

Consensus without Accuracy: Investigating LLM’s Ability to Recover Item Difficulty Using Paired Comparisons

Michael L. Chrzan

Center for Educational Data Science
and Innovation,
University of Maryland

Benjamin W. Domingue

Stanford University *

Abstract

Large language models (LLMs) offer a scalable way to estimate item difficulty before respondent data are available, but it is unclear whether they can provide reliable estimates, either alone or in aggregate. We evaluate seven frontier LLMs on pairwise comparisons of items from seven datasets with respondent-calibrated one-parameter item response theory (1PL IRT) difficulty estimates. We aggregate their judgments using a hierarchical Bayesian Bradley–Terry model that estimates shared item strengths while accounting for model-specific discrimination and presentation-order bias. The pooled rankings recover a meaningful but incomplete difficulty signal: standardized hierarchical Bradley–Terry strengths correlate moderately with IRT difficulty overall (Spearman’s $\rho = 0.400$, $p < .001$), with dataset-level correlations ranging from 0.250 to 0.591. This heterogeneity persists despite high inter-model agreement (Krippendorff’s $\alpha = 0.834$ –0.917) and generally high within-model triplet transitivity, demonstrating that consensus and internal consistency do not guarantee external psychometric validity. The hierarchical model also reveals substantial variation in rater discrimination and position bias, while in-sample predictive accuracy ranges from 0.732 to 0.874 across datasets. These findings support LLM pairwise comparisons as a scalable source of provisional item ordering, but not as a replacement for respondent-based calibration.

1 Introduction

Item calibration - the process of estimating psychometric properties such as item difficulty - is central to educational assessment but resource-intensive, typically requiring large samples of student response data and extended field-testing periods creating barriers for low-resource testing programs, early-stage item development, and rapid test as-

sembly contexts (Jiang et al., 2016). Recent advances in large language model (LLM) capabilities raise an intriguing question: can item difficulty be meaningfully estimated directly from item text, without recourse to any response data? Such an approach would offer immediate practical utility for item writers and test developers seeking a scalable first-pass difficulty signal for item screening, ordering, or rough-cut form assembly before formal calibration. Prior work has explored LLMs for automated item generation and scoring (von Davier et al., 2022), but their deployment as a calibration proxy for estimating the psychometric difficulty of items never yet administered remains less examined as does the extent to which paired comparisons aid in that effort.

Recent work by Wu et al. (2024) introduced concept-guided chain-of-thought (CGCoT) prompting as a generalizable framework for scoring texts on latent constructs. CGCoT pairs a sequence of researcher-crafted prompts, each probing a different facet of the latent construct of interest, with LLM-based pairwise comparisons aggregated into a continuous score via the Bradley-Terry (BT) model (Bradley and Terry, 1952). Originally validated for scaling political affect in social media, CGCoT’s structure maps naturally onto item difficulty estimation: an item’s text can be characterized through prompts targeting cognitive demand and required ability level, with pairs of items compared by an LLM to establish an ordinal difficulty ranking without any examinee data.

Further, pairwise comparison methods offer a natural interface for eliciting difficulty judgments from LLMs. Rather than requiring absolute estimates - which are psychometrically undefined without a response scale and shown to be unreliable for LLMs (Licht et al., 2025) - pairwise comparisons ask a model to make relative judgments, which can then be aggregated into difficulty rankings. This approach also allows reliability to be assessed via ex-

*Correspondence: mlchrzan@umd.edu

tensions of Bradley-Terry methods, such as a Hierarchical Bradley-Terry Model (Böckenholt, 2001), and internal consistency metrics such as triplet transitivity. However, whether LLM-derived BT scores actually align with psychometrically validated difficulty estimates from Item Response Theory (IRT) remains an open empirical question.

This study makes three contributions to answering these questions. First, we provide a multi-model, multi-dataset evaluation of the extent to which LLM pairwise judgments recover respondent-calibrated item difficulty. Second, we aggregate judgments with a hierarchical Bayesian Bradley-Terry model that estimates a shared item-difficulty ranking while separating model-specific discrimination from presentation-order bias, which quantifies uncertainty in the pooled ranking and reveals consequential differences among LLM raters that a simple consensus score would obscure. Lastly, we distinguish three often-conflated properties of LLM judgments: within-model transitivity, inter-model agreement, and external psychometric validity. Despite generally high transitivity and uniformly high inter-model agreement, alignment with 1PL difficulty remains moderate and heterogeneous, showing that consistent LLM consensus is useful for provisional item ordering but cannot substitute for respondent-based calibration.

2 Related Work

The foundational IRT model utilized in large-scale calibration is the 1-Parameter Logistic (1PL) model, commonly known as the Rasch model, introduced by Georg Rasch in 1960 (Rasch, Georg, 1960; Kolesnikova et al., 2026). The Rasch model conceptualizes the probability that a specific student correctly answers a specific item as a logistic function of the difference between the student’s latent ability and the item’s latent difficulty. In generalized terms, if a student’s ability exactly matches the item’s difficulty, the theoretical probability of a correct response is precisely 50%. In traditional workflows, estimating this difficulty parameter requires administering the item to highly calibrated populations whose abilities form a known distribution (Felice and Karaoz, 2024). However, when new assessment items are generated, their parameters are entirely unknown, creating what is known as the cold-start problem (Wainer et al., 2000). If LLMs or machine learning ensembles can accurately predict the difficulty parameter based solely

on item content, these synthetic estimates can be used to jump-start the IRT calibration process. This permits the immediate deployment of new items in live environments and vastly reduces the volume of human response data required to achieve statistical stability. While prior work has examined the ability of other families of machine learning models to predict item difficulty (Felice and Karaoz, 2024; Kapoor et al.), only Kolesnikova et al. (2026) used pairwise comparisons as a methodology and their conclusions were limited by the singular bank of items.

An intuitive methodology for applying LLMs to predict item difficulty is direct absolute estimation. In this paradigm, a language model is provided with the text of an item—including the question stem, correct answer, and distractors—and is instructed via a zero-shot prompt to output a numeric difficulty score or a categorical difficulty level (Kolesnikova et al., 2026). This approach relies on the implicit assumption that the LLM contains a holistic, internally calibrated representation of cognitive complexity that perfectly mirrors human populations. Recent literature indicates that this assumption is fundamentally flawed and suggests that direct absolute judgment by LLMs is fraught with inconsistencies and limitations (Licht et al., 2025; Li et al.; Wang et al., 2026).

To overcome the inherent limitations of direct absolute prediction with LLMs, researchers have achieved state-of-the-art results by altering the LLM’s role within the predictive workflow. Rather than acting as a final psychometric judge, the LLM is repositioned as a highly advanced feature extractor or simulator (Hoyl; Kolesnikova et al., 2026; Scarlatos et al.). In this paradigm, the model is prompted to evaluate specific linguistic, structural, and pedagogical characteristics of the item text. These extracted qualitative features are then fed into traditional supervised machine learning algorithms to predict the final difficulty parameter. By decoupling the semantic extraction (handled by the LLM) from the regression task (handled by ensemble trees like Random Forests and Gradient Boosting algorithms), the predictive system successfully sidesteps the LLM’s inability to globally scale difficulty (Felice and Karaoz, 2024).

While feature-based ensembles and synthetic simulations provide robust solutions, extracting dozens of qualitative features per item can be computationally expensive, highly prompt-intensive, and strictly domain-dependent. An alternative,

highly effective psychometric strategy leverages the language model’s fundamental capacity for relative reasoning: pairwise (paired) comparisons. While assigning an isolated assessment item a difficulty score of 7.8 out of 10 requires a stable internal standard, precise calibration against prior examples, and a consistent, unshifting interpretation of what the numerical scale is meant to represent (DiGiuseppe and Flynn, 2026), pairwise comparisons ask a significantly simpler question: Which of these two items is more difficult? This localized, relative judgment completely removes the need for a global numerical frame. It severely mitigates anchoring bias, allows for high-resolution discrimination between items of similar complexity, and demonstrates remarkable stability across various LLM architectures and model sizes (DiGiuseppe and Flynn, 2026). By transforming the task from an absolute regression to a binary relative choice, the system extracts the model’s highest-confidence judgments and has been shown to be highly effective in recent work examining mathematics tasks (Kolesnikova et al., 2026).

However, when utilizing different LLMs or deploying varying prompts to generate large-scale pairwise comparisons the reliability of the artificial "judges" inherently fluctuates. Inspired by the established Crowd-BT algorithm (Chen et al., 2013)—which automatically estimates the underlying reliability of human crowdsourcing workers and flags annotators with low quality for removal—researchers have introduced the BT- σ model, which extends the traditional Bradley-Terry framework specifically for the "LLM-as-a-jury" setting (Qian et al., 2026). The BT- σ Bayesian model introduces a specific discrimination parameter (σ_k) for each individual LLM judge. This architectural enhancement allows the model to jointly infer the global item difficulty rankings while simultaneously weighting the statistical reliability of each judge using pairwise comparisons alone. If a specific language model frequently outputs comparisons that contradict the consensus of the optimized matrix, its parameter is mathematically penalized, permanently down-weighting its future influence on the latent difficulty scores. This creates a highly resilient, noise-resistant ecosystem capable of safely fusing predictions from advanced, proprietary models with smaller, open-weight models, resulting in a single, cohesive, and highly accurate difficulty continuum. Building on this work, we apply rater-aware Bayesian pairwise aggrega-

tion to educational item calibration, jointly modeling LLM-specific discrimination and presentation-order/position bias and validating the resulting pooled rankings against respondent-calibrated 1PL IRT difficulty across seven datasets.

3 Data

The items we examine were drawn from seven datasets in the Item Response Warehouse (IRW) (Domingue et al., 2025). Using their IRW tags, they are: (1) frac20 - a mathematics/fractions dataset (Tatsuoka, 2002) - and six literacy-focused datasets from three studies: (2) gilbert_meta_2 (Kim et al., 2023), (3) gilbert_meta_7 and (4) gilbert_meta_8 (Kim et al., 2024), as well as (5) gilbert_meta_102, (6) gilbert_meta_103, and (7) gilbert_meta_104 (Relyea et al., 2025). These datasets include items with established 1PL IRT difficulty calibrations, providing a psychometric ground truth for validation. They were selected using the IRW’s filtering capabilities to ensure that the datasets have item text available, the age range is ≤ 18 , the construct is cognitive/educational, the measurement tool is test, the item format is Likert/constructed response, and language is English.

Seven proprietary LLM configurations served as raters: (1) GPT-5.4 Mini, (2) GPT-5.4, (3) GPT-5.6-Luna, (4) GPT-5.6-Sol, (5) Gemini 3 Flash Preview, (6) Gemini 3.7 Flash, and (7) Gemini 3.1 Pro Preview. The panel was selected to vary along three dimensions relevant to the robustness of LLM-based measurement. First, it includes two independently developed model families, OpenAI and Google, reducing the extent to which the findings depend on the idiosyncrasies of a single provider. Second, it includes different tiers and variants within each family, allowing us to examine whether pairwise judgments remain consistent across models marketed for different capability, latency, and cost tradeoffs. Third, the numbered releases within both families broaden the evaluation beyond a single model generation or temporal snapshot. Because the providers do not disclose directly comparable parameter counts, training corpora, or complete architectural details for these proprietary systems, we treat each configuration as a distinct rater rather than interpreting nominal tier or version differences as quantitative measures of model size or capability.

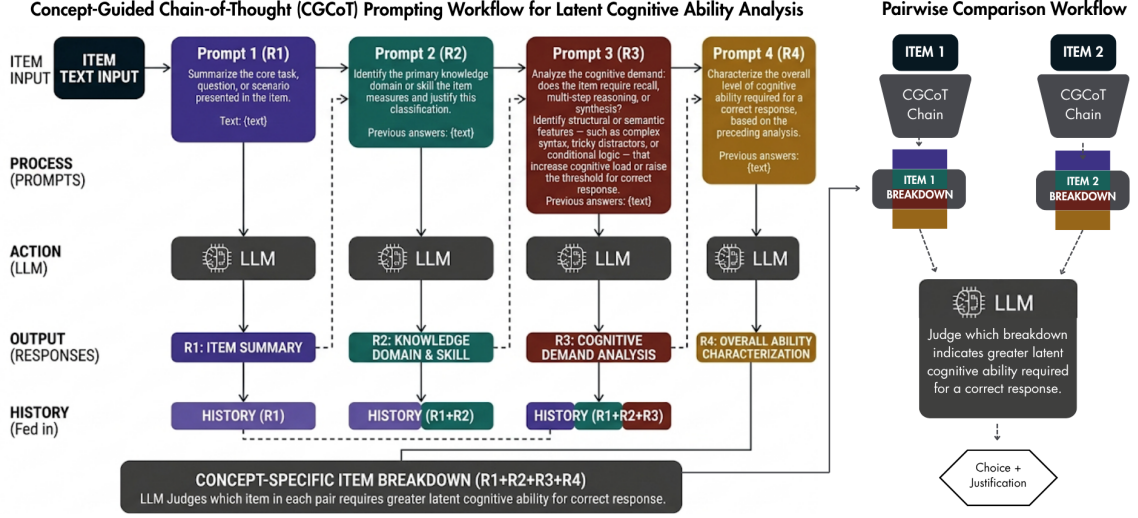


Figure 1: Concept-Guided Chain-of-Thought (CGCoT) Prompting for Item Difficulty Assessment. Each item is processed through a sequence of four prompts, each targeting a different facet of the latent construct of interest, here item difficulty. The LLM’s responses are concatenated and fed into the next prompt, culminating in a comprehensive analysis of the item’s difficulty before pairwise comparisons are made.

4 Methods

4.1 pairadigm and Paired Comparisons

All pairwise comparison workflows—CGCoT prompt chaining, comparison management, and BT model estimation—were implemented using pairadigm (Chrzan, Michael L.), an open-source Python package for using LLMs to measure scalar constructs. The package includes code for generating pairs from the set of items and applying CGCoT prompts to construct item breakdowns. Using evidence from Wu et al. (2024) that beyond 10 pairs per item yielded minimal improvement, where possible we limit the number of possible pairs to 10 per item for compute and cost considerations. The seven configurations listed previously were evaluated as separate comparison judges, enabling rater-specific estimates of consistency, discrimination, and position bias.

For each item, a four-prompt CGCoT chain was administered sequentially to the LLM to target cognitive item difficulty. After the LLM answers each prompt, its response is concatenated to all its previous responses and fed into the next prompt until at the end an entire breakdown of the item is created, as displayed in Figure 1. Prompts are also listed out in Appendix A.

The resulting concept-specific breakdown for each item was used to generate pairwise comparisons across all item pairs within each dataset. Based on the cumulative output from these four steps applied to an item, the LLM judged which

item in a pair required greater latent cognitive ability for correct response.

4.2 (Hierarchical) Bradley-Terry Models

Individual BT models were fit to each LLM’s pairwise preference data for each dataset to produce a continuous BT difficulty score for each item. Triplet transitivity, or the proportion of triplets (A, B, C) in which the model’s pairwise judgments were logically consistent, was computed as an additional consistency check. Inter-rater agreement across LLMs was estimated using Krippendorff’s Alpha (Hayes and Krippendorff) to assess the consistency across models for a given IRW dataset. Finally, convergent validity was assessed via Spearman rank correlations computed between each model’s BT scores and the corresponding 1PL difficulty parameters, stratified by model and dataset (Spearman). Relationships among reliability, inter-rater agreement, and Spearman correlations were examined to assess the relationships between internal consistency and external validity.

A Hierarchical Bradley-Terry model was also fit to assess each LLM’s discrimination and position bias. We formulate the model as follows:

$$\text{logit}P(y_{ijk} = 1) = \alpha_k(\lambda_i - \lambda_j) + \gamma_k$$

where

- $y_{ijk} = 1$ indicates that model k selects the first presented item, i , over the second-presented item, j ,

- λ_i and λ_j are the difficulty parameters for items i and j ,
- α_k is the discrimination parameter for participant k ,
- γ_k is the position bias for participant k .

We structure the data such that every unordered pair is evaluated in both orientations, which balances item identity across first and second positions and makes γ_k estimable from direct orientation reversals. Further, we enforce that (1) the sum of λ_i across all items is zero, and (2) the geometric mean of α_k across all models is one, which identifies the global $\alpha - \lambda$ scale without choosing an arbitrary reference model. We discuss the full model details in the Appendix, including the hierarchical priors for λ_i , α_k , and γ_k in Appendix B as well as model diagnostic checks in Appendix C and sensitivity analysis in Appendix D.

5 Results

5.1 Recovery of Respondent-Calibrated Difficulty

Across the 179 items, standardized HBT item strengths were moderately associated with respondent-calibrated 1PL difficulty (Spearman’s $\rho = 0.400$, $p < .001$). As Figure 2 and Table 1 show, however, this relationship varied substantially by dataset. The strongest association occurred for frac20 ($\rho = 0.591$, $p = .006$). Positive associations were also observed for gilbert_meta_102 ($\rho = 0.449$, $p = .006$), gilbert_meta_8 ($\rho = 0.423$, $p = .022$), gilbert_meta_103 ($\rho = 0.414$, $p = .025$), and gilbert_meta_7 ($\rho = 0.398$, $p = .016$). Estimates were weaker and not statistically distinguishable from zero for gilbert_meta_2 ($\rho = 0.337$, $p = .146$) and the nine-item gilbert_meta_104 dataset ($\rho = 0.250$, $p = .516$). Thus, the pooled ranking recovered a meaningful difficulty signal, but not uniformly across item sets. In particular, the strongest result arose for the mathematics dataset, whereas results varied among the six literacy datasets; the observed heterogeneity therefore does not reduce to a simple mathematics-versus-literacy distinction. The corresponding model-specific correlations, pooled across all items and disaggregated by dataset, are reported in Appendix Tables 2 and 3, respectively.

The HBT ranking also retained substantial information from the model-specific BT rankings.

Across all items, Spearman correlations between each individual model’s BT scores and HBT scores ranged from 0.545 for GPT-5.4 Mini to 0.864 for GPT-5.6 Sol. Hierarchical pooling therefore captured a broadly shared ordering without making any single LLM the reference rater.

5.2 Consistency and Inter-Model Agreement

Most model–dataset combinations exhibited high triplet transitivity, often exceeding 95% and matching transitivity observed in human judgments (Sleboda and Sokolowska). The principal exceptions involved GPT-5.4 Mini, whose transitivity was 76.2% on gilbert_meta_104, 87.9% on gilbert_meta_2, 91.4% on gilbert_meta_7, and 91.9% on gilbert_meta_102; GPT-5.4 also reached only 85.7% on gilbert_meta_104. Complete model-by-dataset estimates and triplet counts appear in Appendix Table 4. Agreement across the seven LLMs was uniformly high by conventional standards, with dataset-level Krippendorff’s α ranging from 0.834 to 0.917. These strong internal-consistency and inter-model-agreement results contrast with the moderate and heterogeneous correlations with 1PL difficulty. Transitivity, agreement among LLMs, and alignment with respondent-calibrated difficulty consequently capture distinct properties of the judgments.

5.3 Rater Effects and Model Fit

Figure 3 shows pronounced variation in both rater discrimination and presentation-order bias. Posterior median discrimination estimates ranged from 0.210 for GPT-5.4 Mini on gilbert_meta_7 (94% HDI [0.137, 0.285]) to 6.013 for Gemini 3.7 Flash on gilbert_meta_104 (94% HDI [2.399, 20.124]); the wide interval for the latter estimate reflects the small number of items. GPT-5.4 Mini generally had low discrimination on the Gilbert datasets. Position effects were likewise consequential. The most negative estimate was for GPT-5.4 Mini on gilbert_meta_104 ($\gamma = -1.927$, 94% HDI [-2.678, -1.254]), indicating a preference for the second-presented item. The largest positive estimate occurred for Gemini 3 Flash Preview on gilbert_meta_8 ($\gamma = 0.858$, 94% HDI [0.599, 1.126]), indicating a preference for the first-presented item. Because each unordered pair appeared in both orientations, these tendencies could be separated from item identity and shared item strength.

The HBT item ordering agreed with observed

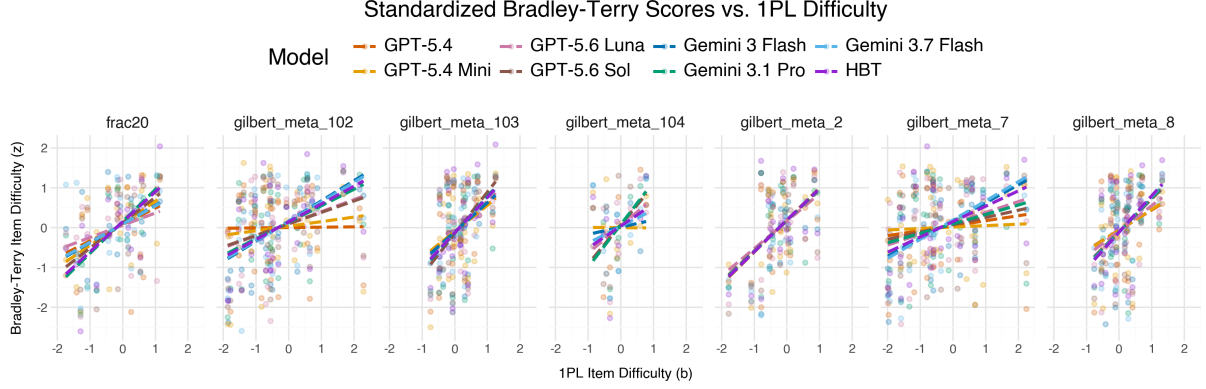


Figure 2: Standardized Bradley–Terry scores versus 1PL IRT difficulty. Each panel represents one dataset; points and dashed linear trends distinguish the seven model-specific BT estimates and the pooled HBT estimate. The pooled HBT ranking is positively associated with respondent-calibrated difficulty overall, but the strength of association varies across datasets and models.

Dataset	Items	Judgments	ρ	p	Krip. α	Rank agr.	Accuracy	Brier
frac20	20	2,324	.591	.006	.839	.732	.732	.174
gilbert_meta_2	20	2,310	.337	.146	.899	.834	.853	.097
gilbert_meta_7	36	4,200	.398	.016	.864	.784	.794	.136
gilbert_meta_8	29	3,430	.423	.022	.902	.848	.854	.101
gilbert_meta_102	36	4,340	.449	.006	.850	.760	.778	.148
gilbert_meta_103	29	3,430	.414	.025	.917	.871	.874	.087
gilbert_meta_104	9	546	.250	.516	.834	.716	.787	.136

Table 1: Dataset-level validity, agreement, and HBT fit. ρ is the Spearman correlation between standardized HBT item strength and 1PL difficulty; α is Krippendorff’s inter-model agreement coefficient. Rank agreement is the posterior mean agreement between the deterministic HBT item ordering and observed winners. Accuracy and Brier scores use the full HBT likelihood and are in-sample. Judgments count model-specific comparisons across both pair orientations.

pairwise winners in 71.6% to 87.1% of judgments across datasets. When rater-specific discrimination and position bias were included through the full likelihood, in-sample predictive accuracy ranged from 73.2% on frac20 to 87.4% on gilbert_meta_103. The clearest gain from incorporating rater effects occurred for gilbert_meta_104, where agreement based on item ordering alone was 71.6%, compared with 78.7% full-likelihood predictive accuracy. These values describe fit to the judgments used for estimation and are not measures of performance on unseen comparisons. Full sampling and predictive diagnostics are reported in Appendix C.

6 Discussion

The central result is qualified but substantively useful: LLM pairwise comparisons contain recoverable information about psychometric item difficulty, yet that information is incomplete. Pooling seven LLMs in the HBT model produced an overall correlation of 0.400 with 1PL difficulty,

demonstrating that judgments based on item text alone capture more than arbitrary preference. At the same time, the dataset-level range from 0.250 to 0.591 is too broad, and the weaker results too consequential, to treat LLM-derived rankings as interchangeable with respondent-based calibration.

The contrast between consensus and validity is particularly important. The LLMs usually produced transitive judgments and agreed with one another, but their shared ranking aligned only moderately with IRT difficulty. Agreement among artificial raters can therefore indicate a stable model-derived construct without establishing that this construct is the one measured in human response data. Models may converge on salient textual cues, curricular content, or solution complexity while missing factors that shape empirical difficulty. Increasing the number of similar raters may sharpen such a shared ordering without correcting a common construct mismatch.

Additionally, the rater parameters we estimate help explain why hierarchical aggregation is prefer-

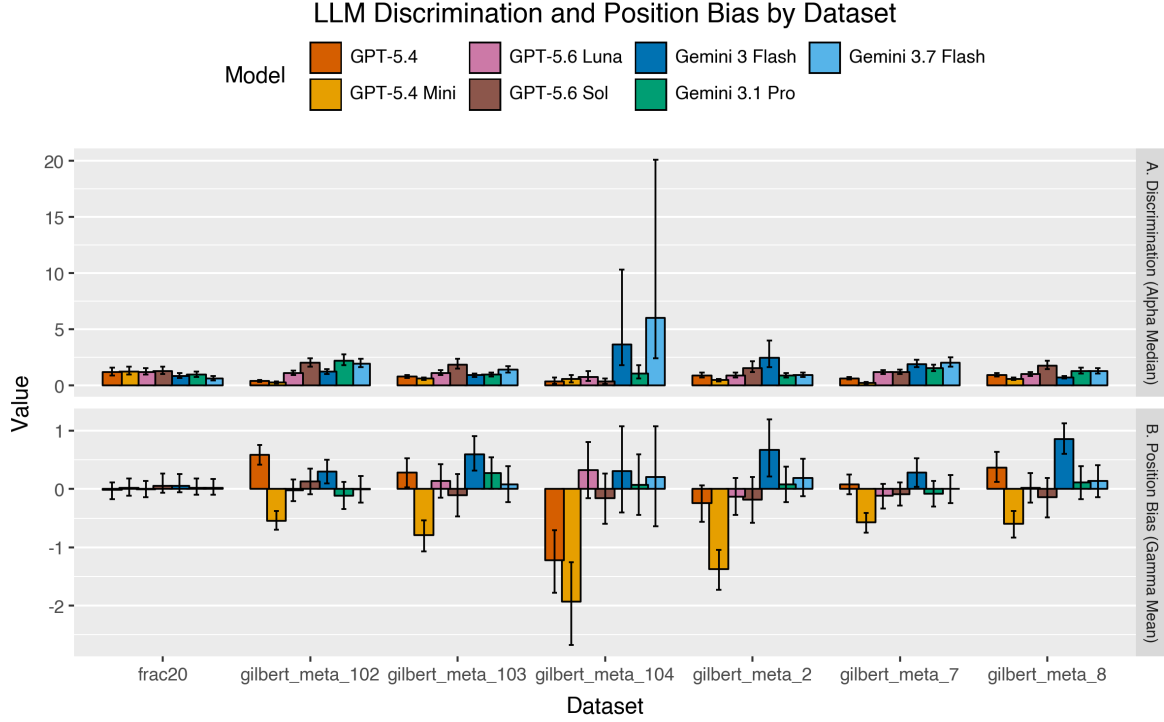


Figure 3: Discrimination and Position Bias of LLMs. The figure shows the discrimination (Alpha) and position bias (Gamma) of different LLM raters across various datasets. Higher alpha values indicate stronger rater ability to distinguish strength between items, while positive gamma values suggest a bias toward the first item shown. Error bars represent 94% highest density intervals (HDI) for each estimate. The results highlight the variability in LLM performance across datasets, with some models exhibiting higher discrimination and lower position bias, suggesting better reliability and less susceptibility to order effects in pairwise comparisons.

able to selecting one model or averaging rankings without adjustment. LLMs differed markedly in how sharply they responded to latent item-strength differences, and several showed presentation-order effects large enough for their 94% intervals to exclude zero. More generally, the results argue for treating LLMs as heterogeneous judges whose decisions require measurement, rather than as exchangeable replications of a single evaluator.

Performance differences across datasets also caution against broad domain-level claims. The one mathematics set in the IRW with item text available at time of writing showed the strongest external alignment, but the literacy datasets ranged from weak to moderate despite their related content. Item count contributes to uncertainty, especially for gilbert_meta_104, but sample size alone cannot explain the heterogeneity. Differences in item format, construct representation, response processes, and the correspondence between textual demand and empirical difficulty are plausible sources that should be tested directly as access to item text grows.

For assessment development, these findings support a limited but practical role for LLM-derived rankings. Considering our results using paired comparisons with those use direct absolute prediction, the evidence suggests LLMs can provide immediate provisional ordering for item review, rough form assembly, or prioritization before respondent data are available but they should not be used as final calibration parameters or as a substitute for field testing. Stronger applications would combine balanced presentation, multiple model families, inspection of rater-specific α_k and γ_k , and anchor items with known IRT parameters. Validation should ultimately be conducted on held-out comparisons and new item sets, with respondent calibration retained as the criterion.

Limitations

This study has several limitations. Our analysis is restricted by design to those IRW datasets with available item text. Although the warehouse’s available item text data is growing, the resulting set remains limited. Additional datasets and domains

are needed to assess the generalizability of the findings. The evaluated LLMs are also proprietary and subject to change, which may limit reproducibility. Future studies should include open-source models, document model versions and configurations, and, as the number of models and datasets grows, examine relationships between model characteristics and performance. This includes examining data that may not have been accessed in model training, the lack of which is another unfortunate reality of current LLM paradigms, even among open-weight models.

Additionally, the hierarchical Bradley–Terry model assumes a particular parametric form for rater discrimination and position bias. Comparing this specification with alternative aggregation models would clarify the robustness of the findings. Finally, we held the prompt design fixed and therefore cannot determine how sensitive LLM judgments are to prompt wording or structure in this specific setting despite evidence CGCoT improves base model performance. Thus, systematically evaluating prompts continues to be an important direction for future work investigating the psychometric properties of LLMs.

7 Conclusion

Paired comparisons elicited from LLMs provide a scalable, nontrivial signal about item difficulty, and hierarchical Bradley–Terry modeling offers a principled way to pool that signal while accounting for differences in rater discrimination and position bias. Nevertheless, high transitivity and strong inter-model agreement did not ensure close correspondence with respondent-calibrated difficulty. LLM consensus is therefore best viewed as evidence for a provisional item ordering, not as psychometric calibration. Respondent data remain necessary when item parameters will support score interpretation or operational assessment decisions.

Acknowledgments

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305D240025 to THE LELAND STANFORD JUNIOR UNIVERSITY. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

References

- Ralph Allan Bradley and Milton E. Terry. 1952. [Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons](#). *Biometrika*, 39(3/4):324–345.
- Ulf Böckenholt. 2001. [Hierarchical modeling of paired comparison data](#). *Psychological Methods*, 6(1):49–66.
- Xi Chen, Paul N. Bennett, Kevyn Collins-Thompson, and Eric Horvitz. 2013. [Pairwise ranking aggregation in a crowdsourced setting](#). In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining, WSDM '13*, pages 193–202. Association for Computing Machinery.
- Chrzan, Michael L. [Pairadigm: A Python Library for Concept-Guided Chain-of-Thought Pairwise Measurement of Scalar Constructs Using Large Language Models](#).
- Matthew R DiGiuseppe and Michael E Flynn. 2026. [Scaling Open-Ended Survey Responses Using LLM-Paired Comparisons](#). *Public Opinion Quarterly*, 90(3):630–656.
- Benjamin W. Domingue, Mika Braginsky, Lucy Caffrey-Maffei, Joshua B. Gilbert, Klint Kanopka, Radhika Kapoor, Hansol Lee, Yiqing Liu, Savira Nadela, Guanzhong Pan, Lijin Zhang, Susu Zhang, and Michael C. Frank. 2025. [An introduction to the Item Response Warehouse \(IRW\): A resource for enhancing data usage in psychometrics](#). *Behavior Research Methods*, 57(10):276.
- Mariano Felice and Zeynep Duran Karaoz. 2024. The British Council submission to the BEA 2024 shared task.
- Andrew F. Hayes and Klaus Krippendorff. [Answering the Call for a Standard Reliability Measure for Coding Data](#). 1(1):77–89.
- Matias Hoyle. [Synthetic Student Responses: LLM-Extracted Features for IRT Difficulty Parameter Estimation](#).
- Shengyu Jiang, Chun Wang, and David J. Weiss. 2016. [Sample Size Requirements for Estimation of Item Parameters in the Multidimensional Graded Response Model](#). *Frontiers in Psychology*, 7.
- Radhika Kapoor, Sang T. Truong, Nick Haber, Maria Araceli Ruiz-Primo, and Benjamin W. Domingue. [Prediction of Item Difficulty for Reading Comprehension Items by Creation of Annotated Item Repository](#).
- James S. Kim, Mary A. Burkhauser, Jackie E. Relyea, Joshua B. Gilbert, Ethan Scherer, Jill Fitzgerald, Douglas Mosher, and Joseph McIntyre. 2023. [Replication Data for: A Longitudinal Randomized Trial of a Sustained Content Literacy Intervention from First to Second Grade: Transfer Effects on Students' Reading Comprehension Outcomes](#).

- James S. Kim, Joshua B. Gilbert, Jackie Eunjung Relyea, Patrick Rich, Ethan Scherer, Mary A. Burkhauser, and Johanna N. Tvedt. 2024. [Time to transfer: Long-term effects of a sustained and spiraled content literacy intervention in the elementary grades](#). *Developmental Psychology*, 60(7):1279–1297.
- Diana Kolesnikova, Kirill Fedyanin, Abe D. Hofman, Matthieu J. S. Brinkhuis, and Maria Bolsinova. 2026. [Estimating Item Difficulty with Large Language Models as Experts](#).
- Ming Li, Han Chen, Yunze Xiao, Jian Chen, Hong Jiao, and Tianyi Zhou. [Can LLMs Estimate Student Struggles? Human-AI Difficulty Alignment with Proficiency Simulation for Item Difficulty Prediction](#). In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 25414–25441. Association for Computational Linguistics.
- Hauke Licht, Rupak Sarkar, Patrick Y. Wu, Pranav Goel, Niklas Stoehr, Elliott Ash, and Alexander Miserlis Hoyle. 2025. [Measuring Scalar Constructs in Social Science with LLMs](#). *Preprint*, arXiv:2509.03116.
- Mengjie Qian, Guangzhi Sun, Mark Gales, and Kate Knill. 2026. [Who can we trust? LLM-as-a-jury for Comparative Assessment](#).
- Rasch, Georg. 1960. *Probabilistic Models for Some Intelligence and Attainment Tests*, volume 1 of *Studies in Mathematical Psychology*. Danmarks Paedagogiske Institut.
- Jackie E. Relyea, Joshua B. Gilbert, Mary Burkhauser, Ethan Scherer, Douglas M. Mosher, Zhongyu Wei, Johanna Tvedt, and James S. Kim. 2025. [Asset-Based Implementation of Structured Adaptations in an Online Third-Grade Content Literacy Intervention](#). *Reading Research Quarterly*, 60(4):e70048.
- Alexander Scarlato, Nigel Fernandez, Christopher Ormerod, Susan Lottridge, and Andrew Lan. [SMART: Simulated Students Aligned with Item Response Theory for Question Difficulty Prediction](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 25071–25094. Association for Computational Linguistics.
- Patrycja Sleboda and Joanna Sokolowska. [Measurements of Rationality: Individual Differences in Information Processing, the Transitivity of Preferences and Decision Strategies](#). 8:1844.
- C. Spearman. [The proof and measurement of association between two things](#). 15(1):72–101.
- Curtis Tatsuoaka. 2002. [Data Analytic Methods for Latent Partially Ordered Classification Models](#). *Journal of the Royal Statistical Society Series C: Applied Statistics*, 51(3):337–350.
- Alina A. von Davier, Robert J. Mislevy, and Jian-gang Hao. 2022. *Computational Psychometrics: New Methodologies for a New Generation of Digital Learning and Assessment: With Examples in R and Python*. Methodology of Educational Measurement and Assessment. Springer International Publishing.
- Howard Wainer, Neil J. Dorans, Ronald Flaugher, Bert F. Green, and Robert J. Mislevy. 2000. *Computerized Adaptive Testing: A Primer*, 2 edition. Routledge.
- Xinyi Wang, Hong Jiao, Ming Li, Sydney Peters, Hanna Choi, Tianyi Zhou, and Qingshu Xu. 2026. [Can LLMs Really Understand Item Difficulty Levels? Implications for Automated Item Generation Using LLMs](#). *Preprint*, arXiv:2607.28634.
- Patrick Y. Wu, Jonathan Nagler, Joshua A. Tucker, and Solomon Messing. 2024. [Concept-Guided Chain-of-Thought Prompting for Pairwise Comparison Scoring of Texts with Large Language Models](#). In *2024 IEEE International Conference on Big Data (Big-Data)*, pages 7232–7241.

A CGCoT Prompts

The prompts used were:

- "Summarize the core task, question, behavioral statement, or scenario presented in this item. Item text: {text}"
- "Identify the primary knowledge domain, skill, clinical symptom, attitudinal belief, or psychological construct this item is attempting to measure and give a concise justification. Text: {text}"
- "Analyze the demand or severity of this item. For cognitive/ability items, does it require recall, multi-step reasoning, or synthesis? For psychological/attitudinal items, does endorsing it require a severe level of a symptom, a strongly held belief, or deep introspection? CONCISELY identify any structural or semantic elements—such as complex phrasing, tricky distractors, conditional logic, or strong emotional wording—that increase the cognitive load or the threshold for endorsement. Text: {text}"
- "Based on your analysis, describe the overall level of the underlying latent trait (e.g., high cognitive ability, severe psychological symptomatology, strong attitude) required for a respondent to successfully answer or strongly endorse this item. Previous analysis: {text}"

B Hierarchical Bradley-Terry Model Details

B.1 Item Difficulty

The model estimates a single λ_i for each item across all LLMs, allowing us to pool information across models while accounting for model-specific discrimination and position bias. These item strengths are centered to remove translation invariance:

$$\begin{aligned}\lambda_i &= \sigma_\lambda(z_i - \bar{z}) \\ u_k &\sim \text{N}(0, 1) \\ \sigma_\lambda &\sim \text{HalfNormal}(1)\end{aligned}$$

Thus $\sum_i \lambda_i = 0$. σ_λ is the standard deviation of the item strengths, z_i is a latent standard-normal random effect, and $\bar{z}$ is the mean of the z_i across all items,

B.2 Rater Position Bias

Another way to conceptualize γ_k is a measure of model k 's tendency to prefer the first item *regardless* of item strength. We model γ_k hierarchically, allowing for partial pooling across models and estimating the overall distribution of position bias:

$$\begin{aligned}\gamma_k &= \sigma_\gamma v_k \\ v_k &\sim \text{N}(0, 1) \\ \sigma_\gamma &\sim \text{HalfNormal}(0.5)\end{aligned}$$

where σ_γ is the standard deviation of the position bias across models and v_k is a latent standard-normal random effect.

B.3 Rater Discrimination

The discrimination parameter α_k captures the ability of model k to distinguish between items of different strengths. We model α_k hierarchically, allowing for partial pooling across models and estimating the overall distribution of discrimination:

$$\begin{aligned}\log(\alpha_k) &= \sigma_{\log \alpha}(u_k - \bar{u}) \\ u_k &\sim \text{N}(0, 1) \\ \sigma_{\log(\alpha)} &\sim \text{HalfNormal}(0.5)\end{aligned}$$

where $\sigma_{\log(\alpha)}$ is the standard deviation of the discrimination across models and u_k is a latent standard-normal random effect.

Centering log discrimination enforces

$$\exp\left(\frac{1}{K} \sum_k \log(\alpha_k)\right) = 1$$

so the geometric mean rater discrimination is one. This identifies the global $\alpha - \lambda$ scale without choosing an arbitrary reference model.

C HBT Model Diagnostics

Each dataset-specific HBT model was estimated with four chains, 500 tuning iterations, 1,000 retained draws per chain, and a target acceptance probability of 0.95. Across item strengths, rater effects, and their scale parameters, the largest $\hat{R}$ was 1.01. The minimum bulk and tail effective sample sizes were 594 and 953, respectively. No divergent transitions were observed, and the minimum chain-level Bayesian fraction of missing information was 0.657. Visual inspection of the trace plots for σ_λ and $\sigma_{\log \alpha}$ showed stable mixing with no evident chain separation. Together, these checks provided no indication that the substantive findings were driven by failed convergence.

We evaluated fitted behavior in two complementary ways. First, for every posterior draw we compared the deterministic ordering of λ_i and λ_j with the observed winner, deliberately excluding α_k and γ_k . Mean agreement ranged from 0.716 for `gilbert_meta_104` to 0.871 for `gilbert_meta_103`. Second, we scored each judgment under the complete likelihood. Mean observed-outcome probabilities ranged from 0.653 to 0.828, mean log scores from -0.516 to -0.272 , Brier scores from 0.087 to 0.174, and classification accuracy from 0.732 to 0.874. `gilbert_meta_103` had the strongest full-likelihood fit, whereas `frac20` had the lowest predictive accuracy and largest Brier score. For `gilbert_meta_104`, full-likelihood accuracy exceeded deterministic ranking agreement by approximately seven percentage points, indicating that rater discrimination and position bias explained a meaningful portion of its observed choices.

These posterior predictive quantities are in-sample diagnostics: the pairwise judgments being scored are the same judgments used to estimate the model. They measure how well the fitted HBT model represents the observed comparison data, not generalization to unobserved pairs, new items, or new LLM raters.

D Sensitivity Analyses

We examined whether the pooled item ranking depended on the unequal number of models from each provider or on any single LLM.

The provider-balanced model assigned weights so that the three Google models and four OpenAI models contributed equal total likelihood mass. Within each dataset, we compared its posterior-mean item ranks with those from the primary equal-judgment model using Spearman correlation. Correlations ranged from 0.956 for `frac20` to 0.997 for `gilbert_meta_103`; mean absolute rank changes were at most 1.20 positions, and no item moved more than four positions. The main ranking was therefore insensitive to provider composition.

We next refit the model seven times per dataset, excluding one LLM at a time. Of the 56 leave-one-model-out fits, 53 correlated at least 0.944 with the primary ranking, and correlations reached 1.000 in one fit. Across all fits, mean absolute rank changes did not exceed 1.56 positions and the largest change for any item was six positions. The three lower correlations all occurred in the nine-item `gilbert_meta_104` dataset: 0.717 after excluding Gemini 3.7 Flash, 0.767 after excluding Gemini 3 Flash Preview, and 0.917 after excluding Gemini 3.1 Pro Preview. By contrast, the provider-balanced ranking for this dataset remained close to the primary ranking ($\rho = 0.983$). This localized sensitivity is consistent with the greater rank volatility expected in a small item set and indicates that the `gilbert_meta_104` ordering depends more heavily on the Gemini raters. It does not alter the broader finding that no single model determined the pooled rankings in the larger datasets.

All sensitivity fits showed satisfactory convergence for the item-strength parameters: the maximum $\hat{R}$ was 1.01 and the minimum bulk effective sample size was 2,277. Taken together, the provider-balanced and leave-one-model-out analyses support the robustness of the primary ranking, with the explicit exception of model-removal sensitivity in the smallest dataset.

E Additional Tables

The following tables report the complete model-specific validity and transitivity results summarized in the main text.

Model	ρ	p	N
GPT-5.4 Mini	.146	.051	179
GPT-5.4	.195	.009	179
GPT-5.6 Luna	.276	< .001	179
GPT-5.6 Sol	.321	< .001	179
Gemini 3 Flash	.377	< .001	179
Gemini 3.1 Pro	.383	< .001	179
Gemini 3.7 Flash	.402	< .001	179

Table 2: Overall model-specific convergent validity across all seven datasets. ρ is the Spearman correlation between each model’s standardized Bradley–Terry score and 1PL IRT difficulty; N is the total number of unique items. The strongest correlation is bolded.

Dataset	Model	ρ	p	N
frac20	GPT-5.4 Mini	.448	.048	20
	GPT-5.4	.325	.162	20
	GPT-5.6 Luna	.292	.212	20
	GPT-5.6 Sol	.644	.002	20
	Gemini 3 Flash	.314	.177	20
	Gemini 3.1 Pro	.594	.006	20
	Gemini 3.7 Flash	.350	.130	20
gilbert_meta_2	GPT-5.4 Mini	.365	.113	20
	GPT-5.4	.435	.056	20
	GPT-5.6 Luna	.287	.220	20
	GPT-5.6 Sol	.352	.128	20
	Gemini 3 Flash	.400	.081	20
	Gemini 3.1 Pro	.358	.121	20
	Gemini 3.7 Flash	.388	.091	20
gilbert_meta_7	GPT-5.4 Mini	.012	.945	36
	GPT-5.4	.163	.342	36
	GPT-5.6 Luna	.191	.264	36
	GPT-5.6 Sol	.214	.209	36
	Gemini 3 Flash	.455	.005	36
	Gemini 3.1 Pro	.262	.123	36
	Gemini 3.7 Flash	.455	.005	36
gilbert_meta_8	GPT-5.4 Mini	.254	.183	29
	GPT-5.4	.227	.237	29
	GPT-5.6 Luna	.303	.110	29
	GPT-5.6 Sol	.400	.032	29
	Gemini 3 Flash	.467	.011	29
	Gemini 3.1 Pro	.452	.014	29
	Gemini 3.7 Flash	.465	.011	29
gilbert_meta_102	GPT-5.4 Mini	.127	.461	36
	GPT-5.4	.019	.911	36
	GPT-5.6 Luna	.324	.054	36
	GPT-5.6 Sol	.289	.087	36
	Gemini 3 Flash	.503	.002	36
	Gemini 3.1 Pro	.419	.011	36
	Gemini 3.7 Flash	.505	.002	36
gilbert_meta_103	GPT-5.4 Mini	.243	.203	29
	GPT-5.4	.259	.175	29
	GPT-5.6 Luna	.284	.136	29
	GPT-5.6 Sol	.464	.011	29
	Gemini 3 Flash	.295	.120	29
	Gemini 3.1 Pro	.431	.020	29
	Gemini 3.7 Flash	.367	.050	29
gilbert_meta_104	GPT-5.4 Mini	-.233	.546	9
	GPT-5.4	.600	.088	9
	GPT-5.6 Luna	.083	.831	9
	GPT-5.6 Sol	.617	.077	9
	Gemini 3 Flash	.067	.865	9
	Gemini 3.1 Pro	.567	.112	9
	Gemini 3.7 Flash	.133	.732	9

Table 3: Model-specific convergent validity by dataset. ρ is the Spearman correlation between each model’s standardized Bradley–Terry score and 1PL IRT difficulty; N is the number of unique items. The strongest correlation within each dataset is bolded; the two Gemini Flash models tie on gilbert_meta_7.

Dataset	Model	Transitivity (%)	Triplets
frac20	GPT-5.4 Mini	99.4	641
	GPT-5.4	98.9	641
	GPT-5.6 Luna	95.9	641
	GPT-5.6 Sol	98.3	641
	Gemini 3 Flash	99.4	641
	Gemini 3.1 Pro	98.4	641
	Gemini 3.7 Flash	99.5	641
gilbert_meta_2	GPT-5.4 Mini	87.9	686
	GPT-5.4	98.3	686
	GPT-5.6 Luna	99.0	686
	GPT-5.6 Sol	99.7	686
	Gemini 3 Flash	99.1	686
	Gemini 3.1 Pro	100.0	686
	Gemini 3.7 Flash	99.6	686
gilbert_meta_7	GPT-5.4 Mini	91.4	706
	GPT-5.4	98.2	706
	GPT-5.6 Luna	97.7	706
	GPT-5.6 Sol	99.6	706
	Gemini 3 Flash	98.4	706
	Gemini 3.1 Pro	99.7	706
	Gemini 3.7 Flash	99.7	706
gilbert_meta_8	GPT-5.4 Mini	96.8	686
	GPT-5.4	99.3	686
	GPT-5.6 Luna	99.4	686
	GPT-5.6 Sol	99.4	686
	Gemini 3 Flash	95.2	686
	Gemini 3.1 Pro	99.9	686
	Gemini 3.7 Flash	99.6	686
gilbert_meta_102	GPT-5.4 Mini	91.9	755
	GPT-5.4	96.3	755
	GPT-5.6 Luna	99.6	755
	GPT-5.6 Sol	99.7	755
	Gemini 3 Flash	98.8	755
	Gemini 3.1 Pro	100.0	755
	Gemini 3.7 Flash	99.5	755
gilbert_meta_103	GPT-5.4 Mini	95.5	719
	GPT-5.4	99.7	719
	GPT-5.6 Luna	99.7	719
	GPT-5.6 Sol	99.7	719
	Gemini 3 Flash	98.1	719
	Gemini 3.1 Pro	99.3	719
	Gemini 3.7 Flash	99.4	719
gilbert_meta_104	GPT-5.4 Mini	76.2	84
	GPT-5.4	85.7	84
	GPT-5.6 Luna	96.4	84
	GPT-5.6 Sol	98.8	84
	Gemini 3 Flash	100.0	84
	Gemini 3.1 Pro	100.0	84
	Gemini 3.7 Flash	100.0	84

Table 4: Observed triplet transitivity by model and dataset. Transitivity is the percentage of evaluated item triplets whose pairwise judgments formed a consistent ordering. Triplets gives the number evaluated for each model.