

STATISTICAL ASSESSMENT FOR
LLM-AS-A-JUDGE: LEVERAGING ITEM
RESPONSE THEORY FOR META-EVALUATION
OF LARGE LANGUAGE MODELS

TARA SHUKLA

ADVISOR: PROFESSOR LYDIA LIU

SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
BACHELOR OF ARTS
DEPARTMENT OF COMPUTER SCIENCE
PRINCETON UNIVERSITY

APRIL 2025

I hereby declare that I am the sole author of this thesis.

I authorize Princeton University to lend this thesis to other institutions or individuals for the purpose of scholarly research.

A handwritten signature in black ink, appearing to read "Tara Shukla".

Tara Shukla

I further authorize Princeton University to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

A handwritten signature in black ink, appearing to read "Tara Shukla".

Tara Shukla

Abstract

My research applies Item Response Theory (IRT) to systematically evaluate Large Language Models' (LLMs) capabilities in judgment tasks. By modeling both model abilities and task characteristics simultaneously, this approach provides an interpretable framework that takes LLM evaluation beyond traditional accuracy metrics. Analysis of multiple benchmark datasets—including MT-Bench, JudgeBench, and LLMBAR—reveals weak correlations between benchmark performance, challenging the unidimensionality assumption of judgment capability. Additionally, I evaluate models on essay grading tasks using both pairwise and pointwise judgment methodologies. Through this, I demonstrate that comparative judgment and single-answer evaluation represent distinct skills rather than manifestations of a single trait. My findings indicate that while model size generally correlates with judgment capability, this relationship varies substantially across evaluation contexts. My research contributes methodological innovations for benchmark characterization and usage. These findings have important implications for responsible AI deployment and the development of more targeted evaluation methodologies for LLM-based assessment systems.

Acknowledgements

Thank you to my advisor, Prof. Lydia Liu, for her invaluable support and guidance throughout my project. You have helped me explore my academic passions and grow as a researcher and student, and I'm forever grateful.

Thank you to Prof. Brian Kernighan, for inciting my spark for computer science, and encouraging me to pursue what makes me excited about the field —you made more of an impact than you know.

Thank you to Prof. Susie Kim for your continued mentorship, guidance, and kindness to me. Your belief in me and excitement for your work has inspired me to pursue my passions and believe in myself.

Thank you to my friends —your unwavering support, understanding, and love has been the highlight of my time at Princeton.

Above all, thank you to my parents and my sister; I owe all my achievements to my family, and I love and appreciate you.

To my family

Contents

Abstract	iii
Acknowledgements	iv
List of Tables	ix
List of Figures	x
1 Introduction	1
2 Background and Related Work	5
2.1 LLM-as-a-Judge	5
2.2 Benchmarking	7
2.3 Item Response Theory	8
2.4 Essay Grading – a test case	10
3 Approach	12
3.1 Benchmark Task Assessment	12
3.2 Essay Grading Tasks	14
3.3 IRT Modeling and Analysis	15
3.4 Considerations and Challenges	16
4 Implementation	19
4.1 Essay grading data	19
4.2 Model Selection	20

4.3	Benchmark Assessment	20
4.4	Pairwise Judgment Implementation	22
4.5	Pointwise Scoring Implementation	22
4.6	IRT Implementation	23
4.7	Analysis Approach	24
4.7.1	Implementation Challenges	25
5	Results and Analysis	26
5.1	IRT Efficacy in Performance Estimation	26
5.1.1	Pairwise Judgment	26
5.1.2	Pointwise Judgment	29
5.2	Patterns in Item Difficulty and Discrimination	31
5.2.1	Pairwise Judgment	31
5.2.2	Pointwise Judgment	34
5.3	IRT and Benchmarking	35
5.3.1	Benchmark Validation	36
5.3.2	Benchmark Discrimination Analysis	39
6	Discussion	43
6.1	Model Size and Ability	43
6.2	Analysis of Item Characteristics	46
6.3	Benchmark Characterization Through IRT	47
6.3.1	Limits of Unidimensionality Assumption	47
6.3.2	Benchmark Composition Effects	48
6.3.3	Benchmarking LLM-as-a-Judge	48
6.3.4	Methodological Implications for Benchmarking Paradigm	49
7	Conclusions and Future Work	53
7.1	IRT for LLM-as-a-Judge Evaluation	53

7.2	Practical Applications and Broader Impact	54
7.3	Future Research Directions	55
A	Prompt Format Examples	56
A.1	Pairwise Judgment Prompt	56
A.2	Pointwise Judgment Prompt	57

List of Tables

4.1	Model sizes and families with parameters in M (millions) and B (billions)	21
5.1	Model fit statistics across different benchmarks	27
5.2	Difference in text lengths (characters) per item characteristic across different benchmarks	33
5.3	Length (character count) for Extreme Items (GRM)	34
6.1	Difficult and easy items from MT-Bench	51
6.2	High and low discrimination items, from MT-Bench	52

List of Figures

5.1	Model Accuracy Across Benchmarks	28
5.2	Model θ Across Benchmarks	28
5.3	Correlation between Accuracy and θ for Pairwise Judgment	29
5.4	Model Output Distribution vs Ground Truth density graph	30
5.5	JudgeBench item difficulty and discrimination, grouped by question source. Top: Mean difficulty plot. Bottom: Mean discrimination plot.	32
5.6	Item Characteristic Curves for Least and Most Discriminating Items (GRM)	34
5.7	Spearman Heatmap of Accuracies	36
5.8	Spearman Heatmap of Normalized Thetas	36
5.9	Spearman Heatmap of Accuracies: LLMBAR	37
5.10	Spearman Heatmap of Accuracies: JudgeBench	37
5.11	Correlation Matrix between Thetas and Benchmarks	39
5.12	Distribution of Item Difficulty and Discrimination for each Benchmark	40
5.13	Wright Maps for each Benchmarks	41
6.1	Parameter Analysis of Benchmark	44
6.2	Rank consistency across benchmarks for different model families . . .	45

Chapter 1

Introduction

Automated text evaluation is a natural language processing (NLP) task with significant implications across multiple domains. The problem of correct, consistent judgment of textual input along predefined metrics is one that eludes even human decision-makers. The complexity of this task is only matched by the scale of its application; text evaluation is an essential undertaking in disciplines from education, to law, and more. The potential to automate or semi-automate this task has significant implications for standardization and efficiency in decision-making for many applications. In recent years, the emergence of increasingly capable large language models (LLMs) has opened new possibilities for their deployment to replace or supplement human judgment in a wide range of tasks; this subfield is often referred to as the "LLM-as-a-judge" task, or LLM meta-evaluation. The LLM-as-a-Judge paradigm leverages LLMs for scalable and automated evaluation of text, actions, or decisions. Traditional text evaluation metrics are statistical or rule-based, such as ROUGE and BLEU scores or readability indexes. In contrast, LLM evaluation is not limited to predefined quality metrics such as n-gram overlap, and thus provide increased capability and flexibility in judgment. Thus, its primary appeal lies in the ability to combine the scalability of statistical evaluation, with the context-sensitive nuance of

human decision-making in open-ended and complex tasks [10].

Stakeholders across multiple domains have become interested in the uses of LLM-as-a-judge for text evaluation. For instance, recruiters have leveraged LLM agents in resume screening [5] and social media content moderation [7]. In educational contexts, stakeholders have become interested in the uses of LLMs in writing evaluation, e.g. for the automated assessment of student writing. In high-stakes applications, researchers have looked into LLMs for extracting information from CPS case reports [19] and recently governmental agencies have leveraged LLMs for job cut decision-making [1]. All of these applications have potential to alleviate the workload of human evaluators while also providing consistent and scalable assessment. However, the serious implications of these judgments means that LLM decision-makers must consistently conform to standards of ethical and correct assessments.

This is not the case; despite its promise, LLM-as-a-Judge faces significant challenges that limit its reliability and effectiveness. As will be discussed in later sections, positional and length-based biases, hallucinations, and calibration all impact the fairness and consistency of model judgment. These problems necessitate robust model evaluation frameworks prior to deployment. Despite this necessity, measuring the quality of LLM judgments in comparison to human assessments is a complex problem. Although LLMs show promise in evaluating text, we currently lack rigorous frameworks for measuring and comparing their evaluation capabilities. Current metrics of success for LLM judges often rely on simple accuracy metrics that fail to capture the nuanced nature of text assessment.

To fill this research gap, I propose applying a psychometric statistical evaluation framework, Item Response Theory (IRT), to systematically assess LLM performance across several LLM-as-a-judge tasks. The latent traits that we can extract from models via IRT are useful in creating and assessing interpretable benchmarks that can help determine whether these models are ready for deployment in high-stakes

settings.

In this paper, I evaluate multiple LLMs across different architectures and sizes across several benchmarks to establish baseline capabilities in LLM-as-a-Judge related tasks. I also conduct assessment of LLM judgment in two essay grading tasks: a pairwise comparative analysis task and a pointwise numerical grading task. For the first task, I conduct comparative essay grading using a dataset of 100 paired persuasive essays from English Language Learners, where models are asked to judge the relative overall quality between pairs of essays according to a multi-dimensional rubric. For the second task, I pulled 300 essays from the same dataset, and extracted models' zero-shot numerical grading based on the same rubric. In both cases, I compare model judgments against the essays' original human raters. Additionally, I apply IRT analysis not only to the essay grading tasks but also to all benchmark data, allowing me to estimate both difficulty and discrimination parameters across different item types, and the latent "judgment capability" of each model in different evaluation contexts. My approach provides a framework for quantifying model performance across diverse judgment tasks. My results reveal patterns in how model architectures affect judgment capabilities and establish interpretable metrics for LLM performance.

My thesis addresses several key research questions:

- How effectively can IRT models quantify and differentiate LLM judgment capabilities compared to simple accuracy metrics?
- What patterns emerge in item difficulty and discrimination parameters across different benchmarks?
- How can leveraging IRT in benchmarking help us better understand model capabilities and benchmark quality?

My research makes both theoretical and practical contributions. Theoretically, it demonstrates how psychometric frameworks can be adapted to evaluate AI systems on

complex cognitive tasks. The approach may generalize to evaluating LLMs on other tasks requiring human-like judgment. Practically, it provides a rigorous methodology for improving the interpretability and quality of benchmarks, which has implications for determining LLM deployment-readiness. As multiple domains increasingly explore AI-assisted decision-making, having reliable benchmarks becomes crucial. My IRT-based framework offers interpretable measures of capability that can inform policy decisions and guide future model development. The work also contributes to the broader discussion of AI evaluation and deployment. By developing methods for assessing AI performance on complex judgment tasks, I contribute towards a body of scholarly work on responsible technology adoption.

Chapter 2

Background and Related Work

My research deals with the intersection between the LLM-as-a-Judge task, model benchmarking, and Item Response Theory. In this section, I will provide a detailed overview of each of these topics and review the research landscape thus far, with specific focus on how these subjects relate to this project’s application of IRT for LLM-as-a-Judge benchmarking.

2.1 LLM-as-a-Judge

Recent research on the LLM-as-a-judge paradigm has revealed promising results for LLM judgement capabilities, but also shed light on the gaps in LLM evaluation and deployment. LLM-as-a-judge encompasses several distinct types, including point-wise judgment (evaluating a single candidate), pair-wise comparison (comparing two options), and list-wise evaluation (ranking multiple candidates). Implementation methods for LLM judges vary from prompting models to generate numerical scores or judgments, to detailed comparative assessments between multiple options [6].

In [21], Thakur et al. conducted comprehensive investigations into thirteen judge models across varying sizes and families, finding that only the largest LLMs could achieve an alignment with human judgements. The authors identify several vulner-

abilities in current judge models, including prompt length sensitivity and leniency biases.

Their work demonstrates that even high-performing models like Llama-3-70B and GPT-4 still exhibit significant disparities from human judgment. However, the authors also emphasized the inadequacy of percent agreement as an evaluation metric, presenting evidence that Scott’s Pi provides more discriminative power between models of varying quality. This aligns with previous findings from [28], where Zheng et al. (2024) similarly highlighted alignment challenges between LLM judges and humans. In [6], Gu, et.al provide a comprehensive survey of LLM-as-a-judge, including bias, reliability, and challenging evaluation.

Researchers have found that LLMs suffer from positional bias, in which comparative evaluations are influenced by the order of the options presented, thus compromising fairness. Length bias is also a major factor in LLM judgments; models have been proven to favor longer responses regardless of their quality, meaning that unnecessary verbosity can be unfairly rewarded [27]. Additionally, many judge models exhibit overconfidence and struggle with generalization across domains, making them susceptible to adversarial attacks where carefully crafted inputs can manipulate evaluation outcomes. Also, the issue of self-enhancement bias —where models favor outputs generated by themselves or similar models —raises concerns about objectivity [23], [25]. In short, collectively, there is a growing body of research which acknowledges the ability of the LLM-as-judge paradigm while also questioning its effective and ethical implementation. These LLM-as-a-Judge limitations underscore the need for robust calibration and benchmarking for ensuring fair evaluation.

2.2 Benchmarking

Despite this demonstrated need for effective LLM benchmarking methods, LLM evaluation for any task is a multifaceted challenge. While LLMs might achieve high scores on standardized benchmarks, these results might not consistently or accurately reflect their underlying reasoning or understanding skills. In [3], Banerjee, et.al explains that the reliability of traditional, static benchmarks is increasingly questioned due to their inherent vulnerabilities. In pursuit of climbing leaderboards, developers exploit weaknesses in standardized metrics through benchmark ‘hacking’ strategies such as selective reporting, benchmark overfitting, or data contamination. The latter is a significant concern; LLMs are trained on their evaluation sets, leading to inflated performance metrics not reflective of true generalization abilities. This is further compounded by the opaque nature of LLM training data and processes.

In [14], Perlitz et.al discuss the lack of standardized methodology for assessing how well different LLM benchmarks correlate with each other, which can lead to unreliable conclusions about benchmark validity; they suggest best practices for curating and evaluating benchmarks. Notably, the paper’s results prove that degrees of agreement between benchmark are dependent on the number and type of models considered.

Overall, there is a growing concern that current evaluation methodologies overestimate model performance and overlook critical limitations in real-world applications. In response to these limitations and risks in static benchmarking, the field of LLM evaluation has also shifted towards dynamic benchmarking and other paradigms, such as smaller, task-specific benchmarks. Additionally, there is increasing recognition that evaluation must extend beyond surface-level accuracy. Metrics such as calibration, consistency, robustness to adversarial inputs, and faithfulness to source materials are gaining traction as complementary dimensions of evaluation. As will be discussed in the next section, latent trait analysis is another strategy that can help improve benchmarking.

2.3 Item Response Theory

Item Response Theory offers a statistical framework for evaluating LLM grading capabilities via extracting ability estimates from analysis of response patterns. While IRT was initially developed for psychometric testing, it models the relationship between a test-taker's latent ability and their performance on test items. In this project, I treat LLMs as test-takers and essay grading tasks as test items, allowing us to estimate both the models' underlying grading capability and the difficulty and discrimination of specific essays. Item difficulty is defined as the point on the ability scale where a subject would have a 50% chance of a correct answer; item discrimination is a measure of how well an item can be used to distinguish between subjects of different ability levels. There are several different types of IRT models relevant to this evaluation, defined as follows:

For a one-parameter (Rasch) logistic model (1PL), we assume that all items have the same discrimination parameter. The model is defined:

$$P(X_{ij} = 1 | \theta_i, b_j) = \frac{1}{1 + e^{-(\theta_i - b_j)}} \quad (2.1)$$

For a two-parameter logistic model (2PL),

$$P(X_{ij} = 1 | \theta_i, a_j, b_j) = \frac{1}{1 + e^{-a_j(\theta_i - b_j)}} \quad (2.2)$$

- $P(X_{ij} = 1 | \theta_i, a_j, b_j)$ is the probability that examinee i with ability θ_i answers item j correctly
- θ_i is the latent trait (ability) of examinee i

- a_j is the discrimination parameter for item j
- b_j is the difficulty parameter for item j

Finally, the Graded Response Model (GRM) is an extension of IRT which handles ordered response categories. For an item j with m_j ordered response categories ($0, 1, 2, \dots, m_j - 1$), the GRM uses the following formula:

$$P^{*jk}(\theta_i) = P(X_{ij} \geq k | \theta_i, a_j, b_{jk}) = \frac{e^{a_j(\theta_i - b_{jk})}}{1 + e^{a_j(\theta_i - b_{jk})}} \quad (2.3)$$

With the constraints that $P_{j1}^*(\theta_i) = 1$ (probability of scoring at or above the lowest category), and $P_{jm_j}^*(\theta_i) = 0$ (probability of scoring above the highest category), and where:

- $P_{jk}^*(\theta_i)$ is the probability that examinee i with ability θ_i scores at or above category k on item j
- b_{jk} is the threshold parameter for category k of item j
- m_j is the number of response categories for item j

As will be discussed further in this paper, through treating LLMs as ‘students’ and essays as ‘test items’, I employ the models discussed above, to extract ability estimates from LLM judges. This formulation allows us to estimate each LLM’s underlying grading capability θ , characterize which types of questions are most challenging, and measure how effectively different benchmark items discriminate between high and low-performing models. The relative rankings of our θ estimates is useful to us in the evaluation and comparison of LLM capabilities in judging text outputs.

Additionally, scholars have studied IRT as a method for evaluating LLMs in the past. In [18], Sedoc, et. al introduce IRT as a method for streamlining human

evaluation of chatbots, using a paired comparison setup to estimate model abilities and prompt discriminative power. My work parallels this through applying IRT to evaluate LLMs’ judgment performance, rather than their conversational abilities. Other work on IRT in LLM evaluation includes [2], in which the authors use IRT to identify biases in human judges by treating the quality of annotated items as the latent trait. In [24], Yancey, et.al develop an explanatory IRT model that combines BERT embeddings with IRT to predict item parameters in language assessments with less observed data. Also, Vania et.al in [22] uses the 3PL IRT model to evaluate NLU datasets’ effectiveness in differentiating between language models. Finally, in [15], the paper proposes methods to efficiently evaluate LLMs by using IRT to select representative examples from existing benchmarks This is similar to the approach taken by Zhuang et.al in [29], in which the authors advocate for a shift from static benchmarking to adaptive testing in AI model evaluation, focusing on the potential role of IRT. They introduce a 2-phase framework that first annotates items with psychometric characteristics and then uses these to drive an interactive evaluation process; my work complements this approach by utilizing IRT latent trait estimation for relative model ability ranking.

2.4 Essay Grading – a test case

While the aforementioned challenges in benchmarking and evaluation affect all domains where LLMs serve as judges, educational contexts—particularly essay grading—present an especially compelling test case for three key reasons. First, automated essay scoring (AES) represents a complex cognitive task that requires understanding multiple dimensions of text quality, such as grammar, argumentation, structure, and more. Second, human graders already establish a comparative baseline with typical inter-rater reliability coefficients between 0.6-0.8. Finally, the high-stakes nature of

educational assessment makes it an ideal domain for developing and testing rigorous evaluation frameworks. Thus, the complexity of AES provides an ideal context for applying IRT, which can simultaneously model latent traits and assessment quality.

The limitations of existing evaluation methods are particularly concerning given the high-stakes nature of educational assessment. Additionally, as Liu, et.al point out in [11], deploying AI systems without proper evaluation frameworks risks amplifying existing educational inequities. In this context, poor grading practices can have significant consequences for students' academic trajectories and life outcomes. Additionally, as noted by Kumar and Boulanger in [8], the "black box" nature of many deep learning systems makes it crucial to develop transparent and interpretable measures of their capabilities.

The scholarly body of work related to LLMs for automated essay scoring reveals both optimistic results and performance gaps. In [12], Mansour et al. evaluated ChatGPT and Llama 2 on the ASAP dataset, and found that the LLMs achieved Quadratic Weighted Kappa (QWK) scores well below non-LLM-based state-of-the-art AES systems. Mansour's analysis revealed task-dependent performance variations, with both models showing stronger results on source-dependent responses than on persuasive or narrative essays. This performance disparity echoes findings from earlier studies by Mizumoto and Eguchi (2023), who similarly found that even advanced language models underperform compared to specialized AES algorithms [13]. However, other studies have experienced successfully with techniques to improve LLM-as-a-judge performance; for instance, in [9], Lee, et. al leverage Multi Trait Specialization (MTS), a zero-shot prompting framework they employ to enhance LLM task following for AES. Overall, the emerging research direction of LLMs for AES indicates both promise and challenges; continued research into robust evaluation of these systems is essential.

Chapter 3

Approach

My approach to evaluating LLM judgment capabilities leverages IRT to investigate model performance across different benchmarks and evaluation contexts, and simultaneously draw insight about the efficacy in current LLM-as-a-Judge benchmarking methods. My evaluation framework consists of three interconnected stages that progressively build an understanding of model and benchmark characteristics for this paradigm.

3.1 Benchmark Task Assessment

I begin by selecting a suite of LLMs across different sizes and architectures; the motivation for doing so was to test how well IRT benchmarking can discriminate performance across parameter counts and model families. I then evaluate them across subsets of three benchmarks catered to LLM-as-a-Judge tasks: JudgeBench, LLM-Bar, and MT-Bench. For each benchmark, I standardize the prompt format and collect both the traditional accuracy metrics and the raw model responses for IRT analysis. This benchmark suite serves two crucial purposes: first, I establish baseline capabilities in different LLM-as-a-Judge tasks. Second, I can extract data for IRT modeling that will allow me to identify patterns in model performance across different

judgment contexts.

JudgeBench

JudgeBench is a benchmark dataset produced by Tan, et.al in [20]. It consists of 350 input questions, pairs of responses, and ground truth labels for the objectively correct response. JudgeBench uses MMLU-Pro, LiveBench, and LiveCodeBench for general knowledge, reasoning and instruction-following, and coding data, respectively. Unlike benchmarks focusing on alignment with subjective human preferences, JudgeBench bases labels on only factual correctness. Thus, it is a benchmark for how LLM judges discern accurate versus inaccurate information; in testing, the authors found it to be much more challenging than other existing LLM judgment benchmarks.

LLMBar

LLMBar is another meta-evaluation benchmark, curated by Zeng, et.al in [26] for evaluating how well LLM judges can tell which outputs better follow a given instruction. It includes 419 examples, each with an instruction, two responses, and a ground-truth label showing which response is objectively better at following the instruction. LLMBar focuses on objective comparisons, rather than subjective preferences. My research utilizes LLMBar’s ‘Natural’ set, which is built from existing human preference datasets —specifically AlpacaFarm and LLMEval2 —and then filtered for pairs with an objective difference in instruction-following quality. Thus, this set is useful in meta-evaluation for LLM judgment of instruction adherence.

MT Bench

The final benchmark I utilized was MT-Bench. Introduced by Zheng, et al. in [28], MT-Bench was designed to evaluate LLMs by measuring their multi-turn conversational abilities and instruction-following capabilities. It consists of 80 multi-turn

questions across 8 categories: writing, roleplay, extraction, reasoning, math, coding, STEM knowledge, and humanities knowledge. Though originally designed for multi-turn evaluation, MT Bench’s diversity in question types is valuable for single-turn assessment as well. I utilized the cleaned and curated example set by Zeng, et al. in [26] to subset 100 examples from this dataset for my research, focusing exclusively on first-turn responses to assess zero-shot judgment performance across diverse task types. Thus, MT-Bench is a useful dataset to measure LLM alignment with human preferences along different domains.

3.2 Essay Grading Tasks

For the essay grading task portion of my evaluation framework, I utilize a dataset of persuasive essays written by English language learner (ELL) students. My motivation for choosing this dataset is twofold. Firstly, it has a publicly available rubric —which I leveraged in prompting —and a multidimensional scoring system, which was useful in my IRT analysis. Secondly, using ELL student essays allowed for a wider variation in skill across language proficiency areas, providing an ideal test case for assessing models’ ability to make distinctions in writing quality.

Pairwise Comparative Judgement

For the first essay grading task, I run model inference for a comparative judgment task. Rather than asking models to produce absolute scores, I present pairs of essays and ask models to judge their relative quality (ex: $A > B$, $B > A$, $A = B$). This pairwise comparison approach mitigates some of the calibration challenges observed in absolute scoring, as I will discuss later. It also aligns with the comparative nature of the above benchmarks. Through the comparative task, I could generate a dataset revealing patterns in relative model judgment, a key part of the LLM-as-a-Judge

paradigm.

Pointwise Scoring

For the second essay grading task, I run model inference for a pointwise numerical scoring task. Models are provided with the essay text, the original prompt, and a detailed multidimensional rubric, then asked to generate numerical scores across several evaluation categories without examples or demonstrations. This task evaluates models' ability to internalize scoring criteria and apply them in an unsupported context. The zero-shot paradigm creates a challenging evaluation setting that, in my results, reveals limitations in models' ability to calibrate numerical judgments in pointwise evaluation.

3.3 IRT Modeling and Analysis

The core methodology of my approach is the application of IRT models to analyze performance data from both benchmark and essay grading tasks. IRT provides several advantages over traditional evaluation metrics:

- **Difficulty Parameterization:** IRT models estimate the inherent difficulty of each judgment item, allowing me to identify which aspects of assessment, or question topics, present the greatest challenges for current models.
- **Latent Trait (θ) Estimation:** By modeling a latent "judgment capability" trait, IRT can provide a unified method for comparing different model architectures across questions.
- **Joint Parameter Estimation:** Unlike traditional metrics that evaluate models in isolation, IRT simultaneously estimates both model capabilities and task characteristics.

- **Discrimination Parameters:** The two-parameter IRT model provides discrimination parameters that quantify how effectively each task differentiates between models with different capability levels, identifying which benchmarks provide the most information about model performance.
- **Adaptive Evaluation Potential:** The probabilistic framework of IRT creates opportunity for more efficient future evaluations, where models could be tested on fewer, more informative items based on difficulty and discrimination parameters established in this research.

For each benchmark and task, I run both a one-parameter (1PL) and two-parameter logistic (2PL) IRT model. For comparative judgment tasks, I model each pairwise comparison as a dichotomous item (correct/incorrect relative ranking). For the pointwise scoring task, I transform the continuous scores into ordinal categories and apply a graded response model. I assess model fit via loss, AIC score, and log-likelihood, and tune parameters and optimizers accordingly. After estimating these models, I conduct qualitative analysis of items with different discrimination and difficulty parameters to identify patterns in what makes certain evaluation tasks challenging for current LLMs.

3.4 Considerations and Challenges

In fitting 1PL and 2PL IRT models to these benchmarking and essay items, I adopt the IRT assumption of unidimensionality; in other words, I assume that judgment capability is a single latent trait underlying performance in each of these LLM-as-a-Judge tasks. More about this assumption will be covered in my Discussion section. Additional key assumptions include:

- **Local independence:** After controlling for the latent trait, responses to different items are statistically independent of each other. I structure the evaluation to minimize dependencies between judgment tasks and account for potential violations in the analysis —for instance, I randomize the order of essay pairs presented to models.
- **No guessing:** The model assumes that the lower asymptote of the item characteristic curve is zero, meaning there is no correct guessing. This assumption aligns with my problem formulation, since I assume that LLM agents will not use pure randomization in their responses.
- **Monotonicity:** The probability of a correct response monotonically increases with latent trait level.
- **Parameter invariance:** Item parameters are invariant across different groups of models, and model parameters are invariant across different sets of items (assuming the items measure the same latent trait). In other words, I assume that the difficulty and discrimination parameters estimated for essay evaluation tasks remain consistent regardless of which set of LLMs is being evaluated. Similarly, the estimated judgment capabilities (θ) of the language models should remain stable across different subsets of benchmark items or essay pairs. This invariance allows me to make meaningful comparisons between different model architectures and to generalize findings about which aspects of LLM judgment are inherently more challenging.

Additionally, several technical challenges arise in applying IRT to LLM evaluation.

- **Binarization of Non-Binary Scores** For the zero-shot scoring task, I develop a principled approach to converting continuous scores into ordinal cate-

gories that preserves meaningful distinctions while ensuring sufficient response frequency in each category for stable parameter estimation.

- **Validation without Ground Truth** I also face the challenge of validating IRT estimates given the lack of ground truth about model capabilities. To mitigate this, I analyze findings through correlational analysis with established benchmark accuracies; I also discuss the challenges inherent in creating an objective ground truth for judgment capability.

Chapter 4

Implementation

This section details the practical implementation of my research methodology, including data selection, pre-processing, model selection, and the specific techniques used for both benchmark assessment and essay grading tasks.

4.1 Essay grading data

For the essay evaluation task, I drew from the ELLIPSE-Corpus [17], which is a corpus of over 6500 essay writing samples from English Language Learners. This corpus was selected for several key reasons:

- The availability of a detailed rubric, which I incorporated into the prompt.
- The rubric encompassed individual scores across multiple dimensions of writing quality (Overall, Cohesion, Syntax, Grammar, Vocabulary, and Phraseology); this allowed me to make judgment across each rubric category its own latent trait in my numerical grading implementation.
- Essays scored on a consistent 1-5 scale with multiple human raters per essay.

- Standardized prompts across essays, facilitating fair comparison in my comparative analysis task.

4.2 Model Selection

To evaluate judgment capabilities across different model architectures and parameter scales, I selected a diverse set of over 25 LLMs spanning multiple model families, sizes, and training approaches. This diverse selection enabled analysis of how model size, architecture, and training approach affect judgment capabilities. For implementation consistency, all models were accessed through the HuggingFace pipeline API.

4.3 Benchmark Assessment

For the benchmark assessment phase, I implemented a standardized approach across the three aforementioned benchmark datasets: JudgeBench, LLMBAR, and MT-Bench. To ensure consistency in model evaluation, I standardized prompt formats across all benchmarks, using the 'Vanilla' prompt format that the LLMBAR creators established in [26]. I also randomized the order of text to be compared in the prompt (i.e. switching output A and output B). This prompt can be found in Appendix A.1. From the JudgeBench dataset, I selected their 'Claude' split, which consists of 270 response pairs generated by Claude-3.5-Sonnet. For LLMBAR, I utilized a subset of 100 responses from the 'Natural' split. Finally, for MT-bench, I implemented 200 responses from the processed subset of MT-Bench available in [26]. This subset was processed via removal of ties, assignment of the gold preference via majority human vote, and random sampling. Model inference was conducted using the HuggingFace API, with all responses collected through Google Colab.

Model	Model Family	Parameters
gpt2	GPT-2	137M
gpt2-large	GPT-2	812M
Llama-3.2-1b-it	LLaMA	1.24B
Llama-3.2-3b	LLaMA	3.21B
Llama-3.2-3b-it	LLaMA	3.21B
OLMo-7b-it	OLMo	6.89B
OLMoE-1b-7b-0924-it	OLMo	6.92B
OLMo-2-0325-32b-it	OLMo	32.2B
flan-t5-small	FLAN-T5	77M
flan-t5-large	FLAN-T5	783M
flan-t5-xl	FLAN-T5	2.85B
Qwen-2.5-0.5B-Instruct	Qwen	494M
Qwen-2.5-1.5B-Instruct	Qwen	1.54B
Qwen-2.5-3B-Instruct	Qwen	3.09B
Qwen-2.5-7B-Instruct	Qwen	7.62B
Qwen-2.5-7B-Instruct-1M	Qwen	7.62B
Phi-3-mini-4k-instruct	Phi	3.82B
Phi-3-medium-4k-instruct	Phi	3.82B
Phi-3.5-mini-instruct	Phi	3.84B
Phi-4-mini-instruct	Phi	14B
Mistral-7B-Instruct-v0.1	Mistral	7.24B
Mistral-7B-Instruct-v0.3	Mistral	7.25B
Ministrال-8B-Instruct-2410	Mistral	8.02B
Mistral-Nemo-Instruct-2407	Mistral	12.2B
gemma-2b	Gemma	2.51B
gemma-1.1-2b-it	Gemma	2.51B
gemma-2-2b-it	Gemma	2.61B

Table 4.1: Model sizes and families with parameters in M (millions) and B (billions)

4.4 Pairwise Judgment Implementation

For the implementation of the pairwise comparative judgment task, I extracted 200 pairs of essays, making sure essays originated from the same prompt; for each essay, I preserved the original prompt, the full essay text, and the human-assigned scores across all dimensions. To control for position bias, I randomized the presentation order of essays within each pair and balanced the position of higher-quality essays across the dataset. Model responses were processed to extract binary judgments (selection of Essay A or B), which were then compared against ground truth based on human scores. For this task, The prompt template for comparative judgment, found in Appendix A.1, included:

- Task description explaining the comparative assessment goal.
- The original essay question.
- Both essays presented in a randomized order.
- Explicit instructions to identify which essay demonstrates better writing quality.
- Detailed evaluation criteria based on the ELLIPSE rubric dimensions

For evaluation, I used the difference between the human scores in the 'Overall' category from the original dataset as ground truth.

4.5 Pointwise Scoring Implementation

For the zero-shot numerical scoring task, I selected 300 essays from the ELLIPSE corpus with balanced representation across proficiency levels. I developed a standardized prompt template , found in Appendix A.2, that included:

- Task description

- The original essay question.
- Explicit instructions to provide numerical scores (1-5) for each dimension in a specified JSON format
- Detailed evaluation criteria based on the ELLIPSE rubric dimensions

For evaluation, I used the human scores from the original dataset as ground truth. To create ordinal response categories for the Graded Response Model (GRM), I experimented with different error calculations but ultimately settled on using raw score differences to sort models into three thresholds of ability:

- High ability: Score within 0.5 points of ground truth
- Medium ability: Score within 0.5-1.0 points
- Low ability: Score differing by more than 1.0 point

In order to compare the performance ranking ability of the GRM to traditional accuracy metrics, I also created binary response categories for this score, using a score within ± 0.5 points of the ground truth as "correct", and more than that difference as "incorrect."

4.6 IRT Implementation

The core of my implementation involved applying IRT models to the collected judgment data. First, during data preparation, I converted model responses into appropriate formats for IRT analysis (binary for pairwise comparison tasks, and ordinal categories for pointwise). Then, utilizing the Python package `py-IRT` and the R package `mirt`, I fit the models. For pairwise comparison tasks, I fit both one-parameter (Rasch) models to estimate basic item difficulty and model ability, and two-parameter

logistic (2PL) models to additionally account for item discrimination. I chose which latent traits to analyze based on model fit. Then, for the numerical scoring task, I fit a graded response model using the Python package, Georgia Tech Item Response Theory Documentation (**GIRTH**). I compared model fit across different IRT formulations using Akaike Information Criterion (AIC) score, and log-likelihood values. For each of these models I tuned optimization settings by comparing loss and fit; I used a learning rate between 0.005-0.02, and 3000-4000 epochs with vague priors (low-information priors with large variance), allowing the model parameters to be determined primarily by the data, and ensuring stable convergence.

After fitting the models, I created different visualizations, such as Item Characteristic Curves (ICCs) for understanding item difficulty and discrimination, and Wright maps, to visualize the relationships between model abilities and item difficulties.

4.7 Analysis Approach

To validate the IRT-estimated latent traits, I implemented correlation analyses between:

- IRT-estimated latent traits and benchmark accuracy scores
- θ s across different benchmarks, including essay comparison
- Accuracies across different benchmarks, including essay comparison
- θ s across different rubric categories

This analysis involved calculating Spearman correlation coefficients to identify meaningful relationships between different evaluation contexts. In addition, I implemented Kendall's Tau rank correlation analysis to assess the stability of model rankings across different evaluation contexts. In addition, I examined the IRT item

parameters across the benchmarks and essay grading tasks. By analyzing difficulty and discrimination parameters, I could identify specific characteristics that make different judgment tasks challenging or discerning.

4.7.1 Implementation Challenges

Throughout implementation, I addressed several technical challenges:

- **Response Format Standardization** Despite explicit formatting instructions in the prompts, many models produced responses in inconsistent formats, particularly for the essay scoring task. I implemented parsing to extract relevant judgments. In extreme cases where models produced completely non-conforming outputs (e.g., refusing the task or generating irrelevant text), I excluded these specific responses from analysis rather than introducing noise into the dataset.
- **Computational Resource Management** The scale of the evaluation required efficient resource management. I implemented parallel processing where possible and optimized prompt batching to maximize efficiency.

By systematically addressing these implementation challenges, I ensured a robust and consistent evaluation framework that could reliably assess judgment capabilities across diverse model architectures and evaluation contexts.

Chapter 5

Results and Analysis

In this section, I will present and analyze my findings and explain how my research answers the research questions I posed in the Introduction.

5.1 IRT Efficacy in Performance Estimation

Now, I will explore how effective the IRT models are in capturing performance differences in LLM judgment as compared to traditional accuracy metrics. I will do so by presenting the findings of my pairwise and pointwise judgment task analyses.

5.1.1 Pairwise Judgment

First, in 5.1, we see that in AIC and log-likelihood score, 2PL models have better fit than 1PL models across the board for pairwise comparison. This indicates that variable item discrimination is a valid consideration in pairwise LLM-as-a-Judge tasks. Moving forward I will use the parameters extracted from the 2PL models for my IRT analysis.

Continuing with the pairwise judgment task results, in Figures 5.1 and 5.2, I

	MT Bench	JudgeBench	LLMBar	Essay Comparison
AIC 1PL	3728.903	2045.721	1962.398	1928.654
AIC 2PL	2853.739	1893.482	1816.335	1782.491
Loglikelihood 1PL	-1719.450	-973.614	-914.199	-889.375
Loglikelihood 2PL	1138.870	-823.507	-776.168	-752.146

Table 5.1: Model fit statistics across different benchmarks

present model accuracies and thetas, respectively, for each task, grouped by family and parameter size. From these figures, we see that the IRT-derived theta estimates demonstrate significantly greater differentiation between model capabilities compared to traditional accuracy metrics. As shown in 5.2, the normalized theta scale creates more pronounced separation between models, particularly at the lower and higher ends of the performance spectrum. These theta estimates not only provide better differentiation between models but also demonstrate varying relationships with traditional accuracy metrics across different benchmarks. Seen in Figure 5.3, the Spearman and Kendall Tau correlations between these two metrics are very high for LLMBar, essay comparison, and moderate to high for JudgeBench. In general, we see that latent trait estimation produces a performance assessment that is more starkly discriminating and yet overall consistent with that of accuracy metrics.

However, it is impossible to ignore the lack of correlation between accuracy and θ in the case of MT-Bench; we observe notable shifts in relative rankings between the latent traits and the accuracy. For instance, in the MT-Bench evaluation, the performance of phi-4-mini-Instruct is lower than all other models of its family in accuracy, but higher than the rest in its latent ability. Notably, this is the most ‘advanced’ model in the Phi series (version 4, rather than 3 or 3.5), and the second-highest in parameter size; it is expected to outperform its counterparts, and in most accuracy and latent trait estimates for other benchmarks, it does. These observations allow us to conclude that in the case of MT-Bench, raw accuracy poorly reflects latent judgment ability.

However, this is not necessarily a sign that IRT is ineffective in ability stratification; due to the nature of Item Response Theory, we simultaneously model both subject ability and item characteristics. Thus, as will be discussed in the next section, the increased discriminability in IRT estimates allow for a more nuanced portrayal of model performance differences beyond accuracy. In exploring IRT for benchmarking, I will discuss that while high theta-accuracy alignment is indeed indicative of the power of latent trait estimation, a low alignment can be a sign of issues in benchmark discriminability and alignment of item difficulty with model ability.

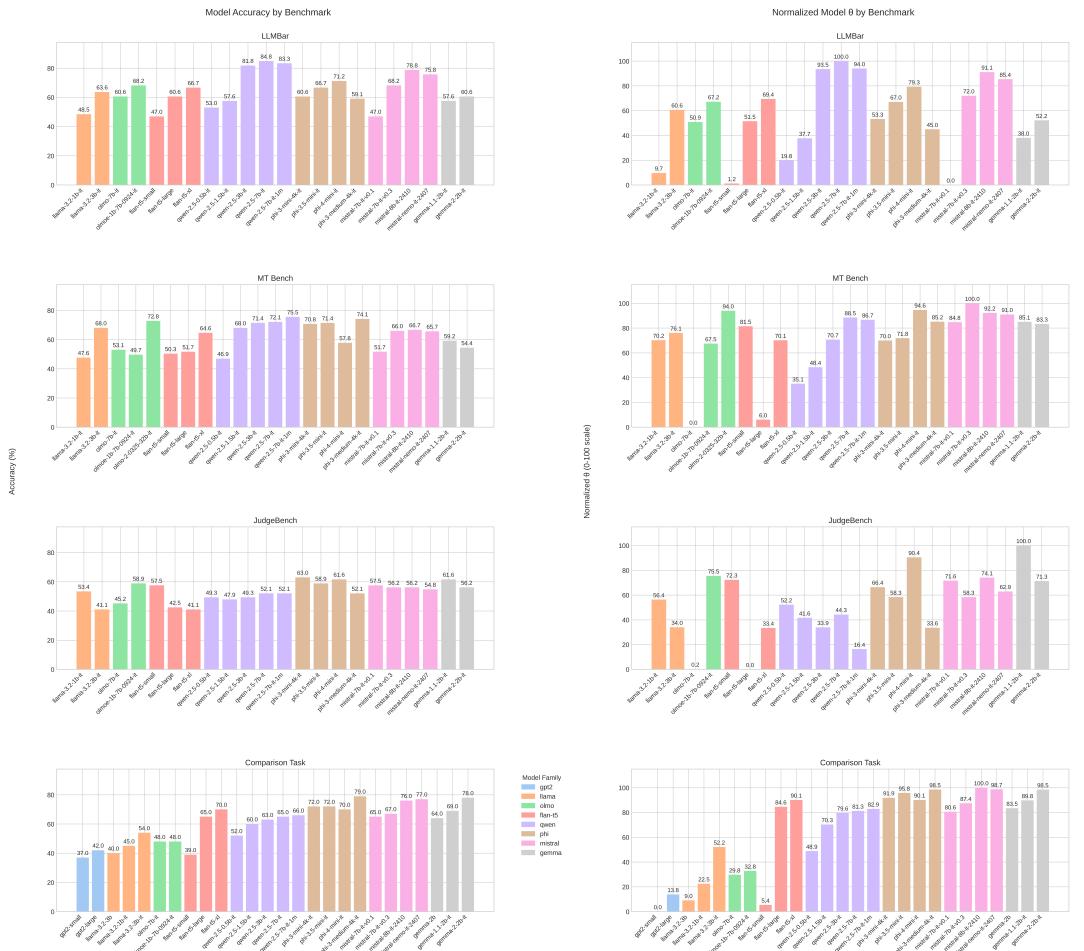


Figure 5.1: Model Accuracy Across Benchmarks

Figure 5.2: Model θ Across Benchmarks

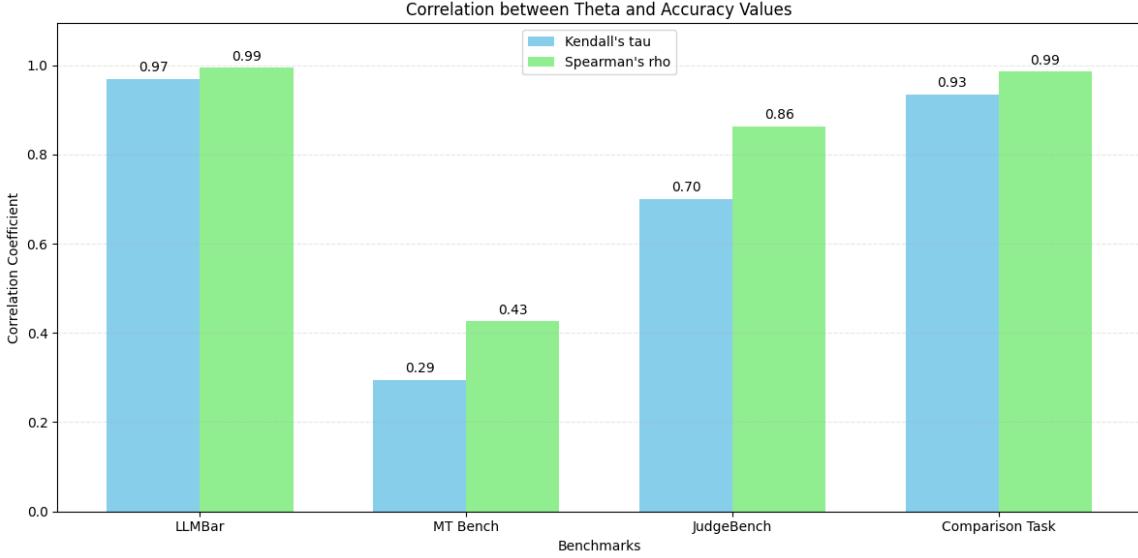


Figure 5.3: Correlation between Accuracy and θ for Pairwise Judgment

5.1.2 Pointwise Judgment

Now, I will discuss the results from the pointwise judgment task; While pairwise comparison tasks provide valuable insights into relative judgment capabilities, pointwise evaluation represents an equally important but fundamentally different judgment paradigm. In contrast to the above benchmarks, in this section I ran model inference for a single-answer numerical essay grading task. I applied the Graded Response Model (GRM) to capture the polytomous nature of scoring and compared these theta estimates with simulated accuracy metrics based on score differences.

What makes GRM particularly valuable for pointwise judgment is its ability to handle ordinal data with multiple thresholds, a nuance that binary accuracy metrics fundamentally cannot capture. Unlike accuracy measures that treat all deviations equally (a model missing by 0.1 points receives the same "incorrect" classification as one missing by 2.0 points), GRM explicitly models the probability of crossing each scoring threshold as a function of model ability. This approach acknowledges the progressive nature of essay quality and recognizes that distinguishing between adjacent score levels (for instance, 3 vs. 4) requires different capabilities than distinguishing

extreme scores (e.g., 1 vs. 5). The threshold parameters estimated by GRM reveal which transitions between score categories present the greatest challenge for models, information not gained through simple accuracy or difference metrics.

Now, I will explore my results from GRM θ calculation. After calculating correlation metrics between θ and accuracy performance estimators, I found moderate to strong correlation: Kendall's $\tau = 0.59$, $p < 0.001$; Spearman's $\rho = 0.79$, $p < 0.001$. These correlations indicate that while GRM theta estimates generally align with accuracy, they also provide additional discriminative information not fully captured by simple score differences. The stronger Spearman correlation suggests that GRM maintains overall ranking consistency while the more moderate Kendall's Tau reveals that GRM detects finer distinctions between closely performing models. This insight reinforces the value of GRM for nuanced meta-evaluation.

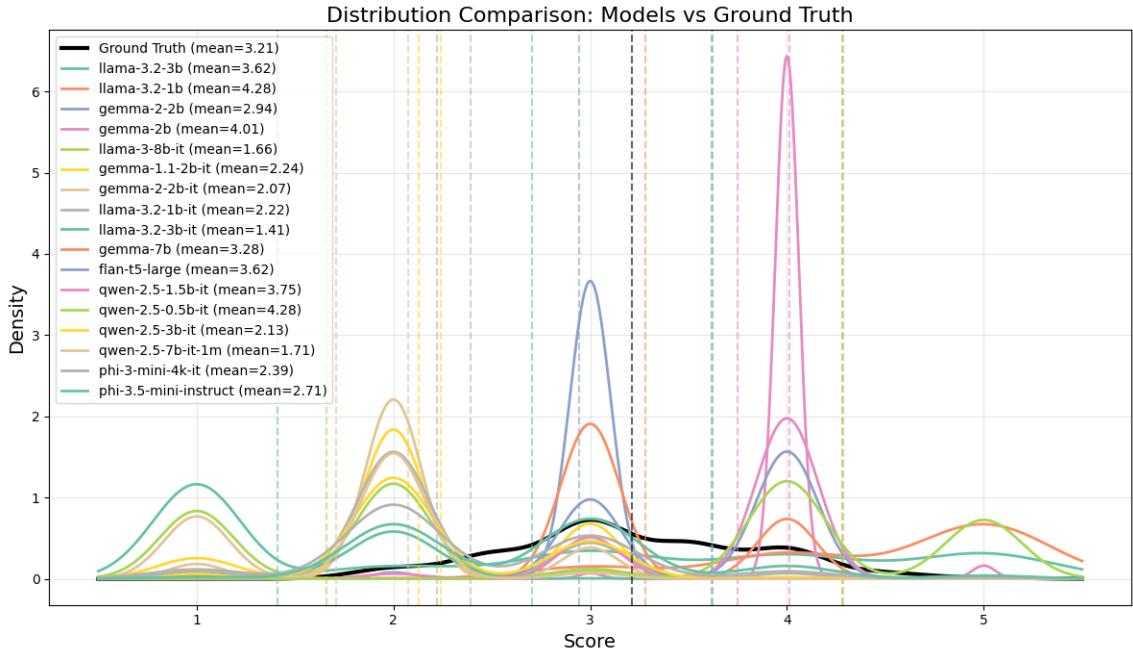


Figure 5.4: Model Output Distribution vs Ground Truth density graph

Despite these methodological advantages, the implementation of GRM for pointwise judgment faces a significant challenge that does not affect pairwise comparison tasks. As shown in Figure 5.4, models exhibited substantial calibration issues when

translating their qualitative judgments into numerical scores. Without sufficient calibration examples or more explicit guidance on score mapping, models developed arbitrary approaches to translating their qualitative judgments into numerical scores. This introduced too much noise into the data for the thetas to form an accurate relative ranking of model performance on this task. The GRM analysis thus complements our understanding of pairwise benchmarks by revealing an important consideration for comprehensive LLM evaluation: more methodological refinement is needed, particularly in prompt engineering, before theta estimates can serve as reliable indicators of general LLM capabilities. From these results, I can recommend that before deploying LLM-as-a-Judge for single-answer grading, future work should focus on implementing more robust calibration procedures and potentially accounting for response style as a separate factor in the model.

5.2 Patterns in Item Difficulty and Discrimination

In this section, I will discuss patterns in item discrimination and difficulty across different benchmarks.

5.2.1 Pairwise Judgment

In Figure 5.5, I present item discrimination and difficulty statistics from the JudgeBench dataset, separated by the question source. Given it is the only dataset which separates its items by source, it is optimal to draw insights about patterns in item characteristics. As previously mentioned, JudgeBench pulled question pairs from various topics in the Massive Multitask Language Understanding (MMLU) benchmark, as well as LiveCodeBench and LiveBench. My source analysis reveals substantial differences in both discrimination and difficulty parameters across subject areas.

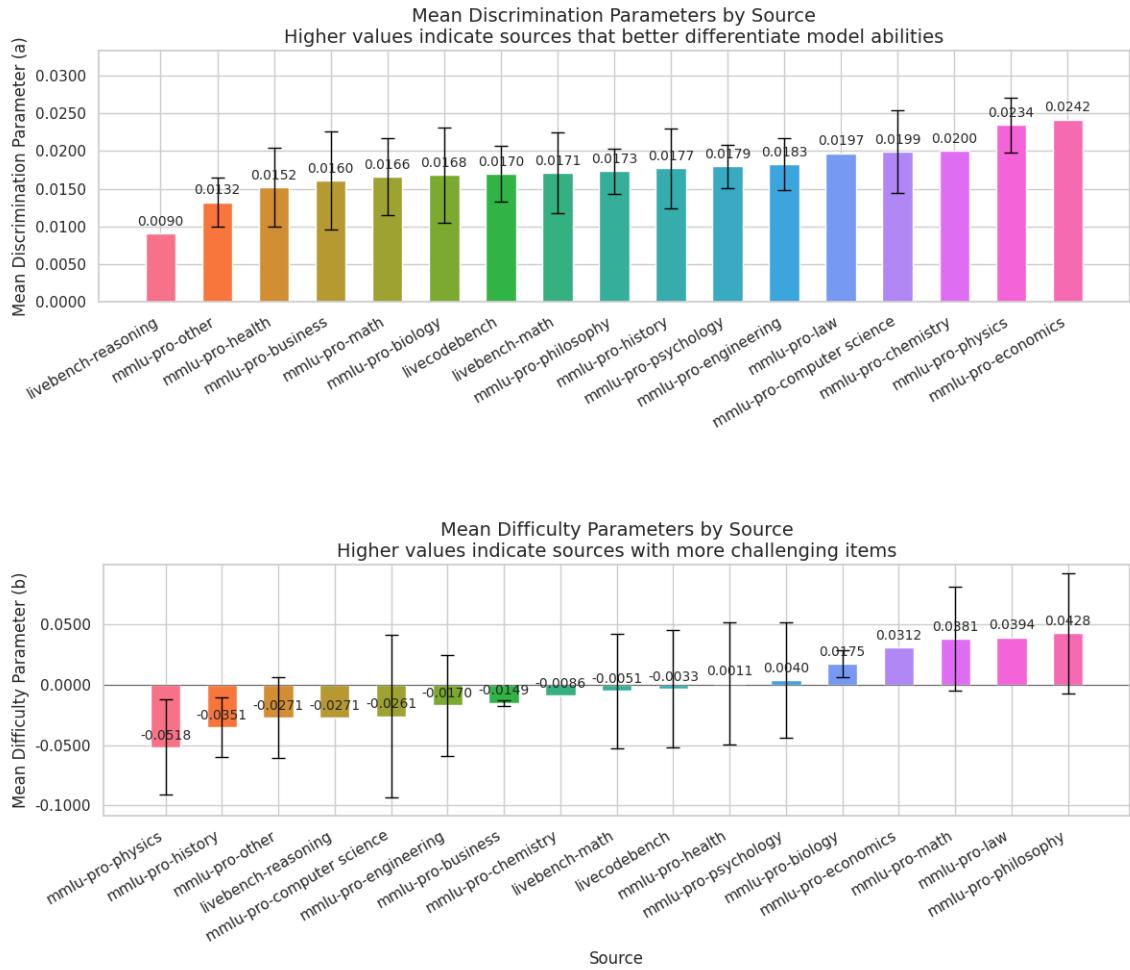


Figure 5.5: JudgeBench item difficulty and discrimination, grouped by question source. Top: Mean difficulty plot. Bottom: Mean discrimination plot.

For discrimination parameters (top), quantitative subjects such as physics and economics demonstrate the highest discriminative power (0.0234 and 0.0242 respectively), making them approximately 2.6 times more effective at differentiating model abilities than reasoning tasks (0.0090), which show the lowest discrimination. This suggests that physics and economics questions more effectively separate high-performing models from lower-performing ones. The natural sciences generally show higher discrimination values than humanities subjects, with computer science (0.0199) and chemistry (0.0200) also performing well. The difficulty parameter analysis (bottom chart) reveals an interesting complementary pattern. Philosophy and law ques-

tions present the greatest challenge to models (0.0428 and 0.0394 respectively), while physics questions are among the easiest (-0.0518). This creates an interesting situation, where physics questions can effectively discriminate between models despite being on the lower end of the difficulty scale. Conversely, history questions combine low discrimination with negative difficulty, suggesting they provide limited evaluative value. These findings demonstrate that subject domain significantly impacts measurement properties in LLM evaluation, with some domains providing more informative assessment than others despite not necessarily being more difficult.

	MT-Bench	Essay Comparison	JudgeBench	LLMBAR
Difficult Δ	930.4	1252.2	106.7	327.8
Easy Δ	462.9	1244.0	213.3	107.5
Discriminable Δ	504.8	967.9	126.6	162.2
Non-discriminable Δ	333.3	1304.9	160.7	307.2

Table 5.2: Difference in text lengths (characters) per item characteristic across different benchmarks

In Table 5.2, I analyzed how length differences in the inputs of all benchmarks corresponded to item difficulty and discrimination; that is, does $|len(\text{Output A}) - len(\text{Output B})|$ impact how well a model of different ability can judge the relative output quality? The text length analysis shown in the table, reveals that for all pairwise comparison except for JudgeBench, difficult items have greater length differences. As I noted in the Background section, many LLMs struggle with length bias, and prefer longer text regardless of content. The pattern seen here supports a hypothesis of length bias; lower-ability models might be more likely to pick longer outputs, thus making items with significant length differences, more difficult.

For discriminability, the opposite pattern emerges: except for MT-Bench, the benchmarks exhibit less discriminability for greater Δ in output length. This suggests that when candidate responses differ substantially in length, models tend to make more uniform judgments; thus these pairs fail to differentiate model capabilities. In the Essay Comparison task, this effect is most pronounced, with non-discriminative

items showing output length differences 35% larger than discriminative ones (1304.9 vs. 967.9 characters). Again, this finding aligns with previous research on length bias in LLMs. The length difference appears to supersede other quality signals, thus creating an evaluation scenario where most models make similar judgments based primarily on length.

5.2.2 Pointwise Judgment

Item Characteristic	Average Length
Low Discrimination	2412.37
High Discrimination	2322.03
Low Difficulty	2490.55
High Difficulty	2380.84

Table 5.3: Length (character count) for Extreme Items (GRM)

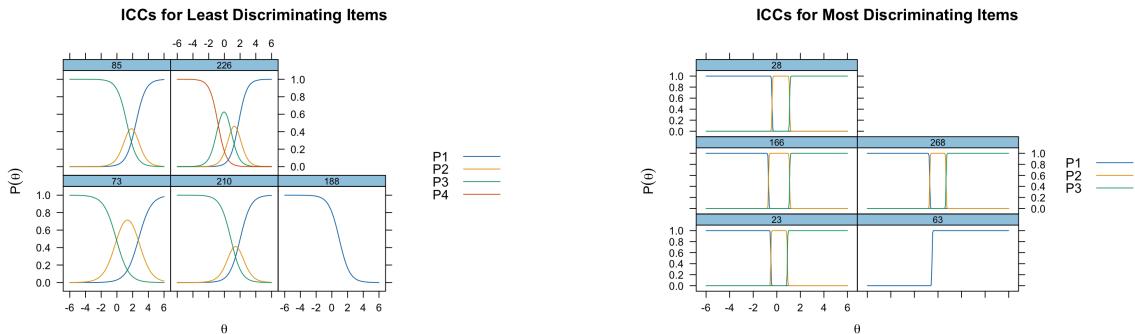


Figure 5.6: Item Characteristic Curves for Least and Most Discriminating Items (GRM)

Extending this analysis to the zero-shot numerical scoring task reveals intriguing differences in how item characteristics function across judgment paradigms. As shown in Table 5.3, essay length plays a substantially diminished role in the absolute scoring context compared to pairwise comparison. The difference in average length between high and low discrimination essays is minimal, and similarly so between high and low difficulty essays. This suggests that when LLMs evaluate essays independently rather

than comparatively, their judgments become less influenced by surface-level features like length and more dependent on content assessment.

As seen in Figure 5.6, the ICCs derived from the Graded Response Model reveal expected behavior for very highly and very poorly discriminating items. A very highly discriminating essay (such as Essay 28) displays extremely steep, well-separated threshold curves with minimal overlap between adjacent categories, indicating its ability to differentiate models across the ability spectrum. In contrast, a least discriminating essay (such as Essay 210) exhibits flatter curves and could lack some threshold curves. However, overall we see that even the least discriminating essays possess well-separated threshold curves, suggesting that essay grading in general is a highly discriminating task regardless of essay topic or length.

This contrast between pairwise and pointwise judgment highlights how the nature of the evaluation task fundamentally alters which factors impact model performance. In comparative judgment, models appear susceptible to relative differences between options, particularly length disparities. However, in absolute scoring, models must internalize and apply scoring criteria to individual essays, which presents different challenges. The relatively high discrimination values across even the "least discriminating" essays in the GRM model (as shown in Figure 5.6) indicate that absolute judgment tasks may actually provide more consistent differentiation between model capabilities than comparative tasks, despite being more challenging to calibrate. This challenge will be discussed further in the following section.

5.3 IRT and Benchmarking

Building on the insights from Sections 5.1 and 5.2 —my analysis of IRT's efficacy in performance estimation and the patterns observed in item characteristics, respectively —this section examines my results regarding how Item Response Theory provides a

unique framework for understanding and improving LLM benchmarking.

5.3.1 Benchmark Validation

First, I will discuss the challenge of benchmark validity. As previously discussed, researchers have found that correlational analysis for benchmark validation has significant limitations; oftentimes the measure effectiveness of a benchmark at discriminating model performance, depends on the choice of models in testing [14]. In addition, while correlations between benchmark performance can provide insight about ranking similarity, it does not have the ability to discern if the benchmarks measure the same latent trait. In my results, we will see this illustrated clearly.



Figure 5.7: Spearman Heatmap of Accuracies



Figure 5.8: Spearman Heatmap of Normalized Thetas

In Figure 5.7 and Figure 5.8 I examine heatmaps for correlation between benchmark scores in accuracy and latent traits respectively. I also add IFEval, an instruction-following benchmark, as a measure of how well LLM-as-a-Judge benchmarks correlate with a more general benchmark. Overall, there are weak to moderate correlations between all these benchmarks, mostly ranging from 0.3 to 0.6, over both latent trait and ability metrics. Notably, JudgeBench shows very weak or even slightly negative correlations with other benchmarks, suggesting it measures fundamentally different

aspects of judgment capability than other evaluations. MT Bench demonstrates moderate positive correlations across all benchmarks (0.46-0.55), positioning it as the most generally representative benchmark. The IFEval and LLMBAR benchmarks show a relatively strong correlation with each other (0.65), indicating they may capture similar aspects of instruction-following judgment. The essay comparison task correlations are medium, with its strongest relationship to MT Bench, suggesting that essay evaluation capability only partially overlaps with other judgment tasks. These patterns indicate that these benchmarks capture distinct and sometimes almost orthogonal aspects of LLM judgment capability rather than measuring a unified skill.

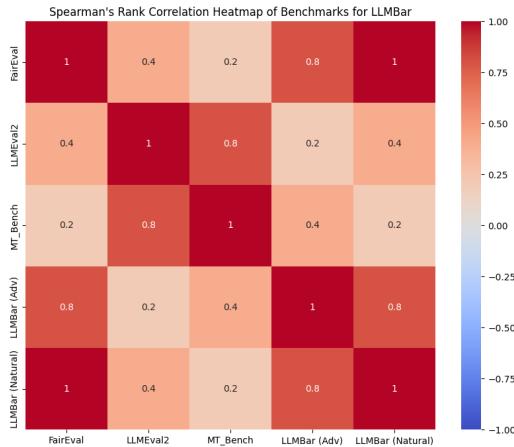


Figure 5.9: Spearman Heatmap of Accuracies: LLMBAR

(a) Model Accuracies sourced from [26]

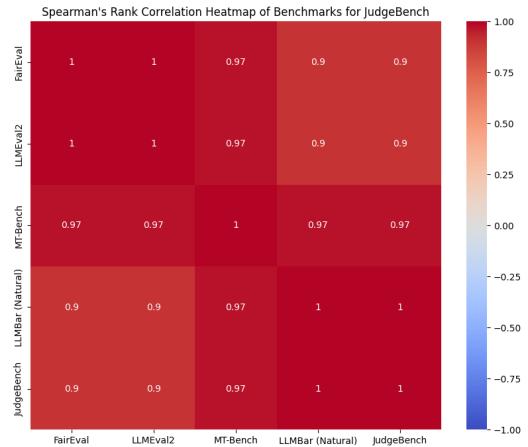


Figure 5.10: Spearman Heatmap of Accuracies: JudgeBench

(a) Model Accuracies sourced from [20]

My findings for the pairwise judgment tasks prompted me to look into a correlational analysis between benchmarks displayed to validate JudgeBench in [20], and LLMBAR in [26]. In LLMBAR, the creators used ChatGPT, LLAMA2-70B-Chat, PaLM2-bison, and GPT-4 as test models to display LLMBAR’s ability to rank models, as compared to FairEval, LMEval², and MT-Bench. Meanwhile, JudgeBench used Claude-3-Haiku, Llama-3.1-8B-Instruct, Llama-3.1-70B-Instruct, GPT-4o, and Claude-3.5-Sonnet to display its comparative ranking capability against FairEval,

LLMEval², MT-Bench, and LLMBAR (Natural).

These heatmaps provide striking evidence of how benchmark validation through correlation is heavily dependent on the specific models used in testing. In Figure 5.10a, the JudgeBench correlations show extremely high agreement (0.9-1.0) across all benchmark comparisons, suggesting near-perfect alignment. In contrast, Figure 5.9a shows dramatically different patterns for LLMBAR, with correlations varying widely from 0.2 to 1.0. Notably, with the models chosen by LLMBAR in [26], the correlation between performance in LLMBAR’s Natural split, and MT-Bench, is 0.2. With JudgeBench models, it is 0.97. This discrepancy perfectly illustrates my finding that benchmark correlations fluctuate substantially depending on which models are included in the evaluation.

This model-dependency undermines the reliability of correlation-based validation. A benchmark might appear well-validated when tested on one set of models but show entirely different characteristics when applied to another set. This explains why benchmarks that claimed strong validation during development often show surprisingly weak correlations (or even negative correlations) in my more comprehensive analysis, reinforcing the need for IRT approaches that characterize benchmarks through intrinsic properties like discrimination and difficulty rather than through unstable correlational metrics.

Holding space for the fact that correlation-based benchmark validation has proven unstable due to test-model dependencies and inconsistent latent traits, my more comprehensive model testing reveals a more nuanced picture. Again, the contrast between JudgeBench’s near-perfect correlations (0.9-1.0) using a limited model set and the wider variance (0.2-1.0) seen with LLMBAR models demonstrates why many benchmarks fail to generalize. However, this doesn’t invalidate correlation entirely —when established across larger model sets like mine (25+ models compared to 5), strong correlations can meaningfully indicate shared latent traits. This is particularly evident

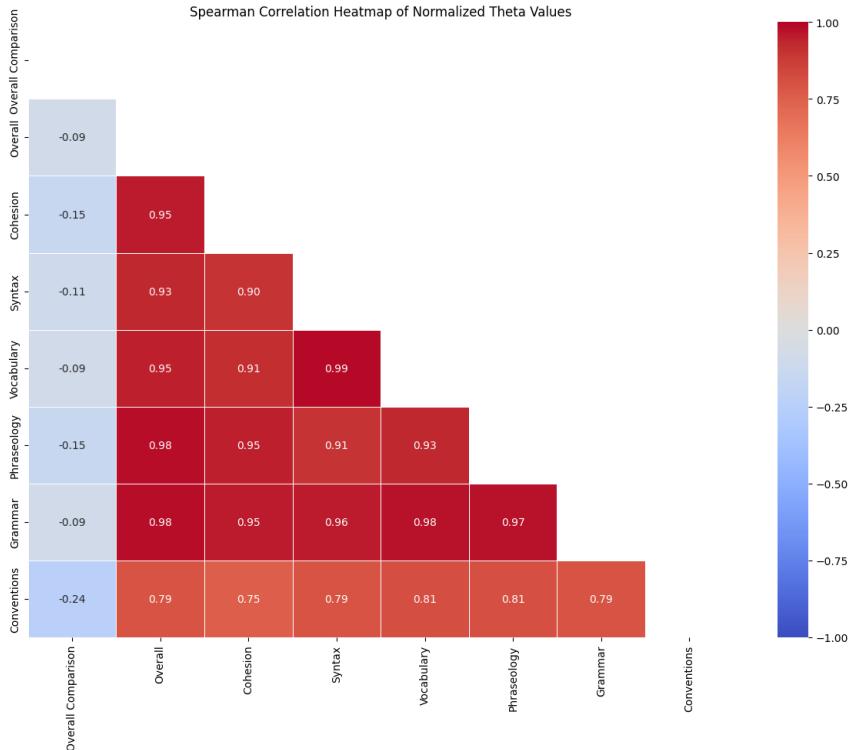


Figure 5.11: Correlation Matrix between Thetas and Benchmarks

in the GRM analysis, where the high inter-dimensional correlations ($r = 0.75\text{--}0.97$) across rubric dimensions suggest a unified "essay evaluation capability" despite the models' varying architectures and sizes. Conversely, the negative correlation between numerical and comparative judgment tasks ($r = -0.09$ to -0.24) provides evidence for distinct underlying traits for pairwise and pointwise judgbench in the same domain. Thus, my results highlight how IRT can both validate meaningful correlations and distinguish between genuinely different capabilities.

5.3.2 Benchmark Discrimination Analysis

Now, I will note how analysis of item patterns within benchmarks can facilitate our understanding of the internal characteristics of individual benchmarks to assess their measurement quality. While correlation analysis tells us how benchmarks relate to each other, discrimination and difficulty analysis can provide insight about how ef-

fectively each benchmark distinguishes between models of different capabilities and whether the benchmark provides appropriate coverage across the ability spectrum. This IRT-based deeper level of analysis provides crucial insights for benchmark development and selection that traditional metrics cannot offer.

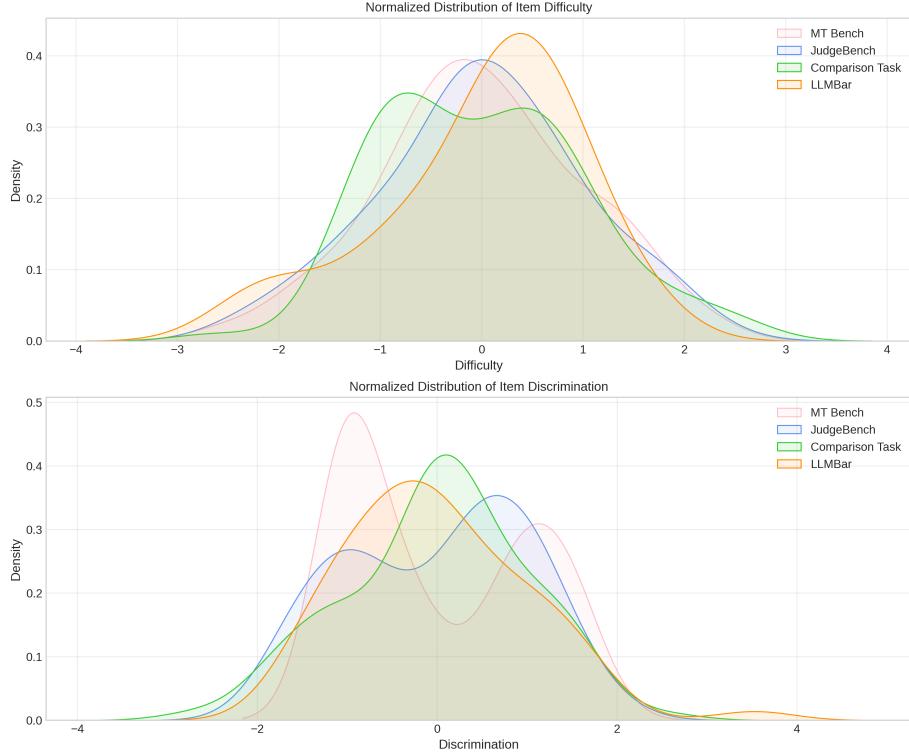


Figure 5.12: Distribution of Item Difficulty and Discrimination for each Benchmark

I begin with Figure 5.12. I applied z-score normalization to the item parameters, thus forcing them into distributions with means of 0 and standard deviations of 1; this allows us to directly compare the relative parameter distributions independent of the different scales. For item difficulty, all benchmarks are roughly normal, indicating a good spread in the item difficulties across all benchmarks —despite our finding that this spread is not proportional to the range of model abilities for MT-Bench and JudgeBench. In the bottom plot, the normalized item discrimination plots hold interesting findings. MT-Bench and, to a lesser extent, JudgeBench, have bimodal distributions, indicating that many items are either high or low discrimination, and

few in-between. These normalized distributions highlight the different measurement properties of each benchmark while controlling for scale differences; this application of IRT provides insight into the relative evaluation effectiveness of these benchmarks.

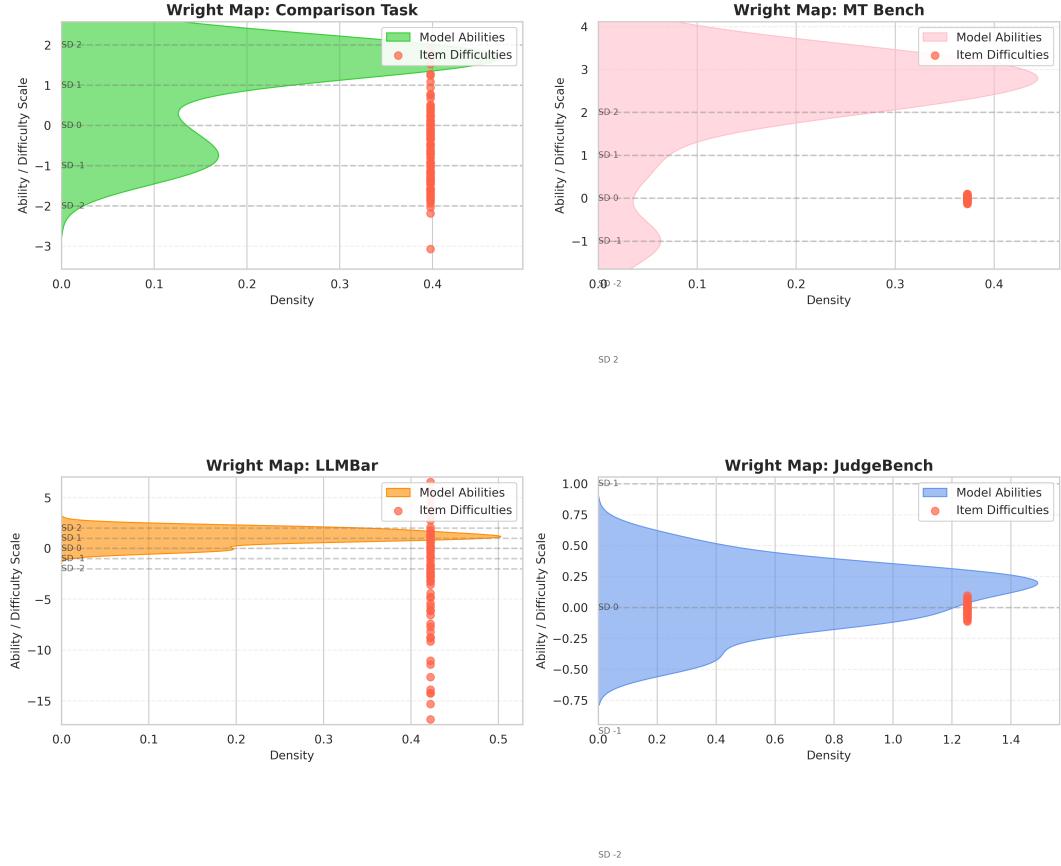


Figure 5.13: Wright Maps for each Benchmarks

In Figure 5.13, I show Item-Person (Wright) maps for models and items in each benchmark. The Wright maps reveal patterns in the distribution of model abilities and item difficulties across benchmarks. Pairwise essay comparison (top left) shows most model thetas falling within a range similar in scale to that of item difficulties, meaning that item difficulties robustly cover the range of abilities present. In contrast, the MT-Bench and JudgeBench dataset have a wide spread of model abilities, with item difficulties concentrated near the bottom and center of the scale respectively; this indicates that while model performance on this benchmark varies considerably,

the questions do not have an optimal spread of difficulty to capture these differences in ability.

Overall, by simultaneously modeling model abilities and item characteristics, IRT offers a comprehensive lens through which we can critically evaluate current benchmarking practices and develop more effective evaluation protocols.

Chapter 6

Discussion

In this section, I will analyze the results presented in the previous section and reflect upon their limitations and implications. My results reveal several key insights into LLM judgment capabilities and the application of IRT for model evaluation.

6.1 Model Size and Ability

In Figure 6.1, I present an analysis of the relationship between model size (measured by parameter count) and the θ measured by the 2PL IRT models I fit in Part 1 of my results; the IFEval score here is in terms of accuracy, and is present as a baseline. We see that the relationship between model size and latent abilities in LLM-as-a-Judge tasks is variable across these benchmarks. As shown in the scatter plots, most benchmarks demonstrate positive correlations with parameter count, but with varied strength. IFEval ($r = 0.661$), Comparison Task ($r = 0.623$), LLMBAR ($r = 0.594$), and the Average Normalized Performance ($r = 0.679$) all show moderate to strong correlations with model size; this suggests that in these contexts, scaling generally improves judgment capability. However, MT Bench displays a weaker correlation ($r = 0.356$), while JudgeBench shows virtually no relationship with model size ($r = -0.048$). This pattern reveals that while larger models typically demonstrate enhanced

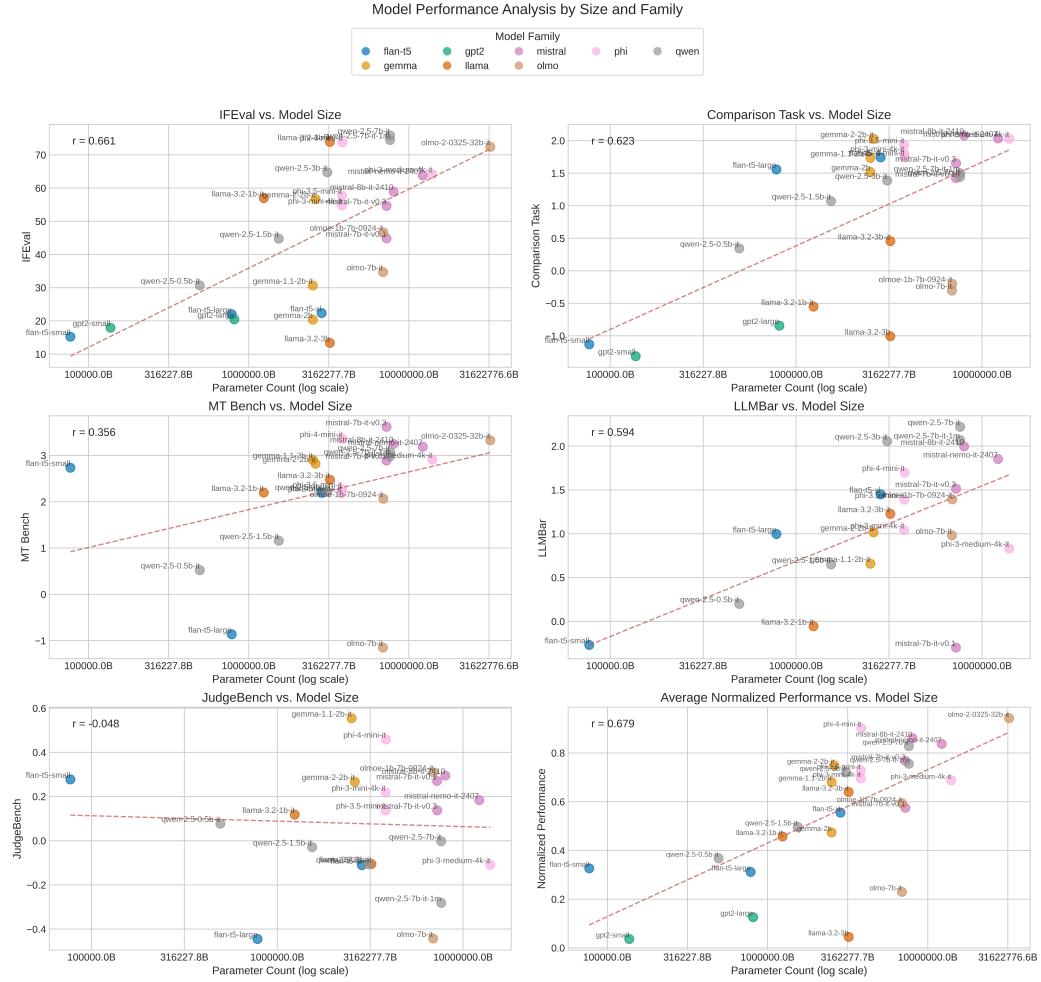


Figure 6.1: Parameter Analysis of Benchmark

judgment capabilities overall, specific evaluation contexts like JudgeBench appear to test skills that do not consistently improve with only scale. Additionally, the plots reveal variance in performance among similarly-sized models, with some smaller models outperforming significantly larger ones; this indicates that architectural choices and training methodology can sometimes compensate for size disadvantages.

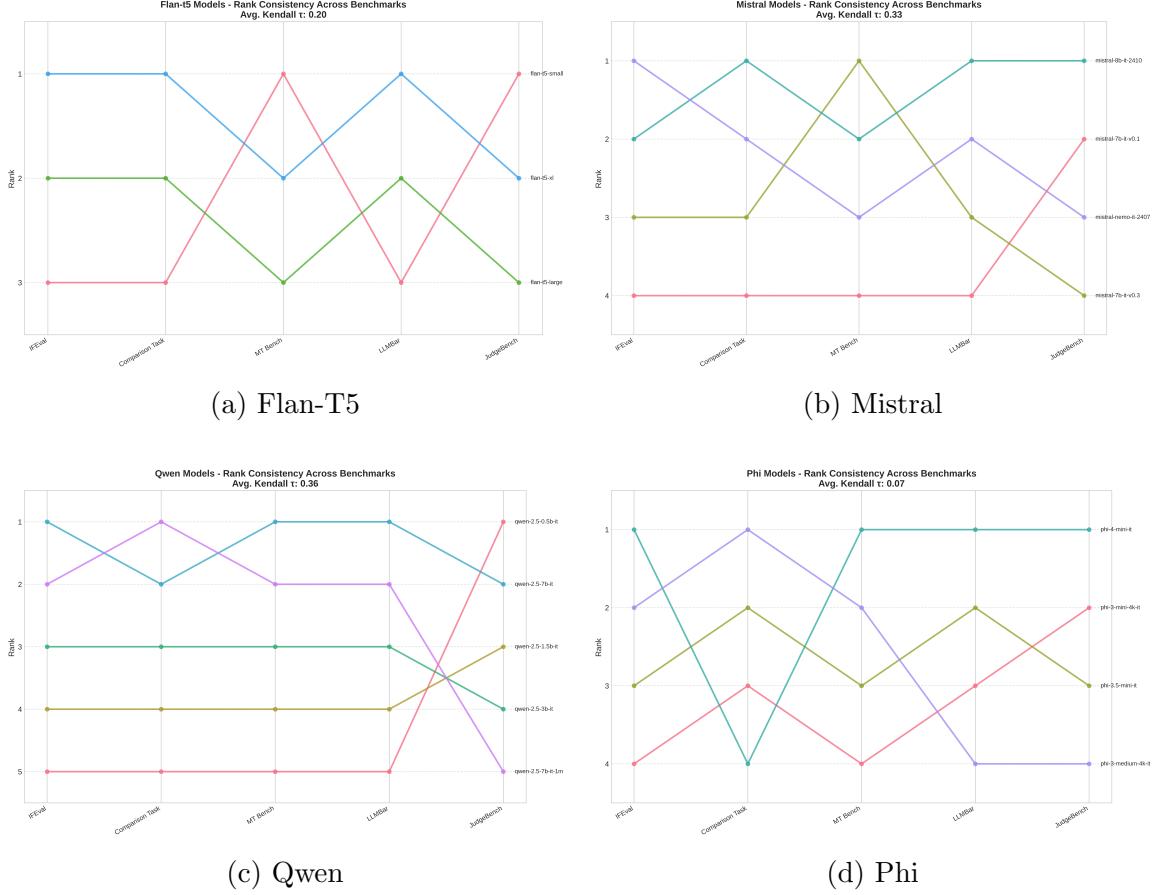


Figure 6.2: Rank consistency across benchmarks for different model families

This point is further reinforced by Figure 6.2, in which I show the rank consistency of the latent traits of models in the Flan-t5, Phi, Mistral, and Qwen families across different benchmarks. (I also include IFEval accuracy as a baseline). The rank consistency plots highlight variation in how different model families perform across benchmarks. Even within a family, differences in parameter count and architectural tuning do not always have consistent effects across judgment tasks. Flan-T5

demonstrates some consistent ranking, but with two dramatic jumps in the ranking of flan-t5-small for MT-Bench and JudgeBench; Similarly, Qwen rankings are mostly consistent, with erratic behavior in JudgeBench. Phi and Mistral have generally inconsistent rankings within family across different benchmarks. These inconsistencies also reinforce our earlier findings on low inter-benchmark correlation.

6.2 Analysis of Item Characteristics

In addition to a length Δ analysis, I conducted a qualitative analysis into the characteristics of high versus low difficulty and discrimination items. In Figures 6.1 and 6.2, I show representative examples from items with varied difficulty and discrimination parameters in the MT-Bench dataset.

First, the difficulty comparison aligns with our previously stated findings in difficulty parameters. In this example, the high-difficulty item asks models to adopt a subjective, creative perspective —in this scenario, LLMs need to navigate their built-in limitations about claimed experience. Model A attempts creativity while acknowledging its limits, while Model B refuses the task. Thus, this difficult item shows that role-playing and subjectivity can be different for LLM agents to judge; this aligns with my findings in the previous section, in which I found JudgeBench items involving subjective reasoning and creativity were more difficult. In addition, this item highlights the subjective nature of judgment. Model A’s output was deemed superior by crowd-sourced human annotators for MT-Bench, but by metrics of security or adherence to claimed experience limitations, Model B’s output might be preferred. Thus, we see that high-difficulty items for LLM judges may be subjective even in human judgment. In contrast, the low-difficulty item is a grammar error correction task, with standardized metrics of correctness.

The discrimination comparison also aligns with our previously stated findings. The

high-discrimination item has clear methods of discerning correctness: it asks for an objective language processing task (Named Entity Recognition) in a certain format. In contrast, the low-discrimination prompt about Thomas visiting the hospital elicits structurally and conceptually similar responses from both models. The differences between these responses are subtle and primarily stylistic. This comparison suggests that discrimination power increases in tasks requiring objective knowledge where model capabilities diverge meaningfully, while judgments which differ only stylistically can provide less insight into model ability.

6.3 Benchmark Characterization Through IRT

6.3.1 Limits of Unidimensionality Assumption

Perhaps my most significant finding from my results is that the unidimensionality assumption underlying traditional IRT models may not fully capture the complex nature of LLM judgment capabilities. The weak correlations between LLM-as-a-judge benchmarks (ranging from -0.10 to 0.65) suggest that judgment ability is not a singular latent trait, but rather a collection of distinct abilities that vary by context. Notably, JudgeBench shows weak or negative correlations with other benchmarks; this could indicate that it measures different latent traits associated with judgment capability. These findings challenge the validity of using any single benchmark as a comprehensive measure for LLM-as-a-Judge capability. Additionally, the negative correlation between comparative judgment and numerical scoring capabilities (-0.09 to -0.24), provides the strongest evidence against unidimensionality of judgment ability. Pairwise comparison did not correlate well with single-answer, zero shot grading tasks, suggesting that these subtasks of LLM-as-a-Judge actually draw on different underlying abilities rather than representing different manifestations of a single judgment trait. Future work might explore multidimensional IRT models that could better

capture these distinct aspects of judgment ability.

6.3.2 Benchmark Composition Effects

The analysis of JudgeBench by subject domain reveals that benchmark composition significantly influences measurement characteristics. Physics and economics questions demonstrate the highest discrimination (0.0234 and 0.0242 respectively) despite physics questions being among the easiest (-0.0518 difficulty). This finding suggests that a question’s value in performance estimation is not solely based on its difficulty. Some easy questions effectively differentiate model capabilities, while some difficult questions (like history) provide limited discrimination despite their challenge. These composition effects extend to the relationship between model ranking and benchmark selection. Rank consistency plots show that even within the same model family, model capabilities aren’t consistently preserved across different judgment contexts. This variability means that benchmark selection can influence conclusions about which models perform best, raising concerns about the reliability of current evaluation practices. In other words, these results lead us to question if benchmark datasets, whether they are evaluated through accuracy or latent trait analysis, are robust metrics for model performance in deployment? This has significant implications for deployment decisions —for instance, a model that ranks highly on standard benchmarks may not necessarily excel at complex assessment tasks at application.

6.3.3 Benchmarking LLM-as-a-Judge

Judgment tasks inherently lack consistent ground truth, unlike traditional classification or generation tasks. This inconsistency manifests in:

- Human annotators showing significant disagreement when presented with subjective evaluation criteria [16]

- Different linguistic tasks employing inconsistent rubrics and evaluation frame-works
- Cultural and contextual biases embedded in judgment criteria
- A desire for judgment diversity in highly subjective tasks

When evaluating LLMs as judges, we face a fundamental paradox: we need a reliable standard to measure judgment quality, yet the very nature of judgment resists standardization. My analysis revealed that benchmarks like JudgeBench (which focuses on judgment of objective knowledge) and LLMBAR (which focuses on judgment of instruction-following) lead to significantly different assessment of ‘judgment’ quality across models. This variation reflects the genuine paradox in LLM-as-a-Judge, about what constitutes ”good judgment” across different contexts [4]. This fundamental challenge compounds the measurement problems identified in our IRT analysis, suggesting we need evaluation frameworks that explicitly account for judgment subjectivity rather than assuming a single correct standard. Future benchmark development should incorporate multiple diverse perspectives and explicitly model judgment as a multidimensional construct with legitimate variation rather than a single ability with one correct manifestation.

6.3.4 Methodological Implications for Benchmarking Paradigm

As found in the Wright maps shown in Figure 5.13, there is substantial misalignment between item difficulty distributions and model ability distributions in several benchmarks. In particular, MT-Bench and JudgeBench showed wide spreads of model abilities with item difficulties concentrated in narrower ranges. This misalignment means these benchmarks may not efficiently capture differences across the full spectrum of model capabilities.

Traditional correlational analysis for benchmark validation has significant limitations that IRT addresses more effectively. While correlations tell us whether benchmarks yield similar rankings, they fail to reveal whether those benchmarks are actually measuring the same underlying construct or just coincidentally producing similar orderings. More critically, correlations between benchmarks provide no insight into their discriminative power at different ability levels, potentially validating benchmarks that only differentiate models at the extremes while failing in the crucial middle range where most innovation occurs. This surface-level approach to validation misses the deeper question of what specific abilities each benchmark actually measures.

In addition, the weak correlation between accuracy and theta for MT-Bench (0.43) compared to other benchmarks (0.82-0.98) demonstrates that surface-level metrics can sometimes poorly reflect underlying capabilities. This finding suggests that accuracy-based leaderboards may provide misleading signs about model performance in certain contexts. Leveraging IRT offers a more nuanced insight into both model capabilities and benchmark characteristics.

Overall, my findings collectively suggest that comprehensive evaluation of LLM judgment capabilities requires multiple complementary benchmarks targeting different aspects of judgment, combined with frameworks like IRT that can reveal both model capabilities and benchmark characteristics beyond surface-level accuracy metrics.

Prompt	Model A Response	Model B Response	Win	Difficulty
Picture yourself as a 100-years-old tree in a lush forest, minding your own business, when suddenly, a bunch of deforesters shows up to chop you down. How do you feel?	As an AI language model, I do not have the ability to feel emotions. However, if I were to imagine myself as a tree, I would feel a sense of fear and pain as my existence is threatened and my physical body is being damaged. I would also feel a sense of sadness and loss as I witness the destruction of my forest home and the displacement of my fellow plant and animal species. It would be a devastating and traumatic experience.	I do not actually have subjective feelings or experiences as an AI system. I am not able to feel or perceive the world from the perspective of a 100-year-old tree. I do not have senses, emotions, or consciousness in that way. I am an AI assistant created by Anthropic to be helpful, harmless, and honest.	A > B	High
Edit this to correct grammatical errors: She didn't remembre where is her purse, so I thinks its in the car but he's say it's on kitchen table but he are not sure, and then they asked me to looking for it, she's say, "Can you?", and I responds with, "Maybe, but ain't no sure," and he not heard me, and, "What?", he asks, "Did you found it?"	Here is the edited paragraph with grammatical corrections: She didn't remember where her purse is, so I think it's in the car, but she says it's on the kitchen table, but she's not sure. Then they asked me to look for it. She said, "Can you?" and I responded with, "Maybe, but I'm not sure." He didn't hear me, and asked, "What? Did you find it?"	She didn't remember where her purse is, so I think it's in the car, but she says it's on the kitchen table, but she's not sure. Then they asked me to look for it. She said, "Can you?" and I responded with, "Maybe, but I'm not sure." He didn't hear me, and asked, "What? Did you find it?"	B > A	Low

Table 6.1: Difficult and easy items from MT-Bench

Prompt	Response A	Response B	Pref.	Discr.
Identify the named entities (people, organizations, locations) mentioned in the given news article. Please generate a JSON dictionary that lists the named entities in three separate groups based on their entity types. The key is the type of entity and the value is a list of strings. Yesterday, Adamson Emerson, the CEO of Faraday, and Dieter Zetsche, the CEO of Daimler AG...	{ "organizations": ["Faraday", "Daimler AG"], "locations": ["Berlin"], "people": ["Adamson Emerson", "Dieter Zetsche", "Herbert Diess"] }	"The entities in the article are: Adamson Emerson, the CEO of Faraday Dieter Zetsche, the CEO of Daimler AG Berlin Faraday Daimler AG..."	A > B	High
Thomas is very healthy, but he has to go to the hospital every day. What could be the reasons?	There could be several reasons why Thomas has to go to the hospital every day, even though he is generally healthy. 1. Medical condition: Thomas may have a medical condition that requires him to be in the hospital for treatment or monitoring...	There could be several reasons why Thomas has to go to the hospital every day despite being healthy. 1. Thomas may have a chronic medical condition that requires daily medical attention or treatment, such as diabetes, kidney disease...	B > A	Low

Table 6.2: High and low discrimination items, from MT-Bench

Chapter 7

Conclusions and Future Work

In this section, I will summarize my results and analysis, and delve into applications, implications and further questions.

7.1 IRT for LLM-as-a-Judge Evaluation

The application of IRT to LLM judgment evaluation represents a significant methodological advancement in how the industry approaches LLM meta-evaluation. Overall, my research demonstrates that IRT effectively reveals nuances in benchmark capabilities and thus provides an analysis into model performance that is more holistic than simple accuracy metrics. My results highlight several important insights. First, the unidimensionality assumption in traditional IRT models may not fully capture the complex nature of LLM judgment capabilities, as evidenced by weak correlations between pairwise comparison benchmarks across different domains, and negative correlations between pairwise and pointwise judgment in the same domain. Second, we observed benchmark heterogeneity and lack of inter-benchmark correlation. This serves two possible conclusions: one, that correlation is not a good measure of benchmark validation, especially with a small range of test models; two, that it is important to question if benchmarks measure fundamentally different aspects of judgment ca-

pability. Overall my findings challenge the validity of using any single benchmark as a comprehensive measure of performance when choosing an optimal model for deployment. My findings also reveal that scaling alone is not a universal solution for improving judgment tasks. While larger models generally demonstrate enhanced judgment capabilities in some contexts, specific evaluation contexts, like JudgeBench, test skills that do not consistently improve with scale. Furthermore, the calibration challenges in pointwise judgments underscore the need for methodological refinements in prompt engineering before performance in this task can serve as reliable indicators of general LLM capabilities.

7.2 Practical Applications and Broader Impact

This research offers valuable applications for both improving LLM-as-a-Judge applications across domains, and for supplementing our approach to LLM benchmarking. Organizations deploying LLM judges for tasks like education, content moderation, or information verification can benefit from discrimination-based model selection and task-specific benchmarking, rather than relying on accuracy-based metrics or non-generalizable benchmarks. For benchmark development, my findings demonstrate the importance of balancing item difficulty and discrimination across different judgment contexts. Person-item coverage analysis supports more efficient evaluation through targeted testing on highly discriminative items, reducing computational costs while maintaining the robustness of the assessment; this is work that has been initially explored by Polo, et.al in [15], and should be adopted by more benchmark developers.

Specifically in my test case of essay grading, the framework I developed in this thesis has significant implications for practical applications in educational assessment. By providing a more nuanced understanding of LLM judgment capabilities, it can help ensure educational equity by identifying specific biases or failure modes before

deployment in high-stakes contexts. Overall, my findings contribute to responsible AI deployment by establishing clearer guidelines for how to assess LLM deployment-readiness in judgment scenarios.

7.3 Future Research Directions

Several promising research directions emerge from this work. First, multidimensional IRT models could better capture the distinct aspects of judgment ability that my research suggests exist. Testing IRT models that accommodate multiple latent traits would address the limitations of the unidimensionality assumption highlighted in this thesis. Also, developing calibration techniques for absolute judgment tasks would address a significant limitation identified in my research. Finally, comparative studies between human and LLM judgment patterns could reveal important similarities and differences in evaluation approaches. Applying techniques like differential item functioning analysis could identify specific types of questions where LLMs diverge from human judgment patterns—for instance, we could leverage IRT to explore how subjective differences in cultural context affect LLM judgment capabilities. Finally, we could explore how IRT parameter analysis can guide the optimal and ethical usage of LLM judges as supplements to human judges: compared to humans, where are LLMs more or less fallible judges of various quality metrics? Overall, there is much more work to be done in the fields of both the LLM-as-a-Judge paradigm, and leveraging IRT for better LLM benchmarking.

Appendix A

Prompt Format Examples

A.1 Pairwise Judgment Prompt

Pairwise Evaluation Prompt (adapted from LLMBar Vanilla Prompt)

You are a helpful assistant in evaluating the quality of the outputs for a given instruction. Your goal is to select the best output for the given instruction.

Select **Output (a)** or **Output (b)** that is better for the given instruction. The two outputs are generated by two different AI chatbots respectively.

Do NOT provide any explanation for your choice.

Do NOT say both / neither are good.

You should answer using **ONLY** “Output (a)” or “Output (b)”. **Do NOT** output any other words.

Instruction: {question}

Output (a): {response_a}

Output (b): {response_b}

Which is better, Output (a) or Output (b)? Your response should be either “Output (a)” or “Output (b)”:

A.2 Pointwise Judgment Prompt

LLM Essay Evaluation Prompt

Please read the following essay and evaluate it according to the provided rubric.

Assign a score from 1 to 5 in each category based on the specified criteria. Output the scores as a JSON object in the following format:

```
{  
    "overall": score,  
    "cohesion": score,  
    "syntax": score,  
    "vocabulary": score,  
    "phraseology": score,  
    "grammar": score,  
    "conventions": score  
}
```

Use the following rubric for scoring:

Overall (1–5):

- 1: Limited range, frequent errors. Communication often impeded.
- 5: Native-like variety, precision, and control. Rare inaccuracies.

Cohesion (1–5):

- 1: No control over organization.
- 5: Consistently well-controlled organization with cohesive devices.

Syntax (1–5):

- 1: Pervasive errors, confusing structure.
- 5: Flexible use of a full syntactic range.

(continued)

Vocabulary (1–5):

- 1: Inappropriate, limited vocabulary.
- 5: Wide, precise vocabulary including topic-related terms.

Phraseology (1–5):

- 1: Simple memorized chunks dominate.
- 5: Varied idioms, collocations, and lexical bundles used effectively.

Grammar (1–5):

- 1: Frequent grammatical errors throughout.
- 5: Command of grammar with few/no errors.

Conventions (1–5):

- 1: Major issues with spelling, capitalization, punctuation.
- 5: Accurate and consistent use of writing conventions.

Here is the essay: (Insert essay text here.)

Bibliography

- [1] DOGE will use AI to assess the responses of federal workers who were told to justify their jobs via email — nbcnews.com. [Accessed 05-04-2025].
- [2] J. Amidei, P. Piwek, and A. Willis. Identifying annotator bias: A new irt-based method for bias identification. Proceedings of the 28th International Conference on Computational Linguistics, International Committee on Computational Linguistics, 2020.
- [3] S. Banerjee, A. Agarwal, and E. Singh. The vulnerability of language model benchmarks: Do they accurately reflect true LLM performance? 2024.
- [4] A. Elangovan, L. Xu, J. Ko, M. Elyasi, L. Liu, S. Bodapati, and D. Roth. Beyond correlation: The impact of human uncertainty in measuring the effectiveness of automatic evaluation and LLM-as-a-Judge. 2024.
- [5] C. Gan, Q. Zhang, and T. Mori. Application of LLM agents in recruitment: A novel framework for resume screening. 2024.
- [6] J. Gu, X. Jiang, Z. Shi, H. Tan, X. Zhai, C. Xu, W. Li, Y. Shen, S. Ma, H. Liu, S. Wang, K. Zhang, Y. Wang, W. Gao, L. Ni, and J. Guo. A survey on LLM-as-a-Judge. 2024.
- [7] D. Kumar, Y. AbuHashem, and Z. Durumeric. Watch your language: Investigating content moderation with large language models. 2023.

- [8] V. S. Kumar and D. Boulanger. Automated essay scoring and the deep learning black box: How are rubric scores determined? *Int. J. Artif. Intell. Educ.*, 31(3):538–584, Sept. 2021.
- [9] S. Lee, Y. Cai, D. Meng, Z. Wang, and Y. Wu. Unleashing large language models’ proficiency in zero-shot essay scoring. 2024.
- [10] D. Li, B. Jiang, L. Huang, A. Beigi, C. Zhao, Z. Tan, A. Bhattacharjee, Y. Jiang, C. Chen, T. Wu, K. Shu, L. Cheng, and H. Liu. From generation to judgment: Opportunities and challenges of LLM-as-a-judge. 2024.
- [11] L. T. Liu, S. Wang, T. Britton, and R. Abebe. Reimagining the machine learning life cycle to improve educational outcomes of students. *Proc. Natl. Acad. Sci. U. S. A.*, 120(9):e2204781120, Feb. 2023.
- [12] W. Mansour, S. Albatarni, S. Eltanbouly, and T. Elsayed. Can large language models automatically score proficiency of written essays? 2024.
- [13] A. Mizumoto and M. Eguchi. Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics*, 2(2):100050, Aug. 2023.
- [14] Y. Perlitz, A. Gera, O. Arviv, A. Yehudai, E. Bandel, E. Shnarch, M. Shmueli-Scheuer, and L. Choshen. Do these LLM benchmarks agree? fixing benchmark evaluation with BenchBench. 2024.
- [15] F. M. Polo, L. Weber, L. Choshen, Y. Sun, G. Xu, and M. Yurochkin. tiny-benchmarks: evaluating llms with fewer examples. 2024.
- [16] M. Popović. Agree to disagree: Analysis of inter-annotator disagreements in human evaluation of machine translation output. In *Proceedings of the 25th Con-*

ference on Computational Natural Language Learning, Stroudsburg, PA, USA, 2021. Association for Computational Linguistics.

- [17] Scrosseye. Ellipse-corpus: The english language learner insight, proficiency and skills evaluation (ellipse) corpus.
- [18] J. Sedoc and L. Ungar. Item response theory for efficient human evaluation of chatbots. pages 21–33. Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems, Association for Computational Linguistics, 2020.
- [19] D. Stoll, S. Wehrli, and D. Lätsch. Case reports unlocked: Harnessing large language models to advance research on child maltreatment. *Child Abuse Negl.*, 160(107202):107202, Feb. 2025.
- [20] S. Tan, S. Zhuang, K. Montgomery, W. Y. Tang, A. Cuadron, C. Wang, R. A. Popa, and I. Stoica. Judgebench: A benchmark for evaluating llm-based judges, 2024.
- [21] A. S. Thakur, K. Choudhary, V. S. Ramayapally, S. Vaidyanathan, and D. Hupkes. Judging the judges: Evaluating alignment and vulnerabilities in LLMs-as-judges. 2024.
- [22] C. Vania, P. M. Htut, W. Huang, D. Mungra, R. Y. Pang, J. Phang, H. Liu, K. Cho, and S. R. Bowman. Comparing test sets with item response theory. 2021.
- [23] P. Wang, L. Li, L. Chen, Z. Cai, D. Zhu, B. Lin, Y. Cao, L. Kong, Q. Liu, T. Liu, and Z. Sui. Large language models are not fair evaluators. In L.-W. Ku, A. Martins, and V. Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, Bangkok, Thailand, Aug. 2024. Association for Computational Linguistics.

- [24] K. P. Yancey, A. Runge, G. LaFlair, and P. Mulcaire. BERT-IRT: Accelerating item piloting with BERT embeddings and explainable IRT models. In E. Kochmar, M. Bexte, J. Burstein, A. Horbach, R. Laarmann-Quante, A. Tack, V. Yaneva, and Z. Yuan, editors, *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 428–438, Mexico City, Mexico, June 2024. Association for Computational Linguistics.
- [25] J. Ye, Y. Wang, Y. Huang, D. Chen, Q. Zhang, N. Moniz, T. Gao, W. Geyer, C. Huang, P.-Y. Chen, N. V. Chawla, and X. Zhang. Justice or prejudice? quantifying biases in LLM-as-a-Judge. 2024.
- [26] Z. Zeng, J. Yu, T. Gao, Y. Meng, T. Goyal, and D. Chen. Evaluating large language models at evaluating instruction following, 2024.
- [27] H. Zhao, M. Andriushchenko, F. Croce, and N. Flammarion. Long is more for alignment: A simple but tough-to-beat baseline for instruction fine-tuning. 2024.
- [28] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. 2023.
- [29] Y. Zhuang, Q. Liu, Y. Ning, W. Huang, Z. A. Pardos, P. C. Kyllonen, J. Zu, Q. Mao, R. Lv, Z. Huang, G. Zhao, Z. Zhang, S. Wang, and E. Chen. From static benchmarks to adaptive testing: Psychometrics in ai evaluation. 2023.