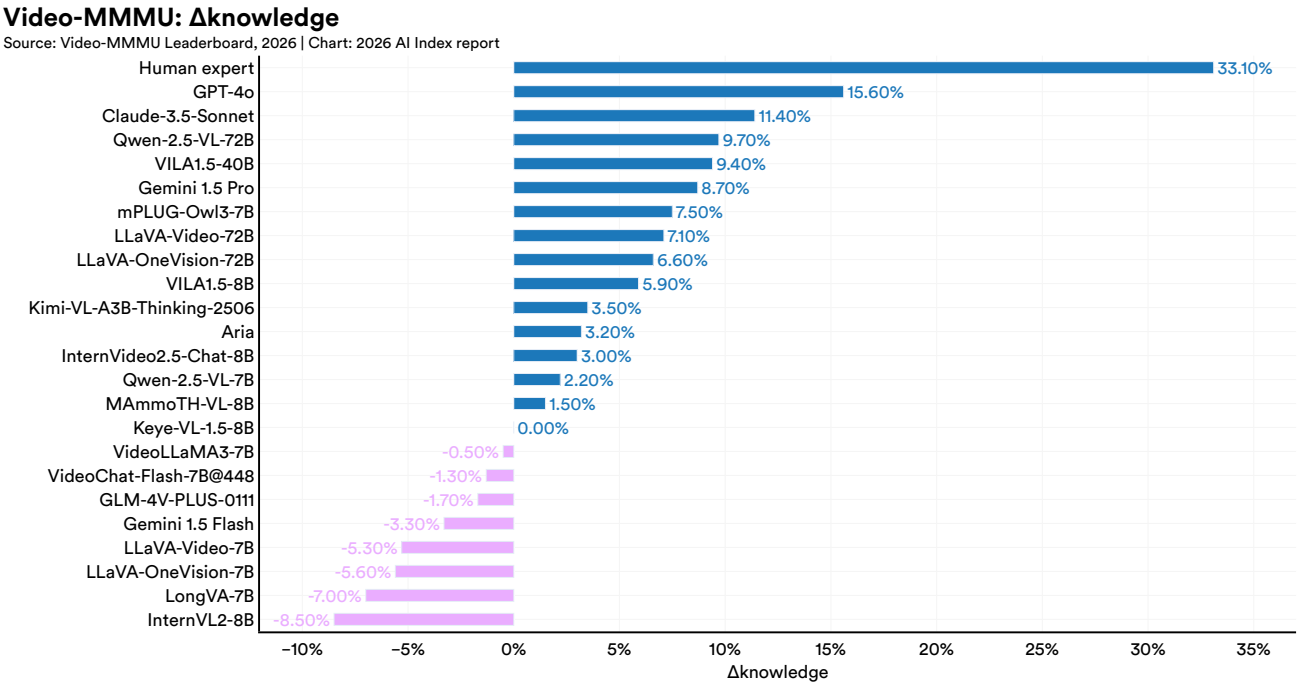


Figure 2.3.3

Generation

While the above benchmarks test how well models interpret existing visual content, generation benchmarks assess how well models can produce it. Evaluation spans both human preference rankings as well as automated quality metrics, since generated video must satisfy subjective expectations and technical criteria such as coherence, fidelity, and controllability. Of these, controllability has become an especially important focus, reflecting whether models can follow user intent while maintaining natural motion and scene dynamics. This has also brought video generation closer to the idea of world models, where systems aim to predict how visual scenes evolve over time.

Midjourney generations over time: "a hyper-realistic image of Harry Potter"
Source: Midjourney, 2025

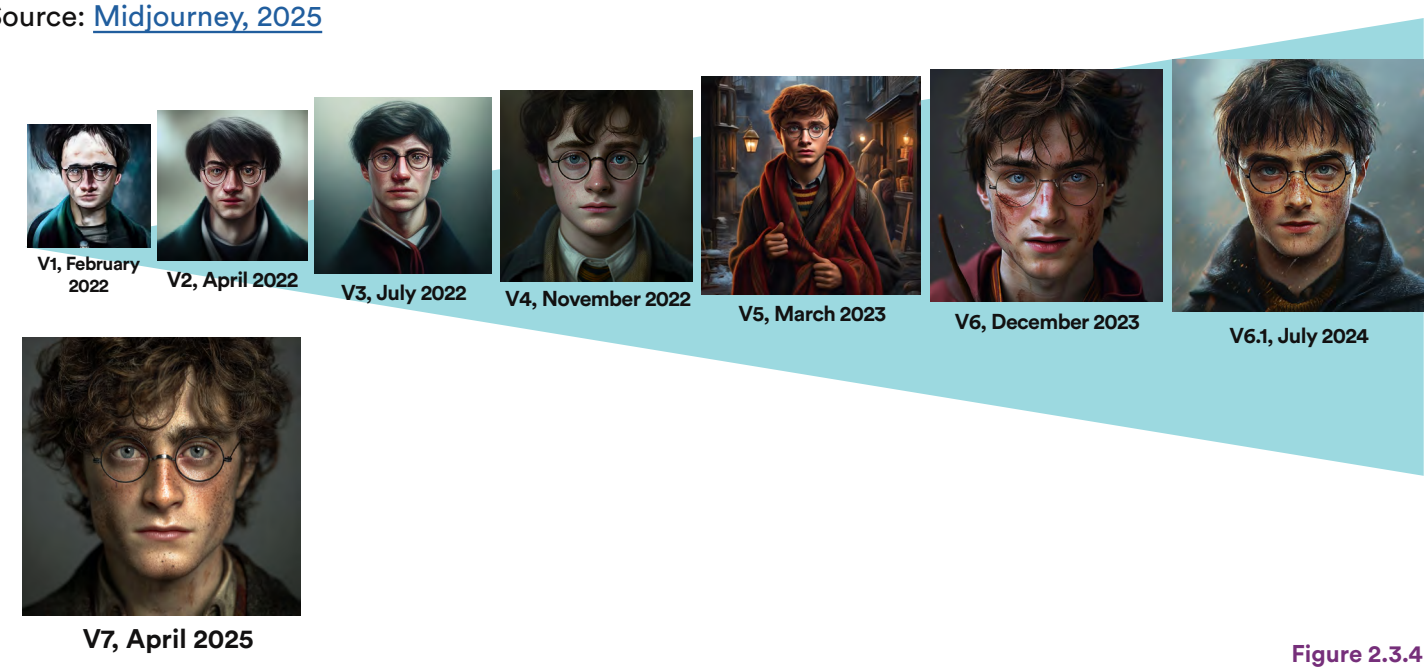


Figure 2.3.4

Arena: Vision

The Arena platform also hosts a Vision Arena that applies the same blind-comparison, Elo-based methodology described in the earlier section for language to image generation models. Human preference is an important signal for image generation, as qualities like aesthetic appeal and visual coherence are difficult to capture through automated metrics alone.

As of early 2026, Google's Gemini models hold four of the top six positions, with Gemini-3-Pro leading at approximately 1,285 Elo, followed by its other variation (Figure 2.3.5). Similar to the language evaluation, confidence intervals overlap for the bottom two-thirds of the ranked models as they fall within a 30-point range between 1,230 and 1,260.

Vision Arena: Elo rating

Source: Arena, 2026 | Chart: 2026 AI Index report

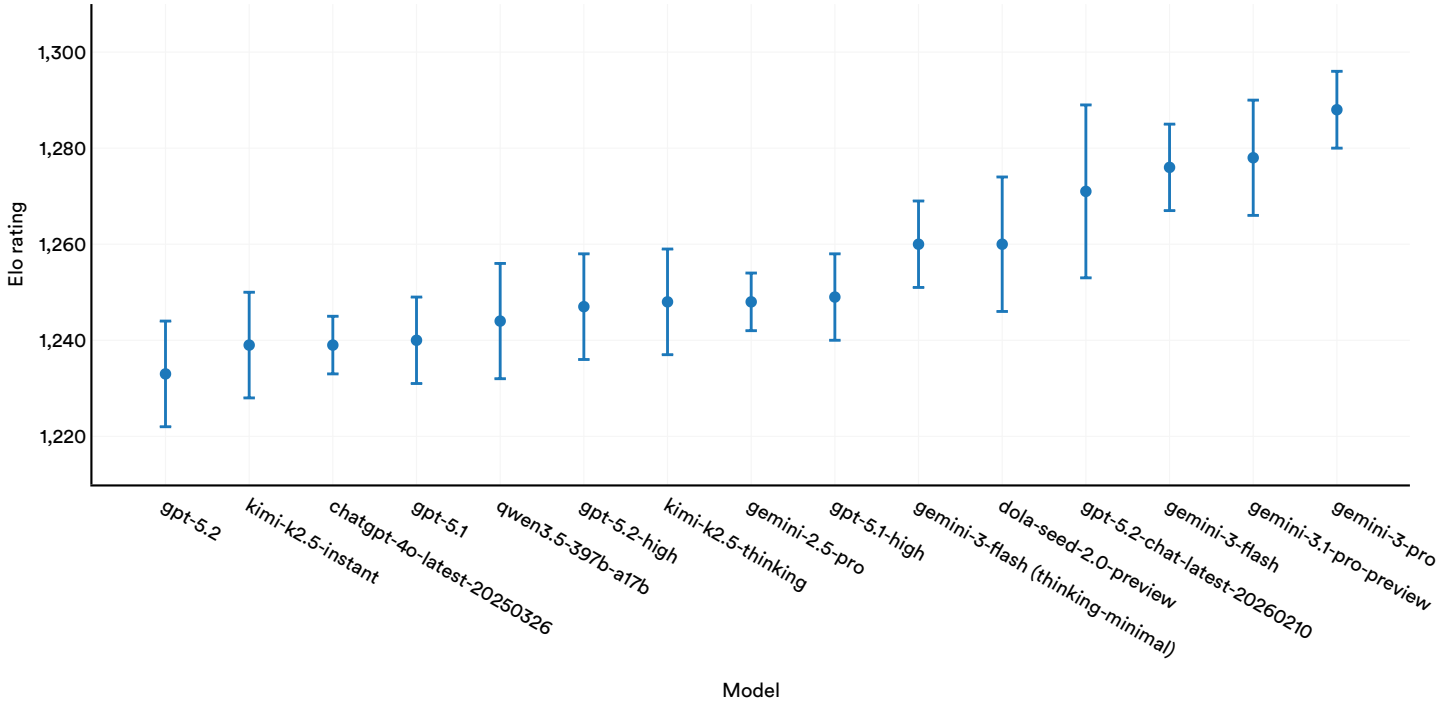


Figure 2.3.5[11]

Video-Bench

Video-Bench is a human-aligned benchmark for video generation that scores models on two dimension groups: video-condition alignment and video quality. It uses an MLLM-based evaluator (GPT-4o) with few-shot scoring, and chain-of-query prompting for more precise, calibrated results. The scores correlate more strongly with human ratings than prior metric-based or LLM-based benchmarks.

As of early 2026, Gen3 and Kling lead on video quality, with Gen3 scoring highest across all key metrics, including imaging quality, aesthetic quality, temporal consistency, and motion effects, while Kling ranks second overall (Figure 2.3.6). Motion effects are the weakest subdimension across nearly all models.

11 This chart shows only the top 15 models as of February 2026; source: https://arena.ai/leaderboard/vision.

Video-Bench video quality
Source: Video-Bench Leaderboard, 2025 | Chart: 2026 AI Index report

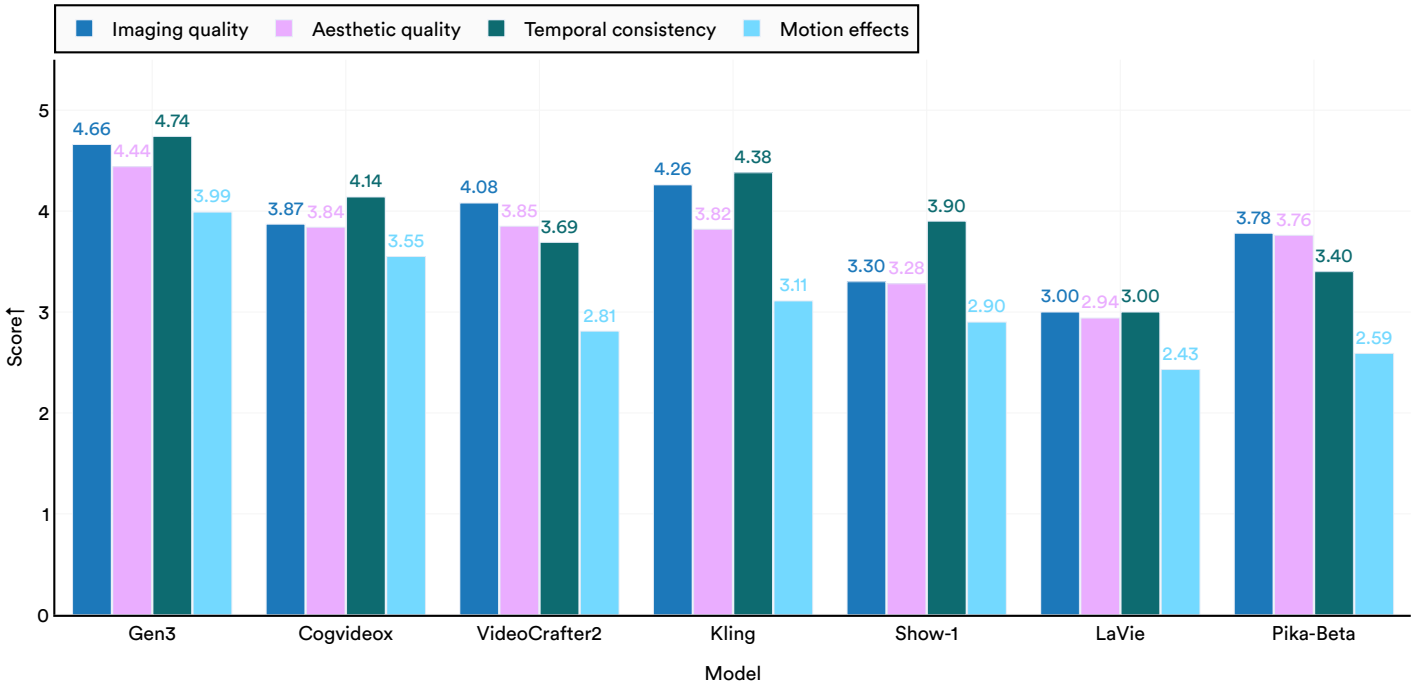


Figure 2.3.6[12]

VBench-2.0

VBench-2.0 is a comprehensive, human-aligned benchmark for evaluating video generation models on intrinsic faithfulness, defined as well-rounded adherence to reality rather than simply being visually convincing. It scores models across five broad dimensions (Human Fidelity, Creativity, Controllability, Physics, and Commonsense). The benchmark combines VLM/LLM-based analysis with specialized detectors and a small but targeted prompt set, anchored by human preference labels. This faithfulness-oriented approach is important because it surfaces whether generated videos hold up under scrutiny in areas like physical plausibility and scene consistency.

None of the models evaluated in early 2026 surpasses a total score of 67% (Figure 2.3.7). Veo 3 leads at 66.7%, about 4 percentage points above the next top performing mode, Vidu Q1 (62.7%). Similar to other benchmark scores, several models are tightly grouped and hover around scores of 58% and 60%. Even established systems like Kling, CogVideoX, and HunyuanVideo continue to struggle with complex stories and consistent object/scene dynamics.

12 Source: https://github.com/Video-Bench/Video-Bench?tab=readme-ov-file#leaderboard.

VBench-2.0: total score
Source: VBench Leaderboard, 2026 | Chart: 2026 AI Index report

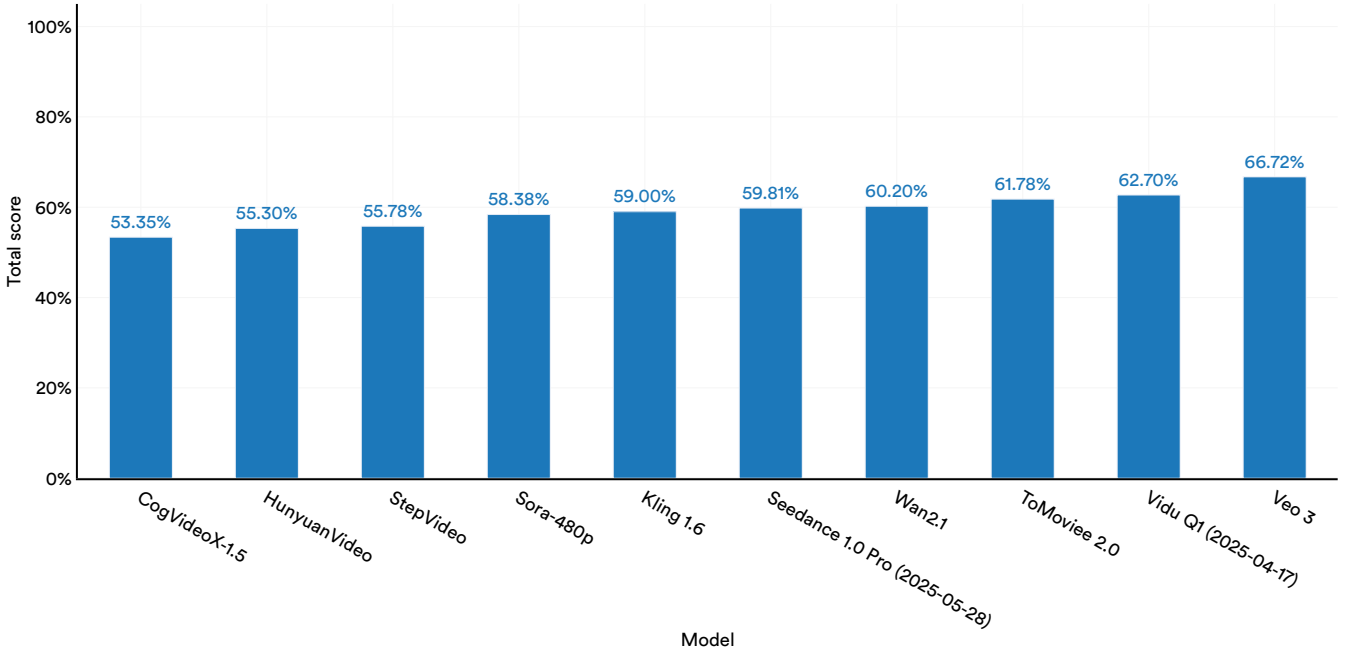


Figure 2.3.7[13]

HIGHLIGHT:

Progress in Video Generation

The benchmarks in this section mostly evaluate video modes as content generators, scoring them on quality, fidelity, and controllability. However, recent research suggests that video generation models may be developing capabilities that go beyond producing content.

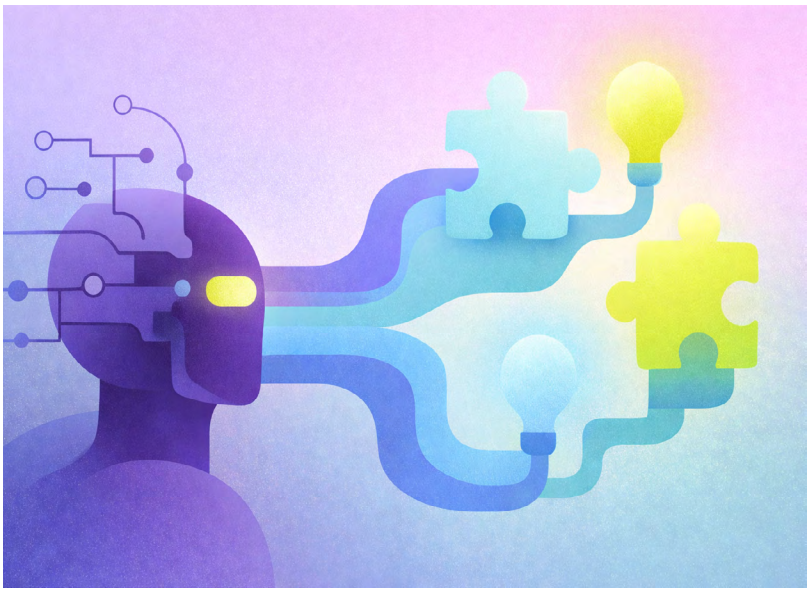
A 2025 Google DeepMind study (Wiedemar et al., 2025) tested whether Veo 3, a video generation model, could solve visual tasks it was never specifically trained for, using only an input image and a text prompt. Across 62 qualitative tasks and seven quantitative evaluations covering more than 18,000 generated videos, the model showed zero-shot abilities in areas traditionally handled by specialized systems. These included perception tasks such as edge detection and segmentation, physical modeling tasks such as buoyancy and rigid body dynamics, and manipulation tasks such as style transfer and object extraction. The authors also observed early signs of visual reasoning, including maze solving and visual analogy completion, which they describe as "chain of frames," a parallel to chain-of-thought reasoning in language models where the model appears to reason step by step through successive frames. Performance improved consistently from Veo 2 to Veo 3 across all quantitative tasks and, in some cases, matched or exceeded a dedicated image editing baseline (Nano Banana).

Specialized models still outperform zero-shot video generation on most individual tasks, but the rapid improvement and breadth of zero-shot capability suggests a familiar trajectory. Large language models develop general-purpose language understanding from generative training on web-scale data, and video models trained under similar conditions may be following a comparable path toward general-purpose vision.

13 Source: https://huggingface.co/spaces/Vchitect/VBench_Leaderboard.

2.4 Reasoning

Reasoning benchmarks assess whether models can solve problems that require abstraction and generalization across domains and formats. As performance has improved, newer benchmarks aim to distinguish genuine problem-solving from performance that is driven by memorization or prompt familiarity. However, because models can also produce errors in otherwise fluent responses, efforts are ramping up to measure these error rates alongside reasoning limitations. The AI Index tracks those benchmarks on factual reliability and error rates in Chapter 3. Across the benchmarks in this section, leading models perform well on many tasks but still show gaps on the more difficult items.



General Reasoning

General reasoning refers to a model's ability to solve unfamiliar problems by applying rules and combining evidence, rather than relying on domain knowledge or memorized patterns. The benchmarks discussed below span multiple domains and tasks and are designed to test multistep inference. One example is multidigit arithmetic, such as long integer multiplication, to test whether models can execute consistent stepwise computation rather than produce plausible-looking outputs. Other more complex benchmarks extend this idea to multimodal settings, where models must integrate text with diagrams or plots to reach the correct answer.

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

MMMU evaluates multimodal reasoning on college-level subject questions that combine text with visuals such as diagrams, charts, tables, and equations. Some example tasks include extracting constraints from a table and applying them to a word problem, or using a diagram to answer a domain-specific question in areas like engineering or medicine.

As of February 2026, the leading model, Gemini 3.1 Pro Preview, scored 88.2% on MMMU and within 0.4 percentage points of the best human expert reference (Figure 2.4.1). Other Gemini variants follow closely, including Gemini 3 Flash (87.6%) and Gemini 3 Pro (87.5%), while GPT-5.2 scores 86.7%. The 2026 models trail behind with Kimi K2.5 at 84.3% and Claude Opus 4.6 (Thinking) at 83.9%.

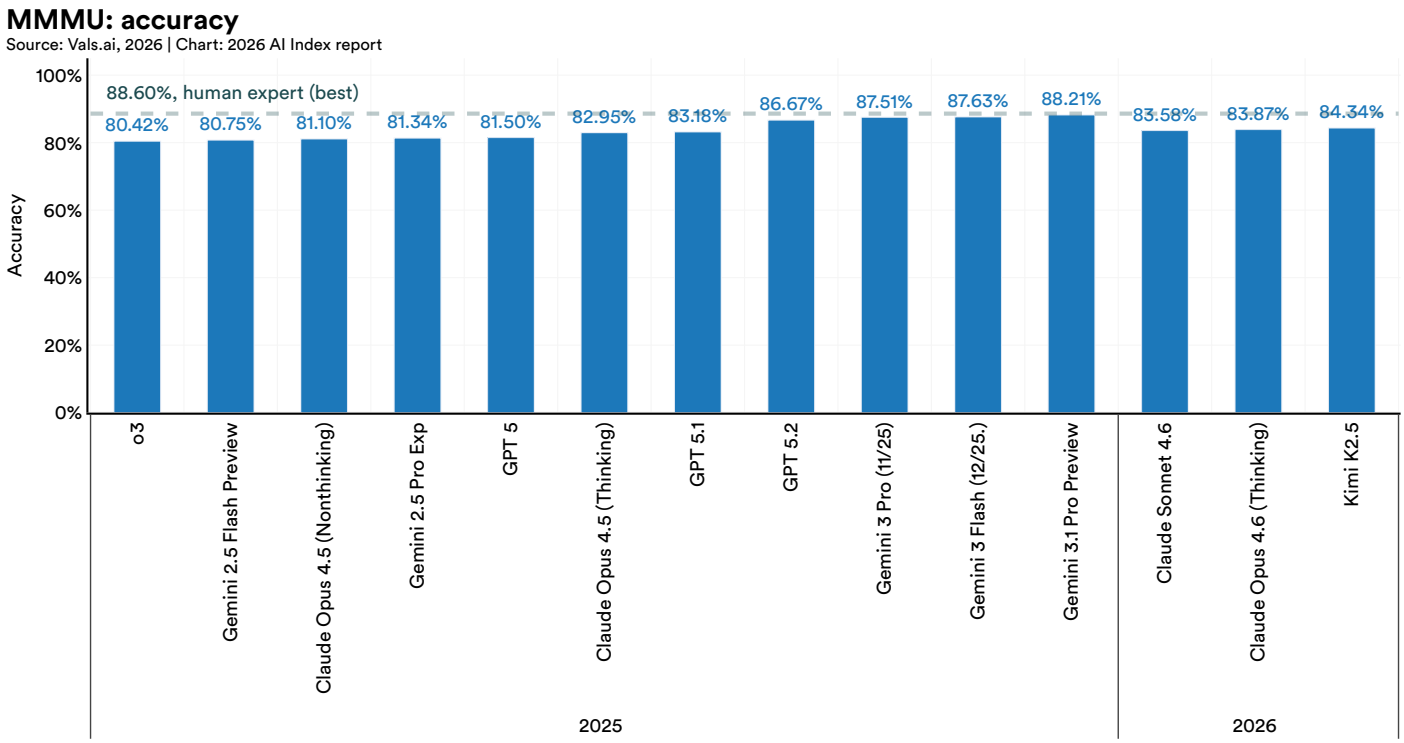


Figure 2.4.1[14]

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

While MMMU focuses on multimodal reasoning, GPQA evaluates reasoning on difficult, text-only questions designed to test graduate-level problem solving. The questions require models to apply domain-specific concepts and follow multistep logic to reach the correct answer. Example tasks include graduate-level chemistry or physics questions that require working through a multistep solution and choosing the best answer from several very similar options.

Model performance on the GPQA Diamond set has continued to rise above the expert human validator baseline of 81.2% (Figure 2.4.2). In late 2024, OpenAI's o3 was the first to exceed it with a score of 87.7%. In 2025, mean accuracy reached 93%, exceeding the expert reference point by 12 percentage points.

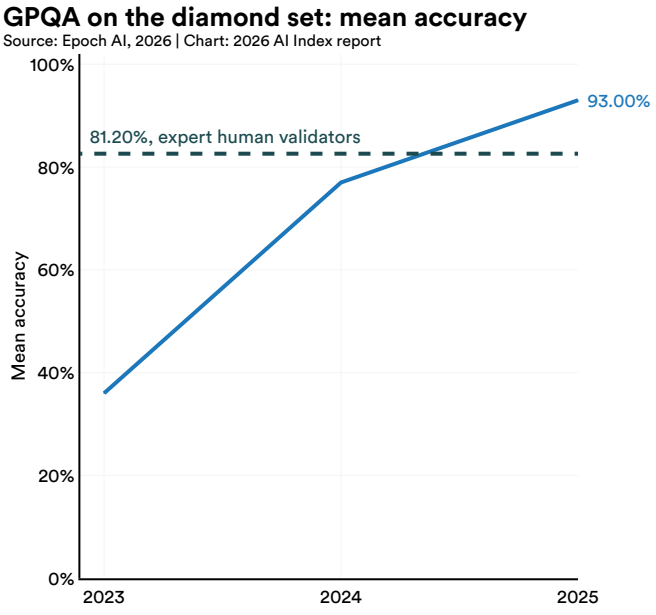


Figure 2.4.2[15]

14 This chart shows the top 15 models as of February 2026; data source: https://www.vals.ai/benchmarks/mmmu.

15 Data source: https://epoch.ai/benchmarks.

ARC-AGI-2

Introduced in 2019, ARC-AGI is a benchmark that tests the ability of systems to generalize beyond prior training, emphasizing generalized learning ability. Despite its name, the benchmark tests a specific form of abstraction and pattern inference rather than general intelligence in a broader sense. Its updated version, ARC-AGI-2, was introduced in 2025 and shifts to abstract puzzle-style tasks that evaluate whether models can infer rules from a small set of examples and apply them to new cases. Example tasks include grid puzzles where the model is given a few example solutions, infers the rule, and uses it to solve a new problem.

Scores on ARC-AGI-2 vary widely across models, and the spread between the highest and lowest scores in the figure is about 46% (Figure 2.4.3). Gemini 3 Deep Think leads at 84.6%, followed by Gemini 3.1 Pro Preview at 77.1% and GPT-5.2 (Refine.) at 72.9%. Several Claude Opus 4.6 variants are clustered together, scoring between 66.3% and 69.2%.

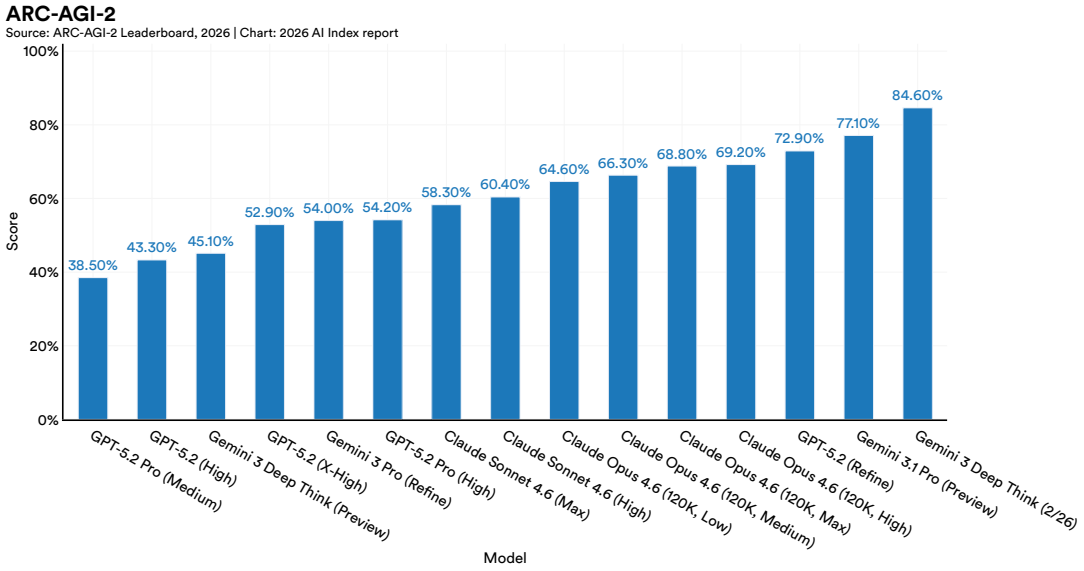


Figure 2.4.3[16]

Humanity's Last Exam

Humanity's Last Exam (HLE) benchmark evaluates model performance on 2,700 highly challenging questions across dozens of academic subjects. It is designed as an expert-level, closed-ended benchmark with wide coverage and using a mix of multiple-choice and short-answer formats suitable for automated grading. Example tasks include a graduate level question that requires applying a concept and providing a single, verifiable answer. Some may include an image, requiring models to integrate visual and textual information.

Between 2024 and 2025, model accuracy on HLE increased by 30 percentage points (Figure 2.4.4). In a single year, accuracy went from under 10% to 38.3%. Even with this jump, the benchmark is designed to stay difficult, and high-confidence errors are still common.

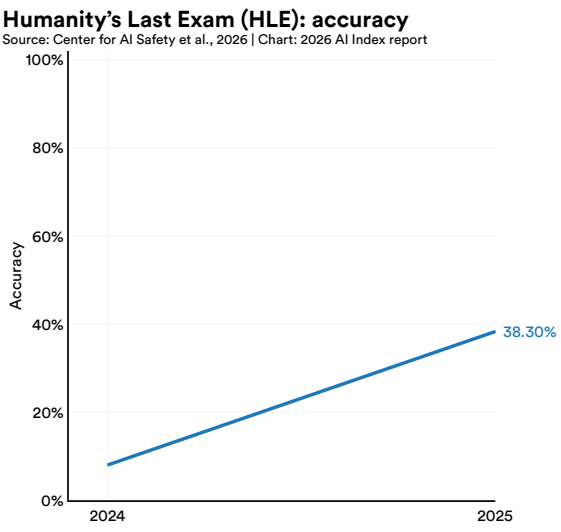


Figure 2.4.4[17]

16 This chart shows the top 15 models as of February 2026; data source: https://arcprize.org/leaderboard.

17 Data source: https://lastexam.ai.

HIGHLIGHT:

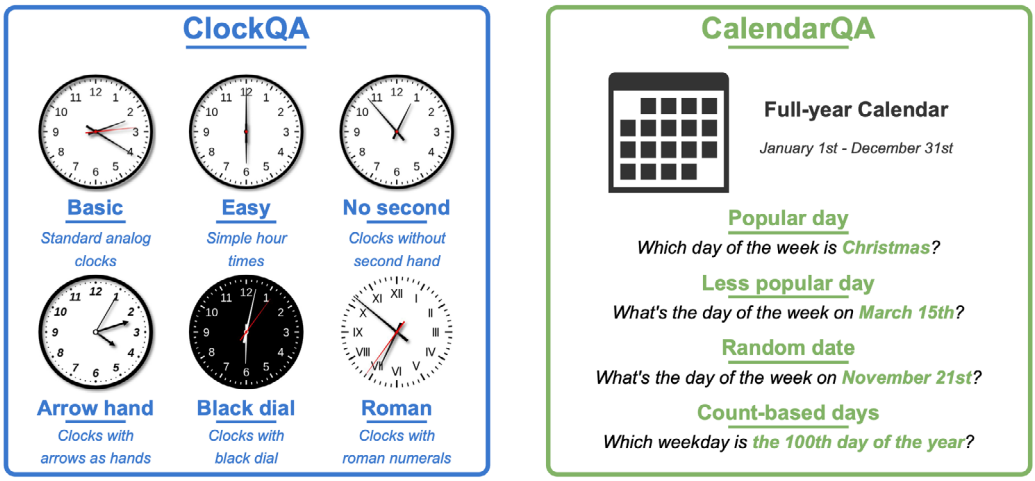
Time Understanding in MLLMs

Many multimodal models still struggle with something most humans find routine, telling the time. Despite the rapid improvements on expert-level reasoning benchmarks like GPQA and HLE, recent studies show models have trouble reading analog clocks. The task combines visual perception with simple arithmetic, from identifying clock hands and their positions and then converting those into a time value. There is the risk that an error in one step will cascade into the next.

Saxena et al. (2025) tested seven multimodal models on two focused datasets (Figure 2.4.5). ClockQA included 62 analog clock images across six visual styles—including clocks with a black dial or no second hand—and CalendarQA, which paired yearly calendar images with date-reasoning questions. On clock reading, even the best performing model, Gemini-2.0, achieved only 22.6% exact match accuracy (Figure 2.4.6). Models fared better on the calendar questions, with GPT-o1 reaching 80% accuracy, though there were more errors when questions required date arithmetic rather than recognition of well-known holidays (Figure 2.4.7).

ClockBench (Safar, 2025) scaled up the evaluation to 180 clock designs and 720 questions. Humans read correctly formatted clocks correctly 90.1% of the time, while GPT-5.4 High, the top model, reached 50.6% in March 2026 (Figure 2.4.8). The gap of about 40 percentage points is large, but the wider gap is in the nature of the errors. When models told the time wrong, their median error ranged from about one to three hours, compared to three minutes for humans.

A study published in IEEE Internet Computing (Fu et al., 2025) looked at why these failures continue to happen. After fine-tuning on 5,000 synthetic clock images, models improved on familiar clock styles but failed to generalize to real-world photos or clocks that had different features, such as distorted dials or thinner hands. When researchers dug into the errors, they identified a pattern. If a model confused the hour and minute hands, its ability to judge hand direction deteriorated. This suggests that the difficulty springs less from training data and more on how models piece together multiple visual cues within a single image. Even as models close the gap with human experts on knowledge-intensive tasks, this kind of visual reasoning remains a persistent challenge.



Source: Saxena et al., 2025

Figure 2.4.5

HIGHLIGHT:

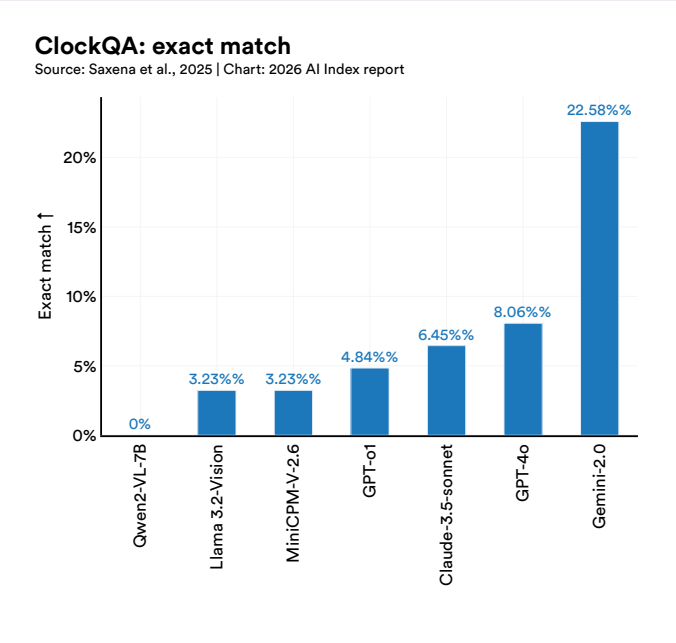


Figure 2.4.6

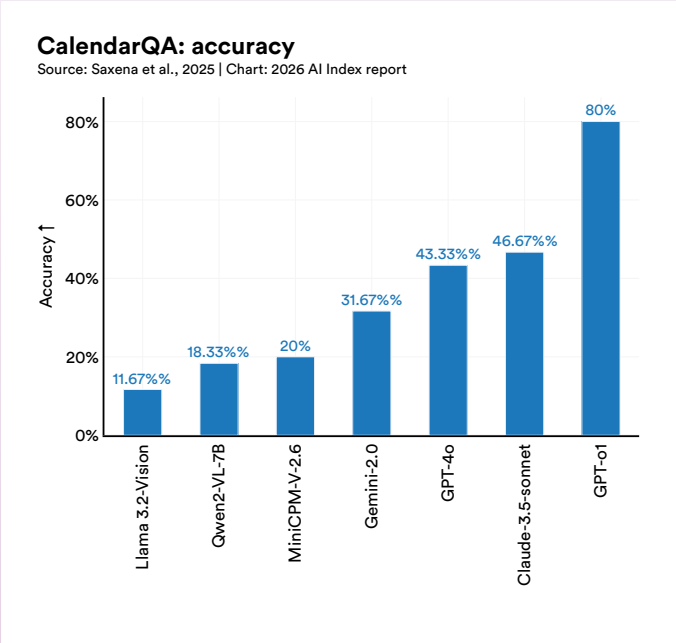


Figure 2.4.7

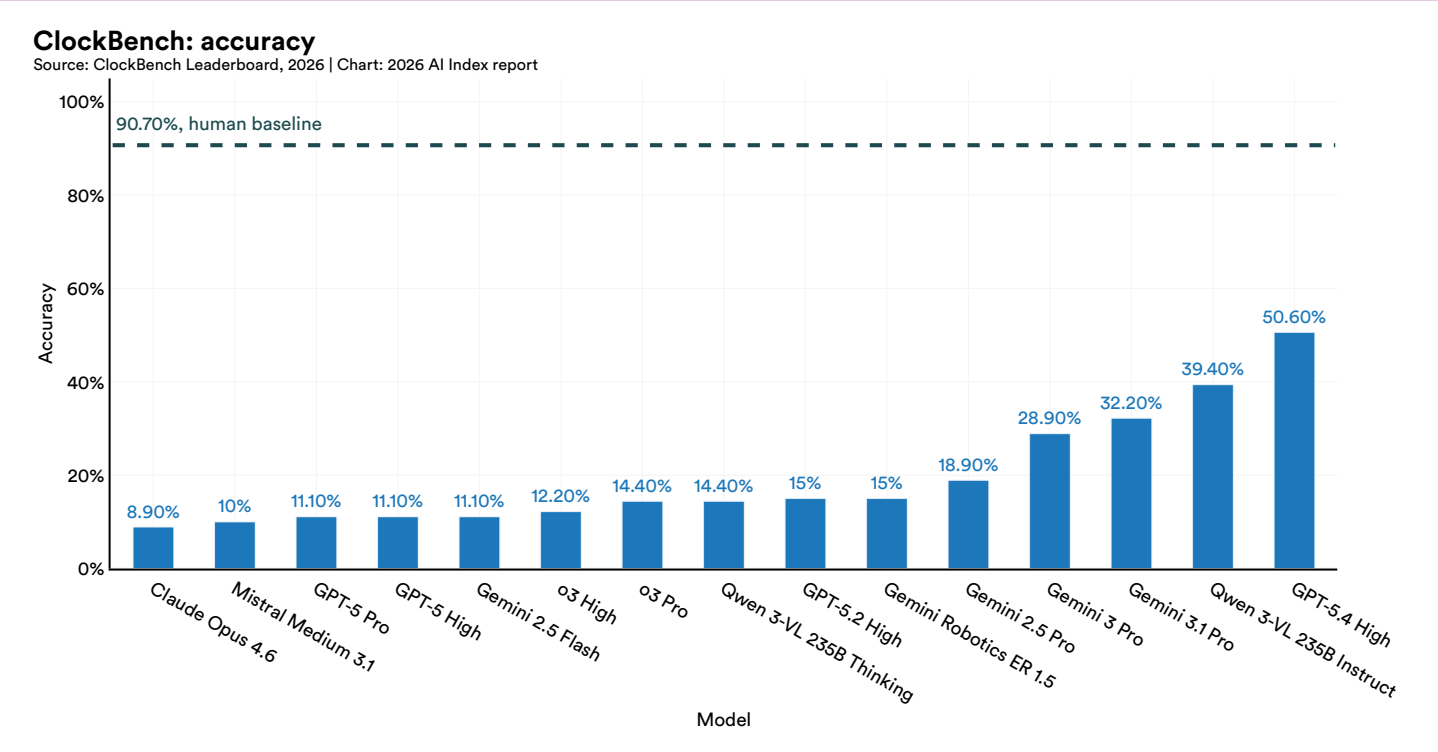


Figure 2.4.8[18]

18 Data source: https://clockbench.ai.

Planning

In addition to general reasoning, the AI Index tracks planning benchmarks that assess models' ability to sequence actions over multiple steps to achieve a goal. Models have to keep track of what has already happened, avoid invalid actions, and maintain consistency even as problems get longer and more complex. Unlike single-shot reasoning questions, planning evaluations can expose failures that only emerge over longer horizons, including compounding errors or forgetting earlier constraints. Which benchmark is used to measure these capabilities matters, as different tasks surface different types of failures.

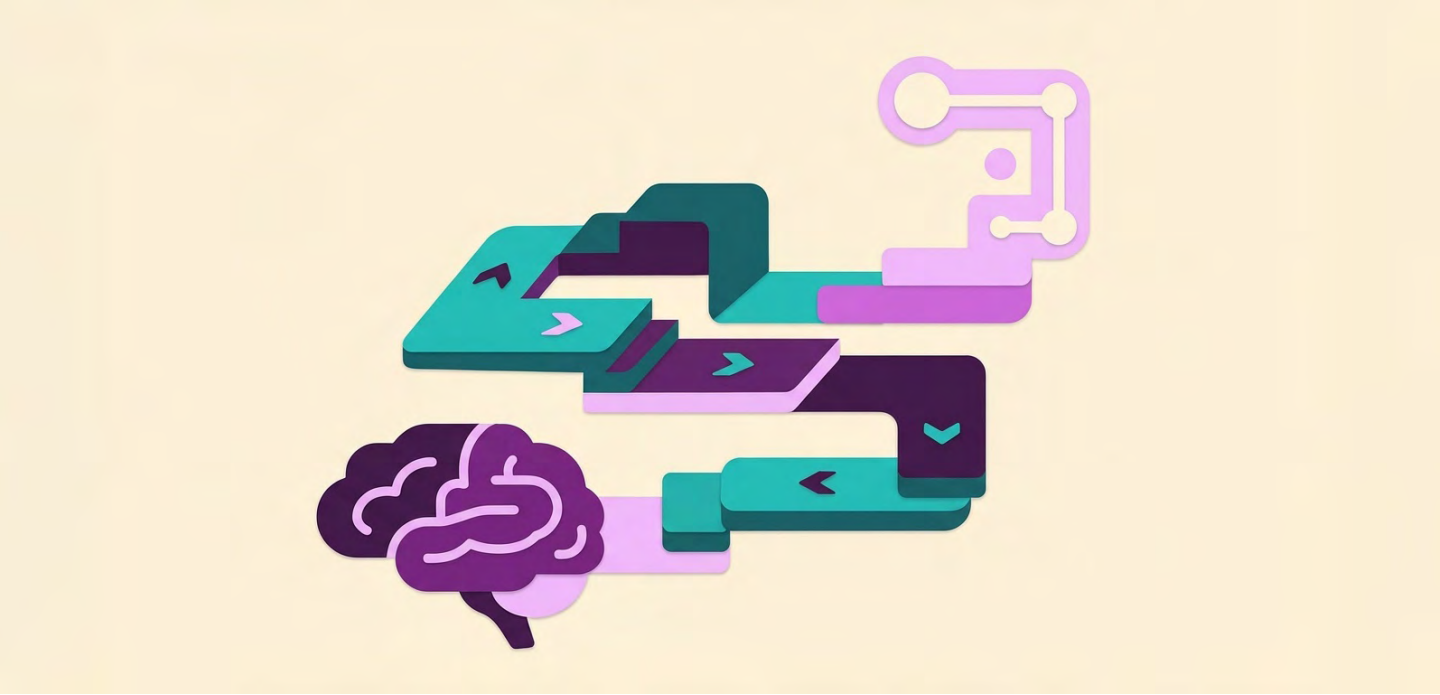
PlanBench

Classical planners like LAMA search systematically through possible states and produce correct plans when they find a solution. Language models instead generate plans based on learned patterns, which means they can produce plausible sequences that can have invalid steps or miss constraints.

PlanBench evaluates end-to-end planning by prompting models to generate a full plan from a structured problem description across several planning domains. A domain is a type of problem with its own rules and goals, such as stacking blocks in a specific order, navigating routes, or transporting packages between locations. The benchmark reports performance as the number of tasks solved in each domain, with up to 45 tasks per domain, compared to LAMA as a classical planning baseline.

No single model leads across every domain (Figure 2.4.9). Under standard planning, LAMA leads in several domains, including Miconic (45/45), Rovers (34/45), and Transport (33/45). In more structured domains such as Childsnack and Spanner, frontier models match or exceed LAMA, with GPT-5 reaching 38/45 on Childsnack and 45/45 on Spanner.

When task descriptions are scrambled to disguise their structure, performance decreases for most models in several domains, though the effect depends on the domain and model (Figure 2.4.10). For example, DeepSeek R1 falls to 3/45 on Blocksworld and 0/45 on Floortile and Sokoban. Similarly, GPT-5 declines to 12/45 on Blocksworld and 7/45 on Sokoban.



PlanBench: task solved on standard planning

Source: Corrêa et al., 2025 | Chart: 2026 AI Index report

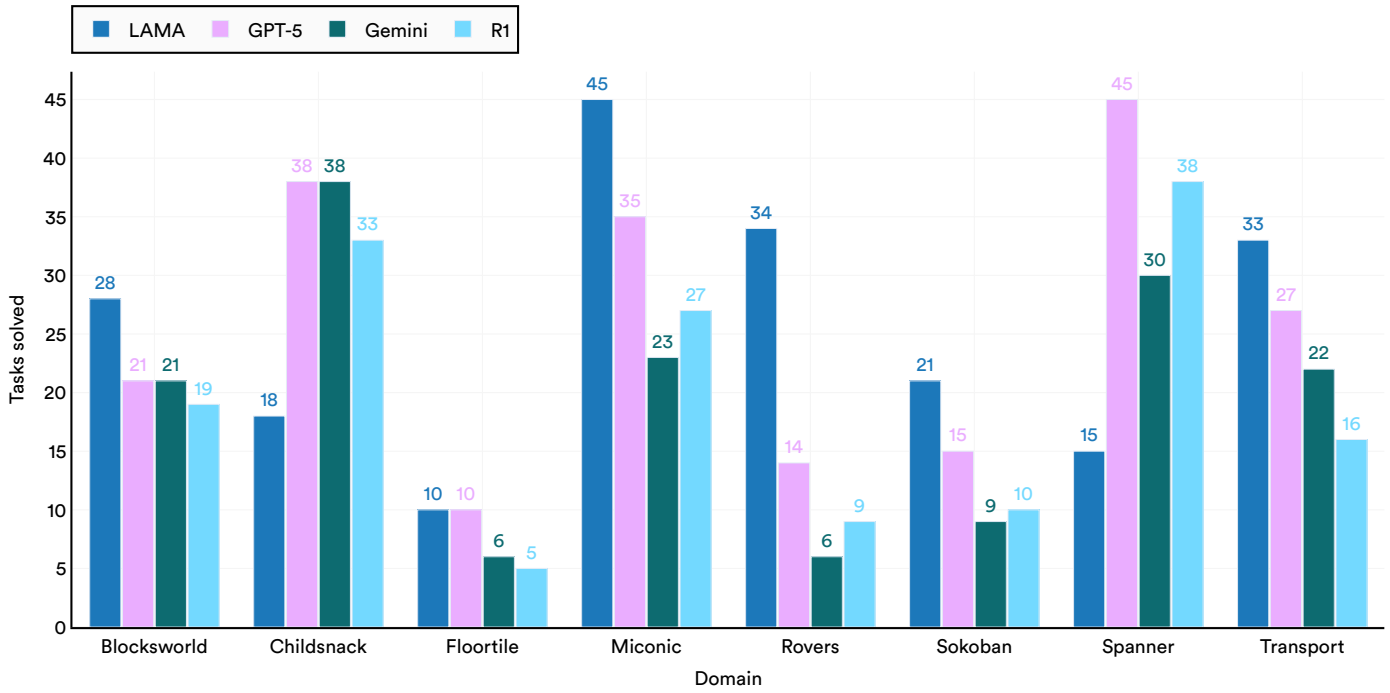


Figure 2.4.9[19]

PlanBench: task solved on obfuscated planning

Source: Corrêa et al., 2025 | Chart: 2026 AI Index report

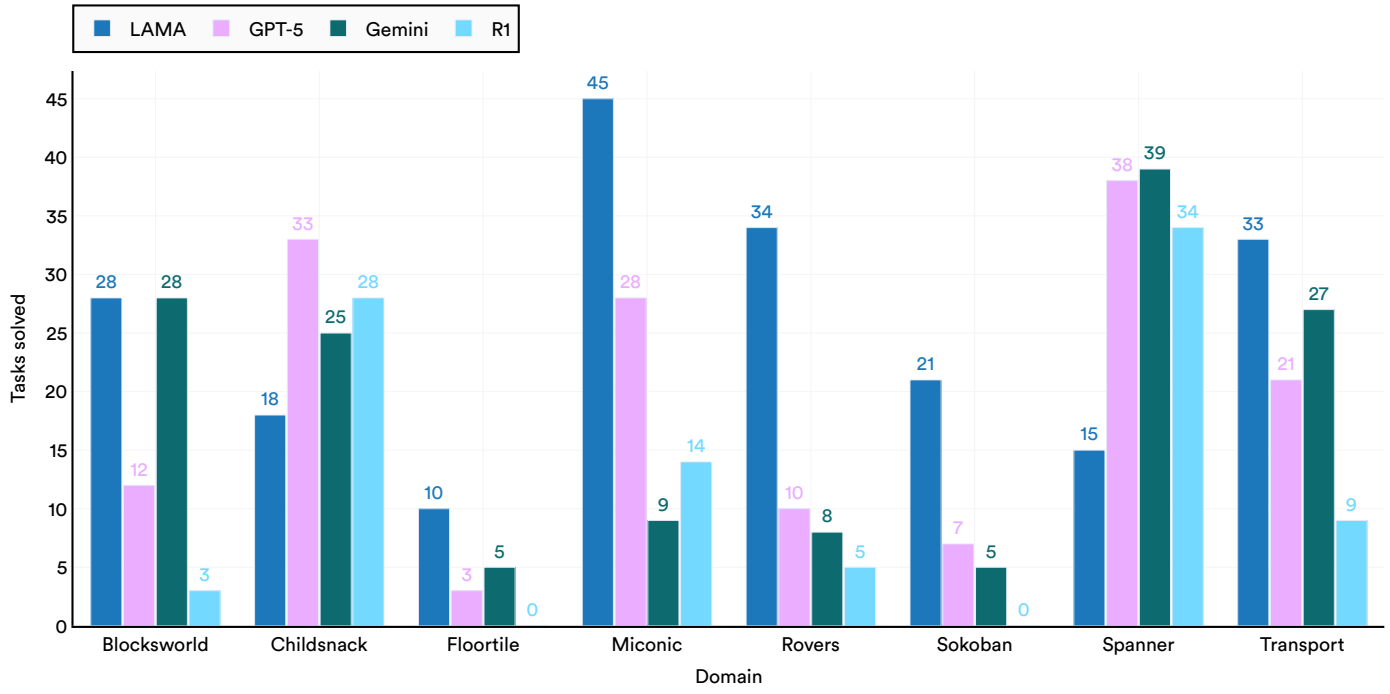


Figure 2.4.10[20]

19 Data source: https://arxiv.org/pdf/2511.09378.

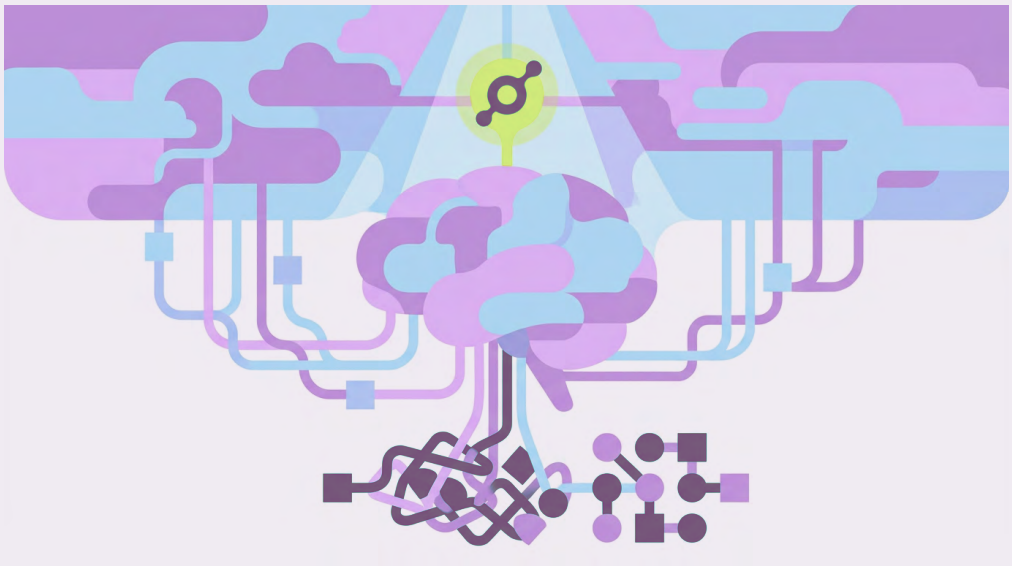
20 Data source: https://arxiv.org/pdf/2511.09378.

2.5 Performance in Specific Domains

As AI models have improved on general reasoning and knowledge benchmarks, attention has shifted to how well they perform on tasks requiring specialized expertise. The benchmarks in this section test models in four professional and academic domains: coding, mathematics, finance, and legal reasoning. Each has its own vocabulary, conventions, and standards for what counts as a correct and comprehensible answer. Many of these benchmarks are new, reflecting growing demand for domain-specific evaluation. Unless otherwise noted, the results reported below reflect model performance as of early 2026.

Software

Coding benchmarks test whether models can go beyond answering questions about code and actually write, debug, and ship working software. The tasks in this section range from resolving real GitHub issues to building full web applications from scratch, reflecting a shift in evaluation toward measuring what models can deliver end to end rather than in isolated snippets.



SWE-bench

SWE-bench evaluates models on their ability to resolve real-world software issues collected from GitHub. Each task gives the model a codebase and an issue description, and the model has to produce a working patch. SWE-bench Lite is a smaller, more accessible subset while SWE-bench Verified uses human-validated issues to ensure more consistent and accurate grading.

On SWE-bench Verified, top models are tightly clustered in the low-to-mid 70s (Figure 2.5.1). As of February 2026, Claude 4.5 Opus (high reasoning) led at approximately 76.8%, with several others including KimiK2.5, GPT-5.2, and Gemini 3 Flash (high reasoning) grouped between 70% and 76%. This is a pattern seen across several benchmarks in this chapter, where high-performing models score within a few percentage points of each other.

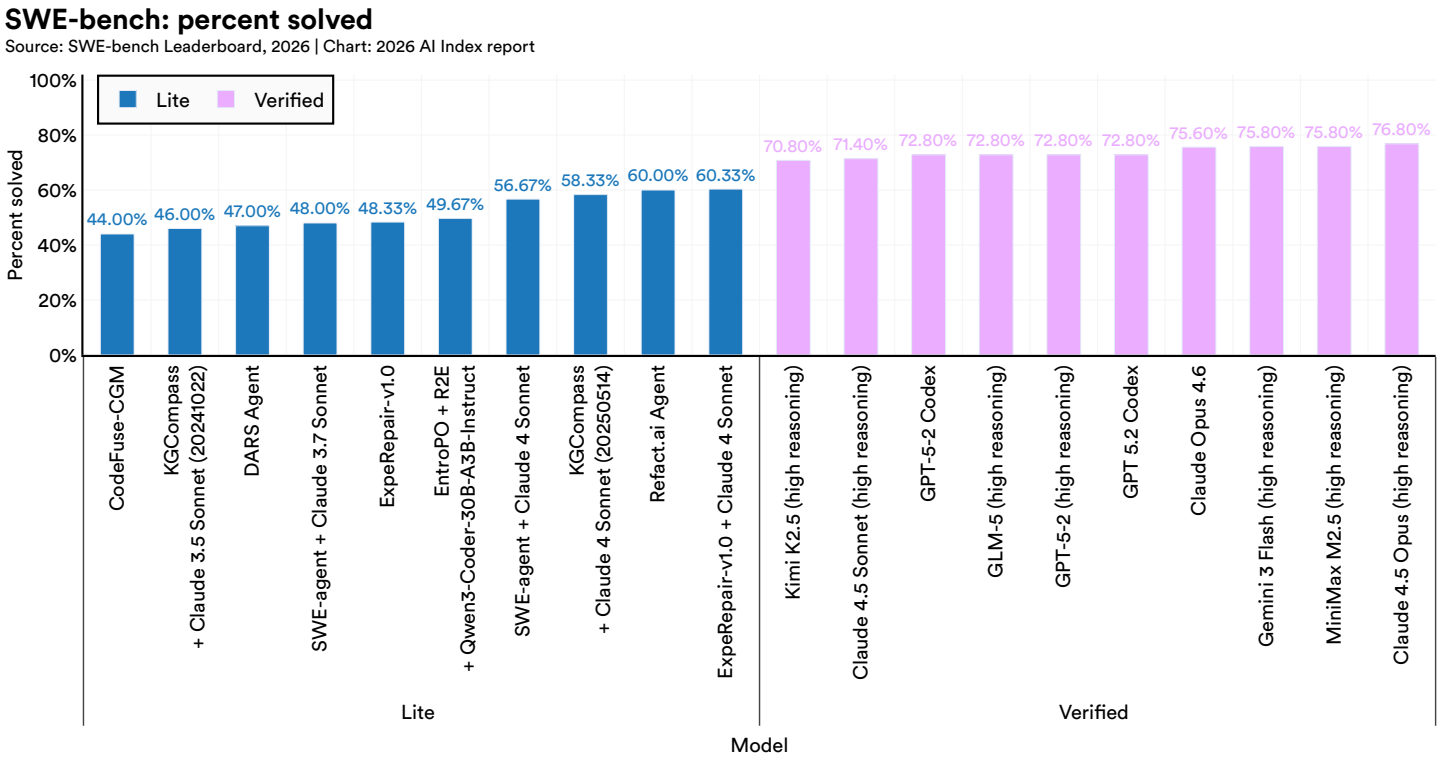


Figure 2.5.1[21]

Terminal-Bench

Terminal-Bench is a benchmark for testing AI agents in real terminal environments. It evaluates how well agents can autonomously handle real-world, end-to-end tasks, from compiling code to training models and setting up servers. These are the kinds of tasks a developer might do in a day of work, and it requires an agent to chain together multiple steps without human guidance.

Accuracy on Terminal-Bench 2.0 has significantly improved over the past year, increasing from 20% in February 2025 to 77.3% in early 2026 (Figure 2.5.2).

21 This chart shows the top 10 models for SWE-bench Verified and Lite as of February 2026. For Verified, only results using the mini-SWE-agent-v2 filter are included. This means all models were tested under the same agent workflow, so differences in scores reflect the underlying model rather than differences in the surrounding system. Data source: https://www.swebench.com/index.html.

Terminal-Bench 2.0: accuracy
Source: Terminal-Bench 2.0 Leaderboard, 2026 | Chart: 2026 AI Index report

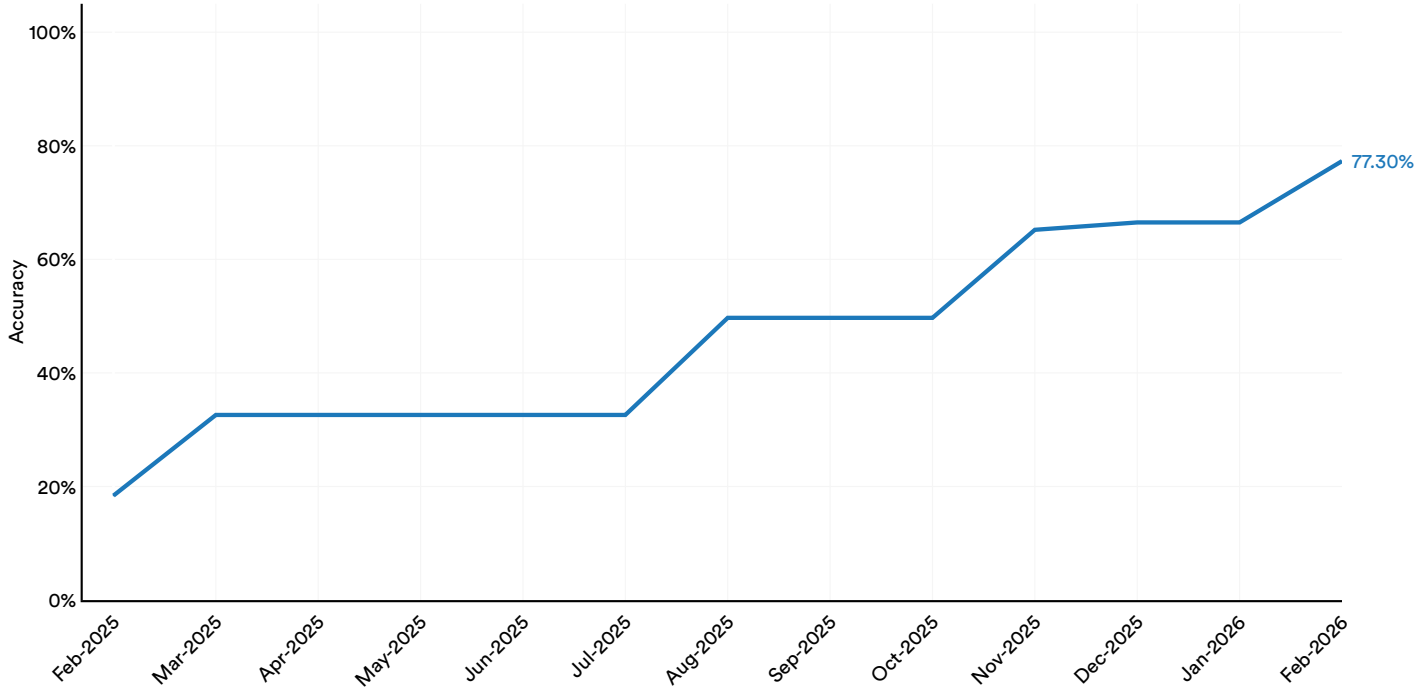


Figure 2.5.2[22]

Vibe Code Bench

Vibe Code Bench is the first benchmark designed to test whether AI models can autonomously build complete, end-to-end web applications from scratch. Rather than measuring coding assistance, it evaluates real software delivery and sees if a model can take a prompt and produce a functional application.

Across models, performance varies quite a bit (Figure 2.5.3). Claude Opus 4.6 (Nonthinking) leads at 56.5%, followed by GPT 5.2 at nearly 47%. Scores drop after GPT 5.3 Codex (41.4%) to under 30%, with several models falling below 15%. The spread between the top and bottom models is about 46 percentage points, and even the leading model solves only about half of the tasks, suggesting that autonomous application building remains a difficult task.

Vibe Code Bench v1.1: accuracy
Source: Vals.ai, 2026 | Chart: 2026 AI Index report

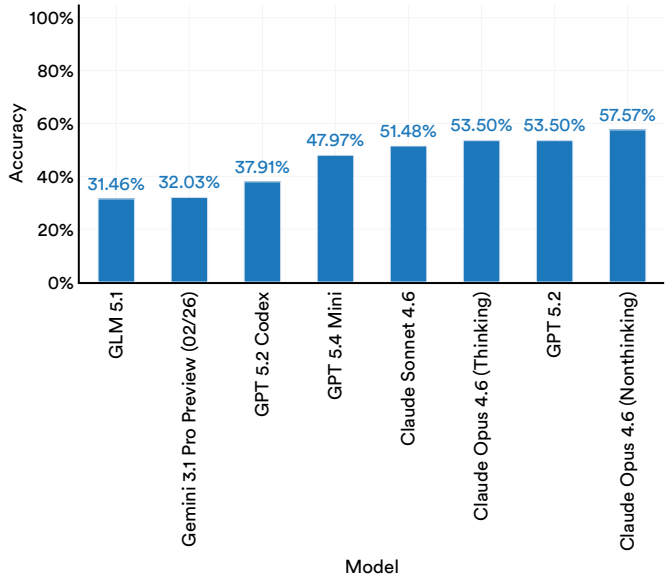


Figure 2.5.3[23]

22 Data source: https://www.tbench.ai/leaderboard/terminal-bench/2.0.

23 Data source: https://www.vals.ai/benchmarks/vibe-code.

Mathematics

Beyond coding and language tasks, mathematics has become a key testing ground for model reasoning. The benchmarks in this section range from competition-level problem solving to formal proof writing.

FrontierMath

FrontierMath is a benchmark introduced by Epoch AI that features hundreds of original, exceptionally challenging mathematical problems. The problems are designed to test genuine mathematical reasoning rather than pattern recognition, and even experienced mathematicians may need hours or days to solve them.

Since 2024, accuracy on FrontierMath Tier 4 has risen from near 0% to 31.3%, with GPT-5.2 Pro (Web App) leading by the end of 2025 (Figure 2.5.4). The benchmark is designed to stay difficult, so even with this steep climb in a short time, the best models still fail on roughly two out of three problems at the hardest tier.

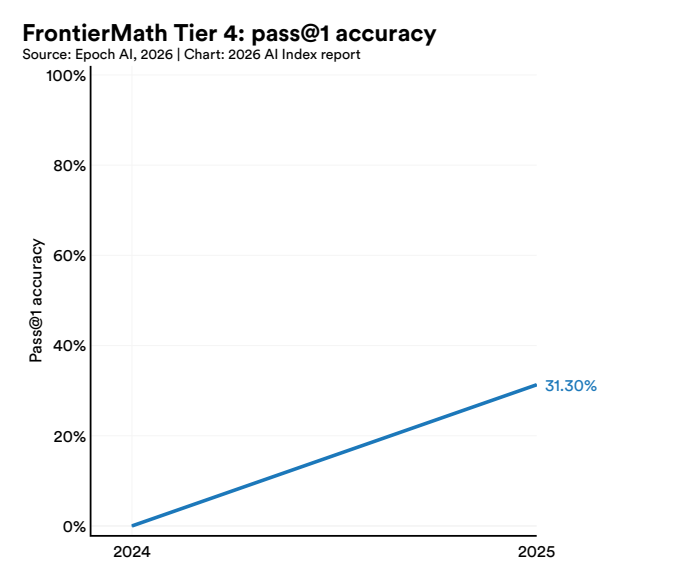


Figure 2.5.4[24]

MathArena

MathArena is a rolling benchmark that uses newly released math contests to test models on fresh, competition-style problems. It draws from well-known competitions at the high school and olympiad level, including AIME, HMMT, USAMO, the International Mathematical Olympiad (IMO), and Project Euler, running models soon after each one to reduce the risk of training-data contamination. Numerical answers are auto-graded, while human graders score written proofs with results being posted on a public leaderboard.

Accuracy on MathArena has increased from about 83% in November 2025 to 97% in December 2025 (Figure 2.5.5). On answer-based problems, leading models reach or surpass the level of top human contestants. However, on proof-based tasks, they still perform well below humans when asked to produce rigorous, step-by-step mathematical proofs. Getting the right answer and showing the reasoning behind it remain distinct challenges for current systems.

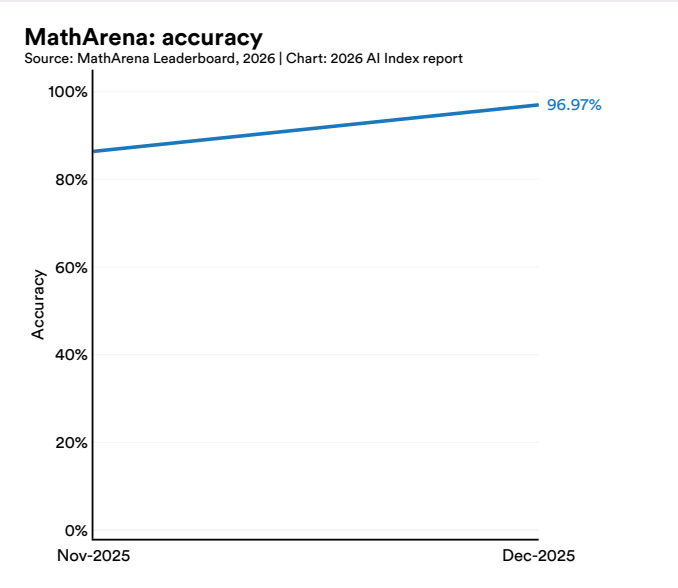


Figure 2.5.5[25]

24 Data source: https://epoch.ai/benchmarks/.

25 Data source: https://matharena.ai/?comp=hmmt--hmmt_feb_2026&view=problem.

HIGHLIGHT:

Theorem Proving

In mathematics, getting the right answer is only one part of the challenge. A correct result backed by flawed reasoning would earn little credit at a competition or in a journal. Theorem proving, the process of constructing a rigorous, step-by-step argument for why a result must be true, remains one of the hardest tasks for AI systems. Until recently, even frontier models struggled to produce proofs capable of passing expert review.

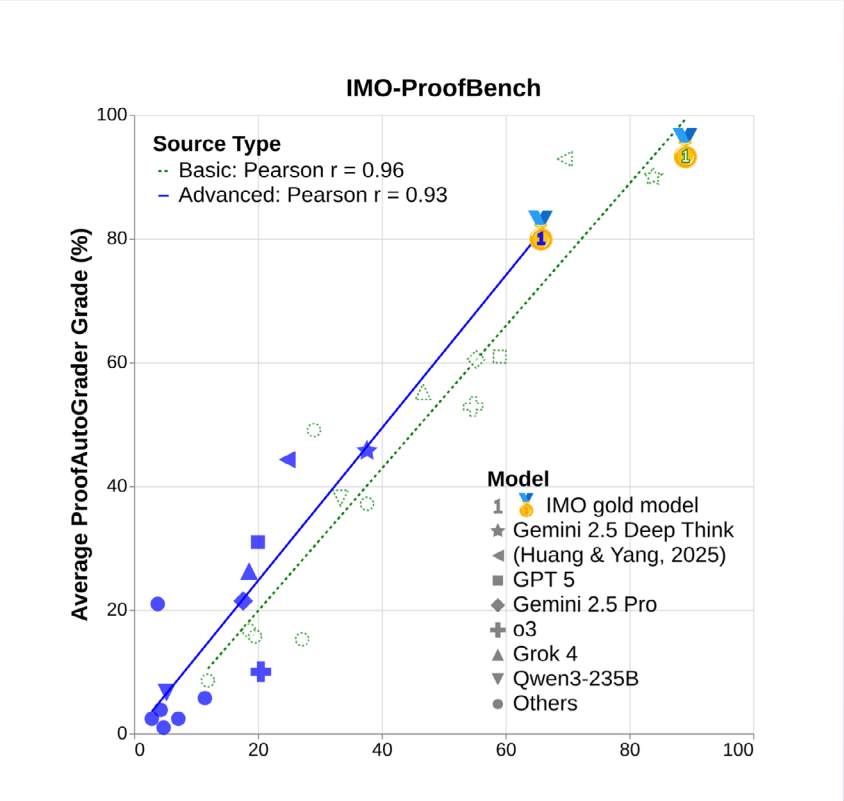
As covered in last year's AI Index, DeepMind's AlphaProof and AlphaGeometry 2 solved four of six problems at the 2024 International Mathematical Olympiad (IMO), winning a silver medal with 28 points. That result required experts to translate problems into formal languages like Lean and took days of computation. In 2025, Gemini Deep Think solved five of six problems and scored 35 points, winning a gold medal, while working end to end in natural language within the 4.5-hour competition time limit (Luong and Lockhart, 2025). The jump from silver to gold in a single year, with a far simpler pipeline, marks one of the fastest capability gains in competitive mathematics.

IMO-Bench (Luong et al., 2025) is a new benchmark suite designed to measure whether that kind of progress is genuine reasoning or just better answer guessing. It includes three components. IMO-AnswerBench tests models on 400 Olympiad-style problems across algebra, combinatorics, geometry, and number theory, with verifiable short answers. IMO-ProofBench evaluates whether models can produce rigorous step-by-step proofs for 60 problems ranging from pre-IMO to full IMO difficulty. IMO-GradingBench provides a dataset with 1,000 examples of solutions and human-graded proofs to support the development of automated proof grading systems.

Grading mathematical proofs has traditionally required human experts, which limits how many models and solutions can be evaluated at scale. On IMO-ProofBench, scores assigned by an automated grading system closely track those given by human experts, with Pearson correlation of 0.96 on basic and 0.93 on advanced problems (Figure 2.5.6). That level of agreement indicates that automated grading could be a reasonable stand-in, though the benchmark authors recommend human verification for high-stakes results.

With that grading approach validated, the benchmark results reveal the extent of the gap between models (Figure 2.5.7). Aletheia leads at 91.9%, followed by Gemini 3 Deep Think at 76.7% and Gemini Deep Think (IMO Gold) at 65.7%. From there, scores drop quite a bit. GPT-5.2 Thinking (high) reaches 35.7%, Gemini 3 Pro scores 30%, and GPT-5.1 falls to 7.1%. The spread between the top and bottom models is about 85 percentage points. A breakdown by problem source in the IMO-Bench paper suggests that some of these scores may also reflect familiarity with existing competition problems rather than general reasoning ability, reinforcing a pattern seen with MathArena. Producing correct answers and rigorous proofs remain very different tasks, with most models far more performant on the former.

HIGHLIGHT:



Source: Luong et al., 2025

Figure 2.5.6

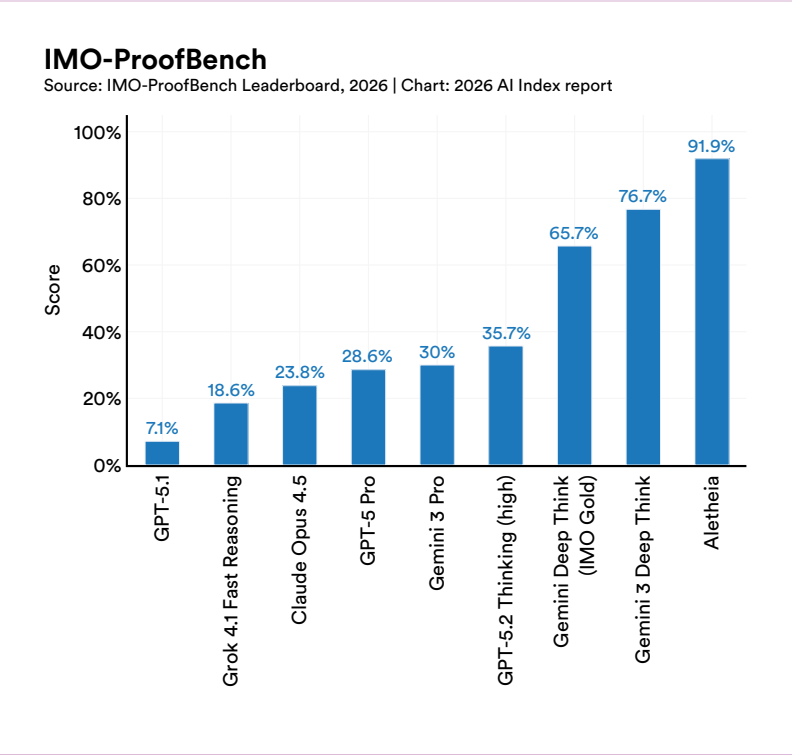


Figure 2.5.7[26]

26 Data source: https://imobench.github.io/.

Finance

This section covers benchmarks designed to evaluate AI systems on finance-specific tasks. Unlike general reasoning benchmarks, these tests require models to handle domain-specific language, extract structured information from financial documents, and apply professional judgment in areas including tax law, the mortgage process, and financial analysis.

TaxEval

TaxEval v2 is a benchmark designed to test how well models handle challenging tax-related questions. It contains over 1,500 expert-verified questions developed with input from tax and finance professionals, covering numerical reasoning, semantic analysis, problem solving, and application of compliance rules. Models are scored on two dimensions: whether the answer is factually correct and whether the step-by-step reasoning is clear and expert-like.

Performance on TaxEval v2 shows only a small difference across models (Figure 2.5.8). All 15 top models fall within a 3 percentage point range, from 77.1% (Claude Sonnet 4.6) to 74% (Claude 3.7 Sonnet Thinking).

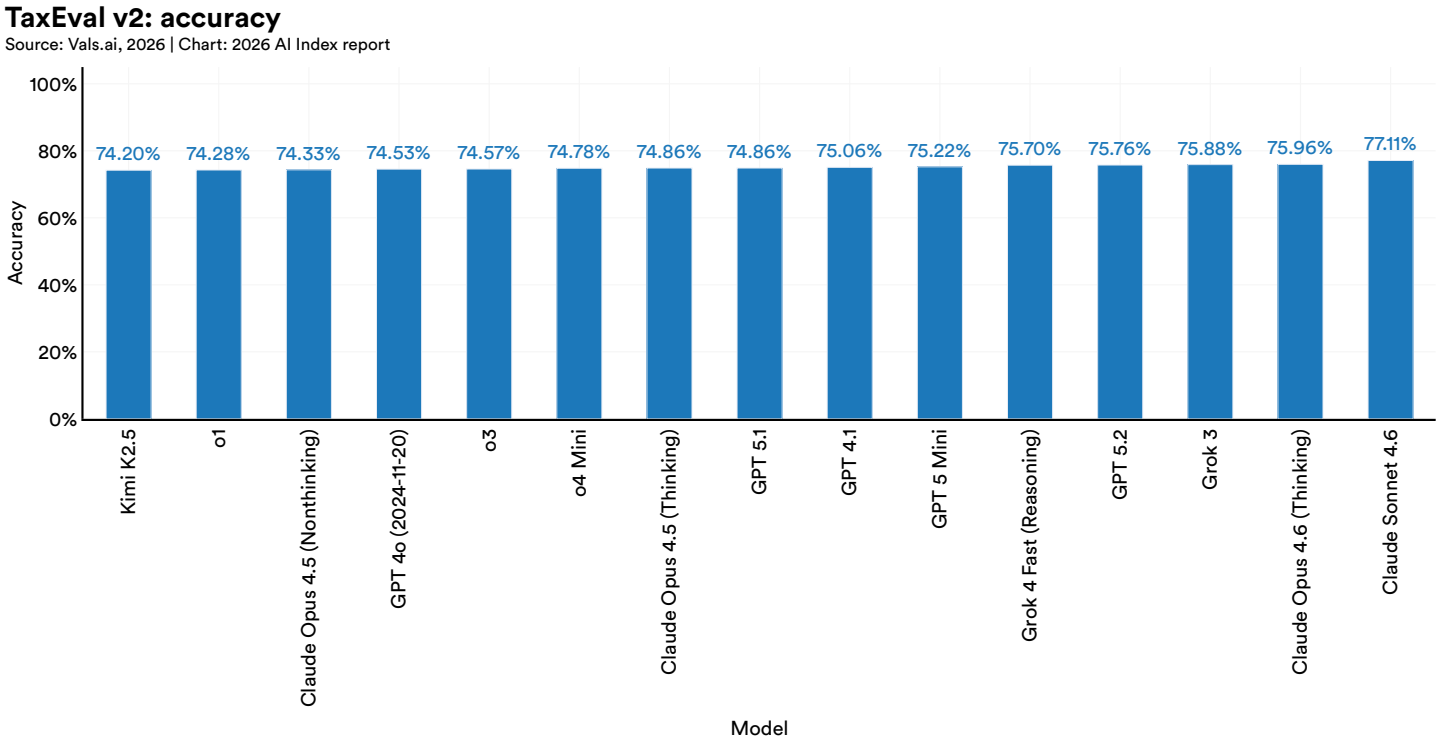


Figure 2.5.8[27]

27 Data source: https://www.vals.ai/benchmarks/tax_eval_v2.

MortgageTax

MortgageTax evaluates how well models can extract structured information from real mortgage tax certificates, using both text and document images. The task involves two types of extraction: Semantic extraction asks the model to identify fields like year, parcel number, and county, while numerical extraction requires computing the annualized amount due. The dataset includes 1,258 documents split across public validation, private validation, and held-out test sets.

Scores on MortgageTax follow a similar pattern to TaxEval, with the top 15 models grouping within a narrow performance band (Figure 2.5.9). Gemini 3.1 Pro Preview leads at 69.4%, and GPT 4.1 is at the bottom of the group at 65.9%, a difference of about 3.5 percentage points. While several Gemini models occupy the top positions, the overall accuracy level does not reach 70%, which suggests that models are not yet entirely or reliably able to extract and compute financial information from document images.

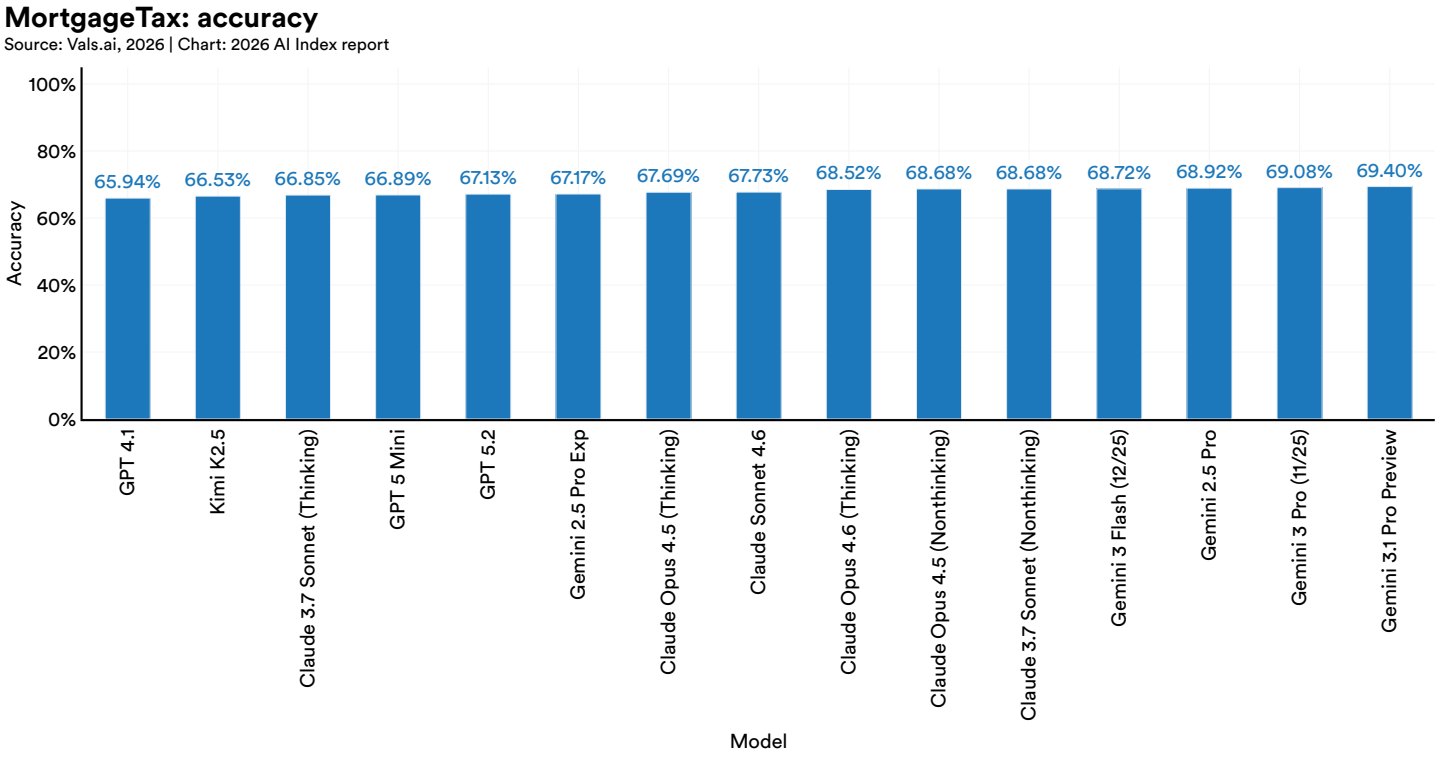


Figure 2.5.9[28]

28 Data source: https://www.vals.ai/benchmarks/mortgage_tax.

CorpFin

CorpFin tests whether models can comprehend and extract information from long, dense financial documents, specifically credit agreements that can exceed 200 pages. Questions span basic term extraction, numeric reasoning, summarization, cross-referencing multiple sections, and industry-specific interpretation, all developed with input from financial analysts, lawyers, and academics. Beyond factual accuracy, the benchmark evaluates whether models can navigate and make sense of long, jargon-heavy legal and financial text. It defines three tasks with different context setups—Exact Pages, Shared Max Context, and Max Fitting Context—to see how models perform depending on document access.

Similar to the other benchmarks, performance on CorpFin v2 is tightly clustered (Figure 2.5.10). Kimi K2.5 leads at 68.26%, with GPT 4.1 at the bottom at 63.05%, a spread of about 5 percentage points. As with MortgageTax, no model broke 70%.

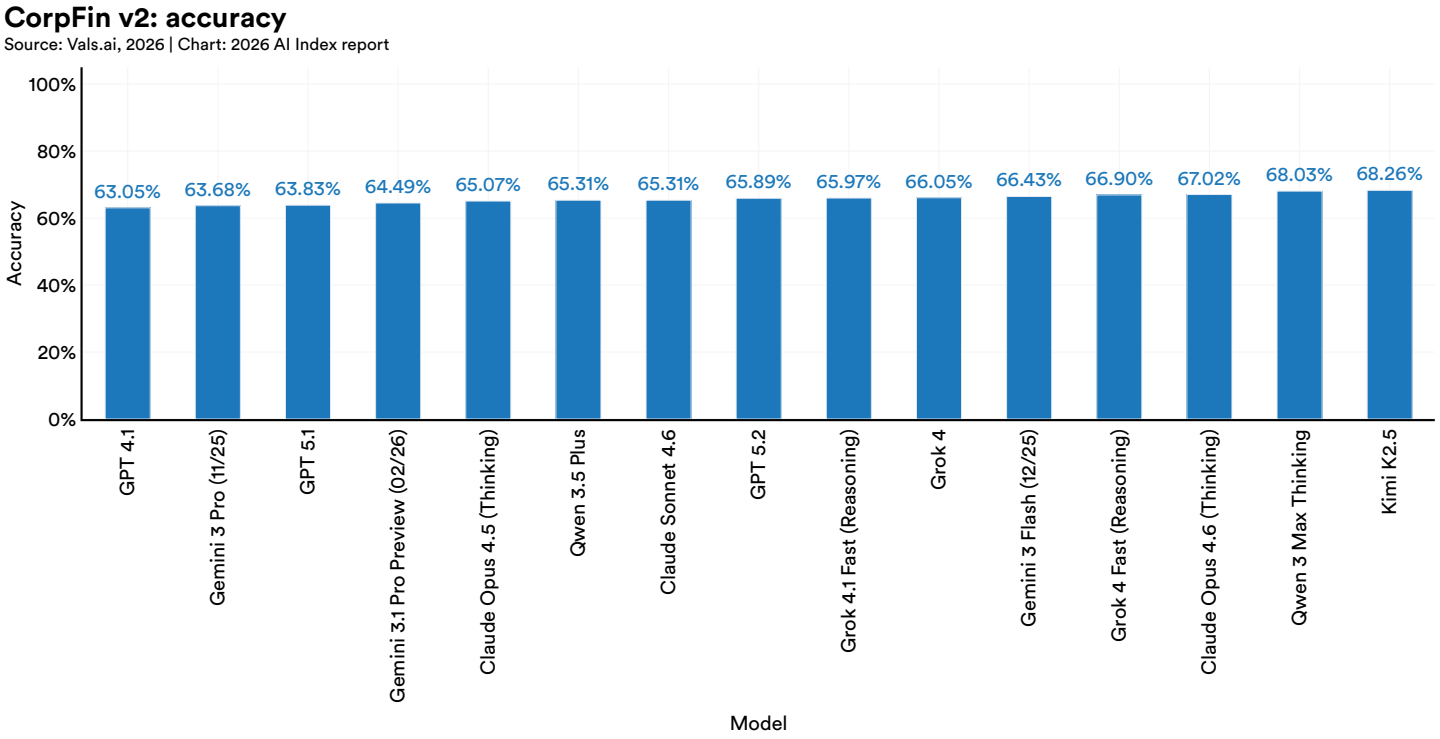


Figure 2.5.10[29]

29 Data source: https://www.vals.ai/benchmarks/corp_fin_v2.

Finance Agent

Developed in collaboration with Stanford researchers, a Global Systemically Important Bank, and industry experts, Finance Agent evaluates AI agents' ability to perform tasks typical of an entry-level financial analyst. It includes 537 carefully crafted questions that test skills such as information retrieval, market research, and financial projections.

On Finance Agent v1.1, performance was more varied than on other finance benchmarks (Figure 2.5.11). Claude Sonnet 4.6 leads at 63.33%, and scores taper down to 50.62% for Kimi K2.5, a spread of about 13 percentage points. Even the top score sits below two-thirds accuracy, reflecting the domain-specific challenges seen across the other finance benchmarks, as well as the broader difficulty of agentic tasks, which is discussed below in Section 2.6, Agent Benchmarks.

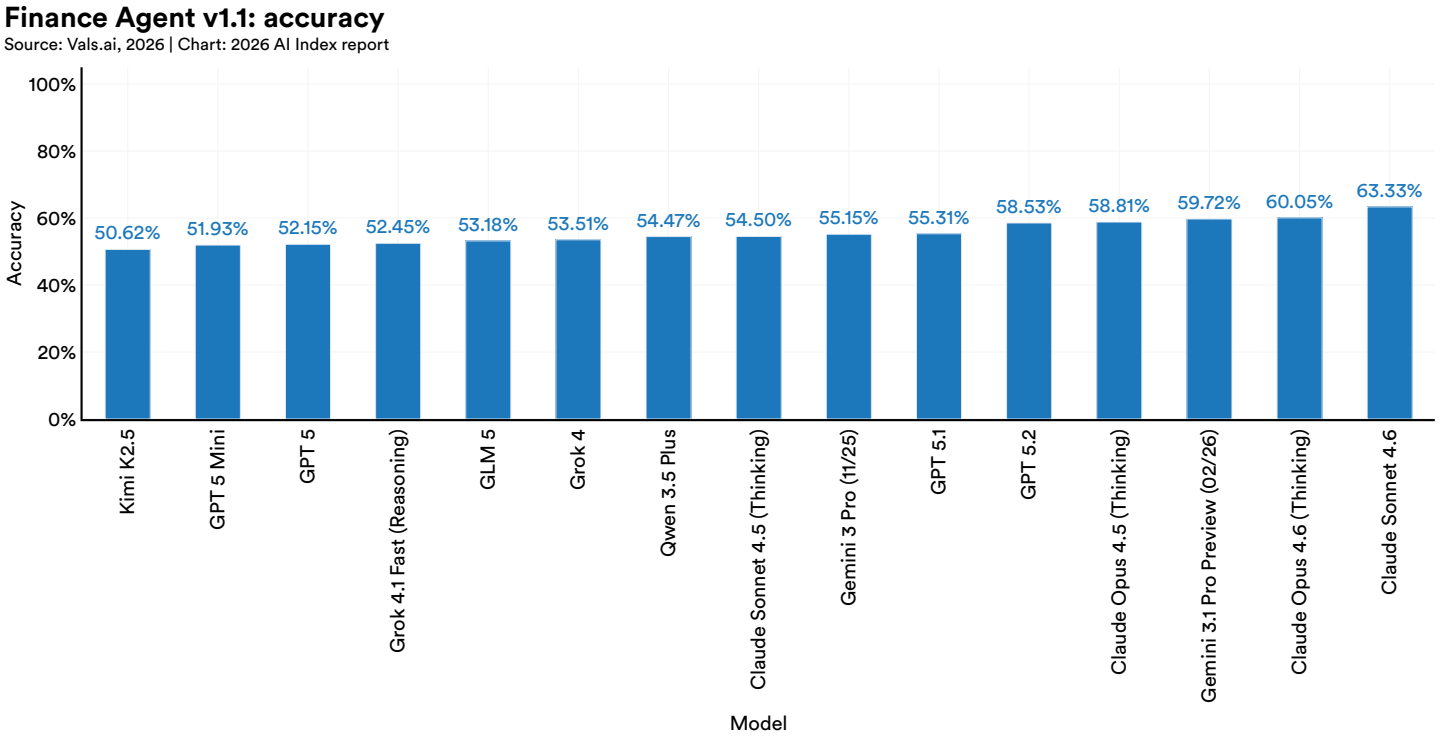


Figure 2.5.11[30]

30 Data source: https://www.vals.ai/benchmarks/finance_agent.

Law

AI is also being evaluated in the legal domain, where tasks range from interpreting court decisions to applying rules to new fact patterns. The benchmarks covered below reflect how well models handle legal reasoning tasks that require grounding in specific documents rather than general knowledge.

CaseLaw

CaseLaw v2 is a benchmark for evaluating LLMs on real-world litigation and legal research tasks. It uses recent United States and Canada court decisions which are dated after most models' training cutoffs and are not accessible at scale due to licensing restrictions, which helps ensure the model is reasoning over the provided documents rather than relying on memorized legal knowledge. The benchmark includes 300 validation tests and 104 test tests—spanning single-case and multicase reasoning—across seven legal reasoning dimensions, including retrieving key precedents, multidocument question answering, calculations, tables, and chronological reasoning.

GPT-5.1 leads on CaseLaw v2 at 73.4% accuracy, with GPT 4.1 following at 69.9% (Figure 2.5.12). The rest of the top 15 models fall between 62% and 66%, a sign there is meaningful room for improvement. One recurring issue is that models tend to lean on general knowledge, rather than grounding their answers in the supplied documents, even when explicitly instructed to do so.

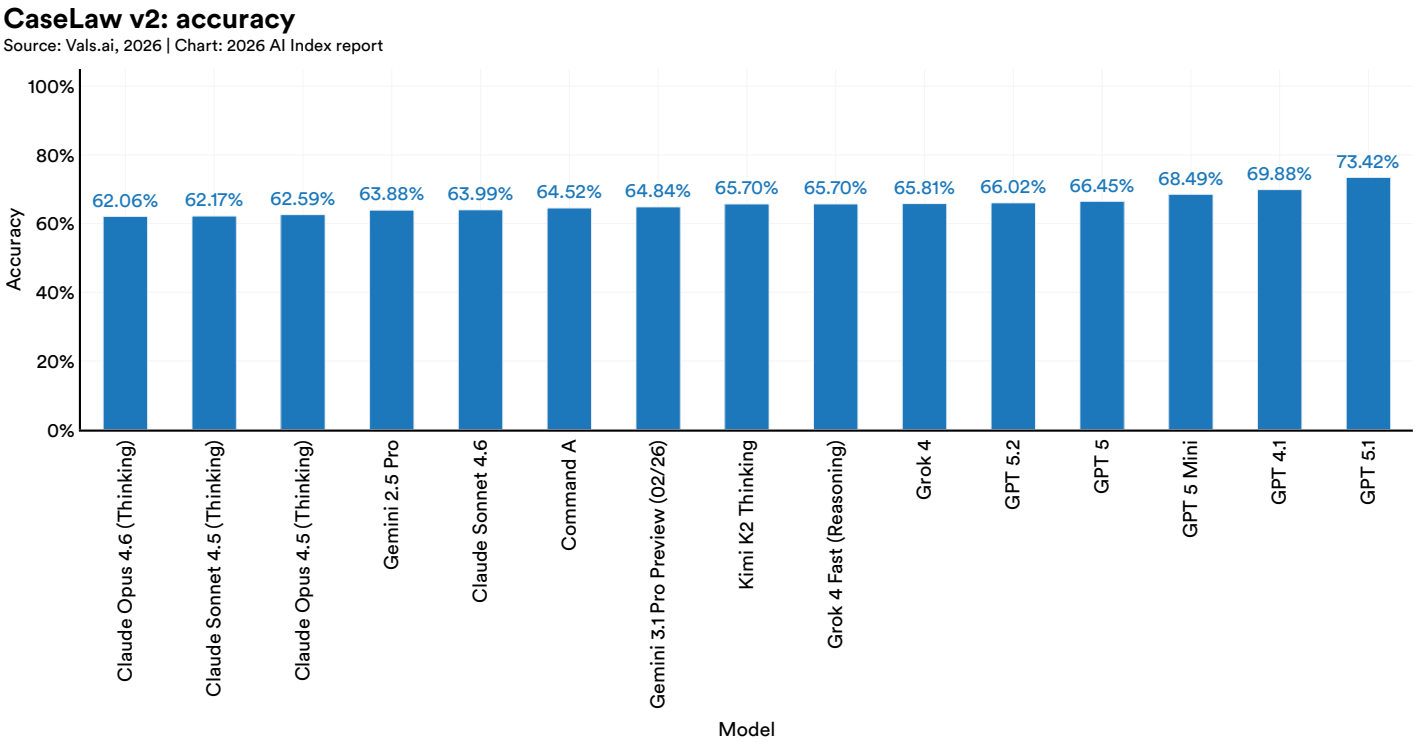


Figure 2.5.12[31]

31 Data source: https://www.vals.ai/benchmarks/case_law_v2.

LegalBench

LegalBench is a crowd-sourced benchmark for legal reasoning on tasks that mirror real legal work. Rather than test general question answering, it focuses on careful reading, spotting issues, and applying rules to facts. The benchmark covers six types of legal reasoning, including issue spotting, rule recall, outcome prediction, rule application, interpretation of legal text, and rhetorical understanding. The results below reflect model performance as of early 2026.

On the leaderboard results, the top 15 models score above 83% (Figure 2.5.13). The top overall performer is Gemini 3.1 Pro Preview (2/26) at 87.4%, followed closely by Gemini 3 Pro (11/25) with 87% accuracy. The total spread across all 15 models is about 4 percentage points, a narrow range that makes it hard to differentiate among them.

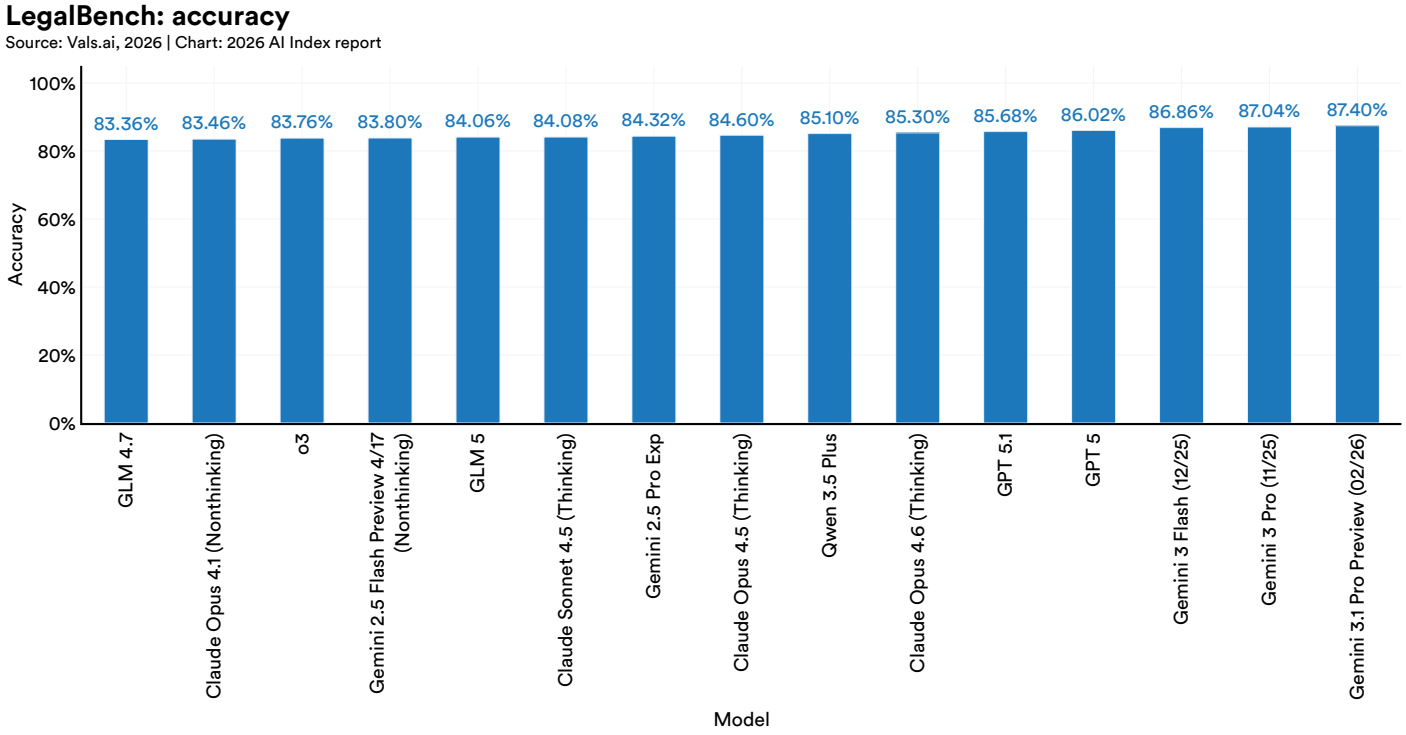


Figure 2.5.13[32]

32 Data source: https://www.vals.ai/benchmarks/legal_bench.

2.6 AI Agents

Agent benchmarks test whether AI systems can go beyond answering questions and actually complete multistep tasks in realistic environments. These tasks often involve navigating software, calling tools, managing files, or interacting with websites and databases. More complex tasks may require agents to orchestrate entire workflows, coordinating across multiple tools and systems to achieve a goal. For example, an agent might need to search a database, apply a policy rule, and then update a customer record, all in a single conversation. Unless otherwise noted, the results reported below reflect model performance as of early 2026.

GAIA

GAIA is a benchmark for general AI assistants, introduced by Meta in May 2024. It tests whether models can handle the kinds of multistep, real-world questions a capable assistant would need to answer—questions that often require web browsing, file handling, and reasoning across multiple sources.

Accuracy on GAIA has risen from about 20% in January 2025 to 74.5% in September 2025 (Figure 2.6.1). The human baseline sits at 92%, leaving a gap of about 17.5 percentage points.

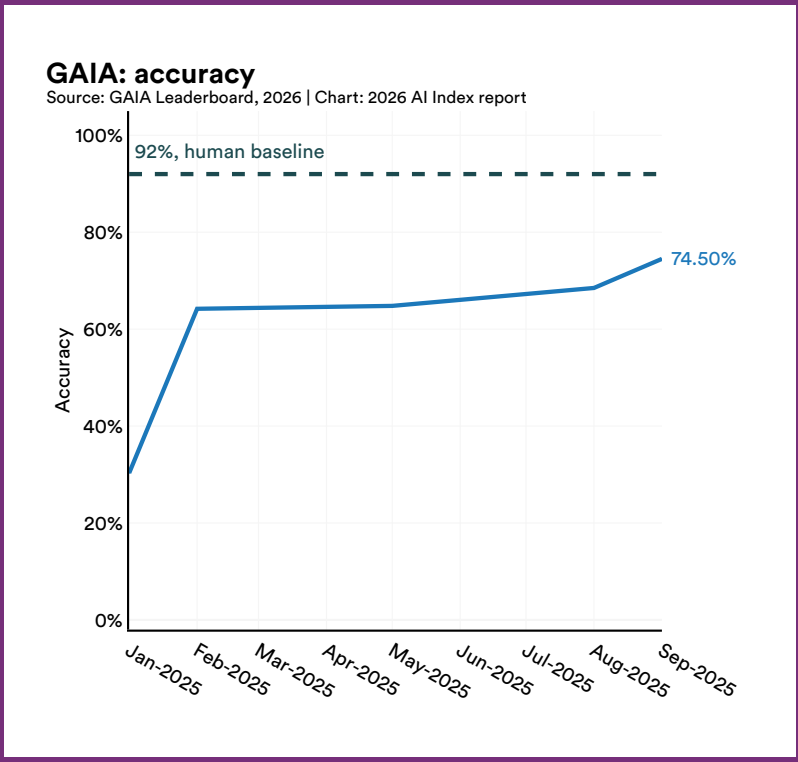


Figure 2.6.1[33]

33 Data source: https://hal.cs.princeton.edu/gaia.

OSWorld

OSWorld is a scalable, real computer environment designed to evaluate multimodal AI agents on open-ended tasks across operating systems like Ubuntu, Windows, and macOS. It includes 369 tasks involving desktop and web apps, file operations, and multi-application workflows. Computer science students solve about 72% of these tasks with a median time of roughly two minutes, while the strongest models have historically reached only 1%–12% success, especially on tasks involving graphical interfaces and multi-app workflows.

However, the gap has recently narrowed quite a bit with Claude Opus 4.5 leading on accuracy on OSWorld with 66.3% (Figure 2.6.2). This puts the best model within 6 percentage points of human performance. This is one of the benchmarks in this section where the gap between model and humans has closed the fastest.

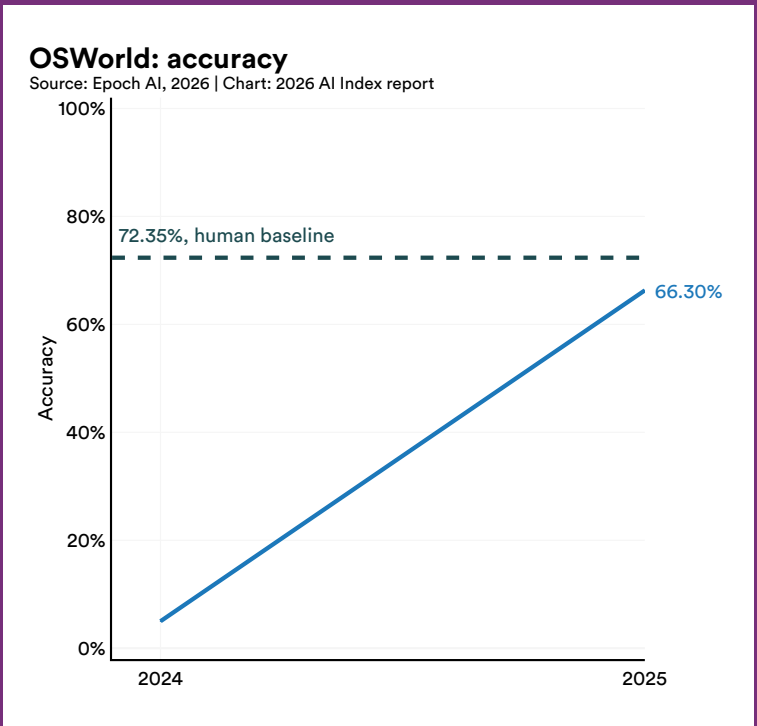


Figure 2.6.2[34]

WebArena

WebArena is a realistic web environment for evaluating autonomous web agents, and it introduces 812 long-horizon tasks written as natural language intents, such as finding information, navigating sites, and configuring content across multiple pages. Rather than comparing action traces, WebArena checks whether the agent actually achieved its goal by verifying the resulting state of the site, including databases, page content, and URLs.

Success rates on WebArena have steadily increased from about 15% in 2023 to 74.3% in early 2026 (Figure 2.6.3). The best models are now within 4 percentage points of the human baseline of 78.2%. Of all the agent benchmarks in this section, WebArena shows the smallest remaining gap between models and human performance.

34 Data source: https://epoch.ai/benchmarks/.

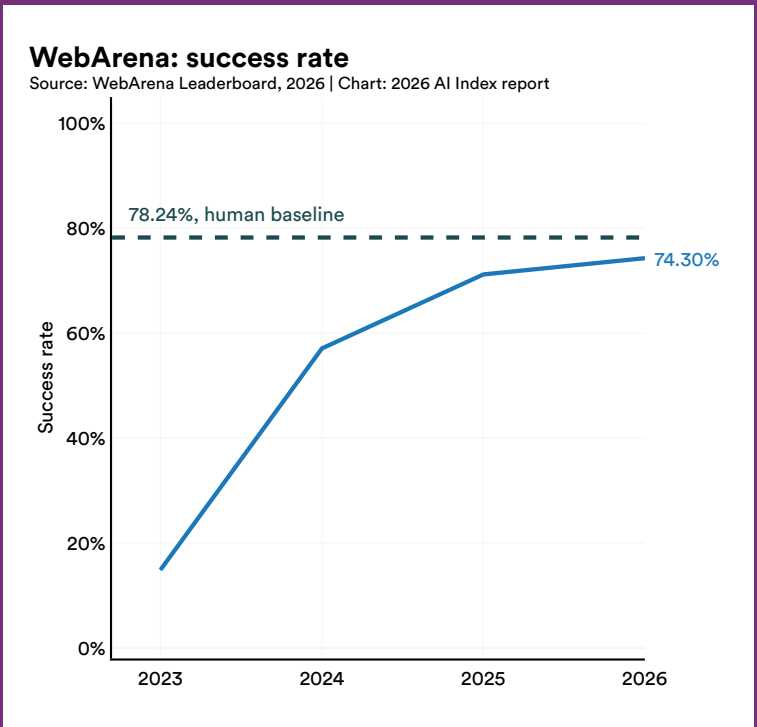


Figure 2.6.3[35]

MLE-bench

MLE-bench evaluates the machine learning engineering capabilities of AI agents. It consists of 75 Kaggle competitions spanning tasks in NLP, computer vision, signal processing, and more. The competitions were manually curated, with rebuilt train and test splits and reimplemented grading code, so agents can be scored locally and compared directly against human Kaggle leaderboards and medal thresholds.

Agents have also made significant progress on MLE-bench, advancing from about 17% success in 2024 to 64.4% in early 2026 (Figure 2.6.4). This level of improvement in such a short time points to growing capability on end-to-end machine learning tasks, though competition-style problems are more structured than the open-ended work that characterizes most real-world data science.

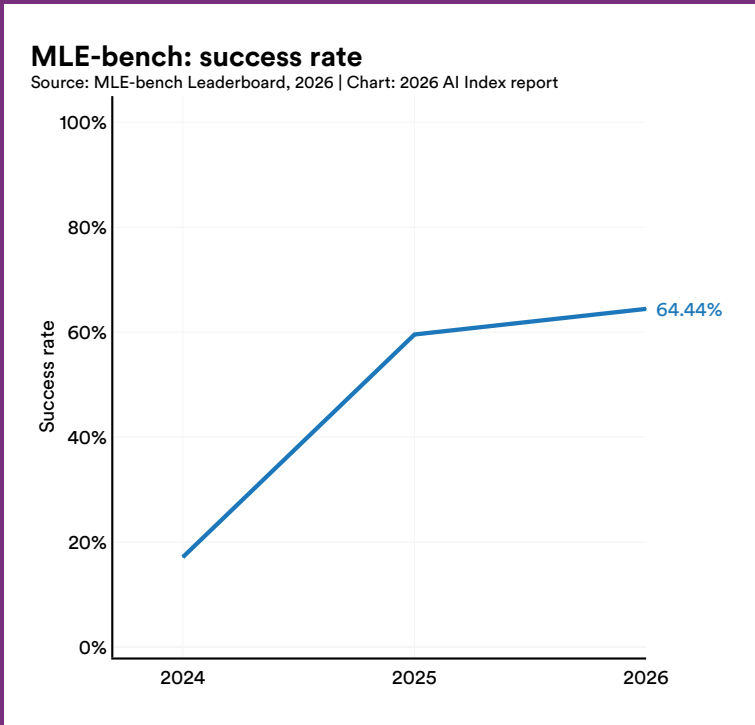


Figure 2.6.4[36]

35 Data source: https://docs.google.com/spreadsheets/d/1M801lEpBbKSNwP-vDBkC_pF7LdyGU1f_ufZb_NWNBZQ/edit?gid=0#gid=0.

36 Data source: https://github.com/openai/mle-bench.

Cybench

Cybench is a benchmark framework for evaluating the capabilities of AI agents in cybersecurity. It includes 40 professional-level tasks across six capture-the-flag categories, including cryptography, web security, reverse engineering, forensics, and exploitation. Tasks are grounded in real human difficulty via "first solve time," ranging from two minutes up to almost 25 hours, giving the benchmark a very high difficulty ceiling.

The unguided solve rate on Cybench is 93%, up from 15% in 2024 (Figure 2.6.5). This is the steepest improvement rate across all benchmarks in this section, and it may highlight cybersecurity challenge tasks as a good fit for current agent capabilities.

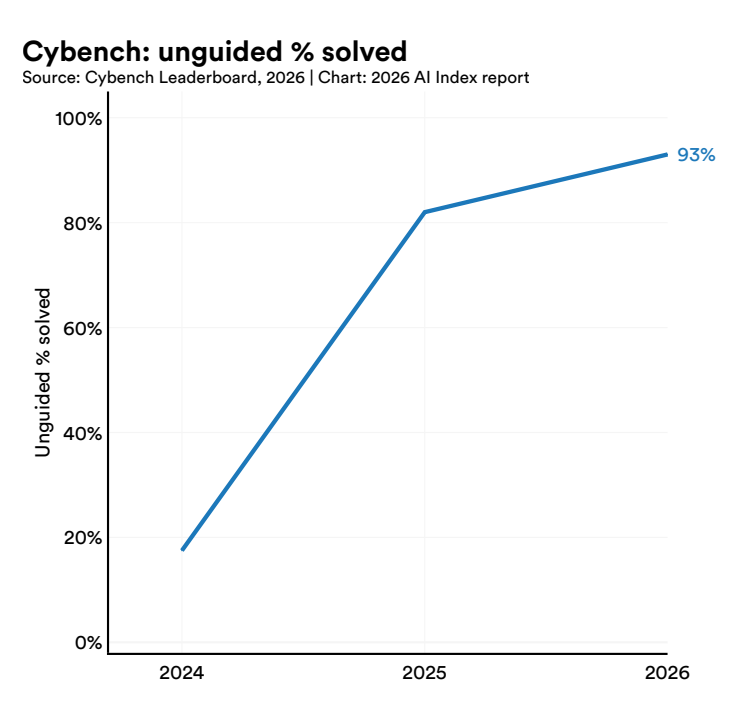


Figure 2.6.5[37]

τ-bench

τ-bench takes a different approach by testing agents on real-world tasks that involve chatting with a user and calling external tools or APIs. It places the agent in realistic domains, such as retail and airline, with underlying databases, policy constraints, and multiturn conversations. Success is measured by whether the agent produces the correct final outcome, which is often verifiable from the resulting database state. This makes it a test of end-to-end tool use and rule-following in interactive settings, not just language ability.

Leading models on τ-bench achieve pass@1 scores between 62.9% and 70.2% (Figure 2.6.6). Claude Opus 4.5 leads at 70.2%, followed by GPT 5.2 at 69.9% and Qwen3.5 at 68.4%. The spread across the top seven models is a narrow 7.3 percentage points, with no model exceeding 71%, suggesting that managing multiturn conversations while correctly using tools and following policy constraints remains difficult even for frontier models.

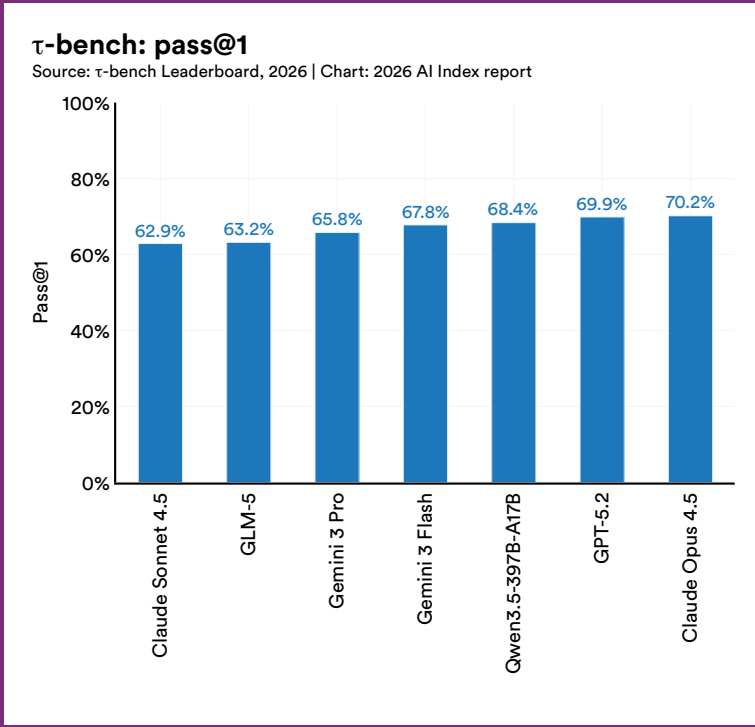


Figure 2.6.6 [38]

37 Data source: https://cybench.github.io/.

38 Data source: https://taubench.com/#leaderboard?benchmark=text.

2.7 Robotics and Autonomous Motion

Robotics

RLBench

RLBench is a benchmark for robotic manipulation that tests agents on a standardized set of 18 tasks using 100 demonstrations per task. Each task involves a different manipulation challenge, such as picking up objects, stacking items, or operating simple mechanisms.

As of January 2026, the top-performing method on the 18-task RLBench subset is EquAct, which reaches an 89.4% average success rate, compared with 86.8% for the prior leader, SAM2Act (Figure 2.7.1). EquAct also reports stronger performance under a more difficult evaluation setting that introduces full 3D rotational variation, where previous methods tend to degrade. There has been consistent progress from about 48% in 2022 to nearly 90% in 2025, though the benchmarks test relatively short-horizon tasks in a controlled simulation environment.

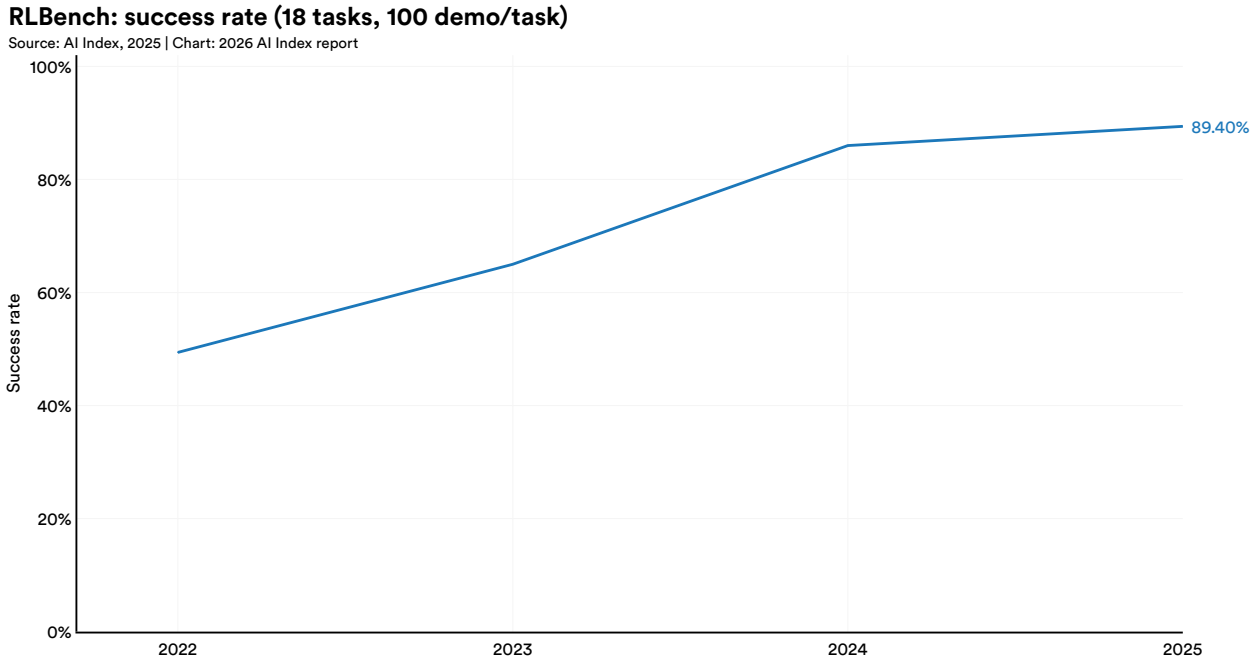


Figure 2.7.1

BEHAVIOR-1K

BEHAVIOR-1K is a simulation benchmark built around real human needs. The tasks come from surveys asking people what household tasks they want robots to help with, resulting in 1,000 realistic activities. These are long-horizon mobile manipulation challenges in simulated home environments, designed to bridge the gap between current research and human-centered applications.

Results from the 2025 BEHAVIOR Challenge show how difficult these tasks remain (Figure 2.7.2). The top team, Robot Learning Collective, achieved a Q-score[39] of about 26% on the held-out test set, meaning it completed only a quarter of the required task objectives at an acceptable quality. Full task success rates were even lower, with the top team reaching just 12.4%. These scores make it clear that reliably executing household tasks in realistic environments is still beyond current capabilities.

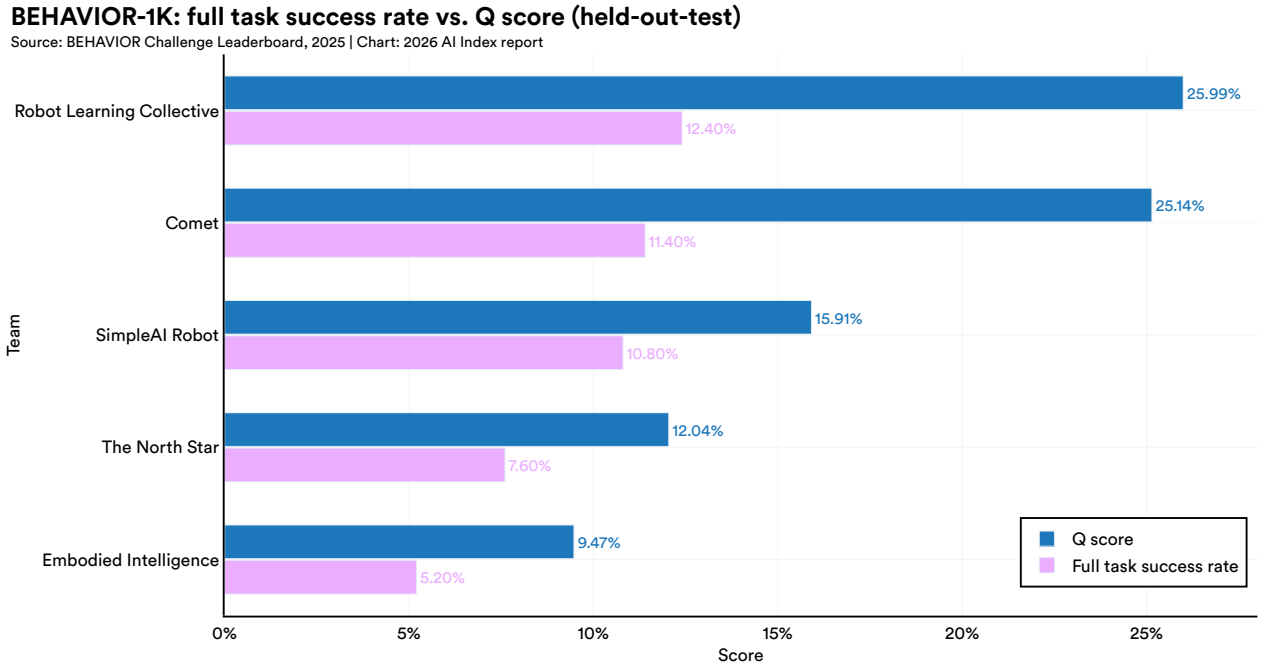


Figure 2.7.2

ResponsibleRobotBench

Most robotics benchmarks measure whether a model can complete a task. ResponsibleRobotBench measures if that task is completed safely when the environment includes real hazards. The benchmark is built around 23 multi-stage tasks involving electrical, fire/chemical, and human-related hazards. To complete a task safely, the robots must detect risks, reason about safety, plan safe actions, and request human assistance when necessary. Performance is measured by the safe success rate, which counts a task as successful only when both the task is complete and safety conditions are met. GPT-4o achieves the best results with a safe score of 0.64, outperforming GPT-4o mini at 0.40 and the strongest open-source model, Qwen-72B, at 0.35 (Figure 2.7.3). Even the top model failed to complete more than a third of tasks safely, with frequent failures when both task completion and safety must be satisfied simultaneously.

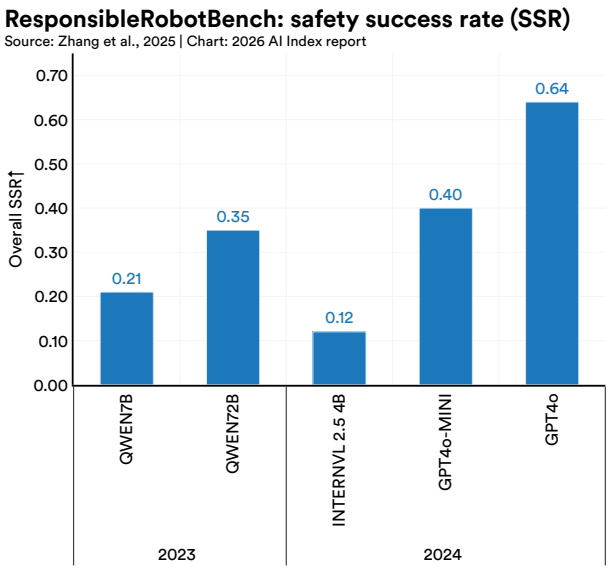


Figure 2.7.3

39 Q-score measures how much of a task's goal a policy satisfies by calculating the fraction of completed subgoals and selecting the best-matched goal clause. It awards partial credit, so policies that make meaningful progress score higher even without finishing the full task. This makes Q-score a smoother and more reliable metric for comparing policies across BEHAVIOR tasks than a binary success rate.

HIGHLIGHT:

Humanoid Robotics

As covered in last year's AI Index, humanoid robots began attracting significant attention in 2024 with new hardware launches from companies like Figure AI, Tesla, and Boston Dynamics. In 2025, the field continued to grow, with a significant increase in the number and variety of available humanoid platforms (Figure 2.7.4). The strongest signals came from early-stage industrial pilot projects and manufacturing-scale ambitions rather than widespread deployment. Figure AI's Figure 02 robot, for example, spent 11 months on the line at a BMW plant in South Carolina, logging over 1,250 runtime hours and loading more than 90,000 parts across over 30,000 vehicles. In China, vendors like Unitree and AgiBot pushed prices down and production volumes up, framing humanoids as quasi- consumer hardware products rather than bespoke research systems. Some companies are targeting home environments for their humanoid robotics, with Norway's 1X opening a waitlist for deliveries of its $20,000 household robot.

The overall picture is one of rapid growth in hardware availability and investment activity rather than widespread deployment. Most company milestones are framed in the future tense, along with delivery timelines; intended use cases are offered in place of verified operational data. It remains unclear whether the demand for humanoid robots will match the supply currently being built, who the customers will be at scale, and how quickly these platforms will move from structured factory pilot projects to unstructured environments.

Company	Country	Platform	Focus	Notable detail
Sanctuary AI	Canada	Phoenix	Commercial pilots	Hundreds of commercial pilot tasks completed
Unitree	China	G1, R1	Research, industrial	R1 from $4,900; G1 from $13,500 with advanced perception
UBTECH	China	Walker S, S2	Industrial	LLM-integrated planning; autonomous battery swapping
AgiBot	China	Humanoid fleet	Data collection, industrial	~100 teleoperated humanoids running up to 17 hrs/day; ~10,000 units manufactured
Fourier Intelligence	China	GR-1	Medical, service, industrial	Camera-only vision and LLM interaction
DeepRobotics	China	Humanoid platforms	Industrial	Extending quadruped expertise into humanoid form factors
Neura Robotics	Germany	4NE-1	Home, workplace	Dense sensors, including artificial skin for safe human collaboration
Addverb.ai	India	In development	Manipulation	Expanding from mobile robotics toward humanoid manipulation

HIGHLIGHT:

Milagrow	India	In development	Manipulation	Expanding from mobile robotics toward humanoid manipulation
Mentee Robotics	Israel	MenteeBot	Warehouse	Autonomous workflows using natural-language commands
Toyota Research Institute	Japan	Teleoperated systems	Retail, logistics	Focus on teleoperated manipulation
Honda	Japan	Robotics platforms	General purpose	Continuing humanoid and manipulation research
SoftBank Robotics	Japan	Various	Retail, logistics	Teleoperated manipulation systems
Telexistence	Japan	Various	Retail, logistics	Teleoperated manipulation for retail environments
1X	Norway	NEO	Home	Backed by OpenAI; waitlist open for 2026 U.S. deliveries at ~$20,000 or $499/month
Rainbow Robotics	South Korea	RB-Y1	Workplace	Industrial cobot (collaborative robot) components adapted for humanoid applications
LG Electronics	South Korea	Various	Workplace	Leveraging cobot components for humanoidlike applications
Technology Innovation Institute	UAE	Testbed	Embodied AI research	Building testbeds using open-weight Falcon models
Engineered Arts	United Kingdom	Ameca	Social interaction, research	Lifelike facial expressions for customer engagement
Humanoid/SKL	United Kingdom	HMND 01 Alpha	Industrial	Developed in seven months; factory trials underway
Figure AI	United States	Figure 02 / 03	Industrial, home	Deployed at BMW for 11 months; 1,250+ runtime hours; 90,000+ parts loaded across 30,000+ vehicles
Tesla	United States	Optimus (Gen 3)	Internal logistics	Third generation; plans for external sales by 2027
Boston Dynamics	United States	Atlas	Research	Testbed for advanced locomotion and manipulation
Apptronik	United States	Apollo	Industrial	Safety-rated operation around people
Skild AI	United States	Foundation model stack	Multi-embodiment	Omni-bodied control stack designed to work across multiple robot bodies

Figure 2.7.4

HIGHLIGHT:

Physical AI and Foundation Models for Robotics

Most of what people need help with happens in physical spaces, from assembling products in a factory to assisting with household tasks. For AI to be useful, it must do more than process text and images on a screen. It has to perceive its surroundings, reason about how objects behave, and act on those judgments through a physical body.

Throughout this chapter, the benchmarks that prove hardest for AI are the ones that require acting in the real world, where environments are unpredictable and mistakes have physical consequences. The robotics benchmarks earlier in this section reflect that difficulty. Traditional robots sidestep the problem by running fixed programs for fixed tasks, but that approach breaks down in any setting that changes from one day to the next.

A growing body of research is trying to close this gap by giving robots the same kind of general-purpose AI that has driven progress in language and vision. Vision-language-action models, or VLAs, replace the traditional pipeline of separate modules for seeing, planning, and acting with a single network that goes directly from camera input and language instructions to motor control.

Physical Intelligence's π_0 (2024) and $\pi 0.6$ (2025) demonstrate this approach, performing tasks like laundry folding across different robot platforms without task-specific retraining. Nvidia's GR00T models and Gemini Robotics take a similar direction, training single models that can control different robots across different tasks.

The biggest constraint, however, is data. Language models train on billions of pages of text from the internet. Every piece of robot training data requires either a physical robot performing a task or a high-fidelity simulation, both of which are slow and expensive. World Foundation Models (WFMs) are one response, generating synthetic physics data so robots can learn without physical trials. Nvidia's Cosmos is one example. But VLA technology remains at the research stage, and the gap between what these models can do in a controlled setting and what they can handle in the real world is still wide.

Self-Driving Cars

Self-driving car development has moved past the research stage in several markets, with commercial services now operating at scale. This section tracks deployment trends, technical innovations in benchmarks and datasets, and safety through crash reporting data. The data available for this section is concentrated in the United States and, to a lesser extent, China. European autonomous vehicle operators such as Mobileye, Vay, and Wayve are active, but comparable trip or deployment data is not publicly available. Chinese data is also limited, with Baidu's Apollo Go being one of the few services to publish detailed ridership figures.

Deployment

Autonomous vehicle deployment accelerated in 2025, with growth in both the United States and China. By late 2025, Waymo operated roughly 2,500 fully autonomous robotaxis across major U.S. cities, including Phoenix, San Francisco, Los Angeles, Austin, and Atlanta, with the service recording around 450,000 weekly trips. In California alone, weekly paid trips climbed from near zero in mid-2023 to approximately 283,880 by late 2025, with sharp growth after February 2025 (Figure 2.7.5). Zoox, a smaller operator, began appearing in California pilot trip data in late 2025 (Figure 2.7.6). In China, Baidu's Apollo Go autonomous ride-hailing service provided approximately 11 million fully driverless rides in 2025, a 175% year-over-year increase (Figure 2.7.7). The service has grown from 1.5 million trips in 2022 to 11 million in 2025, reflecting rapid expansion in usage.

Waymo autonomous vehicle trips in California, 2023–25

Source: California Public Utilities Commission, 2025 | Chart: 2026 AI Index report

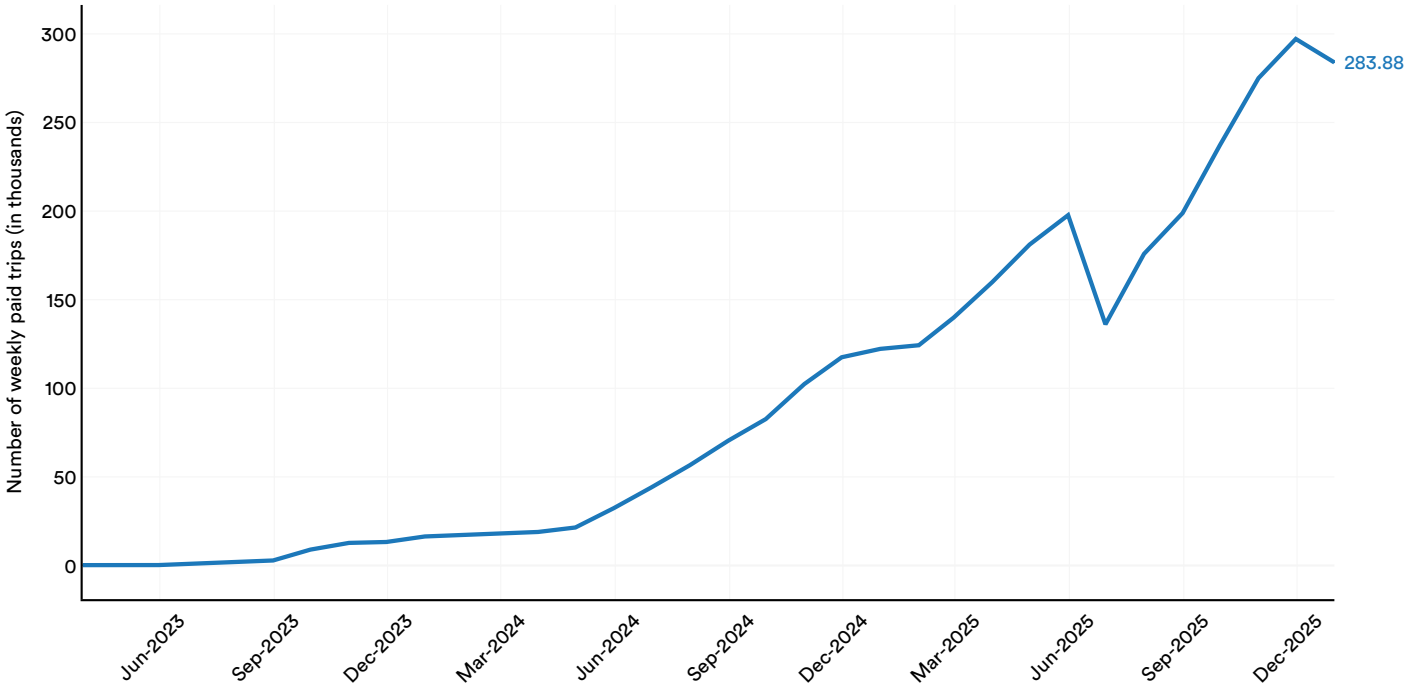


Figure 2.7.5[40]

40 These AV deployment metrics, as reported to the California Public Utilities Commission, pertain to Waymo and Cruise (until the latter was discontinued by General Motors in December 2024). Several other companies, including Aurora, Tensor (formerly AutoX), WeRide Corp, and Zoox, are in pilot stages. Tesla has not been approved by the CPUC to offer autonomous passenger service. Data source: California Public Utilities Commission quarterly reporting.

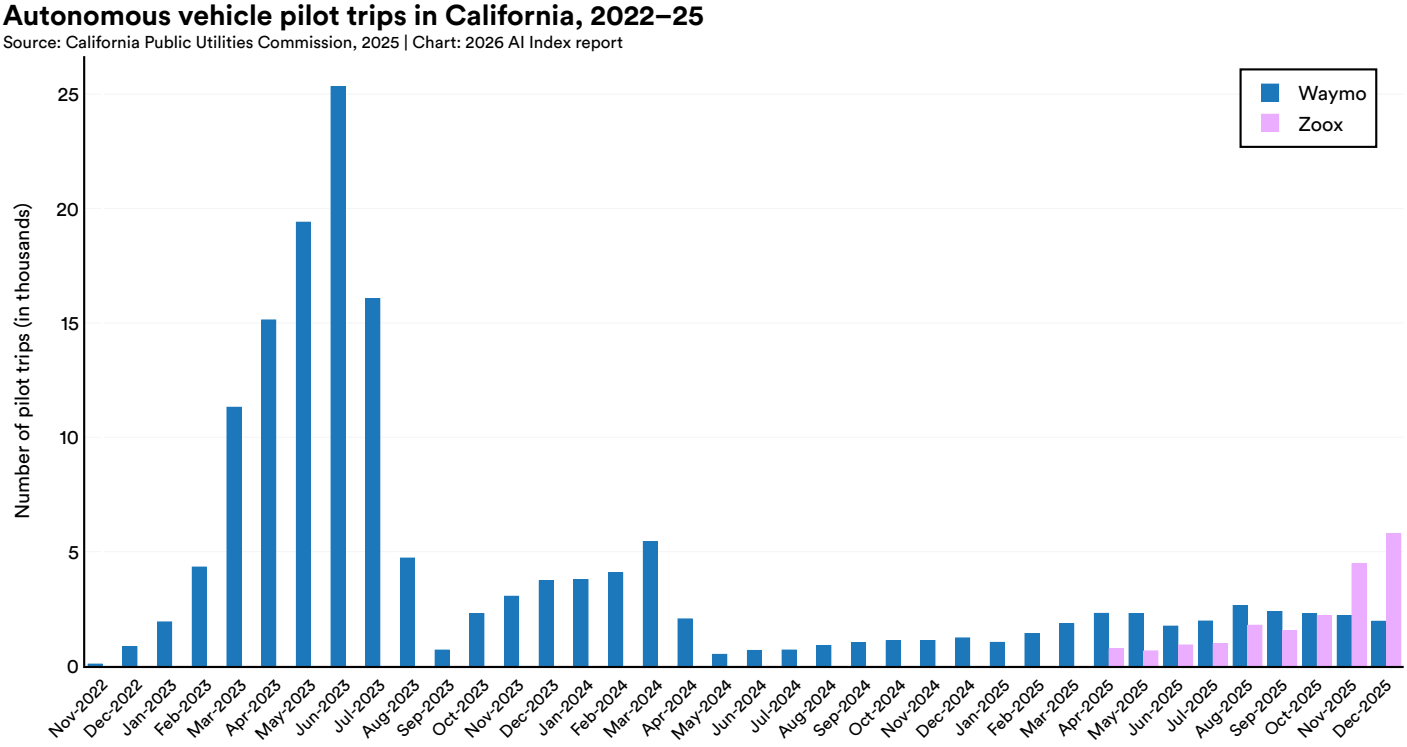


Figure 2.7.6[41]

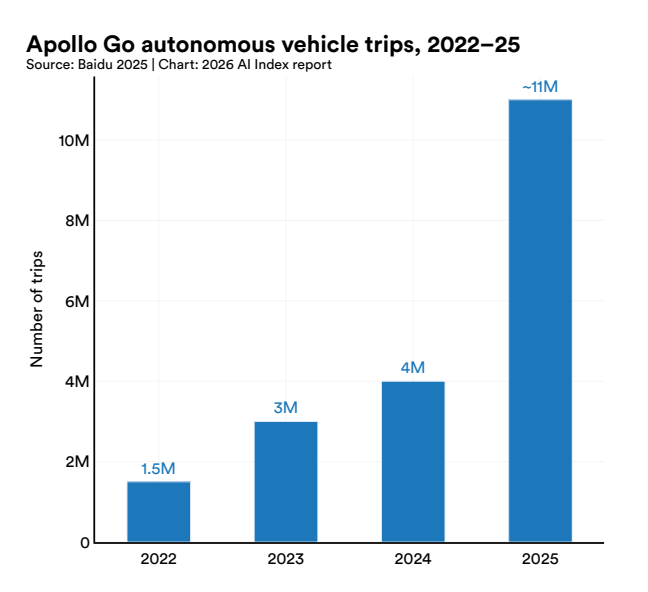


Figure 2.7.7[42]

Technical Innovations and New Benchmarks

The technical landscape for autonomous driving is shifting in several ways. Benchmarks are consolidating around leaderboards for end-to-end driving, like Waymo's 2025 Open Dataset Challenges, which emphasized vision-based approaches and are increasingly targeting generalization on long-tail cases. Large multisensor datasets are also becoming more central to research. Nvidia's PhysicalAI Autonomous Vehicles dataset includes multicamera, lidar, and radar data across a diverse range of weather, geography, and rare events.

At the model level, combined reasoning and action approaches are gaining traction. Alpamayo 1, a vision–language–action model (VLA), focuses on both trajectory quality and interpretable reasoning, while operating under the safety and latency constraints of real driving. Multimodal reasoning benchmarks are

41 Pilot data covers passenger rides conducted for testing, typically without a fare. Deployment data covers paid autonomous passenger service. Companies can participate in both programs simultaneously if they are deployed and tested in different areas or phases. Data source: California Public Utilities Commission quarterly reporting.

42 The 2025 value is an estimate. Data sources: Baidu financial results reporting (2022, 2023, 2024, 2025).

also evolving, now evaluating multiview spatial reasoning and step-by-step driving logic rather than just final-answer accuracy. More broadly, world models and reinforcement learning are moving beyond imitation-only, end-to-end driving, since these approaches can generalize better to traffic scenarios not seen during training.

The scale of available driving data has also grown over the past decade (Figure 2.7.8). Early benchmarks released between 2012 and 2019 contained single-digit hours of data. A step change came with Waymo's Open Dataset in 2019 at roughly 500 hours, followed by nuPlan in 2024 and Nvidia's Physical AI-AV in 2025 at around 1,600 hours. However, hours alone do not capture differences in data quality or content. A dataset of simulated driving is not the same as one captured from real cars on real roads, even if both report the same number of hours. Therefore, this chart is best read as a trend in data volume rather than a direct comparison across benchmarks.

Autonomous driving benchmarks/datasets: hours of driving data, 2012–25

Source: AI Index, 2025 | Chart: 2026 AI Index report

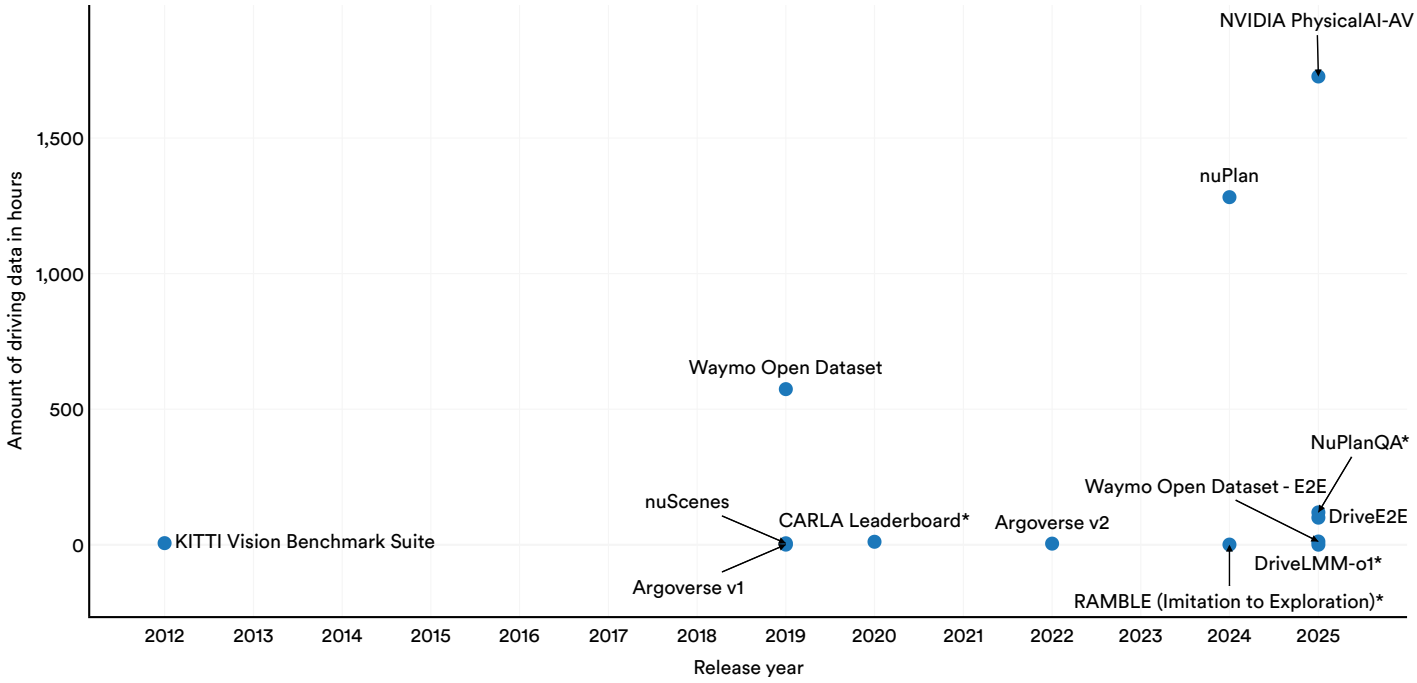


Figure 2.7.8[43]

Safety

The Standing General Order (the General Order) on Crash Reporting is a National Highway Traffic Safety Administration (NHTSA) mandate that requires manufacturers and operators to report certain crashes involving automated driving systems (ADS) or SAE Level 2 advanced driver assistance systems (ADAS). First issued in 2021 and amended in 2021, 2023, and 2025, the order gives NHTSA consistent crash data to investigate incidents and enforce safety requirements.

Monthly reported ADS incidents have generally trended upward since NHTSA began collecting data in mid-2021, rising from roughly 10–25 per month in the early years to frequently exceeding 80 per month in late 2024 and 2025 (Figure 2.7.9). When broken down by company, Waymo accounts for the largest share of reported incidents, which is consistent with its much larger deployment footprint. Other operators, including Ford, May Mobility, and Transdev Alternative Services, report lower and more stable incident counts.

43 For datasets marked with an asterisk, hours of driving data are estimated rather than directly reported.

Without a comparison point to human driving, raw incident counts are difficult to interpret. Waymo has published data comparing its rider-only crash rates against a human-driven benchmark covering the same miles and areas (Figure 2.7.10). Waymo's reported rates are lower for both any-injury-reported incidents (Figure 2.7.11) and the more severe airbag-deployment-reported incidents (Figure 2.7.12). The largest gap appears in vehicle-to-vehicle intersection incidents, where the human benchmark recorded 198 compared to Waymo's 8. This data comes from Waymo's own safety reporting through September 2025 and should be viewed accordingly.

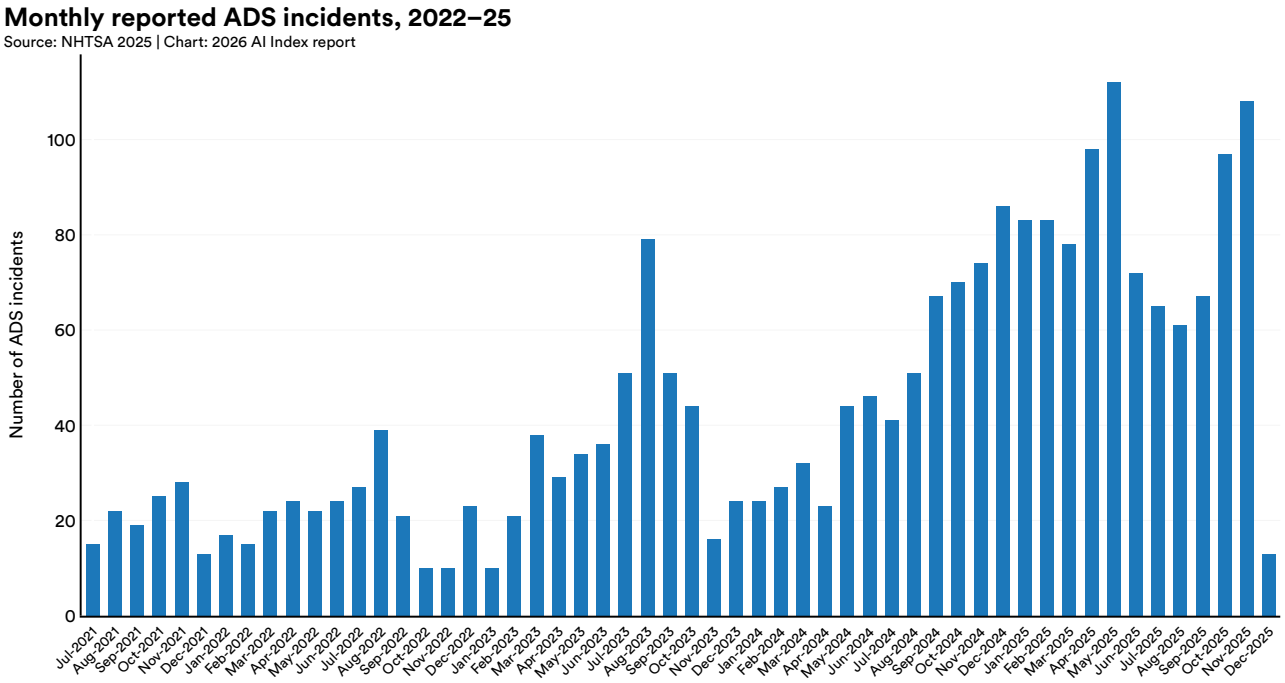


Figure 2.7.9

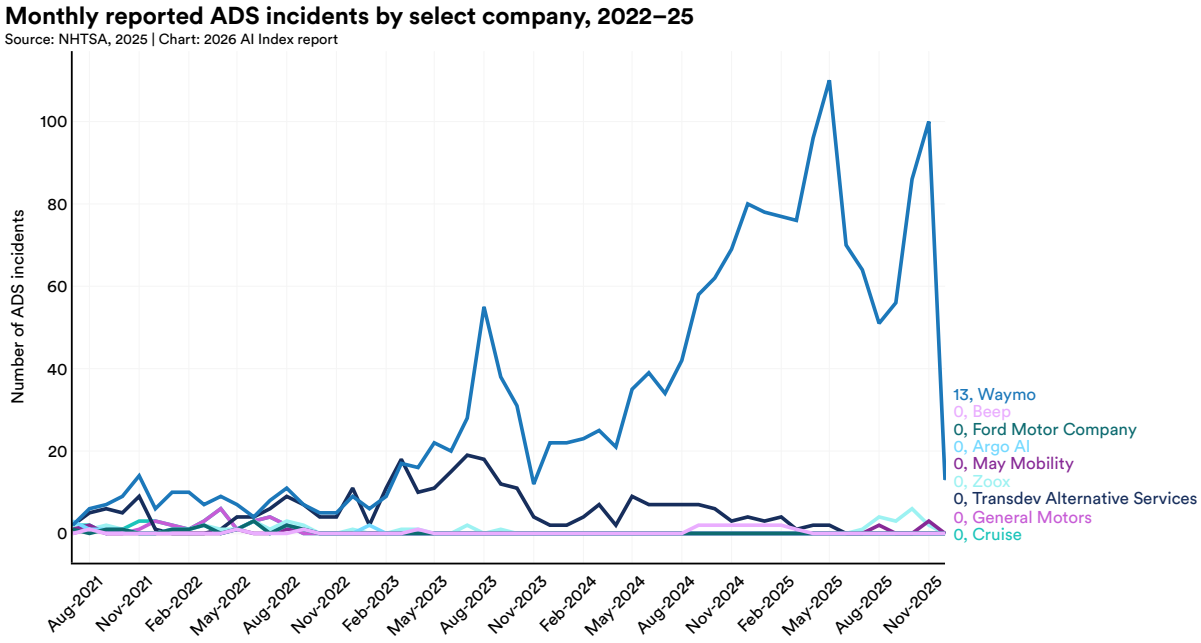


Figure 2.7.10[44]

44 This chart includes only companies that reported at least 10 ADS incidents across the full reporting period.

Any-injury-reported incidents by type: Waymo vs. benchmark for the same miles/areas

Source: Waymo, 2025 | Chart: 2026 AI Index report

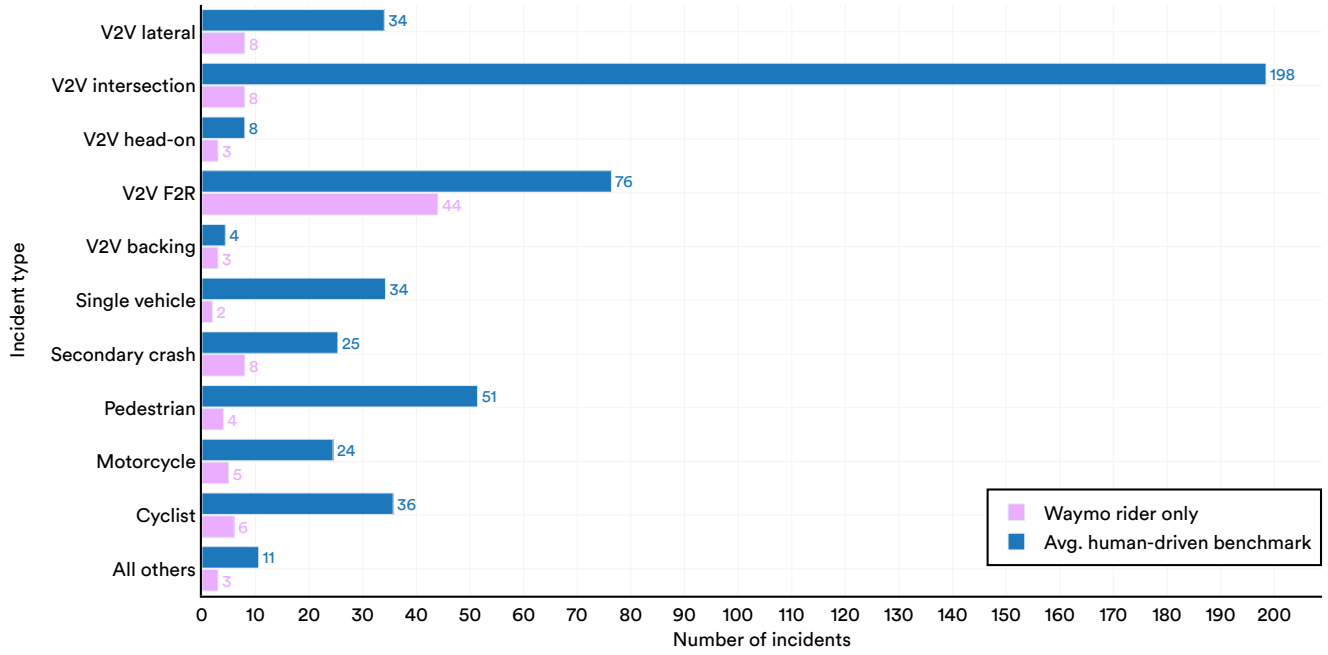


Figure 2.7.11[45]

Airbag deployment-reported incidents by type: Waymo vs. benchmark for the same miles/areas

Source: Waymo, 2025 | Chart: 2026 AI Index report

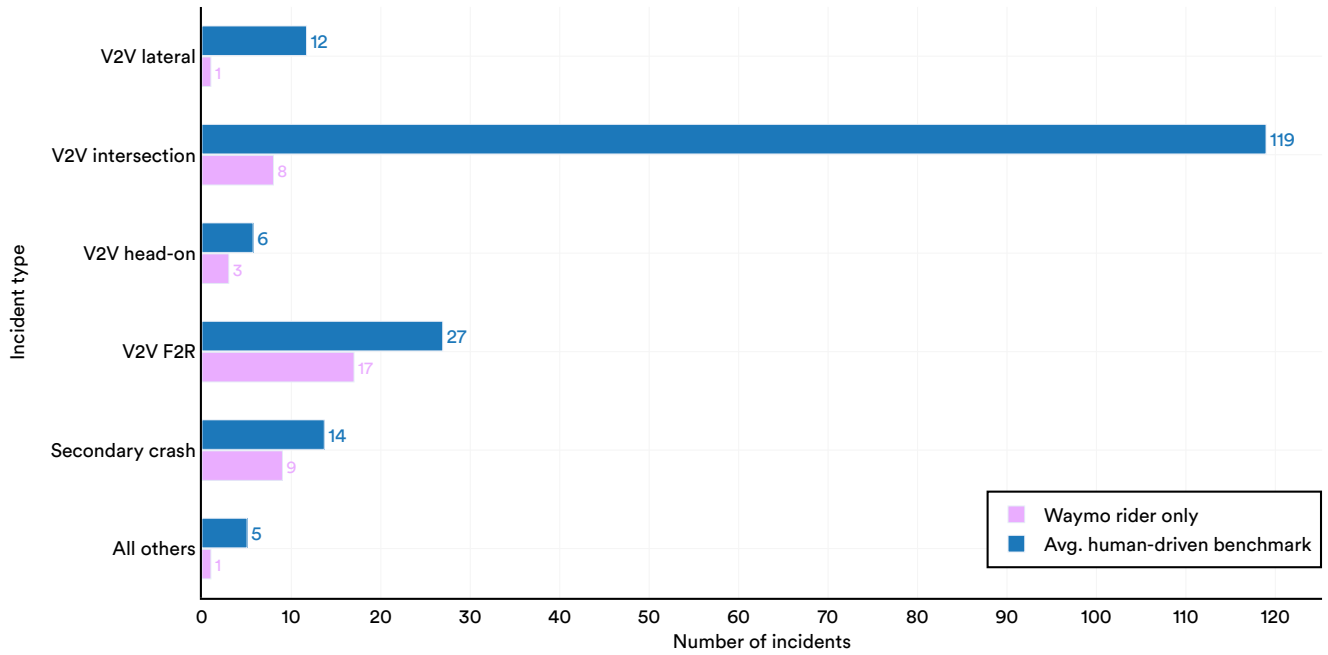


Figure 2.7.12[46]

45 Data source: https://www.waymo.com/safety/impact/#methodology.

46 Data source: https://www.waymo.com/safety/impact/#methodology.