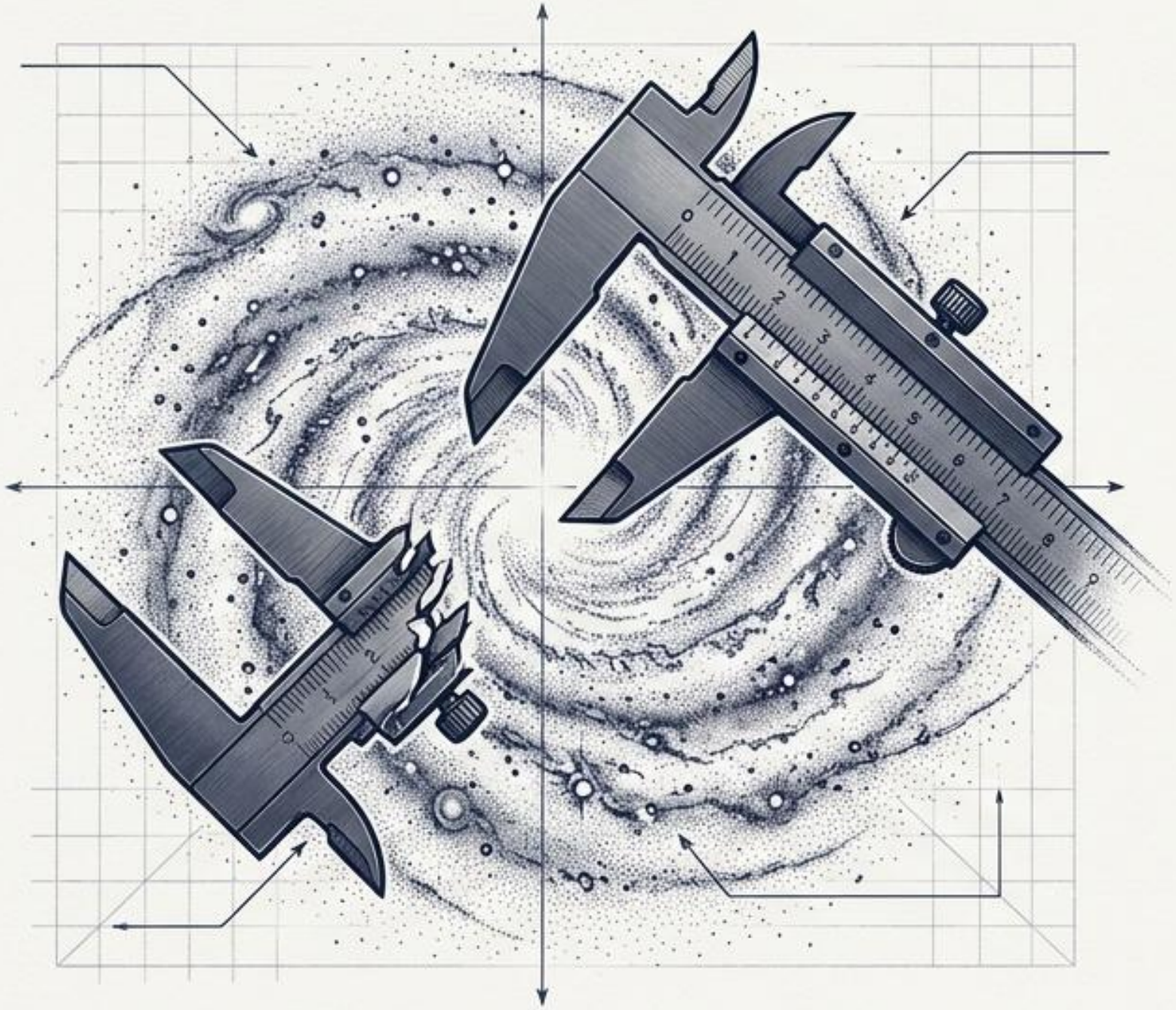
ICML 2026

# HiPhO: How Far Are (M)LLMs from Humans in the Latest High School Physics Olympiad Benchmark?

**Fangchen Yu**[*], Haiyuan Wan[*], Qianjia Cheng[*], Yuchen Zhang, Jiacheng Chen, Fujun Han, Yulun Wu, Junchi Yao, Ruilizhen Hu, Ning Ding, Yu Cheng, Tao Chen, Lei Bai, Dongzhan Zhou[✉], Yun Luo[✉], Ganqu Cui[✉], Peng Ye[✉]

# The Gap in Physics Reasoning

While math Olympiads have been extensively explored, physics remains a 'crown jewel' for AI reasoning.

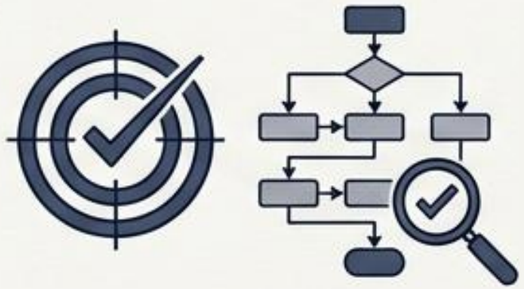



## Outdated Coverage

Datasets like OlympiadBench includes IPhO problems only up to 2021.

## Lack of Modality Diversity

Physics Olympiad problems often involve complex diagrams.



## Coarse Evaluation

Simple accuracy vs. professional step-level gradings

# HiPhO is the first benchmark dedicated to recent physics Olympiads.

## Up-to-date Coverage

Compiles the latest 13 Olympiads from 2024-2025 (IPhO, APhO, EuPhO, etc.)

## Professional Evaluation

Uses official exam rubrics for fine-grained grading at both the answer and step levels.


## Mixed-Modal Content

Complete exams spanning pure text to complex, data-rich diagrams

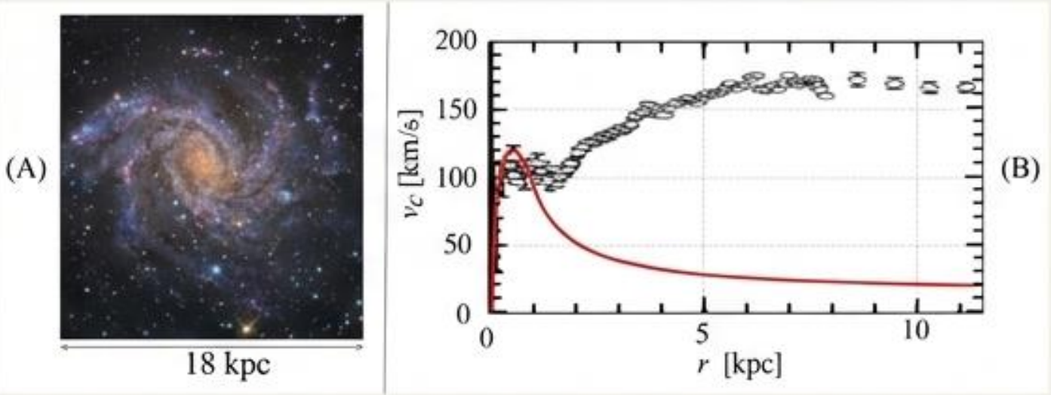
## Human-Level Comparison

Maps AI performance directly against Gold, Silver, and Bronze human cutoffs.

# HiPhO covers 360 expertly curated problems across five physics fields.

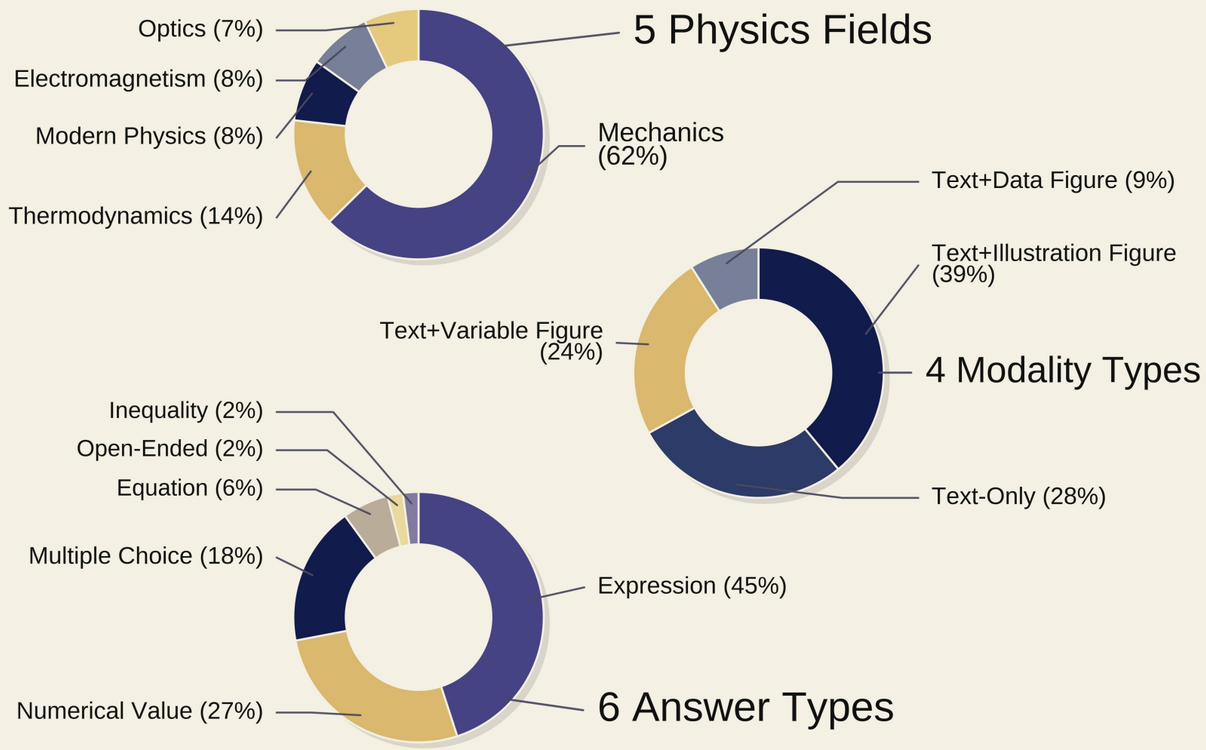
## The Anatomy of a Problem

Context: [Hydrogen and galaxies] This problem ...




| Official Scoring Rubric | |
|---|---|
| - Expression for $g(r)$ | 0.1 pt |
| - Expression for $k_1$ | 0.1 pt |
| - Mass derivation | 0.1 pt |
| **Problem Score = max(answer-level, step-level)** | |

## The Statistics of HiPhO



# Adding variables and data to figures breaks current AI vision.

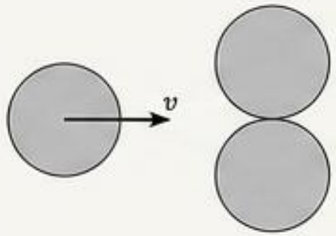
**Massive performance drops:**
**Models fail to extract precise plot coordinates.**

**Text-Only (TO)**

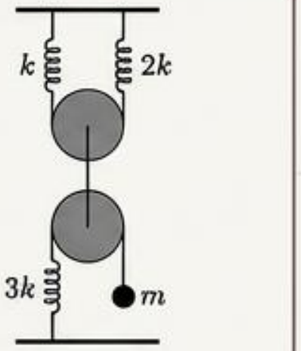
$$E = mc^2;$$
$$F = m\vec{a};$$
$$\Delta\vec{p} = F\Delta t;$$
$$W = \int \vec{F} \cdot d\vec{r}$$

Highest AI Accuracy

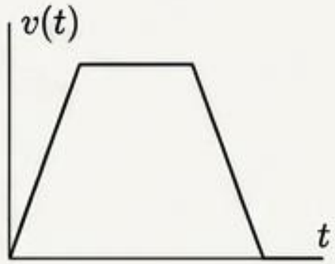
**Text+Illustration Fig. (TI)**




**Text+Variable Fig. (TV)**



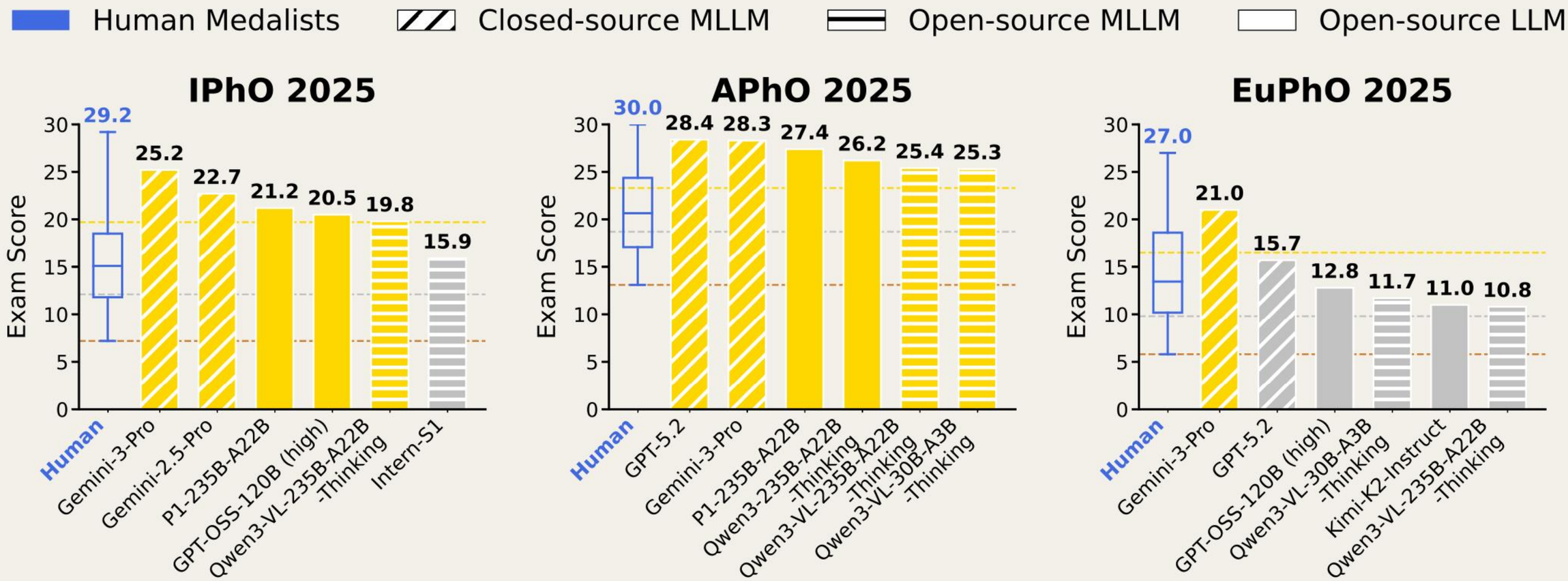

**Text+Data Fig. (TD)**




Visual Complexity Spectrum

FAILURE RATE

# The Leaderboard: AI has reached the podium, but humans still hold the summit.

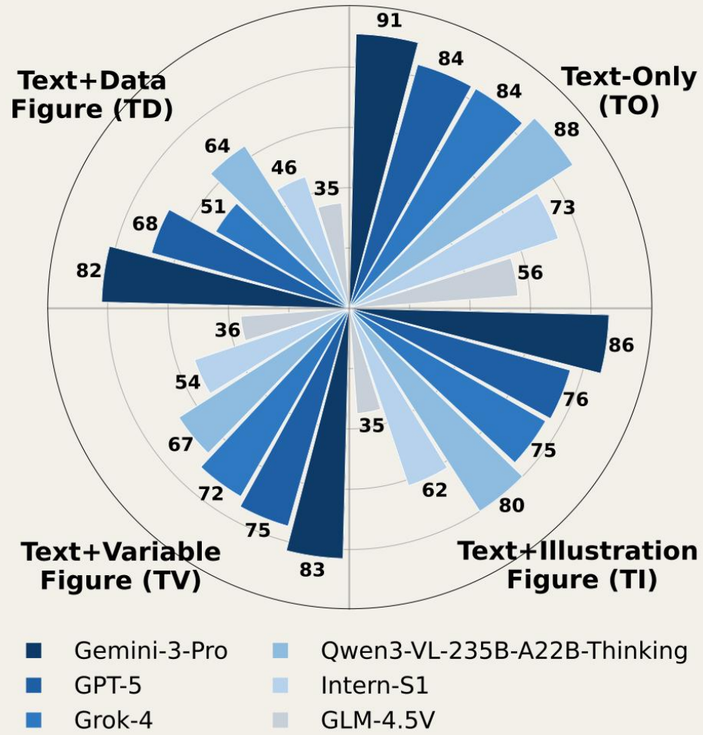



HiPhO: How Far Are (M)LLMs from Humans in the Latest High School Physics Olympiad Benchmark?

# Data-Figures and Optics: Key Bottlenecks in Multimodal and Physics Reasoning

## - Visual Complexity:

Performance drops significantly from TO to TD.




## - Field Difficulty:

Optics: The hardest field (all models < 66%) due to complex geometry and symbolic derivation.

*Table.* Mean normalized scores (%) across five physics fields, reflecting differences in reasoning difficulty.

| Physics Field | Mech. | Elec. | Ther. | Opt. | Mode. |
|---|---|---|---|---|---|
| Gemini-3-Pro | 87 | 88 | 94 | 66 | 87 |
| GPT-5 | 80 | 72 | 82 | 52 | 79 |
| Grok-4 | 74 | 73 | 84 | 51 | 81 |
| Qwen3-VL-235B-A22B-Thinking | 79 | 66 | 86 | 62 | 77 |
| Intern-S1 | 63 | 67 | 64 | 41 | 64 |
| GLM-4.5V | 42 | 35 | 44 | 30 | 49 |

# The Future: Bridging Visual Understanding and Experimental Reasoning.

The State of AI Physics Reasoning: Models have reached the Olympic podium, yet true human-level mastery requires closing the gaps in multimodal grounding, symbolic derivation, and experimental interaction.

Read the Paper

Access Dataset

View Leaderboard