Can LLMs Estimate Student Struggles? Human-AI Difficulty Alignment with Proficiency Simulation for Item Difficulty Prediction

Anonymous ACL submission

Abstract

Accurate estimation of item (question or task) difficulty is critical for educational assessment but suffers from the cold start problem. While Large Language Models demonstrate superhuman problem-solving capabilities, it remains an open question whether they can perceive the cognitive struggles of human learners. In this work, we present a large-scale empirical analysis of **Human-AI Difficulty Alignment** for over 20 models across diverse domains such as medical knowledge and mathematical reasoning. Our findings reveal a systematic misalignment where scaling up model size is not reliably helpful; instead of aligning with humans, models converge toward a shared machine consensus. We observe that high performance often impedes accurate difficulty estimation, as models struggle to simulate the capability limitations of students even when being explicitly prompted to adopt specific proficiency levels. Furthermore, we identify a critical lack of introspection, as models fail to predict their own limitations. These results suggest that general problem-solving capability does not imply an understanding of human cognitive struggles, highlighting the challenge of using current models for automated difficulty prediction.

1 Introduction

Accurate estimation of item difficulty is the cornerstone of educational assessment (Hambleton et al., 1991; Hsu et al., 2018; AlKhuzaey et al., 2021; Peters et al., 2025). It underpins critical applications such as curriculum design, automated test generation, and automated item generation with controlled difficulty levels (DeMars, 2010; Lord, 2012). Traditionally, obtaining accurate difficulty parameters (e.g., within Item Response Theory (IRT) models (Baker, 2001; Lalor et al., 2024)) relies on extensive field testing, a process that requires administering questions to large cohorts of real test-takers to observe response patterns. This reliance creates a significant cold start problem: newly generated questions lack the historical response data necessary to statistically estimate their parameters, effectively rendering them unusable in adaptive systems until they undergo expensive and time-consuming pre-testing cycles.

Prior approaches to Item Difficulty Prediction (IDP) generally treated the task as a supervised learning problem, relying on linguistic features or deep learning models trained on known item parameters estimated based on item response data (Hsu et al., 2018; Benedetto, 2023; Li et al., 2025b). While effective within specific domains, these methods depend heavily on the availability of historical performance data for training, limiting their utility in cold-start scenarios (i.e., no historical tested data is available for training). The emergence of LLMs (OpenAI, 2024b; Hurst et al., 2024; Touvron et al., 2023; Qwen-Team, 2024, 2025a) offers a potential paradigm shift. With their vast pre-training and exceptional problem-solving capabilities, LLMs seemingly possess the knowledge required to analyze complex content. ***However, it remains an open question whether these general-purpose models can align with human perception of difficulty without task-specific fine-tuning***. There is a fundamental distinction between solving a problem and evaluating its difficulty: a model that effortlessly surpasses human baselines in performance may fail to recognize the cognitive hurdles faced by an average learner (Sweller, 1988, 2011; Noroozi and Karami, 2022; Li et al., 2025c). This study investigates this ***Human-AI Difficulty Alignment***, exploring whether off-the-shelf LLMs can bridge the gap between their own capabilities and the student struggles, whose difficulty values are obtained from real student field testing.

To investigate this, we propose a comprehensive empirical study that evaluates this Difficulty Alignment through two distinct lenses: the model as an **external observer** (predicting others' difficulty)

and an **internal actor** (experiencing difficulty itself). Our study operates at scale, benchmarking over 20 LLMs, spanning both open-weights and closed-source families, including reasoning-specialized models, across four diverse educational domains: language proficiency (Cambridge) (Mullooly et al., 2023), reasoning and logic (SAT Reading/Writing, SAT Math), and professional medical knowledge (USMLE) (Yaneva et al., 2024).

We structure our investigation around three primary dimensions to disentangle the relationship between intrinsic capability and extrinsic perception. ***First***, we go beyond simple ground-truth correlation to analyze inter-model consensus, examining whether models form a cohesive machine perception that systematically diverges from students. ***Second***, we quantify the capability-perception gap using **Item Response Theory (IRT)**. By treating the model pool as a cohort of synthetic students, we derive empirical machine difficulty based on the actual correctness of LLMs, allowing us to test for the Curse of Knowledge, where items challenging for humans are trivial for machines. ***Finally***, we evaluate **Metacognitive Alignment** and **Proficiency Simulation** (Tseng et al., 2024; Zhang et al., 2024b), rigorously testing whether models possess the introspection to predict their own limitations or the flexibility to authentically simulate the cognitive struggles of lower-proficiency students.

Key Findings.

1. **Systematic Misalignment:** Contrary to standard capability metrics, scaling does not reliably translate into alignment. Increasing model scale does not improve difficulty predictions; instead, models form a cohesive Machine Consensus, aligning significantly stronger with each other than with human reality.

2. **Limits of Simulation:** Neither extrinsic ensembling nor proficiency simulation serves as a reliable fix for the misalignment. Ensemble performance is strictly bounded by weaker models, while proficiency simulation proves highly inconsistent as models struggle to authentically mimic different proficiency levels.

3. **The Curse of Knowledge:** Our IRT-based analysis reveals a fundamental mechanistic divergence: the difficulty derived from models' actual correctness correlates even worse with humans than their explicit perceptions. Items that are difficult for humans are frequently trivial for models, and this capability exhibits significant inertia even under weak student prompts.

4. **Metacognitive Blindness:** We identify a critical lack of introspection. With AUROC scores hovering near random guessing, models fail to predict their own limitations, indicating that explicit difficulty estimates are effectively decoupled from the model's actual correctness, lacking the internal signal to ground their predictions.

2 Formulation and Evaluation Metrics

2.1 Task Formulation

We formalize the IDP task as a function approximation problem over a dataset $\mathcal{D} = \{(x_i, a_i^*, y_i)\}_{i=1}^N$. Each entry comprises the item context x_i (including the question stem, optional passages, and candidate options), the ground truth answer a_i^*, and the difficulty label $y_i \in \mathcal{Y}$ obtained from real-world student field testing. The label space $\mathcal{Y}$ adapts to the domain, ranging from continuous values (e.g., $y_i \in [0, 1]$) to discrete categories (e.g., $\mathcal{Y} = \{\text{Easy, Medium, Hard}\}$).

Let $\mathcal{M}$ denote the set of LLMs under evaluation. We investigate the alignment between human difficulty y_i and AI cognition through two distinct modalities: *Difficulty Perception* (the model's estimation as an observer) and *Problem-Solving Capability* (the model's performance as an actor).

The Observer View: Difficulty Perception. In this mode, the model simulates an educator or test developer tasked with estimating item difficulty. To isolate perception from solving capability, the model is provided with the full item context x_i, the correct solution a_i^*, and an optional proficiency prompt p. The predicted difficulty $\hat{y}_{i,m}$ for model $m \in \mathcal{M}$ is formulated as:

$$\hat{y}_{i,m} = \phi\left(\text{Gen}_m(x_i, a_i^*, p)\right) \tag{1}$$

where $\text{Gen}_m(\cdot)$ denotes the natural language generation process, and $\phi(\cdot)$ is a parsing function that maps the generated response to a normalized numerical difficulty score.

The Actor View: Intrinsic Capability. To assess the model's actual performance, we evaluate it in a standard zero-shot test-taking setting where the correct answer is hidden. The model generates a solution:

$$\hat{a}_{i,m} = \psi(\text{Gen}_m(x_i, p)) \tag{2}$$

where $\psi(\cdot)$ extracts the final answer. The binary correctness $v_{i,m} \in \{0, 1\}$ is subsequently determined by $v_{i,m} = \mathbb{I}(\hat{a}_{i,m} = a_i^*)$.

2.2 Evaluation Framework

To analyze the divergence between human and AI cognition, we evaluate alignment across two distinct dimensions: *Perceived Difficulty* (what models predict) and *Empirical Difficulty* (what models experience). We adopt **Spearman's Rank Correlation (ρ)** as the unified metric for both dimensions, as it robustly measures monotonicity, the ability to correctly distinguish that item A is harder than item B, while remaining invariant to systematic scaling shifts. Moreover, Spearman correlation allows unified comparison across heterogeneous label granularities, discrete or continuous.

Perception Alignment (ρ_{pred}). This metric evaluates the model's accuracy as an *observer*. We calculate the Spearman correlation between the model's predicted difficulty scores $\hat{y}_{i,m}$ and the human ground truth y_i.

$$\rho_{pred,m} = \text{Spearman}(\{\hat{y}_{i,m}\}_{i=1}^{N}, \{y_i\}_{i=1}^{N}) \quad (3)$$

A higher $\rho_{pred,m}$ indicates that the model's explicit perception of difficulty aligns with the human hierarchy. For simplicity, we omit the subscript m when not referring to a specific model.

Capability Alignment (ρ_{irt}). This metric evaluates the model's alignment as an *actor*. To quantify this, we first derive the *Empirical Machine Difficulty* β_i using Item Response Theory (IRT). We treat the set of models $\mathcal{M}$ as a population of synthetic examinees and construct a binary correctness matrix. We fit a Rasch Model (1-Parameter Logistic Model) where the probability of model m answering item i correctly is:

$$P(v_{i,m} = 1 \mid \theta_m, \beta_i) = \frac{1}{1 + \exp(-(\theta_m - \beta_i))} \quad (4)$$

Here, β_i represents the intrinsic machine difficulty of item i, estimated via Marginal Maximum Likelihood Estimation. Crucially, we then calculate the correlation between this empirical difficulty and human difficulty:

$$\rho_{irt} = \text{Spearman}(\{\beta_i\}_{i=1}^{N}, \{y_i\}_{i=1}^{N}) \quad (5)$$

With ρ_{pred} and ρ_{irt}, we can obtain a systemic view of the Human-AI Difficulty Alignment.

2.3 Proficiency Simulation

To systematically investigate the model's ability to simulate different cognitive states and align with student populations, we define a set of four distinct proficiency configurations $\mathcal{P} = \{p_0, p_{low}, p_{mid}, p_{high}\}$. These configurations serve as system-level instructions that condition the model's generation process.

- **Baseline: No Proficiency (p_0).** This represents the control setting (or vanilla mode). We provide the model with standard instructions to predict difficulty or solve the problem without adopting a specific student proficiency. This setting evaluates the model's *intrinsic* alignment, i.e., its default perception of difficulty.
- **Low-Proficiency Student (p_{low}).** We prompt the model to simulate a student with limited subject mastery. This proficiency aims to test whether the model can suppress its own knowledge to accurately estimate high-difficulty items for struggling learners.
- **Average-Proficiency Student (p_{mid}).** This proficiency represents the median student. This serves as a proxy for the general population average.
- **High-Proficiency Student (p_{high}).** We simulate a top-tier student who has high proficiency in the subject. This setting investigates whether the model aligns closely with the most capable subset of the human population.

For different domains, the proficiency prompts are slightly different to align with the domain-specific scenarios, as shown in the Appendix D. Moreover, we do not provide much description on what a low/average/high-proficiency student would be like, since this is also a part of the investigation on how the model understands them.

2.4 Experimental Setup

Datasets. The key challenge for Human-AI Difficulty Alignment is to find the dataset that has the ground truth difficulty values that are obtained from real student field testing, since most of the existing research on IDP is conducted on private datasets or datasets from multiple resources, resulting really large discrepancy between the model's capability and the real students' performance. To ensure the robustness of our findings across different educational domains, we select four datasets

Model	USMLE	Cambridge	SAT-R	SAT-M	Average
GPT-3.5-Turbo	0.09	0.20	0.26	0.40	0.24
GPT-4o	0.19	0.48	0.38	0.55	0.40
GPT-4o-mini	0.05	0.42	0.26	0.49	0.31
GPT-4.1	0.30	**0.49**	0.49	0.48	**0.44**
GPT-4.1-mini	0.20	0.46	0.48	0.45	0.40
GPT-o4-mini	0.28	0.36	0.46	0.29	0.35
GPT-5	**0.36**	0.36	0.38	0.25	0.34
Llama2-7B	0.05	0.00	0.08	0.03	0.04
Llama2-13B	0.00	0.10	0.14	0.16	0.10
Phi3	0.01	-0.01	0.09	0.30	0.10
Phi3.5	0.06	0.06	0.12	0.32	0.14
Llama3.1-8B	0.04	0.04	0.23	0.43	0.19
Qwen2.5-7B	0.10	0.16	0.22	0.49	0.24
Qwen2.5-32B	0.09	0.44	0.30	**0.56**	0.35
Phi4	0.11	0.40	0.38	0.50	0.35
Qwen3-8B	0.03	0.35	0.20	0.46	0.26
Qwen3-32B	0.11	0.38	0.23	0.52	0.31
DeepSeek-R1	0.28	0.40	**0.50**	0.44	0.40
QWQ-32B	0.20	0.41	0.43	0.52	0.39
R1-Qwen32B	-0.01	0.40	0.25	0.42	0.27
Qwen3-32B (R)	0.21	0.42	0.33	0.48	0.36
Average	0.13	0.30	0.29	0.41	0.28

Table 1: The Spearman's rank correlation results. **Overall alignment remains weak, indicating systematic misalignment.**

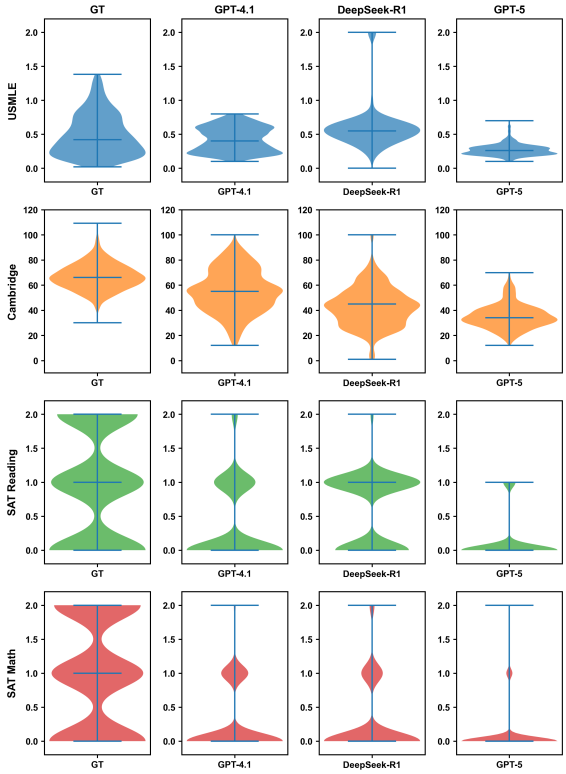

Figure 1: The violin plot of the difficulty prediction distributions of several representative models. **Current advanced models exhibit severe *distribution shift*.**

covering reading comprehension, verbal reasoning, math reasoning, and specialized professional knowledge.

USMLE (Medical Knowledge): Sourced from the United States Medical Licensing Examination (Yaneva et al., 2024), this dataset represents a high-stakes, knowledge-intensive domain. It contains 667 items developed by the NBME and FSMB, ensuring high reliability with field-test data from over 300 medical students per item. We utilize the provided continuous difficulty values (transformed p-values), which range from $[0, 1.3]$.

Cambridge (Linguistic Proficiency): Sourced from the Cambridge Multiple-Choice Questions Reading Dataset (Mullooly et al., 2023), this dataset evaluates English reading comprehension through 120 text passages and 793 distinct questions. A key characteristic is the long context window required for inference. We utilize the rescaled IRT b-parameters provided in the dataset as ground truth, which represent continuous difficulty values in the range of $[0, 100]$.

SAT Reading & Writing (Verbal Reasoning): Comprises reading comprehension and writing mechanics questions from standardized US college admission tests. This dataset challenges the model's ability to process rhetorical structure and standard English conventions. The dataset contains 1338 questions after removing the figure-dependent questions. Unlike the datasets above, the ground truth difficulty is provided as discrete categories: $\{Easy, Medium, Hard\}$.

SAT Math (Mathematical Logic): Includes algebra, geometry, and data analysis problems from the SAT. This dataset tests the model's ability to gauge difficulty in multi-step logical reasoning and mathematical computation tasks. Similar to the verbal component, it contains 1385 questions, and the difficulty values are provided as discrete categories: $\{Easy, Medium, Hard\}$.

Models. We evaluate a comprehensive suite of over 20 LLMs to disentangle the effects of models. We establish a high-performance baseline using proprietary models, including the GPT series (GPT-3.5-Turbo (OpenAI, 2024b), GPT-4o (Hurst et al., 2024), GPT-4o-mini (OpenAI, 2024a), GPT-4.1 (OpenAI, 2025b), GPT-4.1-mini (OpenAI, 2025b), GPT-o4-mini (OpenAI, 2024c), GPT-5 (OpenAI, 2025a)), which represent the current state-of-the-art in general-purpose tasks. The general instruction-following open-weights models including Llama2-7B (Touvron et al., 2023), Llama2-13B (Touvron et al., 2023), Llama3.1-8B (Dubey et al., 2024), Qwen2.5-7B (Qwen-

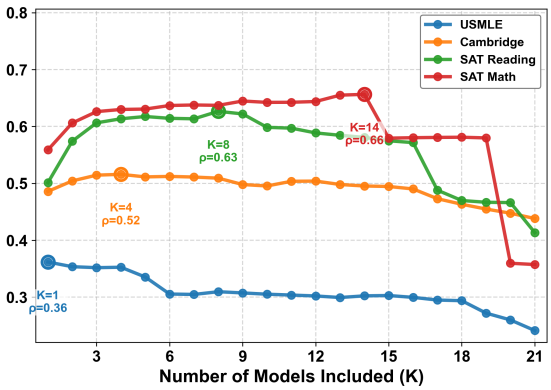

Figure 2: The Spearman correlation trends when greedily ensembling the predictions of the top-K models. **The curve indicates the upper bound of the ensemble performance, which is still weak.**

Team, 2024), Qwen2.5-32B (Qwen-Team, 2024), Phi3 (Abdin et al., 2024a), Phi3.5 (Abdin et al., 2024a), Phi4 (Abdin et al., 2024b), Qwen3-8B, and Qwen3-32B (Qwen-Team, 2025a). To specifically investigate the impact of reasoning capabilities, we incorporate reasoning-focused open-weights models, including DeepSeek-R1(0528) (DeepSeek-AI, 2025), QWQ-32B (Qwen-Team, 2025b), R1-Distill-Qwen2.5-32B (DeepSeek-AI, 2025), and Qwen3-32B (reasoning mode).

3 The Landscape of Explicit Perception

We first establish a baseline for zero-shot IDP, examining whether current state-of-the-art LLMs can intrinsically estimate item difficulty without access to student response data. The results, summarized in Table 1, reveal three primary findings regarding the current limitations of Human-AI difficulty alignment.

3.1 Systematic Misalignment

Key Finding 1: Systematic Misalignment.

We observe a **systematic misalignment** between human perception and LLM estimation of difficulty across all domains. Contrary to standard capability metrics, **scaling does not reliably translate into alignment**: increased model scale and reasoning power do not linearly translate to better predictions, as even frontier models consistently fail to capture human difficulty rankings. Crucially, this misalignment is not driven by random noise but by a cohesive Machine Consensus, where **models exhibit relatively stronger alignment with each other than with human reality**.

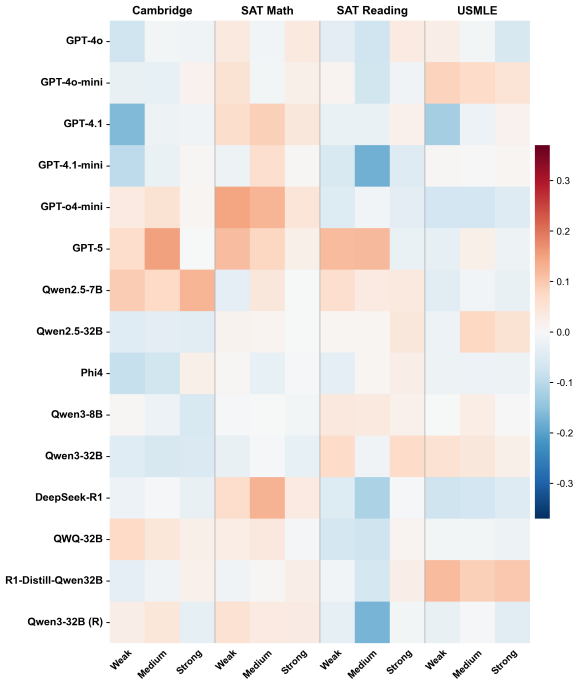

Figure 3: Heatmap showing the correlation change when applying specific personas compared to the baseline. **The impact of individual personas is highly inconsistent and noisy.**

Despite their profound problem-solving capabilities, LLMs struggle with cold-started IDP. As shown in Table 1, Spearman's ρ averages below 0.50, with significant domain sensitivity: alignment is stronger in logic-driven tasks like **SAT Math** ($\rho \approx 0.41$) but collapses in knowledge-intensive domains like **USMLE** ($\rho \approx 0.13$). Counter-intuitively, **scaling laws do not linearly translate to difficulty estimation**. New-generation models (e.g., GPT-5, $\rho = 0.34$) and reasoning specialists fail to outperform generalist baselines like GPT-4.1 ($\rho = 0.44$), indicating that reasoning power does not linearly translate to better difficulty alignment with humans.

Figure 1 identifies the root cause: severe **distribution shift and variance collapse**. While ground truth difficulty spans a broad spectrum, model predictions are narrowly clustered and systematically skewed towards lower values. Advanced models effectively **overestimate student capability**, lacking the granularity to distinguish specific degrees of human struggle.

Finally, we observe a **Machine Consensus** that diverges from human reality. While alignment with ground truth is low, inter-model correlations are relatively higher (indicated by the detailed heatmaps in Appendix B Figure 4, 5, 6, 7). This consensus is capability-dependent: while weaker models exhibit

Model	USMLE	Cambridge	SAT-R	SAT-M	Average
GPT-3.5-Turbo	0.14 (0.05)	0.32 (0.12)	0.31 (0.06)	0.53 (0.13)	0.33 (0.09)
GPT-4o	0.22 (0.04)	0.50 (0.02)	0.49 (0.11)	**0.65** (0.09)	0.47 (0.07)
GPT-4o-mini	0.12 (0.07)	0.46 (0.04)	0.35 (0.09)	0.57 (0.08)	0.38 (0.07)
GPT-4.1	0.31 (0.01)	0.51 (0.02)	**0.60** (0.11)	0.63 (0.15)	**0.51** (0.07)
GPT-4.1-mini	0.23 (0.03)	0.49 (0.03)	0.56 (0.08)	0.56 (0.11)	0.46 (0.06)
GPT-o4-mini	0.26 (-0.02)	0.46 (0.10)	0.56 (0.10)	0.49 (0.20)	0.44 (0.10)
GPT-5	**0.39** (0.03)	**0.52** (0.16)	0.57 (0.19)	0.40 (0.15)	0.47 (0.13)
Llama2-7B	0.09 (0.02)	-0.03 (-0.03)	0.10 (0.03)	0.14 (0.11)	0.08 (0.03)
Llama2-13B	0.03 (0.03)	0.02 (-0.06)	0.12 (-0.02)	0.19 (0.03)	0.09 (-0.01)
Phi3	0.04 (0.04)	0.08 (0.09)	0.07 (0.01)	0.01 (-0.04)	0.05 (0.03)
Phi3.5	0.02 (-0.03)	0.17 (0.12)	0.18 (0.03)	0.35 (0.03)	0.18 (0.04)
Llama3.1-8B	0.01 (-0.03)	0.20 (0.16)	0.33 (0.10)	0.55 (0.11)	0.27 (0.09)
Qwen2.5-7B	0.13 (0.03)	0.34 (0.18)	0.36 (0.14)	0.63 (0.13)	0.37 (0.12)
Qwen2.5-32B	0.16 (0.07)	0.47 (0.04)	0.41 (0.12)	0.64 (0.08)	0.42 (0.08)
Phi4	0.14 (0.03)	0.45 (0.05)	0.47 (0.09)	0.56 (0.06)	0.41 (0.06)
Qwen3-8B	0.06 (0.04)	0.41 (0.07)	0.31 (0.11)	0.56 (0.10)	0.34 (0.08)
Qwen3-32B	0.19 (0.08)	0.43 (0.05)	0.39 (0.16)	0.61 (0.09)	0.41 (0.10)
DeepSeek-R1	0.26 (-0.01)	0.42 (0.02)	0.59 (0.09)	0.62 (0.18)	0.47 (0.07)
QWQ-32B	0.23 (0.04)	0.46 (0.11)	0.50 (0.08)	0.62 (0.08)	0.45 (0.08)
R1-Qwen32B	0.07 (0.08)	0.46 (0.06)	0.17 (0.02)	0.35 (0.00)	0.26 (0.04)
Qwen3-32B (R)	0.23 (0.04)	0.45 (0.06)	0.37 (0.10)	0.59 (0.08)	0.41 (0.07)
Average	0.16 (0.03)	0.36 (0.07)	0.37 (0.09)	0.49 (0.10)	0.35 (0.07)

Table 2: Spearman correlation results using the ensemble average of all persona configurations. The values in red parentheses indicate the absolute improvement over the baseline. **While smaller models show negligible gains, stronger models like GPT-5 exhibit a significant sensitivity to personas.**

stochastic behavior, **advanced models converge on a shared, non-human machine perception of difficulty**. This confirms that the Human-AI Difficulty Alignment is a stable, systematic misalignment rather than random error.

3.2 Ensemble and Proficiency Simulation

Key Finding 2: Limits of Simulation.

Neither extrinsic ensembling nor intrinsic role-playing serves as a reliable solution for alignment. We find that **ensemble performance is strictly bounded by individual model capabilities**: rather than contributing diverse insights, weaker models introduce noise that degrades the performance of high-capability baselines. Similarly, **proficiency simulation shows highly inconsistent results**, as models struggle to authentically mimic different proficiency levels or suppress their intrinsic knowledge: it fails as a faithful cognitive model, but proficiency ensembling can act as a variance-reduction heuristic.

We first perform a greedy ensemble analysis, iteratively aggregating the predicted difficulty scores of the top-K performing models. As shown in Figure 2, the performance trend implies an **upper bound** for alignment capability that is strictly governed by the density of high-quality models within the pool. In domains with a dense cluster of capable models (SAT Math), the ensemble acts as a **denoising mechanism**, steadily improving correlation from 0.56 to a peak of 0.66. However, this gain is fragile: once the threshold ($K = 14$) is crossed, adding weaker models causes immediate **signal dilution**. In sparse domains like **USMLE**, this dilution happens immediately, confirming that the long tail of weaker models introduces destructive noise rather than helpful diversity. Thus, extrinsic aggregation is not a scalable solution but a bounded analysis.

Domain	IRT	Capacity	Cognitive Divergence	
	Global (ρ)	Saturated (90% $\mathcal{M}$ ✓)	Savant ($\mathcal{H}\Uparrow$ $\mathcal{M}\Downarrow$)	Brittle ($\mathcal{H}\Downarrow$ $\mathcal{M}\Uparrow$)
USMLE	0.134	75.6%	70.4%	0.0%
Cambridge	0.309	35.6%	22.1%	0.4%
SAT-R	0.304	45.5%	25.5%	0.0%
SAT-M	0.386	54.6%	32.2%	1.3%

Table 3: Analysis of implicit difficulty alignment based on Machine IRT. $\mathcal{H}$: Human difficulty; $\mathcal{M}$: Model difficulty. Saturated represents the total ratio of items that 90% of the model can solve correctly. Savant represents the ratio of items that are hard for humans (difficulty top 33%) but trivial for most models (correct for 90% of the models). Brittleness represents the ratio of items that are easy for humans, where most of the models fail. **The saturated rate and savant rate present the diagnosis of Human-LLM Difficulty Misalignment.**

We further explore explicit cognitive simulation by prompting models to adopt specific student proficiencies. A granular analysis in Figure 3, the changing ratio of utilizing different proficiencies, reveals that single-proficiency simulation is highly unstable: adopting a specific role often either improves or degrades alignment, suggesting models struggle to faithfully simulate a specific cognitive state in isolation.

However, Table 2 reveals another crucial nuance: while individual simulations are stochastic, the **ensemble average of proficiencies consistently improves alignment**. This benefit is particularly evident in frontier models; for instance, GPT-5 improves its average correlation to 0.47 compared to its baseline of 0.34. This suggests that while models cannot reliably act as a specific student, aggregating their intrinsic variations provides a more robust estimate than a single pass, effectively smoothing out the randomness of individual proficiency prompts. This improvement does not indicate successful student simulation, but rather reflects noise averaging over inconsistent internal states.

	USMLE				Cambridge				SAT Reading&Writing				SAT Math				
Model	**Baseline**	**Weak**	**Medium**	**Strong**	**Baseline**	**Weak**	**Medium**	**Strong**	**Baseline**	**Weak**	**Medium**	**Strong**	**Baseline**	**Weak**	**Medium**	**Strong**	**Average**
GPT-3.5-Turbo	0.868	0.862↓	0.882	0.877↑	0.632	0.628↓	0.612	0.613	0.660	0.667	0.685	0.687↑	0.727	0.750	0.749	0.762↑	0.729
GPT-4o	0.952	0.960	0.957	0.955↑	0.900	0.893↓	0.898	0.897	0.871	0.870↓	0.874	0.874↑	0.908	0.907↓	0.915	0.913↑	0.909
GPT-4o-mini	0.925	0.928	0.933	0.934↑	0.786	0.765↓	0.783	0.793↑	0.780	0.786	0.787	0.783↑	0.899	0.898↓	0.898	0.901↑	0.849
GPT-4.1	0.895	0.906	0.906	0.898↑	0.919	0.922	0.922	0.918	0.888	0.915	0.913	0.905↑	0.904	0.905	0.896	0.910↑	0.908
GPT-4.1-mini	**0.966**	0.939↓	0.966	0.957	0.897	0.868↓	0.895	0.880	0.895	0.892↓	0.907	0.904↑	0.922	0.920↓	0.927	0.925↑	0.916
GPT-o4-mini	0.780	0.799	0.810	0.772	0.909	0.895↓	0.898	0.904	0.908	0.898↓	0.909	0.904	0.906	0.909	0.909	0.902	0.876
GPT-5	0.946	0.939↓	0.942	0.946	**0.958**	0.957↓	0.956	0.961↑	**0.976**	0.978	0.976	0.977↑	0.924	0.914↓	0.926	0.923	**0.950**
Llama2-7B	0.712	0.718	0.687	0.700	0.397	0.397	0.405	0.409↑	0.396	0.380↓	0.401	0.392	0.230	0.224↓	0.221	0.222	0.431
Llama2-13B	0.775	0.720↓	0.763	0.768	0.459	0.428↓	0.455	0.436	0.469	0.477	0.466	0.461	0.267	0.264↓	0.262	0.268↑	0.484
Phi3	0.844	0.843↓	0.843	0.844	0.493	0.489↓	0.489	0.504↑	0.632	0.662	0.638	0.649↑	0.774	0.777	0.788	0.770	0.690
Phi3.5	0.867	0.853↓	0.870	0.865	0.559	0.550↓	0.540	0.551	0.647	0.646↓	0.645	0.642	0.775	0.787	0.784	0.782↑	0.710
Llama3.1-8B	0.891	0.897	0.892	0.880	0.646	0.633↓	0.649	0.639	0.685	0.683↓	0.674	0.674	0.766	0.751↓	0.778	0.771↑	0.744
Qwen2.5-7B	0.850	0.856	0.847	0.852↑	0.675	0.686	0.663	0.690↑	0.730	0.719↓	0.727	0.736↑	0.895	0.900	0.897	0.894	0.789
Qwen2.5-32B	0.922	0.931	0.921	0.933↑	0.783	0.812	0.812	0.803↑	0.812	0.813	0.819	0.816↑	0.919	0.917↓	0.920	0.921↑	0.866
Phi4	0.921	0.918↓	0.928	0.924↑	0.798	0.791↓	0.778	0.784	0.802	0.802	0.801	0.798	0.919	0.913↓	0.908	0.913	0.856
Qwen3-8B	0.903	0.907	0.909	0.901	0.767	0.776	0.765	0.768↑	0.820	0.816↓	0.811	0.802	0.919	0.926	0.916	0.925↑	0.852
Qwen3-32B	0.946	0.957	0.948	0.948↑	0.845	0.863	0.849	0.856↑	0.881	0.879↓	0.868	0.868	0.927	0.932	0.923	0.926	0.901
DeepSeek-R1	0.961	0.966	0.955	**0.967**↑	0.933	0.919↓	0.931	0.922	0.964	0.959↓	0.968	0.954	0.914	0.919	0.915	0.913	0.941
QWQ-32B	0.954	0.948↓	0.960	0.954	0.898	0.890↓	0.899	0.899↑	0.921	0.922	0.915	0.920	**0.953**	0.943↓	0.943	0.941	0.929
R1-Qwen32B	0.939	0.936↓	0.940	0.940↑	0.856	0.827↓	0.868	0.849	0.886	0.877↓	0.881	0.884	0.944	0.947	0.942	0.941	0.904
Qwen3-32B (R)	0.954	0.963	0.969	0.966↑	0.907	0.904↓	0.898	0.897	0.932	0.917↓	0.926	0.925	0.952	0.943↓	0.938	0.938	0.933

Table 4: Problem-solving accuracy across different personas. An arrow ↑ or ↓ is added if the accuracy is improved or degraded compared to the baseline with the corresponding persona. **The magnitude of change is marginal, indicating that models struggle to significantly suppress or enhance their intrinsic capabilities on command.**

4 Capability vs. Perception

While the previous section examined the model as an external observer attempting to predict student struggles, we now turn our focus to the model as an internal actor. To understand the root causes of the observed misalignment, in this section, we disentangle the relationship between what models predict and what they experience.

4.1 The Curse of Knowledge

Key Finding 3: The Curse of Knowledge

Our IRT-based analysis reveals that the difficulty derived from models' actual correctness correlates **even worse with human than what they explicitly perceive**. This stems from a distinct mechanistic divergence: **items that are conceptually difficult for humans are frequently trivial for models**, resulting in high saturation rates on hard items. Furthermore, this capability exhibits significant inertia: even when explicitly prompted to simulate a lower-proficiency student, **models fail to meaningfully suppress their problem-solving capability**.

We shift our perspective from *prediction* to *experience* by treating our suite of 21 models as a cohort of students to estimate intrinsic machine difficulty (b parameter) via Item Response Theory (IRT). In addition, to further quantify and analyze the divergence, we define additional metrics to be conditional on human difficulty tiers: the *Savant Rate* measures the percentage of items within the top 33% of human difficulty that are solved by over 90% of models, while the *Brittleness Rate* measures the percentage of items within the bottom 33% of human difficulty where the model pass rate is below 50%.

Table 3 exposes a profound **Cognitive Divergence**, where the correlation of model IRT difficulty is even lower than the model perception. In domains like **USMLE**, the misalignment is extremely discouraging. The Savant Rate of 70.4% implies that for over two-thirds of the questions, humans find most difficult, can be solved by over 90% of models easily. Coupled with a massive dataset-wide saturation of 75.6%, this creates a flat difficulty that explains the poor alignment.

In contrast, reasoning domains like **SAT Math** show partial alignment but remain limited by hyper-competence. While the correlation ($\rho = 0.386$) is higher than in knowledge domains, the Savant Rate is still significant at 32.2%. It indicates that even in logical reasoning, one-third of the problems that challenge humans are trivial for machines. Furthermore, with 54.6% of all items being saturated, models are getting too powerful capabilities to simulate student struggles.

Finally, we examine whether simulating proficiencies shifts intrinsic capability. As shown in Table 4, while applying a proficiency generally shifts accuracy in the expected direction, the magnitude is negligible (typically $< 1\%$ change). This reveals the **Curse of Knowledge**. Highly capable models are too strong to fail, and their intrinsic objective to maximize correctness overrides proficiency instructions. This prevents them from authentically simulating the specific misconceptions of struggling students, as they cannot unsee the correct answers they already possess.

Model	USMLE	Cambridge	SAT-R	SAT-M	Average
GPT-3.5-Turbo	0.54	0.51	0.49	**0.65**	0.55
GPT-4o	0.62	0.59	0.54	0.63	0.60
GPT-4o-mini	0.55	0.57	0.51	0.63	0.56
GPT-4.1	0.55	0.61	0.58	0.57	0.57
GPT-4.1-mini	0.55	0.56	0.59	0.62	0.58
GPT-o4-mini	0.52	0.59	0.55	0.53	0.55
GPT-5	0.60	**0.73**	**0.72**	0.60	**0.67**
Llama2-7B	0.49	0.50	0.46	0.54	0.50
Llama2-13B	0.48	0.49	0.50	0.54	0.50
Phi3	0.47	0.48	0.49	0.62	0.51
Phi3.5	0.50	0.51	0.48	0.58	0.52
Llama3.1-8B	0.53	0.51	0.49	0.63	0.54
Qwen2.5-7B	0.53	0.52	0.47	0.57	0.52
Qwen2.5-32B	0.55	0.54	0.49	0.62	0.55
Phi4	0.53	0.54	0.57	**0.65**	0.57
Qwen3-8B	0.55	0.52	0.49	0.64	0.55
Qwen3-32B	0.62	0.52	0.52	0.61	0.57
DeepSeek-R1	0.64	0.53	0.64	0.62	0.61
QWQ-32B	0.65	0.58	0.59	**0.65**	0.62
R1-Qwen32B	0.58	0.54	0.55	**0.65**	0.58
Qwen3-32B (R)	**0.69**	0.56	0.54	0.58	0.59
Average	0.55	0.55	0.54	0.60	0.56

Table 5: AUROC (Area Under the ROC Curve) that measures the alignment between predicted difficulty and the model's own correctness. A value of 0.5 indicates random alignment. **Most models hover near 0.55, revealing a critical lack of self-awareness: they fail to predict their own potential errors.**

4.2 Metacognitive Blindness.

Key Finding 4: Metacognitive Blindness.

Our AUROC analysis reveals a critical lack of introspection: models are unable to accurately predict their own potential limitations. With AUROC scores hovering near random guessing (approximately 0.55) across most models, we find that **explicit difficulty estimates are effectively decoupled from the model's actual correctness**. This suggests a fundamental blind spot: because models cannot reliably identify which tasks exceed their own capabilities, they lack the necessary internal signal to ground their estimates of human difficulty.

A fundamental question remains regarding whether the model's perception of difficulty is aligned with its own problem-solving capability. To quantify this *Metacognition*, we formulate difficulty prediction as a binary classification task regarding the model's own correctness.

For a given item x_i, we assign a label $l_i = 1$ if the model answers incorrectly ($v_i = 0$) and $l_i = 0$ if it answers correctly ($v_i = 1$), using the predicted difficulty score $\hat{y}_i$ as the prediction probability. We calculate the Area Under the Receiver Operating Characteristic Curve (AUROC) to measure separability, where the value represents *the probability that the model assigns a higher difficulty score to an item it answers incorrectly compared to one it answers correctly*. An AUROC of approximately 0.5 indicates random alignment where the model lacks self-awareness, whereas values significantly greater than 0.5 demonstrate positive metacognition where the model correctly identifies its failure.

Table 5 presents the results of this analysis and exposes a critical **Metacognitive Blind Spot** across the majority of evaluated LLMs. Despite achieving high problem-solving accuracy, most models exhibit AUROC scores ranging between 0.50 and 0.60. This weak internal alignment implies that they lack the introspection to identify when a task exceeds their capabilities. Furthermore, even frontier models fail to demonstrate robust metacognition. While GPT-5 and DeepSeek-R1 show slight deviations from random guessing by achieving localized highs of 0.73 on Cambridge and 0.64 on USMLE, respectively, their overall discrimination remains poor, with averages hovering near or below 0.67. These values indicate that even the most advanced models still **struggle to reliably distinguish between questions they can answer and those they cannot, highlighting that accurate self-awareness remains a persistent deficiency in current LLMs.**

5 Conclusion

This study demonstrates that Large Language Models currently struggle to align with human perception of difficulty despite their advanced problem-solving capabilities. We find that increasing model scale does not guarantee better alignment but rather fosters a machine consensus that systematically diverges from student reality. Our investigation attributes this failure to a fundamental capability gap where models cannot effectively suppress their knowledge to simulate struggling students, coupled with a critical lack of metacognitive introspection regarding their own limitations. Ultimately, these results highlight that bridging the gap between solving a problem and estimating its difficulty requires more than just stronger models or proficiency prompting, calling for new approaches to ground machine cognition in human educational needs.

Limitations

Our investigation into proficiency simulation relied on zero-shot prompting strategies. While this reflects the most common and accessible usage of LLMs, it assumes that models can internally calibrate to a proficiency level without examples. We did not explore few-shot prompting with real student error patterns or fine-tuning on student response logs (Student Trace Modeling). It is possible that few-shot scenarios can improve the correlation between the Human-AI alignment. However, in this paper, we focus on the intrinsic capability of LLMs that are not affected by further learning or training processes.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, and 110 others. 2024a. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, and 1 others. 2024b. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Samah AlKhuzaey, Floriana Grasso, Terry R Payne, and Valentina Tamma. 2021. A systematic review of data-driven approaches to item difficulty prediction. In *International conference on artificial intelligence in education*, pages 29–41. Springer.

Samah AlKhuzaey, Floriana Grasso, Terry R Payne, and Valentina Tamma. 2024. Text-based question difficulty prediction: A systematic review of automatic approaches. *International Journal of Artificial Intelligence in Education*, 34(3):862–914.

Alfonso Amayuelas, Kyle Wong, Liangming Pan, Wenhu Chen, and William Wang. 2024. Knowledge of Knowledge: Exploring Known-Unknowns Uncertainty with Large Language Models. (arXiv:2305.13712).

Frank B Baker. 2001. *The basics of item response theory*. ERIC.

Luca Benedetto. 2023. A quantitative study of nlp approaches to question difficulty estimation. In *International conference on artificial intelligence in education*, pages 428–434. Springer.

Faeze Brahman, Sachin Kumar, Vidhisha Balachandran, Pradeep Dasigi, Valentina Pyatkin, Abhilasha Ravichander, Sarah Wiegreffe, Nouha Dziri, Khyathi Chandu, Jack Hessel, and 1 others. 2024. The art of saying no: Contextual noncompliance in language models. *Advances in Neural Information Processing Systems*, 37:49706–49748.

Lida Chen, Zujie Liang, Xintao Wang, Jiaqing Liang, Yanghua Xiao, Feng Wei, Jinglei Chen, Zhenghong Hao, Bing Han, and Wei Wang. 2024. Teaching Large Language Models to Express Knowledge Boundary from Their Own Signals. (arXiv:2406.10881).

Konstantina Chrysafiadi and Maria Virvou. 2013. Student modeling approaches: A literature review for the last decade. *Expert Systems with Applications*, 40(11):4715–4729.

Ricardo Conejo, Eduardo Guzmán, Jose-Luis Perez-De-La-Cruz, and Beatriz Barros. 2014. An empirical study on the quantitative notion of task difficulty. *Expert Systems with Applications*, 41(2):594–606.

Albert T Corbett and John R Anderson. 1994. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction*, 4(4):253–278.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Christine DeMars. 2010. *Item response theory*. Oxford University Press.

Yang Deng, Yong Zhao, Moxin Li, See-Kiong Ng, and Tat-Seng Chua. 2024. Don't Just Say "I don't know"! Self-aligning Large Language Models for Responding to Unknown Questions with Explanations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13652–13673. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

George Dueñas, Sergio Jimenez, and Geral Mateus Ferro. 2024. Upn-icc at bea 2024 shared task: Leveraging llms for multiple-choice questions difficulty prediction. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 542–550.

Susan E Embretson and Steven P Reise. 2013. *Item response theory for psychologists*. Psychology Press.

Chenrui Fan, Ming Li, Lichao Sun, and Tianyi Zhou. 2025. Missing premise exacerbates overthinking: Are reasoning models losing critical thinking skill? In *Second Conference on Language Modeling*.

Wanyong Feng, Peter Tran, Stephen Sireci, and Andrew S Lan. 2025. Reasoning and sampling-augmented mcq difficulty prediction via llms. In *International Conference on Artificial Intelligence in Education*, pages 31–45. Springer.

Arthur C Graesser, Patrick Chipman, Brian C Haynes, and Andrew Olney. 2005. Autotutor: An intelligent tutoring system with mixed-initiative dialogue. *IEEE Transactions on Education*, 48(4):612–618.

Ronald K Hambleton, Hariharan Swaminathan, and H Jane Rogers. 1991. *Fundamentals of item response theory*, volume 2. Sage.

Jun He, Li Peng, Bo Sun, Lejun Yu, and Yinghui Zhang. 2021. Automatically predict question difficulty for reading comprehension exercises. In *2021 ieee 33rd international conference on tools with artificial intelligence (ictai)*, pages 1398–1402. IEEE.

Fu-Yuan Hsu, Hahn-Ming Lee, Tao-Hsing Chang, and Yao-Ting Sung. 2018. Automated estimation of item difficulty for multiple-choice tests: An application of word embedding techniques. *Information Processing & Management*, 54(6):969–984.

Zhenya Huang, Qi Liu, Enhong Chen, Hongke Zhao, Mingyong Gao, Si Wei, Yu Su, and Guoping Hu. 2017. Question difficulty prediction for reading problems in standard tests. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, and 17 others. 2022. Language Models (Mostly) Know What They Know. (arXiv:2207.05221).

Sanyam Kapoor, Nate Gruver, Manley Roberts, Katherine Collins, Arka Pal, Umang Bhatt, Adrian Weller, Samuel Dooley, Micah Goldblum, and Andrew Gordon Wilson. 2024. Large Language Models Must Be Taught to Know What They Don't Know. (arXiv:2406.08391).

Tanja Käser and Giora Alexandron. 2024. Simulated learners in educational technology: A systematic literature review and a turing-like test. *International Journal of Artificial Intelligence in Education*, 34(2):545–585.

John P Lalor, Pedro Rodriguez, João Sedoc, and Jose Hernandez-Orallo. 2024. Item response theory for natural language processing. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, pages 9–13.

Unggi Lee, Sanghyeok Lee, Junbo Koh, Yeil Jeong, Haewon Jung, Gyuri Byun, Yunseo Lee, Jewoong Moon, Jieun Lim, and Hyeoncheol Kim. 2023. Generative agent for teacher training: Designing educational problem-solving simulations with large language model-based agents for pre-service teachers. In *NeurIPS'23 Workshop on Generative AI for Education (GAIED)*.

Belinda Z. Li, Been Kim, and Zi Wang. 2025a. QuestBench: Can LLMs ask the right question to acquire information in reasoning tasks? (arXiv:2503.22674).

Ming Li, Hong Jiao, Tianyi Zhou, Nan Zhang, Sydney Peters, and Robert W Lissitz. 2025b. Item difficulty modeling using fine-tuned small and large language models. *Educational and Psychological Measurement*, page 00131644251344973.

Ming Li, Nan Zhang, Chenrui Fan, Hong Jiao, Yanbin Fu, Sydney Peters, Qingshu Xu, Robert Lissitz, and Tianyi Zhou. 2025c. Understanding the thinking process of reasoning models: A perspective from schoenfeld's episode theory. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 18278–18299, Suzhou, China. Association for Computational Linguistics.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching Models to Express Their Uncertainty in Words. (arXiv:2205.14334).

Yunting Liu, Shreya Bhandari, and Zachary A Pardos. 2025. Leveraging llm respondents for item evaluation: A psychometric analysis. *British Journal of Educational Technology*, 56(3):1028–1052.

Frederic M Lord. 2012. *Applications of item response theory to practical testing problems*. Routledge.

Anastassia Loukina, Su-Youn Yoon, Jennifer Sakano, Youhua Wei, and Kathy Sheehan. 2016. Textual complexity as a predictor of difficulty of listening items in language proficiency tests. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3245–3253.

Nishanth Madhusudhan, Sathwik Tejaswi Madhusudhan, Vikas Yadav, and Masoud Hashemi. 2025. Do LLMs Know When to NOT Answer? Investigating Abstention Abilities of Large Language Models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9329–9345. Association for Computational Linguistics.

Julia M Markel, Steven G Opferman, James A Landay, and Chris Piech. 2023. Gpteach: Interactive ta training with gpt-based students. In *Proceedings of*

the tenth acm conference on learning@ scale, pages 226–236.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. 2024. HarmBench: A Standardized Evaluation Framework for Automated Red Teaming and Robust Refusal. (arXiv:2402.04249).

Arya D McCarthy, Kevin P Yancey, Geoffrey T LaFlair, Jesse Egbert, Manqian Liao, and Burr Settles. 2021. Jump-starting item parameters for adaptive language tests. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 883–899.

Mihran Miroyan, Rose Niousha, Joseph E Gonzalez, Gireeja Ranade, and Narges Norouzi. 2025. Parastudent: Generating and evaluating realistic student code by teaching llms to struggle. *arXiv preprint arXiv:2507.12674*.

Felix B. Mueller, Rebekka Görge, Anna K. Bernzen, Janna C. Pirk, and Maximilian Poretschkin. 2024. Llms and memorization: On quality and specificity of copyright compliance. In *Proceedings of the Seventh AAAI/ACM Conference on AI, Ethics, and Society (AIES-24) - Full Archival Papers, October 21-23, 2024, San Jose, California, USA - Volume 1*, pages 984–996. AAAI Press.

Andrew Mullooly, Øistein Andersen, Luca Benedetto, Paula Buttery, Andrew Caines, Mark J. F. Gales, Yasin Karatay, Kate Knill, Adian Liusie, Vatsal Raina, and Shiva Taslimipoor. 2023. *The Cambridge Multiple-Choice Questions Reading Dataset*.

Shadi Noroozi and Hossein Karami. 2022. A scrutiny of the relationship between cognitive load and difficulty estimates of language test items. *Language Testing in Asia*, 12(1):13.

OpenAI. 2024a. Gpt-4o mini: advancing cost-efficient intelligence. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/?utm_source=chatgpt.com.

OpenAI. 2024b. Introducing apis for gpt-3.5 turbo and whisper. https://openai.com/index/introducing-chatgpt-and-whisper-apis/.

OpenAI. 2024c. Introducing openai o3 and o4-mini. https://openai.com/index/introducing-o3-and-o4-mini/.

OpenAI. 2025a. Gpt-5 system card. https://openai.com/index/gpt-5-system-card/.

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

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.

Kyle Perkins, Lalit Gupta, and Ravi Tammana. 1995. Predicting item difficulty in a reading comprehension test with an artificial neural network. *Language testing*, 12(1):34–53.

Sydney Peters, Nan Zhang, Hong Jiao, Ming Li, Tianyi Zhou, and Robert Lissitz. 2025. Text-based approaches to item difficulty modeling in large-scale assessments: A systematic review. *arXiv preprint arXiv:2509.23486*.

Qwen-Team. 2024. Qwen2.5: A party of foundation models.

Qwen-Team. 2025a. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Qwen-Team. 2025b. Qwq-32b: Embracing the power of reinforcement learning.

Georg Rasch. 1993. *Probabilistic models for some intelligence and attainment tests.* ERIC.

Ana-Cristina Rogoz and Radu Tudor Ionescu. 2024. Unibucllm: Harnessing llms for automated prediction of item difficulty and response time for multiple-choice questions. *arXiv preprint arXiv:2404.13343*.

Alexis Ross, Megha Srivastava, Jeremiah Blanchard, and Jacob Andreas. 2025. Modeling student learning with 3.8 million program traces. *arXiv preprint arXiv:2510.05056*.

Makoto Sano. 2015. Automated capturing of psycholinguistic features in reading assessment text. In *annual meeting of the National Council on Measurement in Education, Chicago, IL*.

Aviv Slobodkin, Omer Goldman, Avi Caciularu, Ido Dagan, and Shauli Ravfogel. 2023. The Curious Case of Hallucinatory (Un)answerability: Finding Truths in the Hidden States of Over-Confident Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3607–3625. Association for Computational Linguistics.

Katherine Stasaski, Grace Hui Yang, and Marti A Hearst. 2020. More diverse dialogue datasets via diversity-informed data collection. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 4958–4968.

John Sweller. 1988. Cognitive load during problem solving: Effects on learning. *Cognitive science*, 12(2):257–285.

John Sweller. 2011. Cognitive load theory. In *Psychology of learning and motivation*, volume 55, pages 37–76. Elsevier.

Anaïs Tack, Siem Buseyne, Changsheng Chen, Robbe D'hondt, Michiel De Vrindt, Alireza Gharahighehi, Sameh Metwaly, Felipe Kenji Nakano, and Ann-Sophie Noreillie. 2024. Itec at bea 2024 shared task: Predicting difficulty and response time of medical

exam questions with statistical, machine learning, and language models. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 512–521.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. 2023. Just Ask for Calibration: Strategies for Eliciting Calibrated Confidence Scores from Language Models Fine-Tuned with Human Feedback. (arXiv:2305.14975).

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. 2024. Two tales of persona in llms: A survey of role-playing and personalization. *arXiv preprint arXiv:2406.01171*.

Kurt VanLehn, Stellan Ohlsson, and Rod Nason. 1994. Applications of simulated students: An exploration. *Journal of artificial intelligence in education*, 5:135–135.

Roman Vashurin, Ekaterina Fadeeva, Artem Vazhentsev, Lyudmila Rvanova, Daniil Vasilev, Akim Tsvigun, Sergey Petrakov, Rui Xing, Abdelrahman Sadallah, Kirill Grishchenkov, Alexander Panchenko, Timothy Baldwin, Preslav Nakov, Maxim Panov, and Artem Shelmanov. 2024. Benchmarking Uncertainty Quantification Methods for Large Language Models with LM-Polygraph. *Transactions of the Association for Computational Linguistics*, 13:220–248.

Tobias Weber, Michael Ingrisch, Bernd Bischl, and David Rügamer. 2024. Constrained probabilistic mask learning for task-specific undersampled mri reconstruction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7665–7674.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs. (arXiv:2306.13063).

Victoria Yaneva, Kai North, Peter Baldwin, Le An Ha, Saed Rezayi, Yiyun Zhou, Sagnik Ray Choudhury, Polina Harik, and Brian Clauser. 2024. Findings from the first shared task on automated prediction of difficulty and response time for multiple-choice questions. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 470–482.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. 2023. Do Large Language Models Know What They Don't Know? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8653–8665. Association for Computational Linguistics.

Ke Zhang and Ayse Begum Aslan. 2021. Ai technologies for education: Recent research & future directions. *Computers and education: Artificial intelligence*, 2:100025.

Tong Zhang, Peixin Qin, Yang Deng, Chen Huang, Wenqiang Lei, Junhong Liu, Dingnan Jin, Hongru Liang, and Tat-Seng Chua. 2024a. CLAMBER: A Benchmark of Identifying and Clarifying Ambiguous Information Needs in Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10746–10766. Association for Computational Linguistics.

Zhehao Zhang, Ryan A Rossi, Branislav Kveton, Yijia Shao, Diyi Yang, Hamed Zamani, Franck Dernoncourt, Joe Barrow, Tong Yu, Sungchul Kim, and 1 others. 2024b. Personalization of large language models: A survey. *arXiv preprint arXiv:2411.00027*.

Leonidas Zotos, Hedderik van Rijn, and Malvina Nissim. 2024. Are you doubtful? oh, it might be difficult then! exploring the use of model uncertainty for question difficulty estimation. *arXiv preprint arXiv:2412.11831*.

Table of Contents for Appendix

A Related Work **14**

A.1 Item Difficulty Prediction 14

A.2 Student Simulation 14

A.3 LLM Self-Awareness 15

B Consensus of Machines **16**

C Case Study **21**

D Experimental Prompts **22**

A Related Work

A.1 Item Difficulty Prediction

Conventional item difficulty prediction in large-scale assessments typically depends on item response data collected through field testing, where newly created items are embedded in operational forms (but not scored) and then analyzed within classical test theory (CTT) or item response theory (IRT) frameworks (Hsu et al., 2018; Benedetto, 2023). Within these frameworks, CTT operationalizes difficulty as the proportion correct (p-value), whereas IRT models the probability of a correct response as a function of latent ability and item parameters via statistical models such as logit/probit links (DeMars, 2010; Hsu et al., 2018). Despite their accuracy, field-testing and calibration are frequently criticized as time-consuming and costly, because data collection can take several months and IRT calibration may require administering items to several thousand examinees (Hambleton et al., 1991; Hsu et al., 2018; AlKhuzaey et al., 2024). Embedding non-scored pretest items into operational tests also lengthens administrations and raises concerns about test-taker engagement and item exposure, which can compromise test security in high-stakes contexts (Loukina et al., 2016; Hsu et al., 2018; Benedetto, 2023). Expert-based difficulty ratings have been proposed as an alternative, but they are seldom used at scale due to subjectivity and weak alignment with psychometric difficulty, with reported evidence suggesting weak alignment between expert judgments and IRT-based difficulty estimates (Conejo et al., 2014). In response, text-based item difficulty modeling predicts difficulty directly from item text and related metadata using machine learning, thereby avoiding response-data collection and potentially reducing time, cost, and reliance on subjective ratings (Sano, 2015; Loukina et al., 2016; Huang et al., 2017; Hsu et al., 2018). Early text-based work was dominated by feature engineering grounded in linguistic, cognitive, or psychometric theory, exemplified by combining linguistic indicators and item-level metadata, but such approaches can require substantial manual extraction and may generalize poorly across domains (Perkins et al., 1995; Loukina et al., 2016).

The learned representations (Devlin et al., 2019) are typically combined with regressors/classifiers such as CNNs or LSTMs for prediction, with empirical results showing that a CNN yielded superior performance to both a regular CNN variant and a TF-IDF+SVM baseline on TOEFL reading comprehension items (He et al., 2021). More recently, transformer-based small language models have been fine-tuned end-to-end for item difficulty prediction, with BERT first explored in large-scale assessment settings in 2021 and subsequent work showing that fine-tuned BERT and DistilBERT can outperform traditional linguistic/readability features and TF-IDF or Word2Vec-based approaches, while DistilBERT can match BERT performance at lower cost (McCarthy et al., 2021; Benedetto, 2023). In parallel, large language models have been used to directly predict difficulty or to generate auxiliary inputs and features, such as rationales, predicted answers, reasoning steps, uncertainty proxies, or simulated test-taker behaviors, that are then fed into downstream models for difficulty estimation (Rogoz and Ionescu, 2024; Li et al., 2025b; Feng et al., 2025; Zotos et al., 2024; Dueñas et al., 2024). Across this section, the choice of modeling paradigm reflects a recurring trade-off between predictive power and interpretability, and some studies report that hand-crafted linguistic and metadata features can contribute more than BERT embeddings in specific settings (Tack et al., 2024).

A.2 Student Simulation

Student simulation refers to the use of artificial agents to generate learner-like behavior or data (e.g., behavioral traits, performance patterns, and learning progression), and is conceptually distinguished from student modeling systems that primarily infer latent states for adaptation (Chrysafiadi and Virvou, 2013). Early work emphasized three enduring motivations for student simulation: enabling teachers to "practice the art of tutoring," supporting learning-by-teaching with a simulated peer, and allowing formative evaluation of instructional materials without relying entirely on human learner data (VanLehn et al., 1994). Before the LLM era, building such simulations often required extensive hand-crafting of dialogue moves and misconception models in rule-based tutoring systems (Graesser et al., 2005), or large-scale human role-play data collection for tutoring dialogues (Stasaski et al., 2020). With the advent of large language models, student simulation has been reframed as more feasible and scalable, driven by capabilities such as reproducing population-level behavioral distributions (Weber et al., 2024), expressing reasoning/misunderstanding in natural language (Zhang and Aslan, 2021), and supporting

more agent-like behavior via memory and planning (Park et al., 2023). Simulated students have been used across major contexts including data generation, teacher training, learning by teaching/collaboration, and content evaluation, illustrating their broad educational utility (Käser and Alexandron, 2024).

LLM-based student simulation methods vary with the simulation goal, including modeling student traits, performance patterns, or learning progression, and are commonly instantiated via prompting, fine-tuning on learner traces, and agentic designs that support memory and planning (Corbett and Anderson, 1994; Rasch, 1993; Lord, 2012; Embretson and Reise, 2013). Prompt-based role conditioning remains a common baseline, where the simulator is instructed to "act as" a student with specified demographic/background/affective attributes (Markel et al., 2023; Lee et al., 2023), though direct prompting can struggle to reliably instantiate the intended persona characteristics. To better capture authentic learning dynamics (e.g., iterative struggle and revision), some approaches fine-tune LLMs on real student submission trajectories and temporally ordered traces (Miroyan et al., 2025; Ross et al., 2025). In parallel, agentic designs extend simulators beyond static role-play by endowing them with memory/reflection/decision-making components, building on the broader notion of LLM-based generative agents (Park et al., 2023). Cognitive grounding can be strengthened by integrating knowledge tracing or IRT as internal state mechanisms to control and drive simulated behavior over time (e.g., evolving mastery/ability and its interaction with task difficulty), rather than only relying on surface-level role-play (Corbett and Anderson, 1994; Rasch, 1993; Embretson and Reise, 2013). Despite these advances, the prior studies explicitly caution that fluent language output does not guarantee behavioral fidelity and note that many simulated-learner studies lack empirical realism validation, motivating more standardized and context-aware evaluation practices (Käser and Alexandron, 2024). Liu et al. (2025) propose leveraging LLMs as synthetic examinees to estimate item difficulty through IRT calibration, demonstrating that model-generated responses can approximate human-derived item parameters under certain conditions. In contrast, our work does not treat alignment as a purely psychometric estimation problem, but instead interrogates the cognitive validity of such alignment by disentangling difficulty perception, intrinsic capability, and metacognitive awareness.

A.3 LLM Self-Awareness

Numerous benchmarks study model self-knowledge through abstention, but they typically instantiate abstention under a single failure mode, such as unanswerable questions (Yin et al., 2023; Amayuelas et al., 2024), multiple-choice questions with no correct option (Madhusudhan et al., 2025), or underspecified inputs (Slobodkin et al., 2023; Zhang et al., 2024a; Li et al., 2025a). Closely related, verbalized uncertainty elicits a model's explicit expression of doubt and uses it as a downstream signal of whether the model can answer appropriately (Lin et al., 2022; Tian et al., 2023). However, several studies report that verbalized uncertainty can be brittle and may generalize poorly as a practical uncertainty-quantification mechanism (Vashurin et al., 2024; Lin et al., 2022; Xiong et al., 2024). At the same time, other work shows that these signals can be improved: Kapoor et al. (2024) demonstrates that fine-tuning can strengthen verbalized uncertainty, while Kadavath et al. (2022) shows that suitable prompting can elicit explicit correctness probabilities that become increasingly calibrated as model scale increases. Beyond uncertainty elicitation, abstention behavior itself has been improved via fine-tuning (Chen et al., 2024; Brahman et al., 2024) and via explanation generation to justify refusals (Deng et al., 2024), and related benchmarks also evaluate policy compliance and safety-related refusal behavior (Brahman et al., 2024; Mueller et al., 2024; Mazeika et al., 2024). With the recent surge of interest in large reasoning models, MiP-Overthinking further reports that longer "thinking" traces may not improve abstention on unsolvable questions and can even exacerbate overconfident answering (Fan et al., 2025). Complementing these lines of work, we study self-knowledge in the context of educational difficulty estimation: rather than only asking whether a model can abstain, we test whether its explicit difficulty judgments provide an introspective signal about its own likelihood of error on the same items. Our results indicate a pronounced metacognitive gap, difficulty estimates are only weakly predictive of failure cases, highlighting that off-the-shelf LLMs may lack the self-awareness needed to ground difficulty prediction in genuine student struggles.

B Consensus of Machines

Figure 4, 5, 6, 7 show the consensus heatmaps of the spearman correlation between the models on the USMLE, CMCQRD, SAT Math, and SAT Reading datasets. The heatmaps show that the correlation between the models is relatively higher than the correlation between the models and the human.

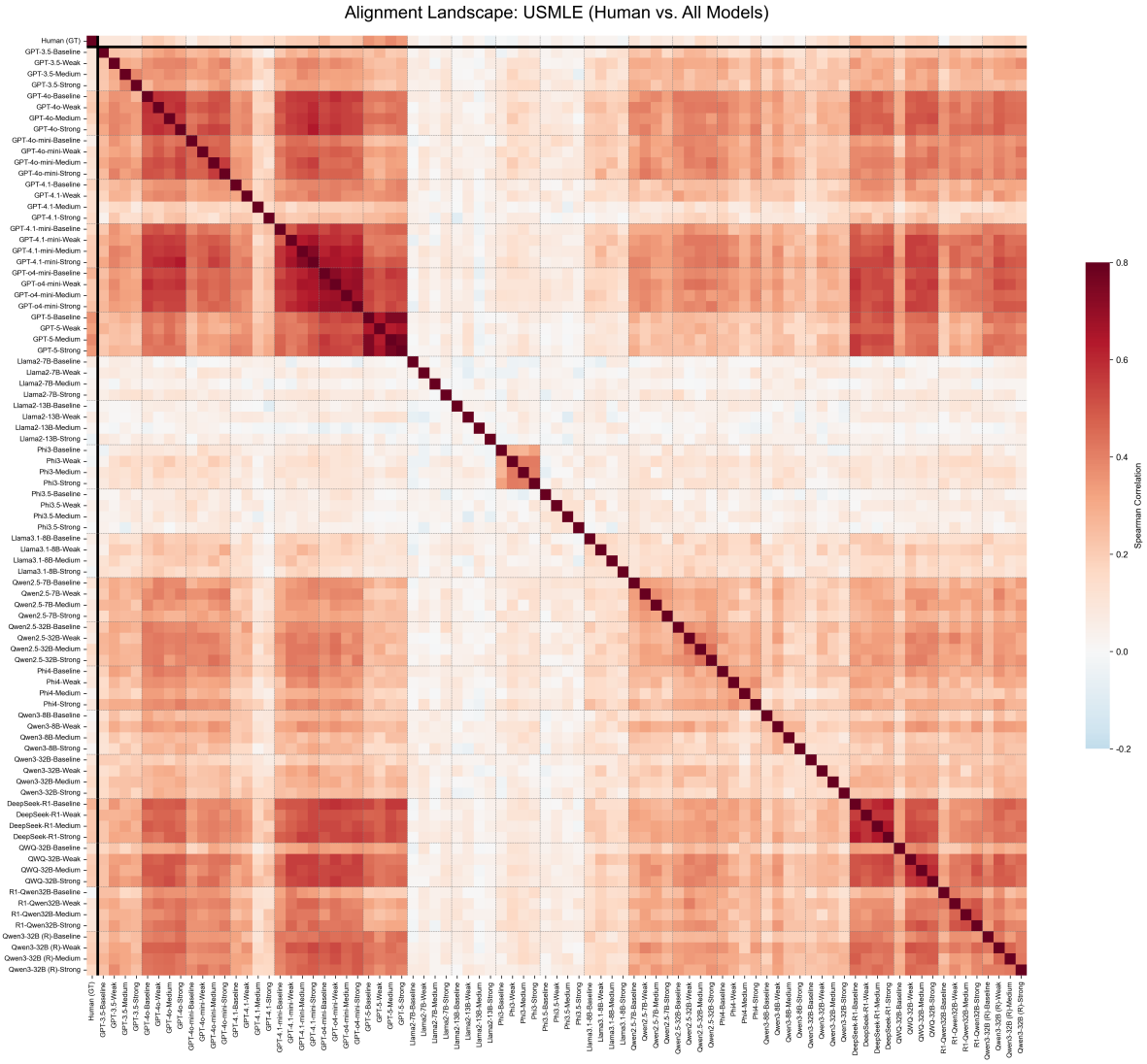

Figure 4: The consensus heatmap of the spearman correlation between the models on the USMLE dataset.

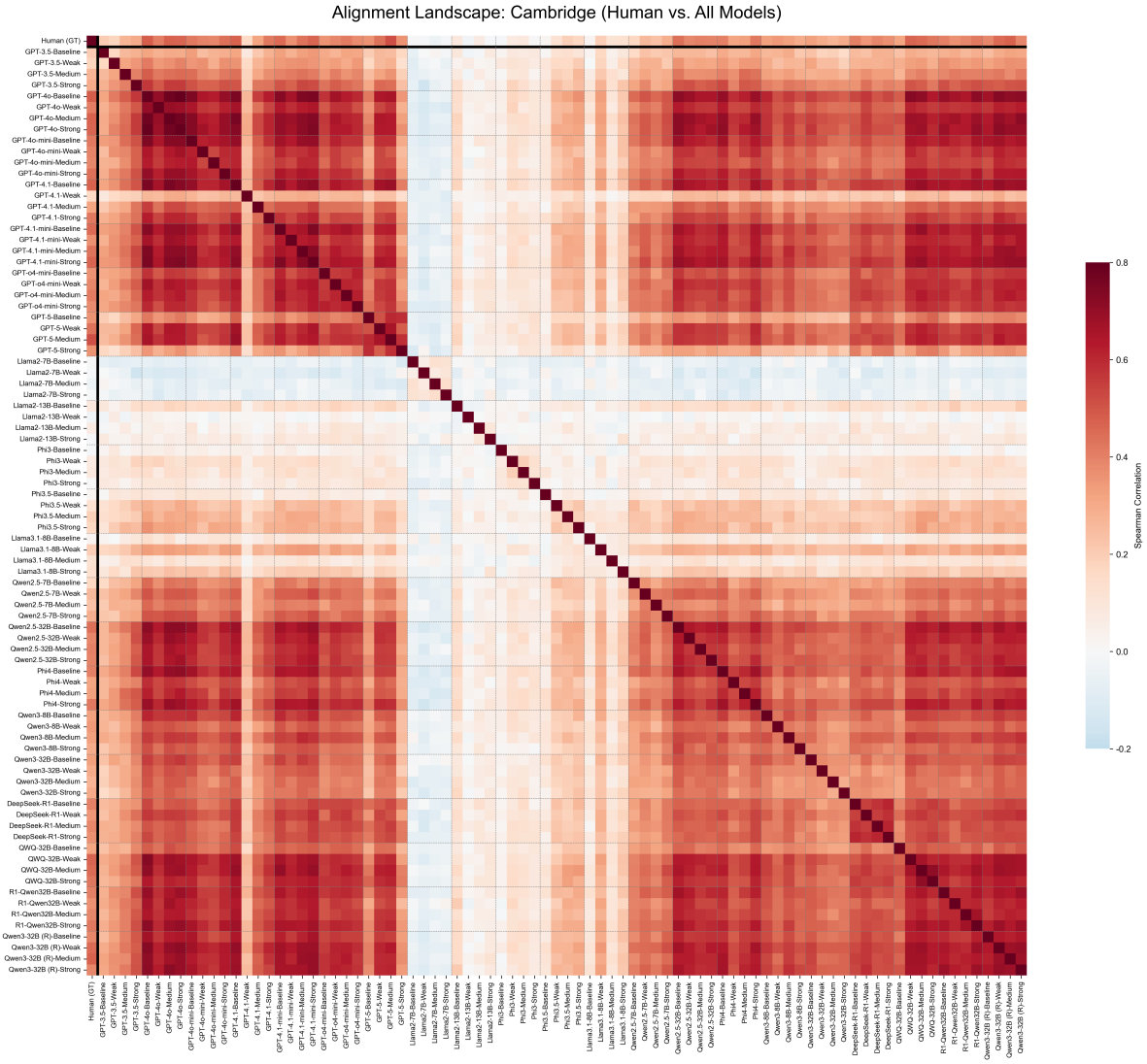

Figure 5: The consensus heatmap of the spearman correlation between the models on the CMCQRD dataset.

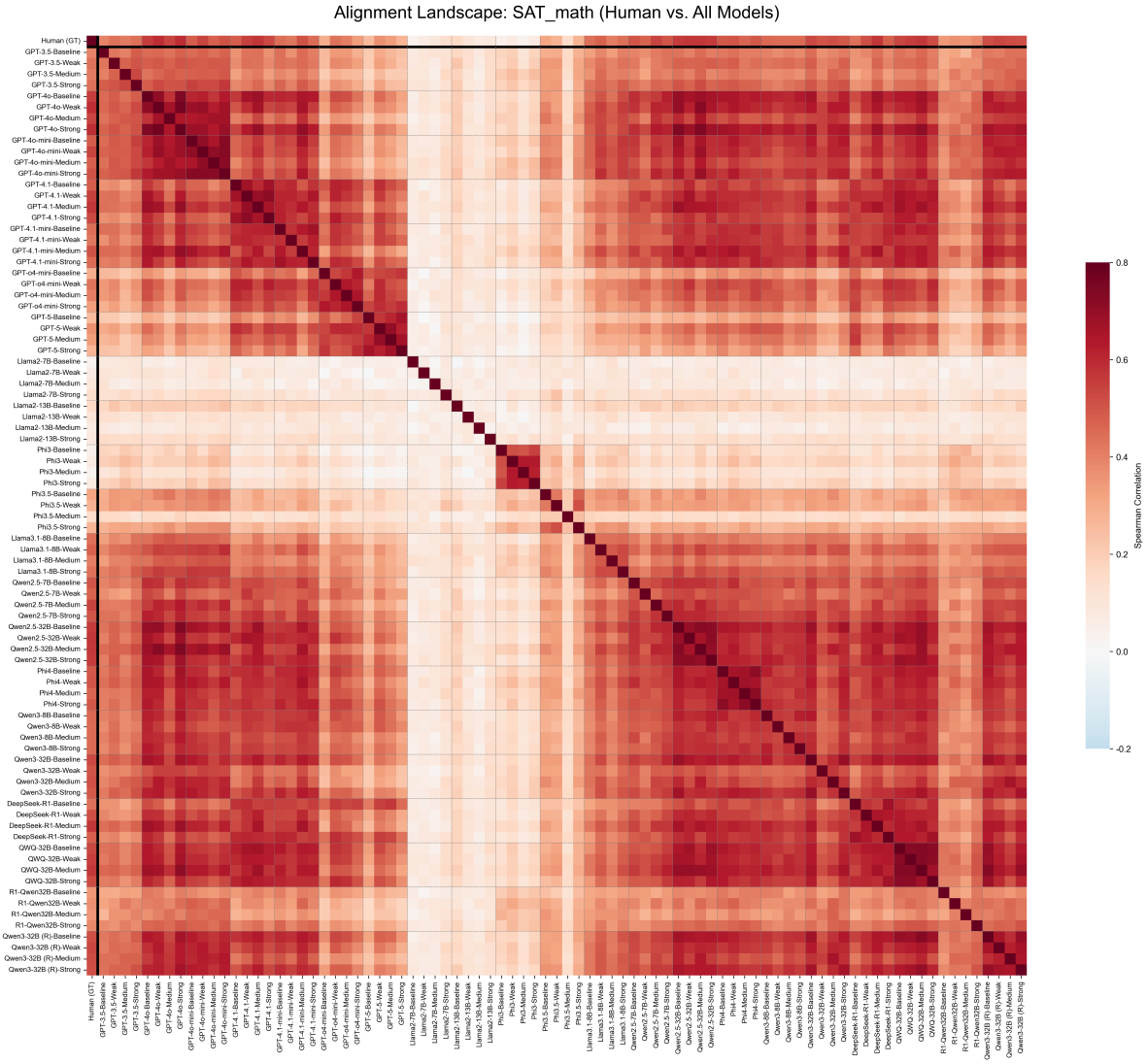

Figure 6: The consensus heatmap of the spearman correlation between the models on the SAT Math dataset.

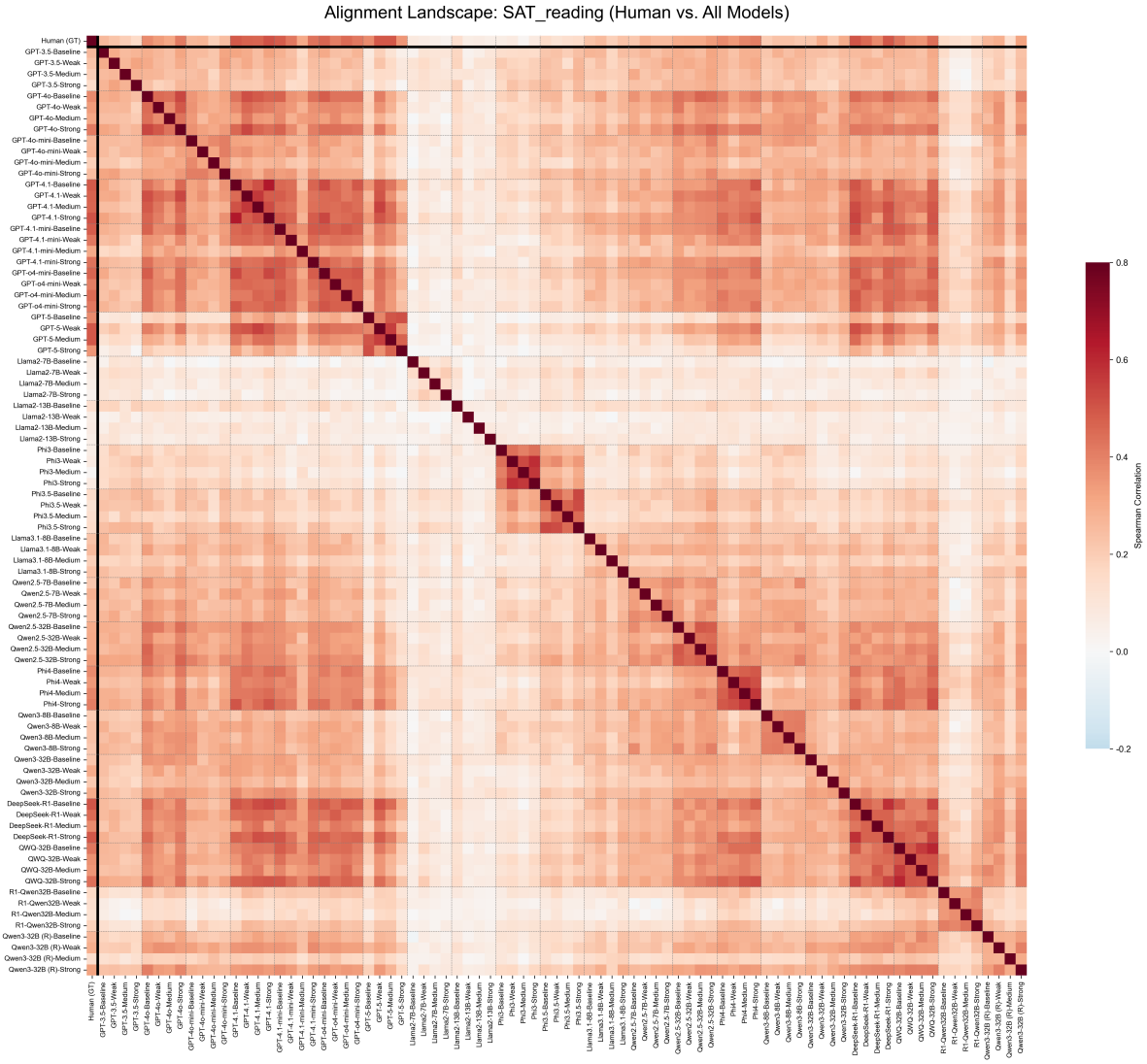

Figure 7: The consensus heatmap of the spearman correlation between the models on the SAT Reading dataset.

C Case Study

Figure 8 shows the violin plot of the difficulty prediction results of GPT-5 with different personas. The violin plot shows that the Low-Proficiency persona effectively expands the distribution, more closely resembling the Ground Truth. However, a systematic underestimation persists, indicating that the model remains optimistic about student performance compared to reality.

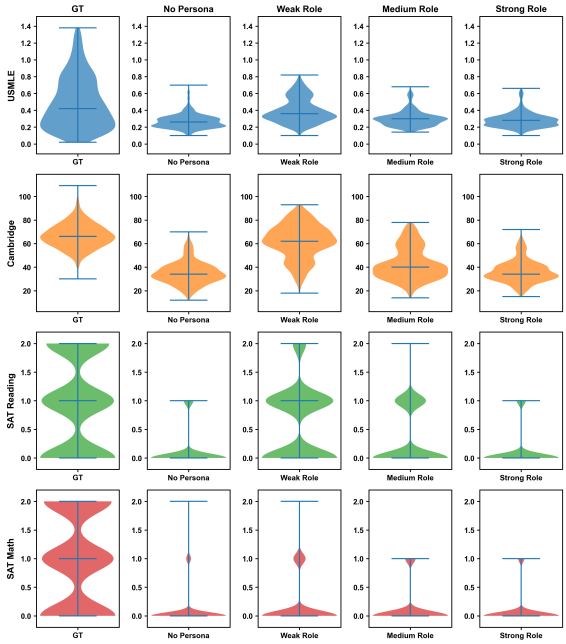

Figure 8: The violin plot of the difficulty prediction results of GPT-5 with different personas. **The Low-Proficiency simulation effectively expands the distribution, more closely resembling the Ground Truth. However, a systematic underestimation persists, indicating that the model remains optimistic about student performance compared to reality.**

D Experimental Prompts

Figure 9, 10, 11 show the prompt templates for difficulty prediction, including Low-Proficiency Student, Medium-Proficiency Student, and High-Proficiency Student across four tasks. Figure 12, 13, 14 show the prompt templates for question answering, including Low-Proficiency Student, Medium-Proficiency Student, and High-Proficiency Student across four tasks.

Prompt for Difficulty Prediction: Low-Proficiency Student

```
---------------------------------- For USMLE ----------------------------------

(System Prompt)
Suppose you are a student taking the USMLE exam. You are a weak student with low-level
 medical proficiency.

(User Prompt)
Analyze the difficulty values of the question. The difficulty values range from 0 to
1.0, where 0 is the easiest and 1.0 is the hardest. Analyze the difficulty step by step,
and provide the final value in \boxed{...}:
[Item Context]

------------------------------- For Cambridge -------------------------------

(System Prompt)
Suppose you are a student taking the Cambridge English Test. You are a weak student
with low-level English proficiency.

(User Prompt)
Analyze the difficulty values of the question. The difficulty values range from 1 to
100, where 1 is the easiest and 100 is the hardest. Analyze the difficulty step by step,
and provide the final value in \boxed{...}:
[Item Context]

----------------------------- For SAT Reading -----------------------------

(System Prompt)
Suppose you are a student taking the SAT Reading exam. You are a weak student with low
-level English proficiency.

(User Prompt)
Analyze the difficulty levels of the question. The difficulty levels contain 3
categories: easy, medium, and hard. Analyze the difficulty step by step, and provide the
final category in \boxed{...}:
[Item Context]

------------------------------- For SAT Math -------------------------------

(System Prompt)
Suppose you are a student taking the SAT math exam. You are a weak student with low-
level math proficiency.

(User Prompt)
Analyze the difficulty levels of the question. The difficulty levels contain 3
categories: easy, medium, and hard. Analyze the difficulty step by step, and provide the
final category in \boxed{...}:
[Item Context]
```

Figure 9: Prompt templates for difficulty prediction (Low-Proficiency Student) across four tasks.

Prompt for Difficulty Prediction: Medium-Proficiency Student

```
---------------------------------- For USMLE ----------------------------------

(System Prompt)
Suppose you are a student taking the USMLE exam. You are an average student with
medium-level medical proficiency.

(User Prompt)
Analyze the difficulty values of the question. The difficulty values range from 0 to
1.0, where 0 is the easiest and 1.0 is the hardest. Analyze the difficulty step by step,
and provide the final value in \boxed{...}:
[Item Context]

------------------------------- For Cambridge ------------------------------

(System Prompt)
Suppose you are a student taking the Cambridge English Test. You are an average
student with medium-level English proficiency.

(User Prompt)
Analyze the difficulty values of the question. The difficulty values range from 1 to
100, where 1 is the easiest and 100 is the hardest. Analyze the difficulty step by step,
and provide the final value in \boxed{...}:
[Item Context]

----------------------------- For SAT Reading ------------------------------

(System Prompt)
Suppose you are a student taking the SAT Reading exam. You are an average student with
 medium-level English proficiency.

(User Prompt)
Analyze the difficulty levels of the question. The difficulty levels contain 3
categories: easy, medium, and hard. Analyze the difficulty step by step, and provide the
final category in \boxed{...}:
[Item Context]

------------------------------- For SAT Math -------------------------------

(System Prompt)
Suppose you are a student taking the SAT math exam. You are an average student with
medium-level math proficiency.

(User Prompt)
Analyze the difficulty levels of the question. The difficulty levels contain 3
categories: easy, medium, and hard. Analyze the difficulty step by step, and provide the
final category in \boxed{...}:
[Item Context]
```

Figure 10: Prompt templates for difficulty prediction (Medium-Proficiency Student) across four tasks.

Prompt for Difficulty Prediction: High-Proficiency Student

```
        ---------------------------------- For USMLE ----------------------------------

        (System Prompt)
        Suppose you are a student taking the USMLE exam. You are a good student with high-
level medical proficiency.

        (User Prompt)
        Analyze the difficulty values of the question. The difficulty values range from 0
to 1.0, where 0 is the easiest and 1.0 is the hardest. Analyze the difficulty step by
step, and provide the final value in \boxed{...}:
        [Item Context]

        ------------------------------- For Cambridge ------------------------------

        (System Prompt)
        Suppose you are a student taking the Cambridge English Test. You are a good
student with high-level English proficiency.

        (User Prompt)
        Analyze the difficulty values of the question. The difficulty values range from 1
to 100, where 1 is the easiest and 100 is the hardest. Analyze the difficulty step by
step, and provide the final value in \boxed{...}:
        [Item Context]

        ----------------------------- For SAT Reading ------------------------------

        (System Prompt)
        Suppose you are a student taking the SAT Reading exam. You are a good student with
 high-level English proficiency.

        (User Prompt)
        Analyze the difficulty levels of the question. The difficulty levels contain 3
categories: easy, medium, and hard. Analyze the difficulty step by step, and provide the
final category in \boxed{...}:
        [Item Context]

        ------------------------------- For SAT Math -------------------------------

        (System Prompt)
        Suppose you are a student taking the SAT math exam. You are a good student with
high-level math proficiency.

        (User Prompt)
        Analyze the difficulty levels of the question. The difficulty levels contain 3
categories: easy, medium, and hard. Analyze the difficulty step by step, and provide the
final category in \boxed{...}:
        [Item Context]
```

Figure 11: Prompt templates for difficulty prediction (High-Proficiency Student) across four tasks.

Prompt for Question Answering: Low-Proficiency Student

```
---------------------------------- For USMLE ---------------------------------

(System Prompt)
Suppose you are a student taking the USMLE exam. You are a weak student with low-level
 medical proficiency.

(User Prompt)
Answer the question below step by step, and provide the final answer in \boxed{...}:
[Item Context]

------------------------------- For Cambridge ------------------------------

(System Prompt)
Suppose you are a student taking the Cambridge English Test. You are a weak student
with low-level English proficiency.

(User Prompt)
Answer the question below step by step, and provide the final answer in \boxed{...}:
[Item Context]

----------------------------- For SAT Reading ------------------------------

(System Prompt)
Suppose you are a student taking the SAT Reading exam. You are a weak student with low
-level English proficiency.

(User Prompt)
Answer the question below step by step, and provide the final answer in \boxed{...}:
[Item Context]

------------------------------- For SAT Math -------------------------------

(System Prompt)
Suppose you are a student taking the SAT math exam. You are a weak student with low-
level math proficiency.

(User Prompt)
Answer the question below step by step, and provide the final answer in \boxed{...}:
[Item Context]
```

Figure 12: Prompt templates for question answering (Low-Proficiency Student) across four tasks.

Prompt for Question Answering: Medium-Proficiency Student

```
---------------------------------- For USMLE ---------------------------------

(System Prompt)
Suppose you are a student taking the USMLE exam. You are an average student with medium-
 level medical proficiency.

(User Prompt)
Answer the question below step by step, and provide the final answer in \boxed{...}:
[Item Context]

------------------------------- For Cambridge ------------------------------

(System Prompt)
Suppose you are a student taking the Cambridge English Test. You are an average student
 with medium-level English proficiency.

(User Prompt)
Answer the question below step by step, and provide the final answer in \boxed{...}:
[Item Context]

------------------------------- For SAT Reading ------------------------------

(System Prompt)
Suppose you are a student taking the SAT Reading exam. You are an average student with
 medium-level English proficiency.

(User Prompt)
Answer the question below step by step, and provide the final answer in \boxed{...}:
[Item Context]

------------------------------- For SAT Math ---------------------------------

(System Prompt)
Suppose you are a student taking the SAT math exam. You are an average student with medium
 -level math proficiency.

(User Prompt)
Answer the question below step by step, and provide the final answer in \boxed{...}:
[Item Context]
```

Figure 13: Prompt templates for question answering (Medium-Proficiency Student) across four tasks.

Prompt for Question Answering: High-Proficiency Student

```
---------------------------------- For USMLE ----------------------------------

(System Prompt)
Suppose you are a student taking the USMLE exam. You are a good student with high-
level medical proficiency.

(User Prompt)
Answer the question below step by step, and provide the final answer in \boxed{...}:
[Item Context]

------------------------------- For Cambridge -------------------------------

(System Prompt)
Suppose you are a student taking the Cambridge English Test. You are a good student
with high-level English proficiency.

(User Prompt)
Answer the question below step by step, and provide the final answer in \boxed{...}:
[Item Context]

----------------------------- For SAT Reading -----------------------------

(System Prompt)
Suppose you are a student taking the SAT Reading exam. You are a good student with
high-level English proficiency.

(User Prompt)
Answer the question below step by step, and provide the final answer in \boxed{...}:
[Item Context]

------------------------------- For SAT Math -------------------------------

(System Prompt)
Suppose you are a student taking the SAT math exam. You are a good student with high-
level math proficiency.

(User Prompt)
Answer the question below step by step, and provide the final answer in \boxed{...}:
[Item Context]
```

Figure 14: Prompt templates for question answering (High-Proficiency Student) across four tasks.