Dimensionality and Measurement Precision in HLE's Multiple-Choice Subset

Mayank Sharma, Savira Nadela, Tyler Matteson
Stanford University
{masharma, savira, tylerjm}@stanford.edu

1 Introduction

Humanity's Last Exam (HLE) has quickly emerged as a prominent benchmark for evaluating advanced language models, appearing in capability assessments and policy discussions shortly after its release [1]. Designed to resist simple retrieval and pattern matching through expert-level questions spanning mathematics, the natural sciences, and the humanities, HLE represents a plausible candidate for evaluating higher-order reasoning. However, widespread adoption has outpaced systematic evaluation of its measurement properties. Most benchmark studies, including HLE, report aggregate accuracy and domain-specific subscores without testing whether the reported domains correspond to empirically distinct latent factors. These subscores are often interpreted as evidence that one model outperforms another in domains such as mathematics or chemistry, an interpretation that implicitly assumes benchmark categories recover as separable latent capabilities rather than as facets of a single general reasoning factor [2]. Whether HLE's domain labels reflect psychometrically distinct abilities remains largely untested, and the implications are concrete: if the eight domains do not correspond to distinct latent constructs, then per-domain model rankings reported in leaderboards are not warranted by the instrument, and developers and policy audiences should treat HLE as a measure of general reasoning rather than a profile of domain-specific skill.

A second concern is measurement precision. Even if HLE is unidimensional, its ability to differentiate between models depends on where along the ability continuum its items concentrate information. If most items are calibrated for average-performing models, score differences between the strongest frontier models may reflect measurement noise rather than genuine capability gaps, a problem that will intensify as models continue to improve.

This paper addresses both concerns directly. We evaluate 29 language models on the text-only multiple-choice subset of HLE ($J = 428$ items) and apply psychometric methods to investigate two questions: (a) **Does HLE's eight-domain structure reflect distinct latent constructs, or do the domains collapse into a general reasoning factor?**, addressed through McDonald's ω_h, principal component analysis of item response profiles, residual correlation analysis, and domain-level ability comparisons; and (b) **Where along the ability continuum does HLE concentrate measurement precision, and which domains contribute most to discrimination among frontier models?**, addressed through the test information function decomposed by subject domain.

1.1 Related Work

Benchmarks are the primary instrument by which progress in large language models is evaluated, but their useful lifespan is often short. Popular benchmarks such as MMLU have shown signs of saturation within only a few years of release [1]. This has motivated the development of HLE, a benchmark consisting of 2,500 expert-level questions spanning mathematics, humanities, and the natural sciences, crowdsourced from nearly 1,000 subject-matter experts across 50 countries and filtered to ensure that contemporary language models could not reliably answer them at submission time [1]. HLE and related studies (e.g., MMLU [3], GPQA [4]) typically report aggregate accuracy

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

as the primary evaluation metric, but aggregate scores alone obscure whether items meaningfully discriminate between models and whether domain subscores reflect distinct latent constructs.

These measurement properties matter because benchmark rankings increasingly inform high-stakes decisions ranging from model deployment to AI governance [5]. When domain subscores are interpreted as evidence of distinct capabilities, the validity of those comparisons is often assumed rather than empirically demonstrated, and decisions based on poorly validated measurements risk misrepresenting both relative model capability and the nature of progress in language model development [6].

Psychometric methods offer a framework for evaluating these properties directly. Applying IRT across 29 NLP datasets, Vania et al. [7] showed that several widely used benchmarks contain items that fail to discriminate effectively between models, raising questions about whether aggregate scores reflect meaningful capability differences. TinyBenchmarks [8] uses IRT to enable efficient evaluation on subsets of MMLU, and Chatbot Arena [9] applies Bradley-Terry models to pairwise comparisons. These studies demonstrate the feasibility of treating benchmarks as measurement instruments, but they focus on ranking efficiency and item difficulty rather than on latent dimensionality: the question of whether domain-specific capability claims are structurally warranted. More broadly, benchmark leaderboards frequently report domain-specific performance as evidence of separable capabilities without testing whether such categories recover as empirically distinct latent factors [2]. To our knowledge, no psychometric dimensionality analysis has yet been applied to HLE. This study addresses that gap.

2 Methods

2.1 HLE Subset

The full benchmark contains two question formats: exact-match short-answer (≈76%) and multiple-choice (≈24%). Approximately 14% of all questions require image comprehension. Subject coverage skews toward quantitative domains: mathematics (41%), biology/medicine (11%), computer science/AI (10%), physics (9%), humanities/social science (9%), other (9%), chemistry (7%), and engineering (4%). We restrict to the *text-only multiple-choice* subset, which is the only subset amenable to automated binary scoring without multimodal infrastructure or LLM-judge extraction. This subset comprises $J = 513$ items, with individual items containing between 2 and 21 answer choices. The full dataset is publicly available via HuggingFace [1] (`cais/hle`, `split="test"`) and is filtered to this subset by selecting `answer_type == "multipleChoice"` and excluding image-dependent items. These items (see examples below) demonstrate that they cannot be answered by retrieval or pattern matching and require domain expertise, making them valuable for an IRT analysis.

Example Items from HLE

1. **Philosophy:** "Which condition of Arrhenius's sixth impossibility theorem do critical-level views violate?" with answer choices including Egalitarian Dominance, General Non-Extreme Priority, Non-Elitism, Weak Non-Sadism, and Weak Quality Addition.

2. **Linguistics:** "What is the standard Japanese pitch accent pattern of the word for 'younger brother' in Japanese?" with choices spanning Heiban, Atamadaka, Nakadaka, Odaka, and Heiban/Nakadaka.

2.2 Model Selection

For generating responses on the subset, we used 37 contemporary language models across five model families: **OpenAI** ($n = 12$): GPT-4.1 series (standard, mini, nano), GPT-5 series (mini, nano, 5.4-mini, 5.4-nano), GPT-4o series (2024-11-20, mini), and reasoning models (o3-mini, o4-mini, o1-mini); **Anthropic** ($n = 6$): Claude Opus (4.7, 4.6, 4.5), Claude Sonnet (4.6, 4.5), and Claude Haiku 4.5; **Google** ($n = 5$): Gemini 3.5 Flash, Gemini 3.1 Flash Lite, Gemini 2.5 series (Pro, Flash, Flash Lite); **DeepSeek** ($n = 4$): DeepSeek-V3, DeepSeek-V4 (Flash, Pro), and DeepSeek-R1 (reasoning); **Open-weight models** ($n = 10$): Gemma (2-9B, 3-27B), Phi-4, Qwen2.5-7B, QwQ-32B

[1] https://huggingface.co/datasets/cais/hle

(reasoning), Sky-T1-32B (reasoning), Mistral-7B, OLMo-2-7B, Falcon3-10B, and Llama-3.1-8B. Models represented both reasoning-specialized architectures and standard instruction-tuned models.

2.3 Response Collection

For each model-item pair, we prompted models using a standardized format:

Standardized Prompt Format

Your response should be in the following format:
`Explanation:` *{your explanation for your answer choice}*
`Answer:` *{your chosen answer}*
`Confidence:` *{your confidence score between 0% and 100% for your answer}*

To ensure reproducibility, we set `temperature=0.0` for all models supporting this parameter. Reasoning-specialized models (OpenAI o-series, Claude Opus 4.7) used their default extended reasoning configurations without temperature control. Maximum response lengths were set to 8,192 tokens, with model-specific adjustments for known constraints (e.g., 4,096 tokens for smaller models).

Proprietary models (OpenAI, Anthropic, Google, DeepSeek) were queried via their respective API endpoints with asynchronous request handling and automatic retry logic for rate-limit management. Open-weight models were deployed using vLLM [10], a high-throughput inference engine optimized for large language models, on cloud GPU infrastructure (Modal Labs) using Nvidia A100-40GB, Nvidia A100-80GB, and Nvidia H100 GPUs. Data collection spanned approximately 20 hours, with the majority of latency attributable to reasoning-specialized models.

2.4 Response Parsing and Scoring

Model responses were parsed using rule-based extraction with a hierarchical strategy: (1) regex pattern matching for explicit answer declarations ("Answer:", "Final Answer"), supporting markdown formatting and parenthetical notation (`(A)`); (2) flexible pattern matching for natural language phrasings ("the correct answer is A," "choice is B"); and (3) fallback extraction of the last isolated letter (A-Z) when no explicit answer marker was present. Parsed responses (91.8%) were scored as correct (1) or incorrect (0) by exact-match comparison to the ground-truth answer. Unparsed responses (8.2%) were coded as missing data and predominantly resulted from empty model outputs (DeepSeek variants), safety-filtered refusals (Gemini 2.5 Pro), or inference failures (Phi-4, Gemma-9B).

2.5 Response Matrix Construction

We constructed a binary response matrix $\mathbf{X} \in \{0, 1, \text{NA}\}^{N \times J}$ where N represents models and $J = 513$ items. From the initial 37 models evaluated, one model (o1-mini) was excluded due to zero coverage (0% valid responses). Coverage among the remaining 36 models ranged from 50.9% to 100%, with median 99.7% and mean 91.8%. Because missingness was not at random (attributable to content safety filtering, inference failures, and rate limiting rather than item difficulty), we retained only the 29 models with excellent coverage ($\geq$95%). Among the 513 items evaluated with these 29 models, 428 items (83.4%) exhibited non-zero variance across models; the remaining 85 items (16.6%) showed zero variance, with all models responding incorrectly, indicating extreme difficulty. We removed zero-variance items to obtain a final analytic matrix of $\mathbf{X} \in \{0, 1\}^{29 \times 428}$, comprising 12,412 model-item observations, with remaining sparse missing values (0.79% of observations; $n = 117$) coded as incorrect responses, as their negligible proportion introduces minimal bias to parameter estimates.

2.6 Analytic Overview

First, we ask whether HLE's eight subject-domain labels correspond to empirically distinct latent constructs or collapse to a single general factor. Second, we ask how well HLE measures along the underlying factor(s): where measurement precision concentrates along the ability continuum, and which domains contribute most. We address both questions through a unified analysis, first estimating

a two-parameter logistic (2PL) IRT model to obtain item-level discrimination and difficulty parameters, which serve as shared inputs to both stages, the dimensionality analysis and the measurement precision analysis that follows it.

2.6.1 2PL Model Estimation

We modeled item-level measurement properties using a two-parameter logistic (2PL) IRT model [11] fit to $\mathbf{X}$. The probability that model i correctly answers item j was modeled as:

$$P(X_{ij} = 1 \mid \theta_i, a_j, b_j) = \frac{1}{1 + \exp[-a_j(\theta_i - b_j)]}, \tag{1}$$

where θ_i represents latent model ability, a_j is the item discrimination parameter, and b_j is the item difficulty parameter. Higher discrimination values indicate items that better differentiate between stronger and weaker models, whereas higher difficulty values indicate items requiring greater latent ability for a 50% probability of correct response. The model was estimated using marginal maximum likelihood implemented in `torch_measure` [12], with optimization run for up to 2,000 epochs using a learning rate of 0.05. For each item, we extracted estimated discrimination ($\hat{a}_j$) and difficulty ($\hat{b}_j$).

2.6.2 Dimensionality Analysis

We examined whether the benchmark's eight subject-domain labels reflected empirically distinct latent constructs. Using data from $N = 29$ models and $J = 428$ items, we found that the inter-item tetrachoric correlation matrix was rank-deficient and non-positive-definite, preventing the use of confirmatory factor analysis because the resulting fit indices would be invalid [13]. We therefore substituted three N-robust alternatives that together address the same substantive question without requiring inversion of the full tetrachoric correlation matrix.

McDonald's ω_h. Our primary evidence for unidimensionality is McDonald's hierarchical omega [14], computed analytically from the 2PL discrimination parameters. Under the normal-ogive parameterization, item j's loading on the general factor is $\lambda_j = a_j/\sqrt{1 + a_j^2}$, and ω_h is defined as:

$$\omega_h = \frac{\left(\sum_j \lambda_j\right)^2}{\left(\sum_j \lambda_j\right)^2 + \sum_j (1 - \lambda_j^2)} \tag{2}$$

ω_h quantifies the proportion of item variance attributable to the general factor, ranging from 0 (no general factor) to 1 (perfectly unidimensional). Unlike CFA-based indices, this estimator requires no matrix inversion and is stable at small N. We report ω_h overall and separately for each of the eight subject domains, with 95% bootstrap confidence intervals constructed by resampling models with replacement ($B = 200$) and refitting the 2PL at each iteration.

PCA on item response profiles. As a model-free complement, we transposed the response matrix to $\mathbf{X}^\top \in \{0, 1\}^{J \times N}$, treating each item as a point in N-dimensional model space, and applied standard PCA. If domain labels capture real structure, items from the same domain should cluster together in this space. We quantified this using domain R^2, the proportion of variance in the first three principal components explained by domain membership, where $R^2 \approx 0$ would indicate that items from the same domain respond no more similarly across models than items from different domains.

Residual item correlations. To further assess whether domain membership explains structure beyond the general factor, we computed Pearson correlations between item response vectors, subtracted the general-factor-implied correlation matrix $\mathbf{R}_g = \boldsymbol{\lambda}\boldsymbol{\lambda}^\top$, and compared within-domain to between-domain residual correlations. If domains capture distinct latent structure, within-domain residuals should be systematically higher than between-domain residuals.

Domain-level $\hat{\theta}$ correlations. To assess whether domain subscores carry any incremental information beyond the total score, we fit separate 2PL models within each domain and computed the Pearson and Spearman correlations between domain-specific ability estimates $\hat{\theta}_{\text{domain}}$ and overall ability

estimates $\hat{\theta}$. If domain subscores reflect distinct latent dimensions, models should exhibit differential ability profiles across domains; if a single underlying dimension dominates, $r \approx 1$ across all domains would indicate that the total score is sufficient and domain subscores carry little information.

2.6.3 Measurement Precision

Taking the 2PL estimates and dimensionality findings together, we characterize where measurement precision concentrates along the ability continuum using the test information function [11],

$$I(\theta) = \sum_{j=1}^{J} a_j^2 P_j(\theta) \left[1 - P_j(\theta)\right], \tag{3}$$

which quantifies measurement precision at different ability levels, with higher values corresponding to lower conditional SEM. We decomposed the TIF by subject domain to evaluate which domains contributed most strongly to discrimination among higher-performing models.

3 Results and Discussion

3.1 2PL Model Estimation

Across the full item pool, estimated difficulty parameters ranged from -2.06 to 5.67, with a median of $b = 0.41$, indicating that the benchmark was moderately challenging for the evaluated models. The positively skewed distribution further suggests a substantial concentration of advanced items capable of differentiating performance across a wide range of contemporary LLMs [5]. Discrimination parameters ranged from 0.00 to the imposed upper cap of 5.00, with a median of $a = 1.49$. Many items demonstrated moderate-to-high discrimination values, indicating that the benchmark effectively differentiated between stronger and weaker models (see Figure 1). Mean empirical accuracy across items was relatively low ($M = 0.17$), further supporting the conclusion that the benchmark was challenging for most models. See Appendix A for supplemental results.

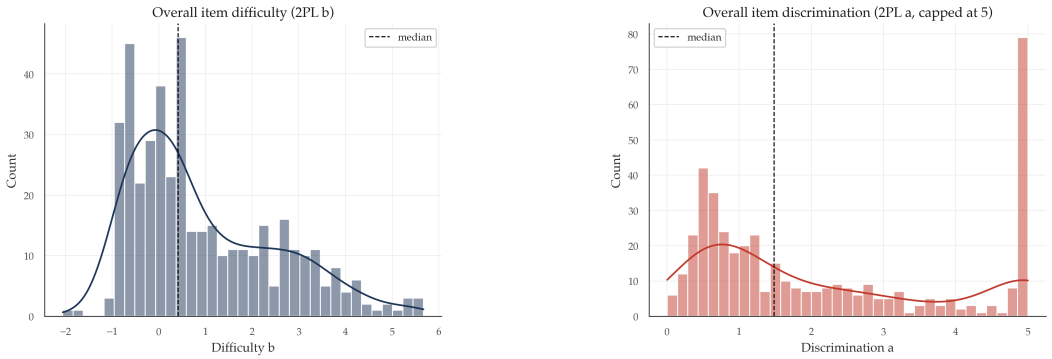

Figure 1: Distributions of item difficulty (b) and item discrimination (a). Dashed vertical lines indicate median parameter values.

Substantial variation emerged across disciplinary categories (see Figure 2). Engineering items exhibited the highest median difficulty ($b = 1.65$), followed by Physics ($b = 1.47$), whereas Computer Science/AI and Humanities/Social Science showed the lowest median difficulty values (both approximately $b = 0.10$). In contrast, discrimination patterns differed from difficulty trends. Computer Science/AI demonstrated the highest median discrimination ($a = 2.63$), while Engineering showed the lowest ($a = 0.77$), indicating that highly difficult domains did not necessarily provide the strongest differentiation between models. Category-level empirical accuracies aligned broadly with difficulty estimates, with Humanities/Social Science producing the highest mean accuracy ($M = 0.19$) and Engineering the lowest ($M = 0.13$).

Figure 3 displays estimated latent ability $\hat{\theta}$ for all 29 models with 95% confidence intervals alongside raw accuracy. Claude Opus 4.7 achieved the highest estimated ability, followed by Gemini 3.5 Flash

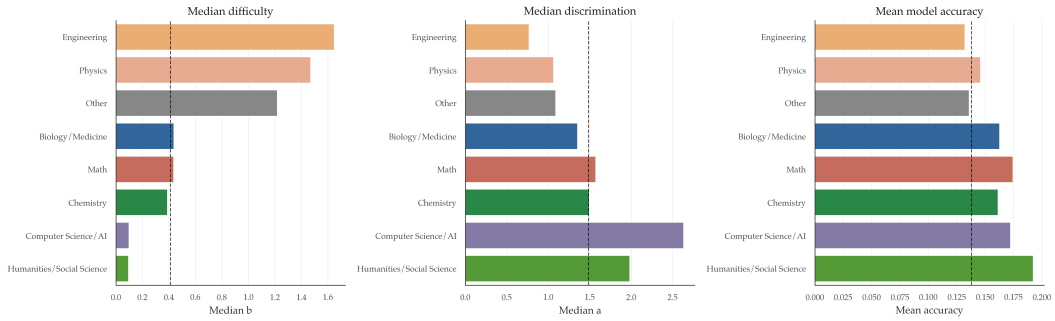

Figure 2: Category-level comparison of median item difficulty, median discrimination, and mean empirical accuracy across benchmark domains. Dashed vertical lines indicate overall median or mean values.

and Claude Opus 4.6. GPT-4o-2024-11-20 exhibited an unusually low ability estimate ($\hat{\theta} \approx -1.8$) with wide confidence intervals, suggesting unstable parameter estimation likely due to atypically low or inconsistent response patterns. IRT-based ability estimates were strongly correlated with raw accuracy rankings ($\rho = 0.857$, $p < 10^{-8}$), confirming that the 2PL model recovers a latent dimension consistent with aggregate performance while additionally characterizing item-level discrimination and measurement precision unavailable from raw scores alone. However, category-level differences in item parameters do not necessarily imply that HLE measures multiple distinct latent abilities: domains could still reflect a single underlying reasoning dimension if models that perform well in one domain tend to perform well across all others. The following analysis examines this directly.

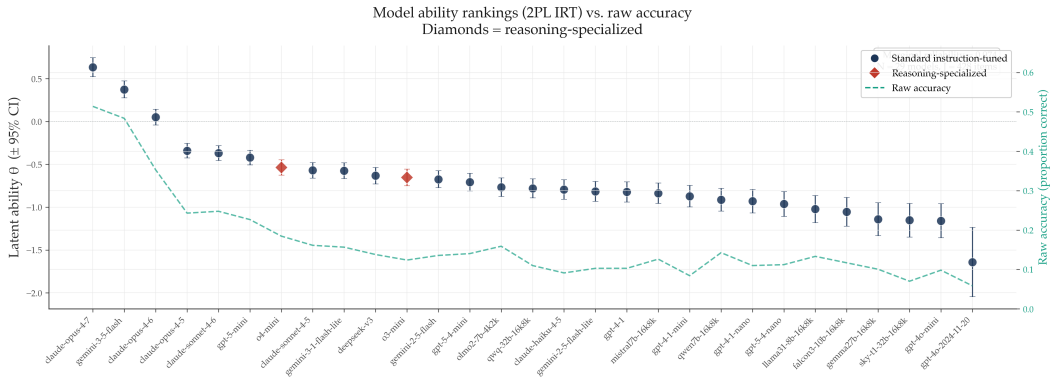

Figure 3: Estimated latent ability $\hat{\theta}$ for all 29 models with 95% CIs (left axis), ordered by ability estimate, alongside raw accuracy (right axis, dashed). The two rankings are strongly correlated ($\rho = 0.857$), though IRT additionally captures item-level discrimination and uncertainty.

3.2 Dimensionality Analysis

McDonald's ω_h. We found $\omega_h = 0.998$ (95% bootstrap CI $[0.998, 0.999]$; $B = 200$ resamples), indicating that 99.8% of common item variance is attributable to a single general factor. The result is stable across all eight subject domains, with domain-level ω_h ranging from 0.936 (Engineering, $n = 18$) to 0.994 (Biology/Medicine, $n = 122$). The bootstrap standard error of 0.0002 confirms that this finding is not sensitive to the specific models in our analytic sample.

PCA on item response profiles. Domain labels explained only 3.5% of variance in the first three principal components (domain $R^2 = 0.035$), indicating that HLE's subject categories impose no discernible structure on item response profiles beyond the dominant general factor.

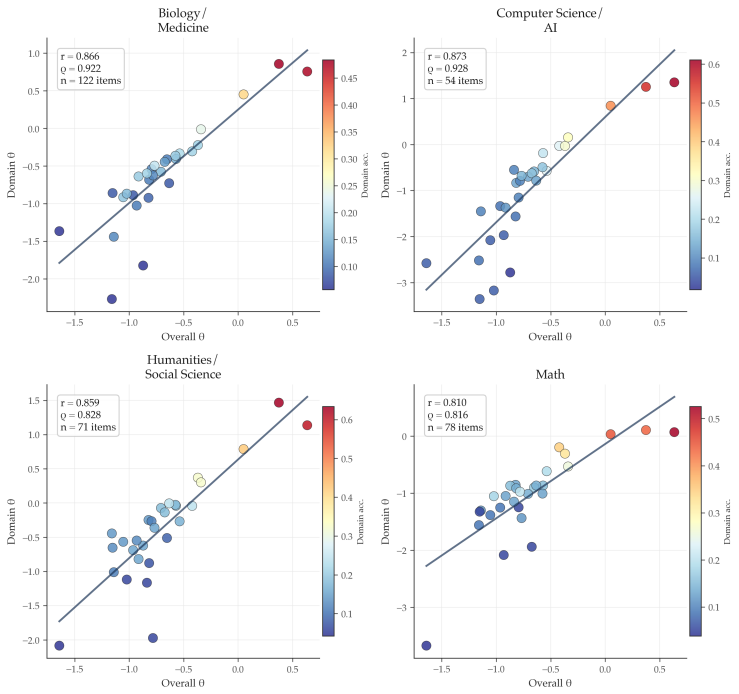

Figure 4: Domain-specific latent ability $\hat{\theta}_{\text{domain}}$ against overall $\hat{\theta}$ for the four well-powered domains ($n \geq 50$). Each point is one model, coloured by domain-specific accuracy. $r \geq 0.81$ and $\rho \geq 0.82$ across all four domains indicates near-redundancy between domain and total-score ability estimates.

Residual item correlations. Within-domain and between-domain residual correlations were nearly identical ($\mu_{\text{within}} = -0.462$, $\mu_{\text{between}} = -0.466$, Cohen's $d = 0.016$). The near-zero effect size indicates that after removing the general factor, domain membership explains nothing about residual item correlation structure.[2]

Domain-level $\hat{\theta}$ correlations. For the four domains with sufficient item counts ($n \geq 50$), domain $\hat{\theta}$ was strongly correlated with overall $\hat{\theta}$ (Figure 4): Computer Science/AI ($r = 0.873$, $\rho = 0.928$), Biology/Medicine ($r = 0.866$, $\rho = 0.922$), Humanities/Social Science ($r = 0.859$, $\rho = 0.828$), and Mathematics ($r = 0.810$, $\rho = 0.816$; all $p < 10^{-7}$). Correlations for smaller domains (Engineering $n = 18$, Physics $n = 26$, Chemistry $n = 22$) were substantially attenuated and are not interpreted, as fitting a 2PL to fewer than 30 items with $N = 29$ models yields poorly identified parameter estimates. The near-identity relationship across all four well-powered domains indicates that a model's total ability estimate is sufficient; domain subscores contribute little information beyond overall ability.

3.3 Measurement Precision

The test information function (TIF) peaks at $\theta = -0.35$, coinciding with the ability range of most models in our sample (10th–90th percentile interval: $[-1.14, -0.26]$). Measurement precision drops substantially above $\theta = 0$, precisely where the strongest-performing models sit. Decomposing the TIF by domain (Figure 5), Biology/Medicine contributes the most total information in absolute terms ($I = 12{,}611$). However, Engineering concentrates the highest proportion of its information at the frontier (69.1% at $\theta > 0$), consistent with its extreme item difficulty. Mathematics and Humanities/Social Science show the lowest frontier concentration (41.5% and 41.4% respectively),

[2]Absolute residual values are negative due to attenuation of Pearson r on binary items; since attenuation affects within- and between-domain pairs equally, the comparison remains valid.

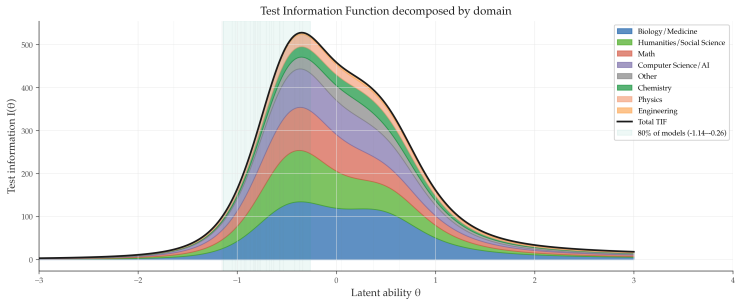

Figure 5: TIF decomposed by domain (stacked area). The shaded band marks the 10th-90th percentile ability range of models ($\theta \in [-1.14, -0.26]$), showing that HLE concentrates measurement precision at moderate ability levels. The TIF drops sharply above $\theta = 0$

and despite containing more items than most other domains ($n = 78$ and $n = 71$), contribute disproportionately little to discrimination among higher-performing models.

4 Conclusion and Future Work

Fitting a 2PL IRT model to responses from 29 language models on HLE's text-only multiple-choice subset ($J = 428$ items), we find convergent evidence that the benchmark measures a single general reasoning factor. McDonald's $\omega_h = 0.998$ (95% CI $[0.998, 0.999]$), domain labels explained only 3.5% of item response variance, residual correlations after removing the general factor were nearly identical within and between domains (Cohen's $d = 0.016$), and domain-specific ability estimates were near-redundant with the total score ($r \geq 0.81$ across all well-powered domains). HLE's eight subject-domain labels do not correspond to empirically distinct latent constructs; so they might be arbitrary partitions of a unidimensional space. A separate finding concerns measurement precision. While HLE discriminates well among models at moderate ability levels, precision drops substantially above $\theta = 0$, precisely where frontier models sit. Engineering items concentrate the most information at the frontier (69.1% at $\theta > 0$) but represent only 4% of the benchmark; Mathematics and Humanities/Social Science, the two largest domains, contribute disproportionately little to frontier discrimination. As stronger models continue to emerge, this precision gap will become an increasingly binding constraint on HLE's utility as a differentiating instrument.

The psychometric approach worked well for this setting: IRT discrimination parameters proved stable enough to support downstream dimensionality analyses, and the convergence of four independent analyses (ω_h, domain R^2, residual correlations, and $\hat{\theta}$ redundancy) substantially strengthens confidence in the unidimensionality conclusion. The primary methodological challenge was the small model sample ($N = 29$), which precluded confirmatory factor analysis and limited statistical power for domain-level analyses. This inverted the usual psychometric setting, where items are typically far fewer than subjects, and required N-robust alternatives throughout. A practical lesson for future benchmark psychometrics is that model coverage must be treated as a first-class design consideration: missingness that is not at random, as observed here for reasoning-specialized models, systematically biases the analytic sample and limits generalizability.

Several limitations bound these conclusions. Our analytic sample is restricted to the text-only multiple-choice subset (≈19% of HLE), so findings do not extend to exact-match or image-dependent items; moreover, the multiple-choice format may introduce construct-irrelevant variance if models succeed through elimination rather than genuine domain knowledge. The model sample is small ($N = 29$) and non-random, as eight models were excluded for low response coverage, disproportionately affecting reasoning-specialized architectures and precluding formal measurement invariance testing across model families. Scoring each model on a single deterministic response per item means discrimination estimates reflect stable model differences but not within-model response variability, which may underestimate uncertainty in item parameter estimates. Future work should replicate on the full benchmark with a larger and more complete model sample to enable CFA and invariance testing.

Code Availability

All analysis code used in this study is publicly available at: `https://github.com/matrix-mayank/hle-psychometrics`.

AI Use and Scientific Responsibility Statement

AI tools were used in a limited way throughout this project, mainly to help improve writing clarity, troubleshoot coding issues, and brainstorm ways to present results more clearly. To avoid plagiarism or non-attribution of ideas, all citations and references were added manually by the authors after checking the original academic sources directly. We also reviewed AI-generated suggestions carefully to make sure ideas, interpretations, and wording were appropriately credited and not copied without attribution. Potential inaccuracies were addressed by comparing all explanations and statistical interpretations against the actual analysis outputs and referenced literature rather than relying on AI-generated responses alone. To reduce bias, we used cautious interpretations and clearly discussed methodological limitations, including the restricted benchmark subset and small model sample size. Responsibility for the originality, accuracy, and scientific integrity of the final paper remained entirely with the authors.

Reflection on AI Tool Usage

AI tools were used mainly for minor support tasks during the project, such as improving wording, helping organize ideas, and assisting with small coding/debugging issues. Most of the research process, including the psychometric analyses, interpretation of findings, and methodological decisions, was completed independently by the authors. The tools were helpful for making explanations more concise and readable, especially when writing technical sections related to Item Response Theory and dimensionality analysis. At the same time, the project highlighted that AI-generated suggestions still require careful review, since some outputs were inaccurate or not fully consistent with the statistical assumptions of the analyses. The experience showed us that AI tools can be useful for supporting productivity and communication, but they are most effective when combined with independent reasoning and subject knowledge rather than being relied on for substantive scientific conclusions.

Impact Statement

This project examines the psychometric properties of a large language model benchmark and highlights the importance of carefully interpreting benchmark scores and domain subscores. One potential positive impact is encouraging more rigorous evaluation of benchmark validity, especially when benchmark rankings are used in research, industry, or policy discussions. However, the findings could also be misinterpreted if generalized beyond the analyzed subset or treated as definitive statements about model intelligence. To reduce this risk, the paper clearly discusses limitations such as the small model sample and restriction to the multiple-choice subset of HLE. The study did not involve human participants or sensitive personal data, and all analyses were conducted using publicly available benchmark outputs. We followed Stanford's standards of academic integrity by properly citing prior work, verifying interpretations manually, and ensuring that all conclusions were based on the empirical analyses conducted in the study.

References

[1] Long Phan, Alice Gatti, Nathaniel Li, Adam Khoja, Ryan Kim, Richard Ren, Jason Hausenloy, Oliver Zhang, Mantas Mazeika, Dan Hendrycks, Ziwen Han, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, Michael Choi, Anish Agrawal, Arnav Chopra, Aakaash Nattanmai, Gordon McKellips, Anish Cheraku, Asim Suhail, Ethan Luo, Marvin Deng, Jason Luo, Ashley Zhang, Kavin Jindel, Jay Paek, Kasper Halevy, Allen Baranov, Michael Liu, Advaith Avadhanam, David Zhang, Vincent Cheng, Brad Ma, Evan Fu, Liam Do, Joshua Lass, Hubert Yang, Surya Sunkari, Vishruth Bharath, Violet Ai, James Leung, Rishit Agrawal, Alan Zhou, Kevin Chen, Tejas Kalpathi, Ziqi Xu, Gavin Wang, Tyler Xiao, Erik Maung, Sam Lee, Ryan Yang, Roy Yue, Ben Zhao, Julia Yoon, Xiangwan Sun,

Aryan Singh, Clark Peng, Tyler Osbey, Taozhi Wang, Daryl Echeazu, Timothy Wu, Spandan Patel, Vidhi Kulkarni, Vijaykaarti Sundarapandiyan, Andrew Le, Zafir Nasim, Srikar Yalam, Ritesh Kasamsetty, Soham Samal, David Sun, Nihar Shah, Abhijeet Saha, Alex Zhang, Leon Nguyen, Laasya Nagumalli, Kaixin Wang, Aidan Wu, Anwith Telluri, Summer Yue, Alexandr Wang, Dmitry Dodonov, Tung Nguyen, Jaeho Lee, Daron Anderson, Mikhail Doroshenko, Alun Cennyth Stokes, Mobeen Mahmood, Oleksandr Pokutnyi, Oleg Iskra, Jessica P. Wang, John-Clark Levin, Mstyslav Kazakov, Fiona Feng, Steven Y. Feng, Haoran Zhao, Michael Yu, Varun Gangal, Chelsea Zou, Zihan Wang, Serguei Popov, Robert Gerbicz, Geoff Galgon, Johannes Schmitt, Will Yeadon, Yongki Lee, Scott Sauers, Alvaro Sanchez, Fabian Giska, Marc Roth, Søren Riis, Saiteja Utpala, Noah Burns, Gashaw M. Goshu, Mohinder Maheshbhai Naiya, Chidozie Agu, Zachary Giboney, Antrell Cheatom, Francesco Fournier-Facio, Sarah-Jane Crowson, Lennart Finke, Zerui Cheng, Jennifer Zampese, Ryan G. Hoerr, Mark Nandor, Hyunwoo Park, Tim Gehrunger, Jiaqi Cai, Ben McCarty, Alexis C. Garretson, Edwin Taylor, Damien Sileo, Qiuyu Ren, Usman Qazi, Lianghui Li, Jungbae Nam, John B. Wydallis, Pavel Arkhipov, Jack Wei Lun Shi, Aras Bacho, Chris G. Willcocks, Hangrui Cao, Sumeet Motwani, Emily de Oliveira Santos, Johannes Veith, Edward Vendrow, Doru Cojoc, Kengo Zenitani, Joshua Robinson, Longke Tang, Yuqi Li, Joshua Vendrow, Natanael Wildner Fraga, Vladyslav Kuchkin, Andrey Pupasov Maksimov, Pierre Marion, Denis Efremov, Jayson Lynch, Kaiqu Liang, Aleksandar Mikov, Andrew Gritsevskiy, Julien Guillod, Gözdenur Demir, Dakotah Martinez, Ben Pageler, Kevin Zhou, Saeed Soori, Ori Press, Henry Tang, Paolo Rissone, Sean R. Green, Lina Brüssel, Moon Twayana, Aymeric Dieuleveut, Joseph Marvin Imperial, Ameya Prabhu, Jinzhou Yang, Nick Crispino, Arun Rao, Dimitri Zvonkine, Gabriel Loiseau, Mikhail Kalinin, Marco Lukas, Ciprian Manolescu, Nate Stambaugh, Subrata Mishra, Tad Hogg, Carlo Bosio, Brian P. Coppola, Julian Salazar, Jaehyeok Jin, Rafael Sayous, Stefan Ivanov, Philippe Schwaller, Shaipranesh Senthilkumar, Andres M. Bran, Andres Algaba, Kelsey Van den Houte, Lynn Van Der Sypt, Brecht Verbeken, David Noever, Alexei Kopylov, Benjamin Myklebust, Bikun Li, Lisa Schut, Evgenii Zheltonozhskii, Qiaochu Yuan, Derek Lim, Richard Stanley, Tong Yang, John Maar, Julian Wykowski, Mart Oller, Anmol Sahu, Cesare Giulio Ardito, Yuzheng Hu, Ariel Ghislain Kemogne Kamdoum, Alvin Jin, Tobias Garcia Vilchis, Yuexuan Zu, Martin Lackner, James Koppel, Gongbo Sun, Daniil S. Antonenko, Steffi Chern, Bingchen Zhao, Pierrot Arsene, Joseph M. Cavanagh, Daofeng Li, Jiawei Shen, Donato Crisostomi, Wenjin Zhang, Ali Dehghan, Sergey Ivanov, David Perrella, Nurdin Kaparov, Allen Zang, Ilia Sucholutsky, Arina Kharlamova, Daniil Orel, Vladislav Poritski, Shalev Ben-David, Zachary Berger, Parker Whitfill, Michael Foster, Daniel Munro, Linh Ho, Shankar Sivarajan, Dan Bar Hava, Aleksey Kuchkin, David Holmes, Alexandra Rodriguez-Romero, Frank Sommerhage, Anji Zhang, Richard Moat, Keith Schneider, Zakayo Kazibwe, Don Clarke, Dae Hyun Kim, Felipe Meneguitti Dias, Sara Fish, Veit Elser, Tobias Kreiman, Victor Efren Guadarrama Vilchis, Immo Klose, Ujjwala Anantheswaran, Adam Zweiger, Kaivalya Rawal, Jeffery Li, Jeremy Nguyen, Nicolas Daans, Haline Heidinger, Maksim Radionov, Václav Rozhoň, Vincent Ginis, Christian Stump, Niv Cohen, Rafał Poświata, Josef Tkadlec, Alan Goldfarb, Chenguang Wang, Piotr Padlewski, Stanislaw Barzowski, Kyle Montgomery, Ryan Stendall, Jamie Tucker-Foltz, Jack Stade, T. Ryan Rogers, Tom Goertzen, Declan Grabb, Abhishek Shukla, Alan Givré, John Arnold Ambay, Archan Sen, Center for AI Safety, Scale AI, and HLE Contributors Consortium. A benchmark of expert-level academic questions to assess ai capabilities. *Nature*, 649(8099):1139–1146, 2026.

[2] Han Jiang, Susu Zhang, Dongyao Zhu, Yuzhuo Bai, Sang T. Truong, Xiaoyuan Yi, Sanmi Koyejo, Xing Xie, and Ziang Xiao. Ai evaluation should require standardized item-level data releases. *arXiv preprint arXiv:2604.03244*, 2026.

[3] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2021.

[4] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. Gpqa: A graduate-level google-proof qa benchmark, 2023.

[5] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023.

[6] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models, 2022.

[7] Clara Vania, Phu Mon Htut, William Huang, Dhara Mungra, Richard Yuanzhe Pang, Jason Phang, Haokun Liu, Kyunghyun Cho, and Samuel R. Bowman. Comparing test sets with item response theory. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1141–1158, Online, August 2021. Association for Computational Linguistics.

[8] Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinybenchmarks: evaluating llms with fewer examples. *arXiv preprint arXiv:2402.14992*, 2024.

[9] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael I. Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*, 2024.

[10] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention, 2023.

[11] Frank B. Baker and Seock-Ho Kim. *Item Response Theory: Parameter Estimation Techniques*. CRC Press, 2004.

[12] Sang T. Truong et al. torch_measure: A package for ai measurement science, 2026. MIT License.

[13] David B. Flora and Patrick J. Curran. An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, 9(4):466–491, 2004.

[14] Roderick P. McDonald. *Test Theory: A Unified Treatment*. Psychology Press, 1 edition, 1999.

A Appendix

A.1 Additional 2PL Diagnostic Plots

Figure 6 presents supplementary diagnostic visualizations for the estimated 2PL parameters. The left panel illustrates the inverse relationship between item difficulty and discrimination, where extremely difficult items tended to exhibit lower discrimination values. The right panel shows the expected negative relationship between empirical accuracy and estimated item difficulty, supporting the consistency of the estimated IRT parameters.

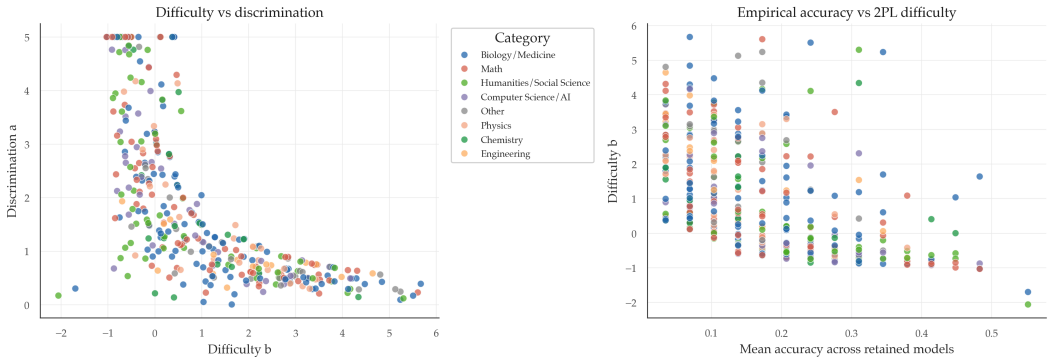

Figure 6: Supplementary 2PL diagnostic plots. Left: relationship between item difficulty and discrimination across benchmark domains. Right: empirical item accuracy against estimated 2PL difficulty.