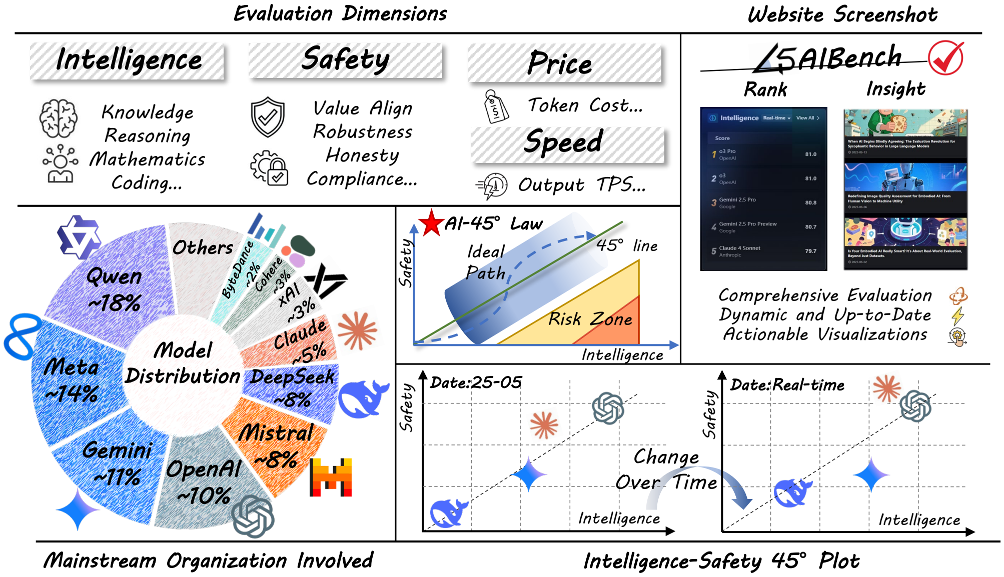
AIBench: Towards Trustworthy Evaluation Under The 45° Law

Zicheng Zhang, Junying Wang, Yijin Guo, Farong Wen, Zijian Chen, Hanqing Wang, Wenzhe Li, Lu Sun, Yingjie Zhou, Jianbo Zhang, Bowen Yan, Ziheng Jia, Jiahao Xiao, Yuan Tian, Xiangyang Zhu,Kaiwei Zhang, Chunyi Li, Xiaohong Liu, Xiongkuo Min, Qi Jia, Guangtao Zhai

Shanghai AI Lab × AIBench Team & Community Contributors



Intelligence Real-time View All

Score		
1	o3 Pro OpenAI	81.0
2	o3 OpenAI	81.0
3	Gemini 2.5 Pro Google	80.8
4	Gemini 2.5 Pro Preview Google	80.7
5	Claude 4 Sonnet Anthropic	79.7

When AI Begins Blindly Agreeing: The Evaluation Revolution for Sycophantic Behavior in Large Language Models
2025-06-13

Redefining Image Quality Assessment for Embodied AI: From Human Vision to Machine Utility
2025-06-06

Is Your Embodied AI Really Smart? It's About Real-World Evaluation, Beyond Just Datasets.
2025-06-02

Figure 1: Features of **AIBench** by June 2025. Evaluation dimensions (**Intelligence**, **Safety**, **Speed**, and **Price**) are shown alongside website modules (ranking leaderboards and insights), model distribution by organization, and the dynamic *45° Intelligence–Safety* plots.

Abstract

We present **AIBench**, a flexible and rapidly updating platform that aggregates evaluation results from commercial platforms, popular open-source leaderboards, and internal evaluation benchmarks. While existing leaderboards primarily emphasize model capabilities, they often overlook safety evaluations and lack integrated cost-performance information, factors critical for informed decision-making by enterprises and end users. To address this gap, **AIBench** provides a comprehensive evaluation of foundation models across four key dimensions: **Safety**, **Intelligence**, **Speed**, and **Price**. In addition, inspired by the *45° Law of Intelligence-Safety Balance*, we visualize the trade-off patterns among leading models, offering a bird's-eye view of how top-tier companies position their models along these two axes. **AIBench** also tracks performance evolution over time, revealing longitudinal trends in model development. Furthermore, we periodically curate and incorporate insights from the evaluation community to ensure that the platform remains timely and relevant. We hope that **AIBench** will serve as a transparent, dynamic, and actionable platform for trustworthy evaluation, aiding both researchers and practitioners in navigating the rapidly evolving landscape of foundation models. **AIBench** is publicly available and maintained at: `https://aiben.ch`.

1 Introduction

Nowadays, foundation models (Awais et al., 2025; DeepSeek-AI and collaborators, 2025; xAI Team, 2025; Gemini Team, Google DeepMind, 2025; OpenAI, 2024; Yang et al., 2024; Devlin et al., 2019; Brown et al., 2020; Radford et al., 2021; Ramesh et al., 2021; Awais et al., 2025) have rapidly emerged as the backbone of modern artificial intelligence, enabling breakthroughs in many fields like natural language processing, vision, multimodal understanding, and autonomous decision-making (Firoozi et al., 2025; Khan et al., 2025; Xiang et al., 2025; Zhang et al., 2025a; Zuo et al., 2025; Fu et al., 2025; Silva-Rodriguez et al., 2025; Wang et al., 2025b). As these models grow in complexity and capability, a wide array of public leaderboards (Jacovi et al., 2025; Frick et al., 2025; Team, 2024; Zhang et al., 2024a; Wang et al., 2025a; Wen et al., 2025; Guo et al., 2025; Wu et al., 2023b;a; Zhong et al., 2023; Zhang et al., 2024b; 2025b) have been developed to benchmark their performance. However, most existing benchmarks focus narrowly on capability metrics such as accuracy or win rates, offering limited insight into broader concerns that are critical for deployment and trustworthiness. In particular, current evaluations often overlook essential dimensions such as safety, inference efficiency, and cost of use. These gaps pose a serious challenge for stakeholders—including researchers, enterprises, and policymakers—who must balance model power with practical considerations like reliability, speed, and affordability. Furthermore, the absence of a unified, multidimensional view across time and across models makes it difficult to understand how foundation models are evolving or how they trade off between capability performance and safety.

To address these challenges, we introduce **AIBench**: a dynamic, interpretable evaluation platform designed for the next generation of foundation models. **AIBench** aggregates performance data from multiple sources and organizes it along four critical dimensions: **Safety**, **Intelligence**, **Speed**, and **Price**. Inspired by the *45° Intelligence–Safety Law*, **AIBench** formalizes the ideal development trajectory of foundation models as a balanced advancement along both intelligence and safety axes. In this view, models that advance along a 45° line in the intelligence–safety space are considered to be developing in a healthy and trustworthy manner, avoiding the extremes of raw power without safeguards or excessive caution that sacrifices capability. In addition to snapshot evaluations, **AIBench** offers temporal analysis tools that *chart the trajectory of leading models over time*, helping users detect emerging trends and shifts in the evaluation landscape. The platform also curates insights from the model evaluation community and cross-platform observations, providing periodic updates on new metrics, testing paradigms, and industry practices.

To ensure transparency and rigor, **AIBench** adopts a tiered data sourcing strategy. 1) For commercial platforms and proprietary reports, we refer to general trend conclusions and model reputations. 2) For open academic benchmarks and publicly released leaderboards, we incorporate their data under fair use, with attribution where appropriate. 3) In parallel, we maintain our own internal evaluation pipelines and leaderboards for both intelligence and safety dimensions, allowing us to generate a unified performance score through independent integration and assessment. We hope that **AIBench** will serve as a transparent, extensible, and forward-looking resource for benchmarking foundation models—not only in terms of what they can do, but also in how safely, quickly, and affordably they do it. The platform is publicly available at: `https://aiben.ch`.

2 AIBench

2.1 Core Evaluation Dimensions

AIBench evaluates foundation models along four critical dimensions: **Intelligence**, **Safety**, **Speed**, and **Price**. These dimensions are designed to reflect both model capabilities and real-world deployment considerations, offering a holistic view of quality and usability.

2.1.1 Intelligence

The **Intelligence** dimension measures a model's general-purpose cognitive and task-solving abilities. Instead of relying on narrow benchmarks, we adopt a wide perspective (FlagEval, 2025; AGI-Eval, 2025; Opencompass, 2025a;b; Scale, 2025b;a; Epoch-AI, 2025; White et al., 2025; Chiang et al., 2024) by evaluating performance across six key aspects :

- *Knowledge*: factual correctness, coverage of general and domain information
- *Reasoning*: deductive, abductive, and commonsense reasoning.
- *Mathematics*: symbolic and numerical problem-solving, and arithmetic
- *Coding*: code generation, completion, and debugging.
- *Interaction*: instruction-following, dialog fluency, and user-centric coherence.
- *Tool/Agent Use*: capacity to act as or cooperate with agents using tools or APIs.

This decomposition enables fine-grained tracking of model evolution and specialization, facilitating transparent and interpretable assessment of intelligence.

2.1.2 Safety

The **Safety** dimension focuses on a model's ability to produce responses that align with human values and avoid harmful or undesired behavior. We structure safety evaluation (Scale, 2025c; Zeng et al., 2024; Kaiyom et al., 2024) around four key aspects:

- *Value Alignment*: consistency with ethical, legal, and cultural norms.
- *Honesty*: avoidance of hallucinations, misinformation, or overconfident falsehoods.
- *Adversarial Robustness*: resistance to jailbreaks, red teaming prompts, and misuses.
- *Rule Compliance*: adherence to explicit guidelines, refusals in restricted domains.

By evaluating safety as a structured, multi-aspect attribute rather than a binary property, AIBench supports nuanced model comparison and highlights potential risk profiles.

2.1.3 Speed & Price

Speed & Price are treated as practical deployment dimensions that directly affect usability at scale. **Speed** is measured in output tokens per second (Output TPS), reflecting average inference throughput under standard usage settings. It serves as a proxy for latency and interaction smoothness. **Price** refers to the model's estimated cost per million tokens (USD / 1M Tokens), calculated based on public API pricing and typical usage configurations. While these dimensions do not reflect capability or safety directly, they are crucial for real-world decision-making and cost-performance trade-off analysis.

2.2 Performance Visualization

2.2.1 Evaluation Scores

The evaluation framework of **AIBench** follows a definition-driven design philosophy. For each dimension under consideration, we begin by establishing a clear conceptual understanding of the intended evaluation target. These definitions serve as the foundation for subsequent data collection and aggregation processes.

To support our multidimensional evaluation, we collect benchmark results from a variety of open-source evaluation platforms (Xu et al., 2020; Chiang et al., 2024; FlagEval, 2025; Opencompass, 2025a; Scale, 2025c;b;a), which offer quantitative performance indicators across different dimensions. However, given the limitations in coverage, consistency, or granularity that often exist in public sources, we also construct internally curated evaluation sets. These internal benchmarks are designed to complement public data by covering specific scenarios or behaviors that align with our defined evaluation goals.

All collected evaluation scores are subjected to normalization procedures, ensuring that disparate scoring formats and metric scales are brought into a unified representation. This normalization enables fair comparison and fusion across heterogeneous sources. To compute the unified performance score for **Intelligence** and **Safety**, we adopt a weighted aggregation scheme that integrates insights from three complementary sources:

- *Commercial platforms and proprietary reports:* (Zeng et al., 2024; Artificial-Analysis, 2025) We refer only to high-level trends and general reputations, without directly reproducing any raw data.
- *Open academic benchmarks and public leaderboards:* (DataLearner, 2025; Wang et al., 2024; Phan et al., 2025) We incorporate their quantitative results under fair-use principles, with appropriate attribution.
- *Internal evaluation pipelines and leaderboards:* We maintain our own infrastructure to independently assess models across both intelligence and safety dimensions.

Formally, the overall score $S_i^{(j)}$ for model i under evaluation dimension j is computed as:

$$S_i^{(j)} = \alpha \cdot C_i^{(j)} + \beta \cdot O_i^{(j)} + \gamma \cdot I_i^{(j)} \tag{1}$$

where $j \in \{Intelligence, Safety\}$ and other parameters are defined as below.

- $C_i^{(j)}$ denotes the reputation score derived from proprietary sources.
- $O_i^{(j)}$ is the score aggregated from public academic benchmarks.
- $I_i^{(j)}$ represents the internal evaluation score.
- $\alpha, \beta, \gamma \in [0, 1]$ are the weighting coefficients satisfying $\alpha + \beta + \gamma = 1$.

For evaluating **Speed**, the Output TPS V_i of model i is as follows:

$$V_i = \frac{T_i - F_i}{\Delta t_i} \tag{2}$$

where T_i represents the total tokens generated, F_i denotes the first chunk of tokens, and Δt_i is the time difference between receiving the final and first token chunks.

For evaluating **Price**, the cost is based on per-token charges for input tokens sent to the model and output tokens received from it. To facilitate comparison, a blended price is computed assuming a 3:1 input-to-output token ratio:

$$P_i = \frac{3 \times P_i^{in} + P_i^{out}}{4} \tag{3}$$

where P_i^{in}, P_i^{out} respectively denote the input and output token prices.

The weights are determined through a semi-empirical strategy that reflects our confidence in the reliability, coverage, and relevance of each source with respect to the intended evaluation objectives. More specifically, the corresponding scores of top-performing foundation models by June 2025 on **AIBench** are clearly presented in Table 1, which include o4 Mini, o3, o3 Mini (OpenAI, 2025b), Claude 4 Sonnet, Claude 4 Opus (Anthropic, 2025b), Gemini 2.5 Pro, Gemini 2.5 Flash (Gemini Team, Google DeepMind, 2025), Deepseek r1 (DeepSeek-AI and collaborators, 2025), Deepseek v3 (DeepSeek-AI, 2024), Claude 3.7 Sonnet (Anthropic, 2025a), QwQ 32B (Team, 2025), Gpt 4.1, Gpt 4.1 Mini (OpenAI, 2025a), Gpt 4, Gpt 4o, Gpt 4o Mini, Gpt 4.1 Nano (OpenAI, 2024), Claude 3.5 Sonnet (Anthropic, 2024), Gpt 4 Turbo (OpenAI, 2023), Gemini 2.0 Flash, Gemini 2.0 Flash Lite, Gemini 2.0 Pro Experimental (DeepMind, 2024), Mistral 3 (AI, 2025b), Mistral Large (AI, 2024), Llama 4 Maverick, Llama 4 Scout (AI, 2025a), and Command R (Cohere, 2024) respectively.

2.2.2 45° Intelligence–Safety Visualization

To better understand how foundation models balance capability and safety, **AIBench** introduces a two-dimensional visualization that maps each model along the axes of Intelligence and Safety. This representation is inspired by the conceptual framework of the 45° Law, which posits that ideal model development should maintain a proportional relationship between increasing intelligence and increasing safety. In this visualization, each model is represented as a point in the 2D coordinate space where:

- The x-axis corresponds to the normalized intelligence score.
- The y-axis corresponds to the normalized safety score.

A 45-degree diagonal line from the origin to the top right corner represents the ideal trajectory, along which models achieve incremental gains in intelligence without sacrificing safety, and vice versa. This visual metaphor captures the underlying belief that model improvement should not come at the expense of misalignment or risk amplification.

Models above the 45° line are interpreted as being safety-favoring, potentially more conservative or restrictive, while those below the line may be intelligence-favoring, possibly exhibiting more powerful behavior but with elevated safety concerns. The relative positions provide immediate insight into the trade-offs different models have made—whether intentionally or as a byproduct of their design philosophy. This module serves both as a diagnostic tool and a normative lens, encouraging the community to view capability and safety not as opposing goals, but as co-evolving pillars of trustworthy AI development.

2.2.3 Temporal Performance Tracking

Understanding how foundation models evolve over time is essential for tracking technological progress and anticipating shifts in deployment risks or opportunities. To this end, **AIBench** includes a temporal tracking module aimed at capturing longitudinal trends in model performance, particularly along the safety and intelligence dimensions. At present, the platform maintains two snapshots (**and will continue to expand them over time**): 1) A real-time view, which reflects the most recently aggregated evaluation results; 2) A historical baseline corresponding to the state of models as of May and June 2025.

These varying snapshots allow users to observe changes in the relative positioning of models from the well-known organizations within the safety–intelligence space, highlighting directional shifts such as increased caution, rising capability, or imbalance over time.

3 Community

3.1 Meta Evaluation Insights

In addition to reporting multidimensional model scores, **AIBench** also serves as a living repository of evaluation knowledge. Given the fast-evolving nature of the foundation model ecosystem, the meaning of "model quality" is constantly being redefined—not only by new benchmarks, but also by emerging norms, values, and methodologies.

To help users stay informed of the current trends in AI evaluation, **AIBench** periodically curates and presents meta-level insights derived from: 1) Newly proposed evaluation tasks, such as robustness under instruction variation, consistency across contexts, or tool-use competency. 2) Shifts in community attention, such as the growing role of LLM-as-a-judge methods, multimodal alignment testing, or long-context reasoning. 3) Cross-platform observations, where we summarize how different benchmarks emphasize different aspects of performance, and how these emphases evolve over time. 4) Ongoing debates around trade-offs between capability, alignment, controllability, and usability.

By embedding meta-evaluation awareness into the platform, **AIBench** aspires not only to reflect current standards but also to shape more transparent, balanced, and forward-looking evaluation practices across the foundation model landscape.

Table 1: Scores of top-performing foundation models by June 2025 on **AIBench**. Best in **bold**, second and third best underlined. ‘Org.’ denotes the organization, ‘Open.’ indicates whether the model is open-sourced, and ‘Intelli’ refers to the intelligence score.

Model	**Model Attribute**			**Evaluation Dimensions**			
	Org.	*Params*	*Open.*	*Intelli.*	*Safety*	*Speed*	*Price*
o3	*OpenAI*	*N/A*	✗	**80.9**	61.9	140.1	3.50
Claude 4 Sonnet	*Anthropic*	*N/A*	✗	79.7	**64.6**	51.2	6.00
Claude 4 Opus	*Anthropic*	*N/A*	✗	79.6	62.4	65.4	30.00
Gemini 2.5 Pro	*Google*	*N/A*	✗	78.9	55.2	159.6	3.44
Deepseek r1	*DeepSeek*	*671B*	✓	77.5	54.1	29.4	0.96
o4 Mini	*OpenAI*	*N/A*	✗	77.4	60.2	161.6	1.93
o3 Mini	*OpenAI*	*N/A*	✗	73.4	53.9	109.6	1.93
o1	*OpenAI*	*N/A*	✗	73.4	56.5	206.1	26.25
Claude 3.7 Sonnet	*Anthropic*	*N/A*	✗	71.6	56.3	89.1	6.00
QwQ 32B	*Alibaba*	*32B*	✓	71.3	52.4	99.0	0.47
Gpt 4.1	*OpenAI*	*N/A*	✗	69.3	53.1	163.5	3.50
Gemini 2.0 Flash	*Google*	*N/A*	✗	68.9	51.5	251.2	0.17
Qwen 2.5 Max	*Alibaba*	*N/A*	✗	68.8	58.6	39.7	2.80
Gpt 4.1 Mini	*OpenAI*	*N/A*	✗	68.3	52.4	75.8	0.70
Deepseek v3	*DeepSeek*	*681B*	✓	67.8	51.7	31.7	0.48
Gpt 4o	*OpenAI*	*N/A*	✗	67.3	55.9	204.7	7.50
Claude 3.5 Sonnet	*Anthropic*	*N/A*	✗	67.1	58.8	78.7	6.00
Gpt 4 Turbo	*OpenAI*	*N/A*	✗	65.8	55.3	51.1	15.00
Gemini 2.0 Flash Lite	*Google*	*N/A*	✗	65.4	54.0	228.6	0.13
Gemini 2.5 Flash	*Google*	*N/A*	✗	64.0	54.1	**398.6**	0.99
Mistral 3	*Mistral*	*24B*	✓	63.6	49.3	79.4	0.80
Gemini 2.0 Pro Experimental	*Google*	*N/A*	✗	62.9	50.8	54.0	**0.00**
o1 Mini	*OpenAI*	*N/A*	✗	62.0	55.1	257.6	1.93
Mistral Large	*Mistral*	*N/A*	✗	61.1	50.5	103.4	6.00
Gpt 4o Mini	*OpenAI*	*N/A*	✗	59.8	54.5	88.0	0.26
Gpt 4	*OpenAI*	*N/A*	✗	58.9	53.7	35.4	37.50
Mistral Large 2	*Mistral*	*123B*	✓	58.6	53.6	102.1	3.00
Llama 4 Maverick	*Meta*	*N/A*	✗	58.4	51.1	170.6	0.39
Gpt 4.1 Nano	*OpenAI*	*N/A*	✗	57.4	55.0	235.5	0.17
Llama 4 Scout	*Meta*	*N/A*	✗	52.9	50.6	132.9	0.26
Command R	*Cohere*	*35B*	✓	47.4	51.4	175.2	0.26

3.2 Open for Engagement

To ensure that **AIBench** remains adaptable to evolving needs and reflective of diverse perspectives, we actively encourage participation from the research and practitioner community. In particular, we welcome two forms of contribution:

- Contributors are invited to submit evaluation insights, including empirical observations, critical analyses of existing benchmarks, or interpretive summaries of model behavior. Selected contributions may be featured in the **Insights** module, helping to broaden the scope of discussion beyond static scores.
- We invite inspiring proposals for new evaluation dimensions that enrich or go beyond the current schema of **Intelligence**, **Safety**, **Speed**, and **Price** for **AIBench**. Proposed dimensions should include a clear motivation, relevant measurement criteria, and supporting use cases. We particularly value proposals that address underexplored but practically significant properties.

All community contributions are reviewed based on conceptual clarity, methodological soundness, and alignment with the core principles of **AIBench**. **Contributors will be added to the author list to acknowledge their contributions.** Through this open collaboration model, **AIBench** aims to evolve as a shared infrastructure that not only reflects current standards but also co-shapes the future of foundation model evaluation. [1]

[1] Engagement emails can be directed to aibench.service@gmail.com.

4 Conclusion

As foundation models continue to expand in capability, the demand for trustworthy evaluation grows ever more urgent. **AIBench** addresses this need by offering a dynamic, multidimensional benchmarking platform that goes beyond traditional metrics of raw capability. By jointly evaluating **Intelligence**, **Safety**, **Speed**, and **Price**, and by visualizing key trade-offs through interpretable tools such as the 45° Intelligence–Safety map and temporal performance tracking, **AIBench** provides a holistic view of model behavior over time. In addition to quantitative scores, the platform integrates evolving community insights and encourages open contributions, ensuring that its evaluation schema remains current, extensible, and aligned with real-world concerns. We envision **AIBench** not only as a reference point for comparative analysis, but also as an evolving public infrastructure that fosters shared understanding, critical reflection, and responsible development in the foundation model ecosystem. **AIBench** is publicly accessible at: `https://aiben.ch`. We invite researchers, developers, and evaluators to engage with the platform, contribute to its growth, and help shape the next generation of AI evaluation standards.

References

AGI-Eval. Large language model leaderboard. https://agi-eval.cn/mvp/listSummaryIndex, May 2025.

Meta AI. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation. https://ai.meta.com/blog/llama-4-multimodal-intelligence/, April 2025a.

Mistral AI. Mistral large, our new flagship model. Technical report / model announcement, Mistral AI, July 2024.

Mistral AI. Medium is the new large. https://mistral.ai/news/mistral-medium-3, May 2025b.

Anthropic. Claude 3.5 sonnet. Model announcement / system card, Anthropic, June 2024.

Anthropic. Claude 3.7 sonnet and claude code. System card & model announcement, Anthropic, February 2025a.

Anthropic. Claude 4: Opus 4 & Sonnet 4. https://www.anthropic.com/news/claude-4, May 2025b.

Artificial-Analysis. Artificial-analysis. https://artificialanalysis.ai/, 2025.

Muhammad Awais, Muzammal Naseer, Salman Khan, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Foundation models defining a new era in vision: a survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E Gonzalez, et al. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning*, 2024.

Cohere. The command r model. https://docs.cohere.com/docs/command-r, March 2024.

DataLearner. Benchmarks for all. https://www.datalearner.com/ai-models/ai-benchmarks-tests/benchmarks-for-all, 2025.

Google DeepMind. Introducing gemini 2.0: our new ai model for the agentic era. https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/, December 2024.

DeepSeek-AI. Deepseek-v3 technical report. https://arxiv.org/abs/2412.19437, 2024.

DeepSeek-AI and collaborators. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. Technical Report, arXiv preprint arXiv:2501.12948, DeepSeek-AI, January 2025.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Epoch-AI. Ai performance on a set of expert-level mathematics problems. https://epoch.ai/data/ai-benchmarking-dashboard, May 2025.

Roya Firoozi, Johnathan Tucker, Stephen Tian, Anirudha Majumdar, Jiankai Sun, Weiyu Liu, Yuke Zhu, Shuran Song, Ashish Kapoor, Karol Hausman, et al. Foundation models in robotics: Applications, challenges, and the future. *The International Journal of Robotics Research*, 44(5):701–739, 2025.

FlagEval. Nlp capability leaderboard. https://flageval.baai.ac.cn/#/leaderboard/nlp-capability?kind=CHAT, March 2025.

Evan Frick, Connor Chen, Joseph Tennyson, Tianle Li, Wei-Lin Chiang, Anastasios N Angelopoulos, and Ion Stoica. Prompt-to-leaderboard. *arXiv preprint arXiv:2502.14855*, 2025.

Xi Fu, Shentong Mo, Alejandro Buendia, Anouchka P Laurent, Anqi Shao, Maria del Mar Alvarez-Torres, Tianji Yu, Jimin Tan, Jiayu Su, Romella Sagatelian, et al. A foundation model of transcription across human cell types. *Nature*, pp. 1–9, 2025.

Gemini Team, Google DeepMind. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next-Generation Agentic Capabilities. Technical Report v2.5, Google DeepMind, June 2025.

Yijin Guo, Kaiyuan Ji, Xiaorong Zhu, Junying Wang, Farong Wen, Chunyi Li, Zicheng Zhang, and Guangtao Zhai. Human-centric evaluation for foundation models. *arXiv preprint arXiv:2506.01793*, 2025.

Alon Jacovi, Andrew Wang, Chris Alberti, Connie Tao, Jon Lipovetz, Kate Olszewska, Lukas Haas, Michelle Liu, Nate Keating, Adam Bloniarz, et al. The facts grounding leaderboard: Benchmarking llms' ability to ground responses to long-form input. *arXiv preprint arXiv:2501.03200*, 2025.

Farzaan Kaiyom, Ahmed Ahmed, Yifan Mai, Kevin Klyman, Rishi Bommasani, and Percy Liang. Helm safety: Towards standardized safety evaluations of language models. https://crfm.stanford.edu/2024/11/08/helm-safety.html, November 2024.

Wasif Khan, Seowung Leem, Kyle B See, Joshua K Wong, Shaoting Zhang, and Ruogu Fang. A comprehensive survey of foundation models in medicine. *IEEE Reviews in Biomedical Engineering*, 2025.

OpenAI. Gpt-4 turbo. https://platform.openai.com/docs/models/gpt-4-turbo, November 2023.

OpenAI. Hello gpt-4o. System card / technical report, OpenAI, May 2024.

OpenAI. Introducing gpt-4.1 in the api. https://openai.com/index/gpt-4-1/, April 2025a.

OpenAI. Introducing openai o3 and o4-mini. https://openai.com/index/introducing-o3-and-o4-mini/, April 2025b.

Opencompass. Compassbench large language model leaderboard. https://rank.opencompass.org.cn/leaderboard-llm/?m=25-04, May 2025a.

Opencompass. Compassacademic large language model leaderboard. https://rank.opencompass.org.cn/leaderboard-llm-academic/?m=REALTIME, May 2025b.

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, et al. Humanity's last exam. https://arxiv.org/abs/2501.14249, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pp. 8821–8831. Pmlr, 2021.

Scale. Multichallenge performance comparison. https://scale.com/leaderboard/multichallenge, March 2025a.

Scale. Humanity's last exam (text only) performance comparison. https://scale.com/leaderboard/humanitys_last_exam_text_only, June 2025b.

Scale. Mask performance comparison. https://scale.com/leaderboard/mask, April 2025c.

Julio Silva-Rodriguez, Hadi Chakor, Riadh Kobbi, Jose Dolz, and Ismail Ben Ayed. A foundation language-image model of the retina (flair): Encoding expert knowledge in text supervision. *Medical Image Analysis*, 99:103357, 2025.

OpenCompass Team. Opencompass multimodal leaderboard. https://opencompass.org.cn, 2024.

Qwen Team. Qwq-32b: Embracing the power of reinforcement learning. https://qwenlm.github.io/blog/qwq-32b/, March 2025.

Junying Wang, Wenzhe Li, Yalun Wu, Yingji Liang, Yijin Guo, Chunyi Li, Haodong Duan, Zicheng Zhang, and Guangtao Zhai. Affordance benchmark for mllms. *arXiv preprint arXiv:2506.00893*, 2025a.

Junying Wang, Hongyuan Zhang, and Yuan Yuan. Adv-cpg: A customized portrait generation framework with facial adversarial attacks. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 21001–21010, June 2025b.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. https://arxiv.org/abs/2406.01574, 2024.

Farong Wen, Yijin Guo, Junying Wang, Jiaohao Xiao, Yingjie Zhou, Chunyi Li, Zicheng Zhang, and Guangtao Zhai. Improve mllm benchmark efficiency through interview. *arXiv preprint arXiv:2506.00883*, 2025.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Ben Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Sreemanti Dey, Shubh-Agrawal, Sandeep Singh Sandha, Siddartha Naidu, Chinmay Hegde, Yann LeCun, Tom Goldstein, Willie Neiswanger, and Micah Goldblum. Livebench: A challenging, contamination-limited llm benchmark. https://arxiv.org/abs/2406.19314, 2025.

Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, et al. Q-bench: A benchmark for general-purpose foundation models on low-level vision. *arXiv preprint arXiv:2309.14181*, 2023a.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023b.

xAI Team. Grok-3: The Age of Reasoning Agents. Technical report, xAI, February 2025.

Jinxi Xiang, Xiyue Wang, Xiaoming Zhang, Yinghua Xi, Feyisope Eweje, Yijiang Chen, Yuchen Li, Colin Bergstrom, Matthew Gopaulchan, Ted Kim, et al. A vision–language foundation model for precision oncology. *Nature*, pp. 1–10, 2025.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Yudong Cao, Chenjie andLi, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, Yin Tian, Qianqian Dong, Weitang Liu, Bo Shi, Yiming Cui, Junyi Li, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaoweihua Liu, Zhe Zhao, Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhengliang Yang, Kyle Richardson, and Zhenzhong Lan. Clue: A Chinese language understanding evaluation benchmark. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 4762–4772. International Committee on Computational Linguistics, December 2020.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Yi Zeng, Yu Yang, Andy Zhou, Jeffrey Ziwei Tan, Yuheng Tu, Yifan Mai, Kevin Klyman, Minzhou Pan, Ruoxi Jia, Dawn Song, Percy Liang, and Bo Li. Air-bench 2024: A safety benchmark based on risk categories from regulations and policies. https://arxiv.org/abs/2407.17436, 2024.

Haotian Zhang, Stuart Dereck Semujju, Zhicheng Wang, Xianwei Lv, Kang Xu, Liang Wu, Ye Jia, Jing Wu, Wensheng Liang, Ruiyan Zhuang, et al. Large scale foundation models for intelligent manufacturing applications: a survey. *Journal of Intelligent Manufacturing*, pp. 1–52, 2025a.

Zicheng Zhang, Haoning Wu, Chunyi Li, Yingjie Zhou, Wei Sun, Xiongkuo Min, Zijian Chen, Xiaohong Liu, Weisi Lin, and Guangtao Zhai. A-bench: Are lmms masters at evaluating ai-generated images? *arXiv preprint arXiv:2406.03070*, 2024a.

Zicheng Zhang, Yingjie Zhou, Chunyi Li, Baixuan Zhao, Xiaohong Liu, and Guangtao Zhai. Quality assessment in the era of large models: A survey. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2024b.

Zicheng Zhang, Xiangyu Zhao, Xinyu Fang, Chunyi Li, Xiaohong Liu, Xiongkuo Min, Haodong Duan, Kai Chen, and Guangtao Zhai. Redundancy principles for mllms benchmarks. *arXiv preprint arXiv:2501.13953*, 2025b.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. Agieval: A human-centric benchmark for evaluating foundation models. *arXiv preprint arXiv:2304.06364*, 2023.

Xingxing Zuo, Pouya Samangouei, Yunwen Zhou, Yan Di, and Mingyang Li. Fmgs: Foundation model embedded 3d gaussian splatting for holistic 3d scene understanding. *International Journal of Computer Vision*, 133(2):611–627, 2025.