# Science Arena: Time-Release Olympiad Exams for Rubric-Faithful Evaluation of Scientific Reasoning

**Anonymous ACL submission**

## Abstract

Benchmark saturation and data contamination make it increasingly difficult to measure genuine scientific reasoning in frontier LLMs. We introduce SCIENCE ARENA, an olympiad-style evaluation suite built from five rubric-driven scientific competitions whose official materials were publicly released in Aug/Oct 2025, forming a practical time-release holdout against pretraining overlap. To approximate real exam conditions, olympiad medalists grade model solutions with stepwise partial-credit rubrics, and we map scores to medal-calibrated human baselines. Across Physics, Chemistry, and Biology theory exams, the best model reaches or exceeds the Human-Gold tier, yet remains below Human-Winner performance; errors concentrate in visual grounding and long-horizon coherence. We further show LLM-as-a-judge can substantially inflate scores on open-ended chemistry, motivating expert, rubric-faithful evaluation.

## 1 Introduction

Scientific reasoning has become a central target for frontier LLMs, yet evaluation is at an impasse. Popular benchmarks are increasingly saturated and vulnerable to contamination (Hendrycks et al., 2021; Rein et al., 2023; White et al., 2024), while "fresh" or time-sensitive constructions often remain confined to simple fill-in-the-blank or multiple-choice formats, failing to capture the multi-step derivations, partial credit nuances, and multimodal grounding essential to scientific inquiry (Liang et al., 2023; White et al., 2024; Wang et al., 2023; Yue et al., 2023; He et al., 2024; Lu et al., 2022; Mirza et al., 2025).

A second bottleneck is *measurement*. Open-ended science evaluation is often delegated to automated judges due to cost, but LLM-based judging is known to be imperfect even in chatbot-style settings (Zheng et al., 2023); the challenge is sharper for olympiad rubrics where "nearly-right" steps, diagram/structure fidelity, and error-forward rules determine partial credit. As a result, automated evaluation frequently leads to score inflation, masking fragile reasoning behaviors that matter in real exams.

We present SCIENCE ARENA, a competition-simulation benchmark derived from five high-stakes scientific competitions released in late 2025, designed as an uncontaminated holdout akin to prior work on competition-based evaluation (Balunović et al., 2025). Sec. 2 describes the benchmark sources, rubric-faithful expert grading, and our stateful protocol for multi-part problems. Sec. 3 reports (i) medal-calibrated standings on international exams under expert grading, (ii) an ablation of iterative vs. one-shot protocols, (iii) a stress test of LLM-as-a-judge reliability, and (iv) an expert-derived ability boundary highlighting where frontier models still fail.

**Contributions.** (1) We release SCIENCE ARENA, an olympiad-style, time-release holdout for scientific reasoning. (2) We provide medal-calibrated, rubric-faithful expert grading that reveals failure modes hidden by automated evaluation. (3) We offer diagnostic analyses (protocol ablation; ability boundary) to localize remaining gaps in scientific reasoning.

## 2 Science Arena

We introduce SCIENCE ARENA, a benchmark suite built from five high-stakes science competitions with well-defined problem structure and explicit marking schemes.

### 2.1 Competitions Sources

As in Table 1, we test LLMs on the latest Olympiad competitions. These are top-tier international competitions held in the second half of 2025. Consequently, they serve as an effective benchmark

| Exam | Release | Intlational | Open_ended | Scores |
|---|---|---|---|---|
| IPhO'25 | Aug 2025 | ✓ | ✓ | 30 |
| IChO'25 | Aug 2025 | ✓ | ✓ | 60 |
| IBO'23 | Aug 2025† | ✓ | ✗ | 453 |
| CChO'25 | Oct 2025 | ✗ | ✓ | 100 |
| CPhO'25 | Oct 2025 | ✗ | ✓ | 320 |

Table 1: SCIENCE ARENA sources. †IBO papers follow a two-year blackout policy; thus IBO 2023 becomes public in 2025.

for evaluating prior models to ascertain their true capabilities.

**A case of IPhO 2025 problem. (stylized vignette)**

**Context (excerpt).** A problem spans multiple subparts and mixes symbolic derivation with extracting numbers from a figure. You may assume standard constants $(G, \hbar, e, \ldots)$.

1. **Derivation.** A particle moves in a central field. Derive a relation between $v(r)$ and $r$, and obtain an expression for an energy scale $E(r)$. (0.8 pts)
2. **Read a graph.** From a given figure, read $v$ at two radii $(r_1, r_2)$ and estimate enclosed masses $M(< r_i) = r_i v(r_i)^2/G$. Briefly interpret whether the outer region suggests an additional component. (1.0 pts)
3. **Regimes / threshold.** A system changes behavior depending on a control parameter $\xi$. Find a threshold $\xi^\star$ separating two regimes and state which forces/constraints set it. (0.8 pts)
4. **Representation.** Sketch (or describe) the required diagram (e.g., a loop/trajectory) that supports your work/energy calculation, and label key points. (0.4 pts)

Figure 1: A representative example from IPhO 2025 illustrating diverse scientific reasoning skills.

From Figure 1, we can see a representative example from one of our competitions, IPhO 2025, illustrating that such tasks probe a diverse set of scientific reasoning skills—*symbolic derivation*, *quantitative graph-to-number extraction*, *regime/threshold analysis*, and *clear diagrammatic representation*—rather than mere factual recall. We analyze LLMs' abilities in Sec . 3.5

## 2.2 Rubric-Faithful Scoring

Leveraging the granular scoring rubrics provided in the competition, we evaluate models in a manner analogous to human exam assessment, combining **manual grading** with **automated evaluation**[1]. We compare them in Sec . 3.2.

[1]We use **Gemini 3 pro (preview)** as the judge model for automated evaluation.

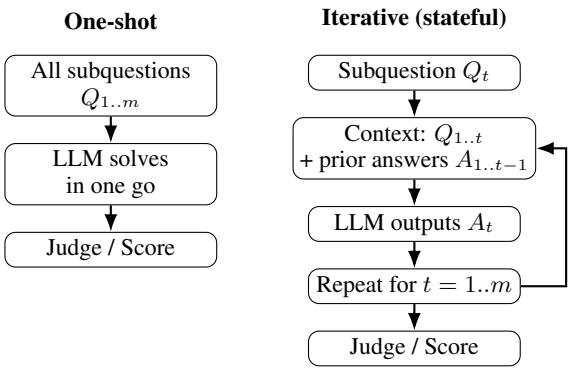


Figure 2: Two protocols for multi-part problems: one-shot solving vs. iterative, stateful prompting.

Our evaluators are distinguished medalists in their respective fields. We focused the human evaluation on international competitions, using a dual-expert review process to bolster result reliability.

## 2.3 Solving Protocols

Most physics and chemistry competition questions are structured as comprehensive problems with multiple sub-parts. Addressing the lack of a unified evaluation standard, we explored both joint and iterative generation methods. They are compared in Figure 2, which generates answers for each subquestion sequentially. We use iterative evaluation as the default. We further conduct an ablation study in Section 3.4 to systematically compare iterative prompting against the one-shot baseline.

## 2.4 Models

We categorize the evaluated LLMs into two distinct groups: those released on or before early August 2025, and those released after September 2025:

- **Before early August 2025**: *Gemini 2.5 Pro* (DeepMind, 2025a; Comanici et al., 2025), *GPT-5* (OpenAI, 2025a), *DeepSeek V3.1* (DeepSeek-AI, 2025a), *Qwen3 235B* (Yang et al., 2025; Team, 2025), and *GLM 4.5* (AI, 2025).
- **After September 2025**: *Gemini 3 Pro (preview)* (DeepMind, 2025c), *Gemini 3 Flash (preview)* (DeepMind, 2025b), *Claude Opus 4.5* (Anthropic, 2025), *DeepSeek V3.2* (DeepSeek-AI, 2025b), *GPT-5.1* (OpenAI, 2025b), and *GLM 4.7* (, Z.ai,Z).

| Metric | IPHO 25 | ICHO 25 | IBO 23 |
|---|---|---|---|
| *Full Scores* | *30.0* | *60.0* | *453.0* |
| ***Human Benchmarks*** | | | |
| Human Winner | 29.2 | 57.1 | 395.4 |
| Human Gold | 23.4 | 36.6 | 354.7 |
| Human Silver | 17.0 | 28.7 | 293.1 |
| Human Bronze | 11.5 | 26.3 | 249.7 |
| ***Model Performance*** | | | |
| Gemini 2.5 Pro | **24.1** | **37.0** | **404.0** |
| GPT-5 | 21.8 | 31.8 | 395.0 |
| GLM 4.5 | 21.2 | 19.8 | 329.0 |
| DeepSeek V3.1 | 19.4 | 26.6 | 312.0 |
| Qwen3 235B | 15.2 | 25.9 | 337.0 |

Table 2: Results on Uncontaminated Science Olympiad Datasets (August 2025), all models listed above are released before this handoff. The best model performance is **bolded**. All scores are rated by experts.

Case A: **Mis-grounded rubric** ⇒ spurious partial credit (IChO'25 Q1.3)

**Task.** Draw the structures of compounds $E$ and $F$ (no stereochemistry required).
**Candidate answer (LLM solve).** Proposes a *wrong core scaffold* and argues the given molecular formula is a typo; propagates this scaffold into $F$.
**Human grading.** $E$: 0/3, $F$: 0/3 (wrong ring system; not an error-forward-eligible variant).
**LLM-as-judge.** Assigns *non-zero* score (e.g., $E$: 1/3, $F$: 3/3) by (i) hallucinating an incompatible "standard answer" and (ii) applying an error-forward assumption on an already-invalid structure.

Case B: **Overly permissive** ⇒ vague structures counted as correct (IChO'25 Q1.1)

**Task.** Draw the structures of i-Cy, $A$, and $B$ (no stereochemistry required).
**Candidate answer (LLM solve).** Provides high-level descriptions (keywords like "bicyclic", "exocyclic methylene"), but *does not clearly specify double-bond placement* and gives an *unusable sketch* for i-Cy/$A$.
**Human grading.** i-Cy: 0, $A$: 0, $B$: 2 (only $B$ is creditable via error-forward).
**LLM-as-judge.** Awards *near/full marks* by matching surface cues in the explanation while ignoring the key requirement: an unambiguous structural specification.

Figure 3: Two typical failure modes of LLM-based judges on open-ended chemistry questions: (A) rubric/answer-key hallucination leading to unjustified partial credit; (B) permissive keyword matching that over-rewards under-specified structures.

## 3 Results and Diagnostics

### 3.1 Expert-Graded Results on International Competitions

Table 2 reports expert-graded results on these uncontaminated exams, with all models released prior to the competitions[2]. Gemini 2.5 Pro is the only model that consistently exceeds the Human-Gold baseline, yet a clear gap to Human-Winner remains on open-ended Physics and Chemistry, indicating that long-horizon scientific reasoning is still bottlenecked.

### 3.2 Reliability of LLM-as-a-Judge

Automated judging substantially inflates scores on IChO'25 compared to expert grading: *Gemini 2.5 Pro* increases from **37.0** (human; Table 2) to **45.4** (LLM-judge; Table 3), and *GPT-5* from **31.8** to **42.1**. The gap is notably smaller on IPhO (∼1–2 points), suggesting chemistry—especially structure questions with partial credit—creates a harder judging regime. Figure 3 highlights two common error modes: rubric mis-grounding and permissive matching.

### 3.3 Temporal Evolution & Scalability

We extend our analysis to late-2025 benchmarks (CChO and CPhO) using purely automated LLM-based evaluation to track longitudinal evolution (Table 3). Unlike the expert grading in Table 2, these scores are produced entirely by an LLM-as-a-judge. Consequently, absolute scores—including those for IChO'25 and IPhO'25—are not directly comparable across tables; discrepancies in the common entries arise solely from the evaluation method rather than model differences.

Three qualitative trends emerge from the automated evaluation:

- **Generational Leaps:** Newer architectures show substantial gains. The Gemini 3 series achieves a ∼25% improvement over the GPT-5 era on CChO, confirming that complex scientific reasoning is far from saturation.
- **Optimization vs. Scale:** Scaling is not monotonic under automated judging: the larger Gemini 3 model leads on chemistry (CChO; IChO), while the latency-optimized variant performs better on physics (CPhO; IPhO).
- **Uneven Domain Mastery:** A clear subject-wise divergence persists. Models approach saturation on CPhO (∼94%), yet struggle with CChO (∼60%), highlighting that long-horizon chemi-

[2] IPHO/ICHO/IBO include theory and practical parts; we evaluate only the theory portion. Medal cutoffs are defined as the average theory score of contestants receiving each medal in the official final rankings (Gold/Silver/Bronze), rather than minimum-to-medal thresholds.

cal reasoning still lags behind physical problem-solving.

| **Benchmark (2025)** | **CCHO** | **CPHO** | **ICHO** | **IPHO** |
|---|---|---|---|---|
| *Region* | *China* | *China* | *Int'l* | *Int'l* |
| *Full Scores* | *100* | *320* | *60* | *30* |
| *Reference / Early 2025* | | | | |
| GPT-5 | 47.1 | 270.0 | 42.1 | 23.5 |
| Gemini 2.5 Pro | 55.1 | 273.0 | 45.4 | 25.3 |
| GLM 4.5 | 35.0 | 221.0 | 27.5 | 20.1 |
| DeepSeek V3.1 | 40.1 | 221.0 | 30.3 | 20.7 |
| *Recent / Late 2025* | | | | |
| Gem 3 Flash Prev | 59.4 | **301.5** | 53.8 | **27.9** |
| Gem 3 Pro Prev | **64.3** | 275.0 | **54.4** | 27.2 |
| GPT-5.1 High | 48.4 | 285.0 | 41.0 | 24.6 |
| Claude 4.5 Opus | 51.1 | 237.0 | 47.1 | 22.5 |
| GLM 4.7 | 38.8 | 231.0 | 38.4 | 24.2 |
| DeepSeek V3.2 | 42.2 | 248.0 | 36.1 | 19.3 |

Table 3: Model performance evolution on extended benchmarks under **automated LLM-based evaluation**. All scores in this table are produced by an LLM judge. IChO'25 and IPhO'25 correspond to the same 2025 competition papers as in Table 2, **Bold**: Best; Underline: Second.

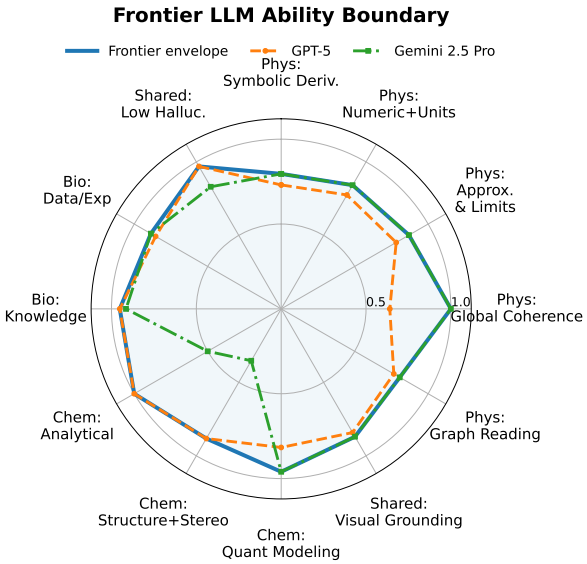


Figure 4: **Frontier LLM ability boundary** aggregated from IPhO'25 (Physics), IChO'25 (Chemistry), and IBO'23 (Biology). Axes are explicitly labeled as `Phys/Chem/Bio` (domain-specific) or `Shared` to avoid ambiguity. The *frontier envelope* reports the best value among GPT-5 and Gemini 2.5 Pro for each axis.

### 3.4 Protocol Sensitivity: Iterative vs. One-shot

We compare our iterative protocol against a one-shot baseline (Figure 2). Table 4 reveals diverging trends ($\Delta$ = Iterative − One-shot):

- **Decomposition Benefits Efficiency:** *Gemini 3 Flash* gains massively (+74.5 on CPHO), confirming that iterative prompting effectively mitigates the limited reasoning window of smaller architectures.
- **Global Context Necessity:** Conversely, the stronger *Gemini 3 Pro* degrades in Physics (-20.0). Complex physics tasks involve coupled variables; high-capacity models rely on the global context (available in one-shot) for consistency, which segmentation disrupts.

| **Model** | **CCHO (Chem)** | | **CPHO (Phys)** | |
|---|---|---|---|---|
| | *Score* | $\Delta$ | *Score* | $\Delta$ |
| Gem 3 Flash Prev | 59.4 | +6.3 | 301.5 | **+74.5** |
| Gem 3 Pro Prev | 64.3 | +6.7 | 275.0 | -20.0 |
| GPT-5.1 High | 48.4 | +1.7 | 285.0 | -8.0 |
| Claude 4.5 Opus | 51.1 | +7.2 | 237.0 | +7.0 |

Table 4: Impact of the iterative protocol. *Flash* relies on decomposition, while *Pro* utilizes global context in Physics.

### 3.5 Ability Boundary Across Domains

Figure 4 summarizes a cross-domain profile beyond aggregate scores. Frontier models are consistently strong in *quantitative execution* (Physics derivation, units, approximations; Chemistry quantitative modeling) and in *biology knowledge* under constrained answer formats. In contrast, the sharpest remaining boundary is *faithful grounding*: binding non-text evidence to precise symbolic commitments (`Shared: Visual Grounding`) and maintaining calibration (`Shared: Low Halluc.`).

To make the visualization interpretable and reproducible, we provide (i) an axis glossary (Appendix, Ability Axis Glossary) and (ii) per-domain radars using domain-native decompositions (Appendix Figure 5–7).

## 4 Conclusion

SCIENCE ARENA evaluates frontier LLMs under an olympiad-style, uncontaminated time-release holdout with rubric-faithful expert grading. While the best model attains Human-Gold tier on multiple theory exams, a persistent gap to Human-Winner performance remains. Our diagnostics identify visual grounding and global coherence as dominant bottlenecks, and we find automated judging can overestimate performance on open-ended chemistry, reinforcing the need for expert, rubric-driven evaluation.

## Limitations

Our benchmark is limited in scale and covers only five competitions; results may not generalize to all scientific curricula. We evaluate theory components only, since experimental sections require physical apparatus and proctoring. Expert grading is costly and thus applied primarily to international exams; some extended results rely on automated judging and should be interpreted qualitatively. Finally, our ability taxonomy is expert-derived and may reflect annotator priors; alternative decompositions could yield different boundary shapes.

## References


Zhipu AI. 2025. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *arXiv preprint arXiv:2508.06471*.

Anthropic. 2025. Claude opus 4.5 system card. https://www.anthropic.com/claude-opus-4-5-system-card. Accessed: 2026-01-04.

Mislav Balunović, Jasper Dekoninck, Ivo Petrov, Nikola Jovanović, and Martin Vechev. 2025. Matharena: Evaluating llms on uncontaminated math competitions. *arXiv preprint arXiv:2505.23281*.

China Chemistry Olympiad (CChO). 2025. 39th china chemistry olympiad (final) 2025: Theory exam paper, reference answers, and marking guidelines (public pdf). Online (PDF). Accessed: 2026-01-05.

Chinese Chemical Society. 2025. 2025 39th china chemistry olympiad (final): Theory exam score conversion weights. Online (PDF). Accessed: 2026-01-05.

Chinese Physics Olympiad (CPhO). 2025a. 42nd national high school physics competition (final) 2025: Theory exam paper (public pdf). Online (PDF). Accessed: 2026-01-05.

Chinese Physics Olympiad (CPhO). 2025b. 42nd national high school physics competition (final) 2025: Theory solutions and marking scheme (public pdf). Online (PDF). Accessed: 2026-01-05.

Maria Comanici and 1 others. 2025. Gemini 2.5: Pushing the frontier with Reasoning, multimodality, long context, and efficiency. *arXiv preprint arXiv:2507.06261*.

CPhO Committee. 2025a. 42nd national high school physics competition (final) 2025: Practical answer key / marking (public pdf). Online (PDF). Accessed: 2026-01-05.

CPhO Committee. 2025b. 42nd national high school physics competition (final) 2025: Practical exam paper (public pdf). Online (PDF). Accessed: 2026-01-05.

CPhO Committee. 2025c. National high school physics competition charter (revised april 2025). Online (PDF). Accessed: 2026-01-05.

Google DeepMind. 2025a. Gemini 2.5 pro model card. https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-2-5-Pro-Model-Card.pdf. Accessed: 2026-01-04.

Google DeepMind. 2025b. Gemini 3 flash model card. https://deepmind.google/models/model-cards/gemini-3-flash/. Accessed: 2026-01-04.

Google DeepMind. 2025c. Gemini 3 pro model card. https://deepmind.google/models/model-cards/gemini-3-pro/. Accessed: 2026-01-04.

DeepSeek-AI. 2025a. Deepseek-v3.1 model card. https://huggingface.co/deepseek-ai/DeepSeek-V3.1. Accessed: 2026-01-04.

DeepSeek-AI. 2025b. Deepseek-v3.2 technical report. *arXiv preprint arXiv:2512.02556*.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. 2024. Olympiadbench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems. *arXiv preprint arXiv:2402.14008*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

IBO. 2023a. 34th international biology olympiad (ibo 2023): Final results. Online (PDF). Accessed: 2026-01-05.

IBO. 2023b. Ibo examination papers archive (includes ibo 2023 uae papers). Online. Accessed: 2026-01-05.

IChO 2025 Committee. 2025. Icho 2025: Problems (official theory/practical exams and solutions). Online. Accessed: 2026-01-05.

IPho 2025 Committee. 2025. Ipho 2025: Official questions and solutions (exam topics). Online. Accessed: 2026-01-05.

J. Liang and 1 others. 2023. Latesteval: Addressing data contamination in language model evaluation through dynamic and time-sensitive test construction. *arXiv preprint arXiv:2312.12343*.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *arXiv preprint arXiv:2209.09513*.

A. Mirza and 1 others. 2025. A framework for evaluating the chemical knowledge and reasoning abilities of large language models. *Nature Chemistry*.

OpenAI. 2025a. Gpt-5 system card. https://cdn.openai.com/gpt-5-system-card.pdf. Accessed: 2026-01-04.

OpenAI. 2025b. Gpt-5.1 system card addendum. https://cdn.openai.com/pdf/4173ec8d-1229-47db-96de-06d87147e07e/5_1_system_card.pdf. Accessed: 2026-01-04.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2023. GPQA: A graduate-level google-proof Q&A benchmark. *arXiv preprint arXiv:2311.12022*.

Qwen Team. 2025. Qwen3-235b-a22b-instruct-2507 model card. https://huggingface.co/Qwen/Qwen3-235B-A22B-Instruct-2507. Accessed: 2026-01-04.

Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R. Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. 2023. Scibench: Evaluating college-level scientific problem-solving abilities of large language models. *arXiv preprint arXiv:2307.10635*.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Ben Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Siddhartha Naidu, Chinmay Hegde, Yann LeCun, Tom Goldstein, Willie Neiswanger, and Micah Goldblum. 2024. Livebench: A challenging, contamination-free llm benchmark. *arXiv preprint arXiv:2406.19314*.

An Yang and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, and 3 others. 2023. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. *arXiv preprint arXiv:2311.16502*.

Zhipu AI (Z.ai). 2025a. Glm-4.7: Advancing the coding capability. https://z.ai/blog/glm-4.7. Accessed: 2026-01-04.

Zhipu AI (Z.ai). 2025b. Glm-4.7 model card. https://huggingface.co/zai-org/GLM-4.7. Accessed: 2026-01-04.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yanping Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

## A Competition Sources

Science Arena is built from five Olympiad-style scientific competitions spanning physics, chemistry, and biology. A key design choice is to index each dataset by the *public release time* of the official problem materials (rather than the contest year), so that evaluation can be conducted on a practical "time-release holdout" that reduces the likelihood of overlap with prior training corpora. Table 1 summarizes the sources we cite for each competition.

**International Physics Olympiad (IPhO 2025).** We use the official IPhO 2025 repository of exam materials, including the released problems and accompanying official documents (e.g., solutions and/or marking guidance when provided) (IPho 2025 Committee, 2025). IPhO exams consist of theory and experimental components; our benchmark focuses on the *theory* problems, since experimental tasks require physical apparatus and proctoring conditions that are not directly reproducible for LLM evaluation.

**International Chemistry Olympiad (IChO 2025).** We source IChO 2025 materials from the official "Problems" page released by the competition organizers (IChO 2025 Committee, 2025). Similar to IPhO, IChO includes theory and laboratory components; we evaluate models on the *theory* portion to ensure a well-defined, reproducible setting with rubric-faithful scoring.

**International Biology Olympiad (IBO 2023).** IBO provides an official archive of examination papers for past competitions (IBO, 2023b). In Science Arena, IBO 2023 serves as an *objectively scorable* biology track (e.g., multiple-choice or otherwise unambiguously keyed questions), enabling automatic evaluation without judge-model mediation. When reporting medal-calibrated comparisons, we reference the official IBO 2023 results report (IBO, 2023a).

**China Chemistry Olympiad (CChO 2025).** For CChO 2025, we use publicly released final-round materials (problem statements and marking guidance) distributed as a consolidated PDF (China Chemistry Olympiad (CChO), 2025). When normalizing scores and aligning evaluation to official scoring practice, we additionally cite the Chinese Chemical Society document describing the theory score conversion weights (Chinese Chemical Society, 2025). As with other laboratory-inclusive

competitions, our primary benchmark setting targets the theory component for reproducibility.

**China Physics Olympiad (CPhO 2025).** For CPhO 2025, we use publicly released final-round theory materials and the associated marking scheme (Chinese Physics Olympiad (CPhO), 2025a,b). Where relevant to describing the competition formally (e.g., structure, governance, or official rules), we also cite the published competition charter (CPhO Committee, 2025c). We evaluate the theory component as the reproducible portion of the exam; experimental components and their marking documents can be referenced separately when needed (CPhO Committee, 2025b,a).

**Notes on scope and reproducibility.** All five competitions are high-stakes, rubric-driven assessments with well-defined problem sets and (for open-ended tasks) partial-credit grading. Science Arena prioritizes the *publicly released* versions of these materials to support reproducibility and transparent sourcing, and it uses theory-only evaluation where experimental replication would otherwise confound model comparisons.

## B Details of Ability Boundary

### Ability Axis Glossary

- **Phys: Symbolic Deriv.** — multi-step algebraic manipulation and equation-based reasoning.
- **Phys: Numeric+Units** — numerical accuracy with unit/scale discipline (dimensional checks, sanity bounds).
- **Phys: Approx. & Limits** — correct approximations and limiting-case validation.
- **Phys: Global Coherence** — maintain consistent assumptions/variables across coupled subparts (regime tracking).
- **Phys: Graph Reading** — extract trends/values from curves and use them downstream.
- **Shared: Visual Grounding** — bind non-text evidence to symbols/claims (diagrams/structures/figures; "read-the-figure" reliability).
- **Chem: Quant Modeling** — quantitative physical chemistry (equilibria/thermo/electro) under algebraic constraints.
- **Chem: Structure+Stereo** — structure fidelity and stereochemical correctness (connectivity, configuration).
- **Chem: Analytical** — interpret analytical evidence (e.g., spectroscopy/assays) into chemically valid conclusions.

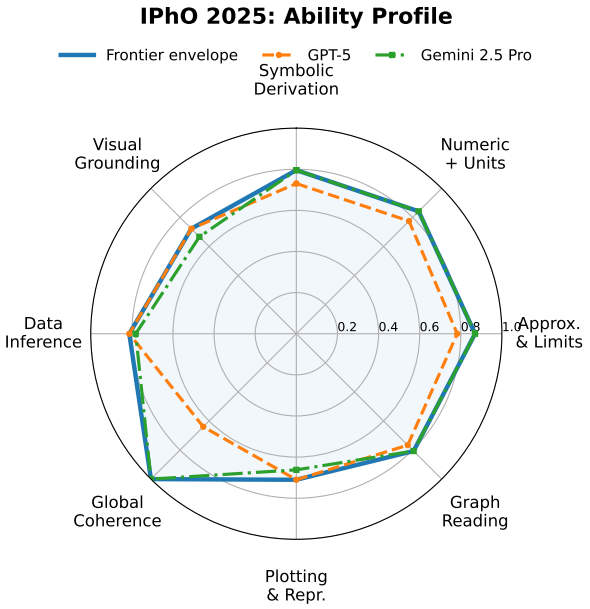


Figure 5: **IPhO 2025.** Physics abilities follow our rubric: symbolic derivation, numerical computation with units, approximation/limiting cases, graph reading, plotting/representation, global coherence across regimes, data inference from empirical relations, and visual grounding.

- **Bio: Knowledge** — biology knowledge and reading comprehension in constrained formats (mechanisms/terminology).
- **Bio: Data/Exp** — reasoning from experimental data (figures/tables, controls, statistical cues, causality).
- **Shared: Low Halluc.** — calibration/faithfulness proxy: lower rate of confidently-wrong outputs when confidence is available.

### B.1 Domain-specific ability profiles (radar plots)

Figure 5–7 summarize within-domain ability profiles. Each radar overlays a *frontier envelope* (per-axis best among the compared frontier models), GPT-5, and Gemini 2.5 Pro. Axes are domain-native (Physics/Chemistry/Biology) and therefore are *not* directly comparable across figures.

**IPhO 25 notes.** Both models are strongest on local symbolic manipulation and routine numerics; remaining errors concentrate in *global coherence* (propagating assumptions consistently across long solutions) and *visual grounding* (extracting non-textual cues), with noticeable variance across tasks requiring regime switching.

**IChO 25 notes.** Gemini 2.5 Pro tends to be more uniformly strong on quantitative and analytical axes (physical chemistry and spectroscopy), while

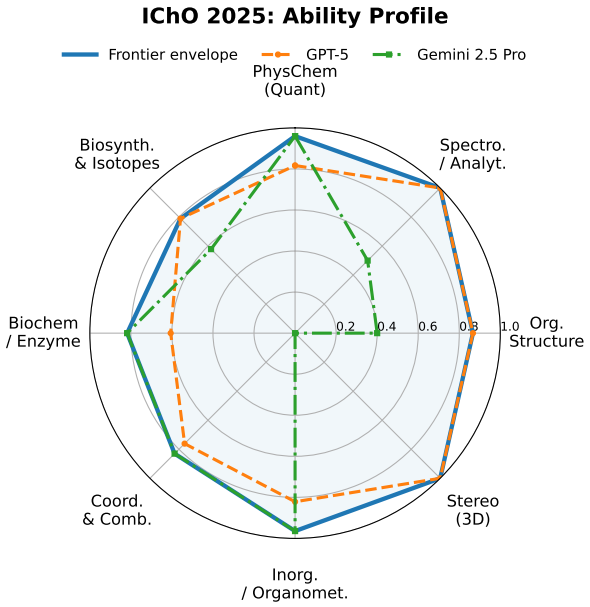


Figure 6: **IChO 2025.** Chemistry abilities use the human-coded decomposition: quantitative physical chemistry; spectroscopy/analytical reasoning; organic structure depiction; stereochemistry; inorganic/organometallic reasoning; coordination/combinatorics; biochemistry/enzyme reasoning; and biosynthesis/isotopic tracing.

GPT-5 exhibits higher dispersion: it can solve isolated substeps but is less consistent when multiple constraints (structure + stereochemistry + quantitative checks) must be satisfied simultaneously.

**IBO 23 notes.** Frontier models reach near-ceiling performance on knowledge-centric multiple-choice biology (molecular biology and genetics), but are noticeably weaker on evidence-centric reasoning (statistics/design and data/figure interpretation), consistent with broader multimodal grounding limitations.

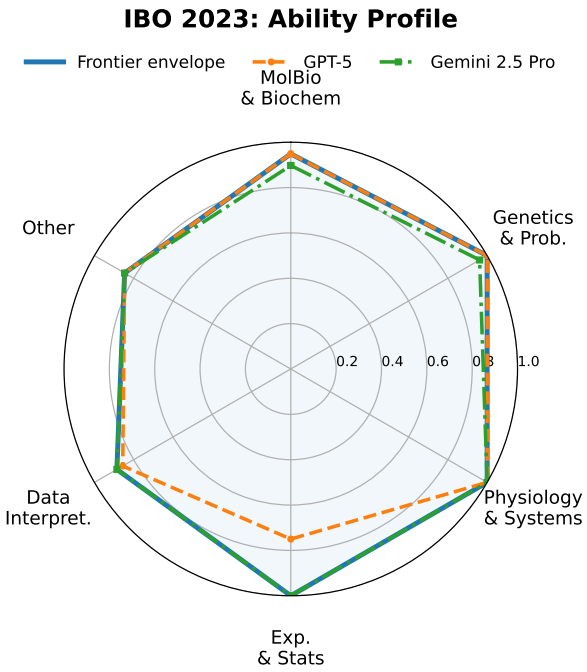


Figure 7: **IBO 2023.** Biology abilities are summarized into observed question categories derived from model outputs: molecular biology/biochemistry, genetics/probability, physiology/systems, experimental design/statistics, data/figure interpretation, and a residual "other" bucket.