TRIEVAL: A THREE-FACET FRAMEWORK FOR MODELING CAPABILITY, TASK DIFFICULTY, AND PROMPT SENSITIVITY IN LLM EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Average benchmark accuracy remains the dominant metric in LLM evaluation. However, it conflates model capability, task difficulty, and prompt-dependent variation across models and tasks. In addition, characterizing prompt sensitivity can require costly evaluation across thousands of configurations. To address these challenges, we introduce TRIEVAL, a three-facet framework that jointly estimates baseline model capability under a reference prompt, task difficulty, and prompt effects on models and tasks. A regularized Gaussian variational expectation-maximization (RGVEM) algorithm with closed-form updates is developed for scalable estimation. We also introduce prompt sampling to reduce the number of configurations in the evaluation while preserving estimation quality. We evaluate TRIEVAL through simulations and an empirical analysis of five open-weight LLMs across $6{,}005$ prompt configurations and six benchmarks. Simulations show that TRIEVAL more accurately recovers reference-prompt model rankings than conventional methods. Empirical results show that (i) a small set of tasks exhibits anomalous difficulty shifts across prompts; (ii) model rankings vary across prompts; (iii) prompt effects vary systematically across models; and (iv) estimates from sampled configurations closely match those from the full configuration set.

1 INTRODUCTION

Reliable evaluation of large language models (LLMs) is essential for their continued development and safe deployment. Current practice often ranks models by average accuracy across benchmark tasks under fixed or varying prompt configurations. However, this average accuracy collapses several sources of variation into a single score, including model capability, task difficulty, prompt effects, and their interactions. In fact, benchmark tasks vary substantially in difficulty, and model performance can be highly sensitive to prompting (Sivarajkumar et al., 2024). Furthermore, prompt effects are model-dependent: the same prompt may improve some models while disadvantaging others to varying degrees (Sclar et al., 2024). As a result, differences in average accuracy may reflect prompting advantages rather than differences in model capability, potentially distorting model rankings and obscuring variation in prompt sensitivity.

Recent work has applied item response theory (IRT) to LLM evaluation to separate model capability from task difficulty (Lalor & Yu, 2020; Vania et al., 2021; Hofmann et al., 2025; Zhou et al., 2026). While this addresses one source of conflation, existing methods still assume a static model-task relationship across prompts. Therefore, prompt effects remain entangled with estimated capability and difficulty, making it hard to distinguish a model's baseline capability under a reference prompt from model-prompt interactions.

We introduce TRIEVAL, a measurement framework that treats models, tasks, and prompts as three interacting facets of LLM evaluation. It estimates baseline model capability under a reference prompt and task difficulty while capturing how prompts elicit a model's capability and shift a task's difficulty. This framework extends the psychometric concept of measurement invariance to prompt-dependent evaluation. In addition, prompts are often varied along multiple design dimensions (e.g., choice ordering). Because fully crossing these dimensions can yield thousands of configurations, we

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

introduce prompt sampling to preserve estimation quality with substantially fewer configurations, reducing evaluation cost.

Our contributions are as follows: **C1 (Framework).** We propose TRIEVAL to jointly estimate model capability, task difficulty, and their interactions with prompt configurations from task-level completion outcomes. To our knowledge, it is the first framework that jointly estimates these sources of variation in LLM evaluation. **C2 (Efficient estimation and evaluation).** We develop a regularized Gaussian variational expectation-maximization (RGVEM) algorithm with closed-form updates for scalable estimation. A log-penalty on residual interaction variances encourages parsimony while retaining higher-order interactions beyond additive prompt effects. We also show that prompt sampling yields estimates close to those from exhaustive evaluation, reducing future response-collection costs. **C3 (Model characterization and comparison).** Empirical analyses show how TRIEVAL characterizes differences in model capability and prompt sensitivity. Simulations further show more accurate recovery of reference-prompt model rankings than conventional methods. **C4 (Task diagnosis).** Task-level sensitivity estimates identify tasks with unusually large difficulty shifts across prompts. We use a median absolute deviation (MAD)-based criterion to flag such tasks for further inspection.

2 BACKGROUND AND RELATED WORK

2.1 LLM EVALUATION AND ITEM RESPONSE THEORY

Standard LLM benchmarks such as MMLU-Pro (Wang et al., 2024), ARC (Clark et al., 2018), HELM (Liang et al., 2023), and BIG-Bench (Srivastava et al., 2023) provide a fixed task set on which models are ranked by average accuracy. As noted, this aggregation does not account for differences in task difficulty. As a result, average scores can depend strongly on the particular tasks included in an evaluation (Rodriguez et al., 2021), yielding leaderboards that are sensitive to task sampling and potentially unstable across evaluation setups (Biderman et al., 2024; Siska et al., 2024).

IRT addresses this by modeling task completion as a function of model capability and task properties. This allows model capability to be estimated on a stable latent scale, independent of the specific set of tasks used, a property known as measurement invariance. Prior work has shown that IRT-based evaluations produce more stable and well-calibrated scores (Vania et al., 2021; Hofmann et al., 2025).

However, existing IRT methods generally treat the model-task relationship as static across prompts. They either assume a fixed prompt throughout or treat performance variation across prompts as noise to be averaged away rather than as a structured effect to be modeled. As prompt sensitivity has emerged as a primary source of evaluation instability, this assumption is increasingly untenable.

2.2 PROMPT SENSITIVITY AND MEASUREMENT INVARIANCE

Recent studies quantify prompt sensitivity through changes in log-likelihood or accuracy across prompt configurations (Chatterjee et al., 2024; Zhuo et al., 2024). However, these descriptive measures do not separate model capability from task difficulty. PromptEval (Polo et al., 2024) fits a separate IRT model for each LLM to estimate its performance distribution across prompts while accounting for task difficulty. Although these distributions support model comparisons, PromptEval does not jointly estimate baseline capability and prompt sensitivity across LLMs, and it assumes task difficulty is invariant across prompts. TRIEVAL estimates reference-prompt capability and model-specific prompt sensitivity on a common latent scale while allowing task difficulty to vary across prompts, enabling direct model comparisons and task diagnosis.

We formalize prompt-induced variation through the lens of measurement invariance, treating prompt configurations as conditions across which model and task properties may vary. TRIEVAL decomposes two forms of prompt-dependent variation. Model-prompt interactions describe how a model's performance changes across prompts, whereas task-prompt interactions capture idiosyncratic shifts in task difficulty across prompts, which may arise from task features such as ambiguous phrasing or reasoning traps.

Note that adapting measurement invariance methods to LLM evaluation faces two challenges. First, unlike human test-takers who respond to each task once, the same LLMs are evaluated repeatedly across prompts, introducing a repeated-measures structure. Second, classical methods typically

consider measurement invariance across a few discrete groups (e.g., demographic groups), whereas LLM evaluation may involve thousands of configurations formed by crossing prompt design dimensions (e.g., choice ordering). This scale makes both estimation and response collection very costly. TRIEVAL accommodates repeated evaluations through its hierarchical model design and enables scalable estimation through RGVEM. We further evaluate whether TRIEVAL can recover full-data estimates from fewer prompt configurations, thereby reducing response-collection costs.

3 THE TRIEVAL FRAMEWORK

3.1 PROBLEM FORMALIZATION

We consider J LLMs evaluated on a benchmark of I tasks under S prompt configurations. Let $y_{ijs} \in \{0, 1\}$ indicate whether model j successfully completes task i under prompt s. A prompt s represents a unique combination crossing D prompt dimensions (e.g., choice ordering). With K_d levels in dimension d, there are $S_{\max} = \prod_{d=1}^{D} K_d$ possible configurations, of which $S \leq S_{\max}$ are evaluated. We represent each prompt s by a dummy-coded vector $\mathbf{x}_s \in \{0, 1\}^P$, where $P = \sum_{d=1}^{D}(K_d - 1)$ is the total number of dummy-coded features. Adding an intercept gives the augmented feature vector $\tilde{\mathbf{x}}_s = [1, \mathbf{x}_s^\top]^\top \in \{0, 1\}^{P+1}$.

The inputs of TRIEVAL are the observed response tensor $\boldsymbol{Y} = \{y_{ijs}\}$ and the $S \times (P+1)$ augmented design matrix $\tilde{\boldsymbol{X}}$, whose s-th row is $\tilde{\mathbf{x}}_s^\top$. The goal of TRIEVAL is to recover three sets of quantities, formally introduced in the next section: (i) model capability θ_j, (ii) model-specific prompt sensitivity $\boldsymbol{\alpha}_j$ and $\sigma^2_{\theta_j}$, and (iii) task-level prompt sensitivity $\sigma^2_{b_i}$ after accounting for model-level effects.

3.2 MEASUREMENT MODEL

To disentangle these three sources of variance, we extend a standard IRT model into the three-facet TRIEVAL framework. The probability of a correct response is modeled via the logistic function $\sigma(\cdot)$:

$$P(y_{ijs} = 1 \mid \theta_{js}, b_{is}) = \sigma(\theta_{js} - b_{is}), \tag{1}$$

where θ_{js} and b_{is} denote the effective capability of model j and the effective difficulty of task i under prompt configuration s, respectively. Both are modeled as random effects with the following priors.

Model capability. The effective capability θ_{js} is modeled as:

$$\theta_{js} \sim \mathcal{N}(\tilde{\boldsymbol{\alpha}}_j^\top \tilde{\mathbf{x}}_s,\ \sigma^2_{\theta_j}), \tag{2}$$

where $\tilde{\boldsymbol{\alpha}}_j^\top = (\theta_j, \boldsymbol{\alpha}_j^\top)$ is the augmented sensitivity vector. The intercept θ_j represents the baseline capability of model j at the reference prompt configuration, and $\boldsymbol{\alpha}_j \in \mathbb{R}^P$ captures the model's additive sensitivity to prompt dimensions. The variance $\sigma^2_{\theta_j}$ captures residual prompt-by-model interactions beyond this additive structure. Together, these quantities characterize the full prompt effect on model capability elicitation, enabling the model characterization of contribution C3.

Task difficulty. The effective difficulty b_{is} is modeled as:

$$b_{is} \sim \mathcal{N}(\beta_{0i},\ \sigma^2_{b_i}), \tag{3}$$

where $\beta_{0i} \sim \mathcal{N}(0, \sigma^2_\beta)$ is the baseline difficulty of task i, with σ^2_β reflecting heterogeneity in task difficulty across the benchmark. The task-specific variance $\sigma^2_{b_i}$ reflects deviations in the difficulty of task i from its mean (i.e., β_{0i}), attributable to variation in prompt configurations. If a task is invariant to prompt variations, $\sigma^2_{b_i}$ will be 0. Tasks with large $\sigma^2_{b_i}$ are flagged as exhibiting high task-level prompt sensitivity, forming the benchmark diagnosis of contribution C4.

Two things are worth noting. First, the zero-centered prior on baseline task difficulty anchors the scale of this latent variable model for model identification, so that it does not require fitted model capabilities to average zero. Second, residual variances $\sigma^2_{\theta_j}$ and $\sigma^2_{b_i}$ accommodate higher-order prompt-model and prompt-task interactions beyond additive prompt effects, while the regularization still encourages model sparsity. This balances model flexibility with interpretability. Overall, TRIEVAL consists of a latent variable set $\mathcal{Z} \equiv \{\theta_{js}, \beta_{0i}, b_{is}\}$ and a global parameter set $\{\tilde{\boldsymbol{\alpha}}_j, \sigma^2_{\theta_j}, \sigma^2_\beta, \sigma^2_{b_i}\}$ to be estimated.

3.3 REGULARIZED VARIATIONAL INFERENCE

Estimation of the global parameter set $\{\tilde{\boldsymbol{\alpha}}_j, \sigma^2_{\theta_j}, \sigma^2_{\beta}, \sigma^2_{b_i}\}$ proceeds by maximizing the log marginal likelihood $\ell(\boldsymbol{Y}) = \log p(\boldsymbol{Y})$, which requires marginalizing the joint distribution $p(\boldsymbol{Y}, \mathcal{Z})$ over the latent variable set $\mathcal{Z}$. This is approximated by its evidence lower bound (ELBO)

$$\mathcal{L}(q) = \mathbb{E}_q[\log p(\boldsymbol{Y}, \mathcal{Z})] + H(q) \leq \ell(\boldsymbol{Y}), \tag{4}$$

where $p(\boldsymbol{Y}, \mathcal{Z}) = p(\boldsymbol{Y} \mid \mathcal{Z})p(\mathcal{Z})$, with $p(\boldsymbol{Y} \mid \mathcal{Z})$ being the Bernoulli likelihood whose success probability is defined in Equation 1 and $p(\mathcal{Z})$ the hierarchical priors in Equations 2 and 3, and $H(q) = -\mathbb{E}_q[\log q(\mathcal{Z})]$ is the entropy of q.

As described in Section 3.2, prompt effects are primarily captured through the interpretable additive sensitivity parameters $\tilde{\boldsymbol{\alpha}}_j$, with residual variances $\sigma^2_{\theta_j}$ and $\sigma^2_{b_i}$ retained only when supported by the data. To ensure this parsimony, we augment the ELBO with a log-penalty on the residual variances,

$$\mathcal{L}_{\text{reg}}(q) = \mathcal{L}(q) - \lambda \left[\sum_{j=1}^{J} \log \sigma^2_{\theta_j} + \sum_{i=1}^{I} \log \sigma^2_{b_i} \right], \tag{5}$$

where $\lambda > 0$ controls the strength of regularization and is selected via the Bayesian Information Criterion (BIC) or Generalized Information Criterion (GIC) (Ma et al., 2023; Ren et al., 2026).

However, $\mathcal{L}(q)$ and $\mathcal{L}_{\text{reg}}(q)$ remain intractable due to two challenges: the high-dimensional latent variable structure and the logistic likelihood in $p(\boldsymbol{Y} \mid \mathcal{Z})$. Following recent work (Cho et al., 2021; Ren et al., 2026), we address the first challenge through the mean-field approximation $q(\mathcal{Z}) = \prod_l q(z_l)$ (Bishop, 2006). We address the second one by a local variational bound (Jaakkola & Jordan, 2000), which introduces auxiliary parameters ξ_{ijs} to construct a tractable lower bound on each logistic term. Together, these yield the tractable regularized objective $Q'_{\text{reg}}(\boldsymbol{Y})$ derived in Appendix A. Maximization of $Q'_{\text{reg}}(\boldsymbol{Y})$ follows an Expectation-Maximization scheme, which iterates between the following two steps until convergence.

E-step. With global parameters fixed, the variational posteriors $q(\mathcal{Z})$ are updated. The optimal posteriors $q(\theta_{js})$, $q(b_{is})$, and $q(\beta_{0i})$ take the forms $\mathcal{N}(\mu_{\theta_{js}}, \sigma^2_{\theta_{js}})$, $\mathcal{N}(\mu_{b_{is}}, \sigma^2_{b_{is}})$, and $\mathcal{N}(\mu_{\beta_{0i}}, \sigma^2_{\beta_{0i}})$, respectively, with sufficient statistics updated in closed form, as detailed in Appendix A.

M-step. With variational posteriors fixed, $Q'_{\text{reg}}(\boldsymbol{Y})$ is maximized with respect to the global parameters $\{\tilde{\boldsymbol{\alpha}}_j, \sigma^2_{\theta_j}, \sigma^2_{\beta}, \sigma^2_{b_i}\}$, all of which admit closed-form updates. As representative examples, the augmented prompt sensitivity vector $\tilde{\boldsymbol{\alpha}}_j$ and residual variance $\sigma^2_{\theta_j}$ for model j are updated as

$$\tilde{\boldsymbol{\alpha}}_j = \left(\sum_s \tilde{\mathbf{x}}_s \tilde{\mathbf{x}}_s^\top \right)^{-1} \left(\sum_s \mu_{\theta_{js}} \tilde{\mathbf{x}}_s \right), \tag{6}$$

$$\sigma^2_{\theta_j} = \frac{\sum_s \left[(\mu_{\theta_{js}} - \tilde{\boldsymbol{\alpha}}_j^\top \tilde{\mathbf{x}}_s)^2 + \sigma^2_{\theta_{js}} \right]}{S + 2\lambda}. \tag{7}$$

The denominator $S + 2\lambda$ in Equation 7 demonstrates how λ shrinks the residual variance $\sigma^2_{\theta_j}$ relative to the unpenalized estimator. This algorithm, combining Gaussian variational posteriors, the EM optimization scheme, and the log-penalty, is referred to as the RGVEM algorithm. Its complete derivation is provided in Appendix A.

4 EMPIRICAL EXPERIMENT

4.1 EXPERIMENTAL SETUP

To demonstrate the utility of TRIEVAL on empirical data, we analyze responses from the dataset of variation evaluation (DOVE) (Habba et al., 2025), a large-scale collection of model responses under

diverse prompt perturbations. This analysis includes five LLMs, Llama-1B, Llama-3B, and Llama-8B (Grattafiori et al., 2024), OLMoE-7B (Muennighoff et al., 2024), and Mistral-7B (Jiang et al., 2023). These open-weight models provide a controlled setting for systematic prompt manipulation at scale. In addition, we apply the framework to 100 randomly sampled tasks from each of six benchmarks, including ARC Challenge, ARC Easy, MMLU Abstract Algebra, MMLU College Computer Science, MMLU Human Sexuality, and MMLU Moral Disputes (Clark et al., 2018; Hendrycks et al., 2021). For brevity, the following results refer to ARC Challenge unless otherwise specified.

The full-data analysis uses all $S = 6{,}005$ prompt configurations available in DOVE, crossing five intent-preserving dimensions with $S_{\max} = 6{,}552$ possible combinations. Table 1 shows the five dimensions and their levels, which yield $P = \sum_{d=1}^{5}(K_d - 1) = 29$ dummy-coded prompt features in $\mathbf{x}_s$. We also assess prompt sampling by comparing estimates from subsets of configurations with those from the full data. Unless otherwise stated, the empirical results below use all available configurations. DOVE also provides detailed metadata, including temperature settings, sampled exemplars, prompt and generated-token log probabilities, and answers. Further details are available in Habba et al. (2025).

Table 1: Prompt dimensions and number of levels in the DOVE dataset (Habba et al., 2025).

Dimension	Possible Values	K_d
D1: Enumerator	capitals, greek, keyboard, lowercase, numbers, roman	6
D2: Separator	|, or, OR, \n, ,, ;, \s	7
D3: Choice order	alphabetical, correct_first, longest_to_shortest, none, reverse_alphabetical, shortest_to_longest	6
D4: Phrasing	13 distinct templates (see Appendix B)	13
D5: Demonstrations	Zero-shot, five-shot	2

4.2 Heterogeneity in task difficulty

Figure 1 illustrates task difficulty heterogeneity in ARC Challenge. Panel A shows the distribution of accuracy rates for each task across all configurations and models, with tasks ordered by their mean accuracy rate (displayed as dots). The dark and light gray bands span the 25th-75th and 5th-95th percentiles, respectively. The wide variation in accuracy across tasks motivates modeling task difficulty explicitly, as argued in Section 1. TriEval quantifies this heterogeneity through the baseline difficulty variance, estimated at $\hat{\sigma}^2_\beta = 1.26$ for the ARC Challenge and ranging from 0.44 to 1.26 across the six benchmarks.

Accuracy also varies across prompt configurations within each task. Panel B retains the same task ordering and the task means (i.e., black dots), highlights two representative configurations in color, and displays the remaining configurations in gray. While the accuracy under the two representative prompts broadly follows the mean trend, their profiles diverge for individual tasks. Note that these differences alone do not directly establish task-prompt interactions because the accuracy averaged across models combines model-level and task-level prompt interactions. We use TriEval to separate these sources of variation and examine task-specific prompt sensitivity in the next section.

4.3 Task-level prompt sensitivity

Figure 2 presents the estimated residual variance $\hat{\sigma}^2_{b_i}$ for all tasks, capturing task-specific difficulty shifts across prompts after accounting for model-prompt interactions. Note that most tasks exhibit negligible residual variance, but several tasks stand out with much larger values. Similar patterns also appear in other benchmarks, as depicted in Appendix Figure 6. To screen for prompt-sensitive tasks, let $m = \text{median}(\hat{\sigma}^2_{b_i})$ and $\text{MAD} = \text{median}\,|\hat{\sigma}^2_{b_i} - m|$. When MAD is positive, we flag tasks whose MAD-based modified z-score exceeds 3.5.

Such sensitivity may seem unexpected because the prompt variations should not change the tasks' meaning. To understand its potential sources, Table 2 shows two sensitive tasks. The correct answer to task 70 is debatable under the broader scientific consensus, and task 88 contains the typo "43 40 m/s," suggesting that task-level ambiguity may contribute to sensitivity.

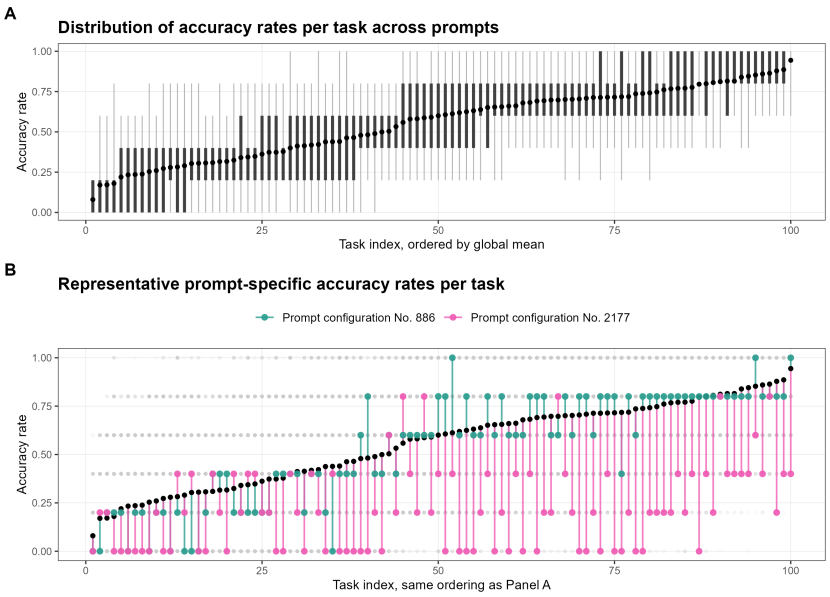


Figure 1: Heterogeneity in item passing rate.

These prompt-sensitive tasks can conflate task-level variation with model-level prompt effects, especially in accuracy-based analyses. We examine this by comparing prompt effect estimates before and after filtering these tasks, defining the effect as the accuracy change between a target prompt and the reference prompt. Filtering changes the estimated effects by -0.03 to 0.05. These shifts are small in absolute terms but substantial relative to the unfiltered estimates (Appendix Table 5).

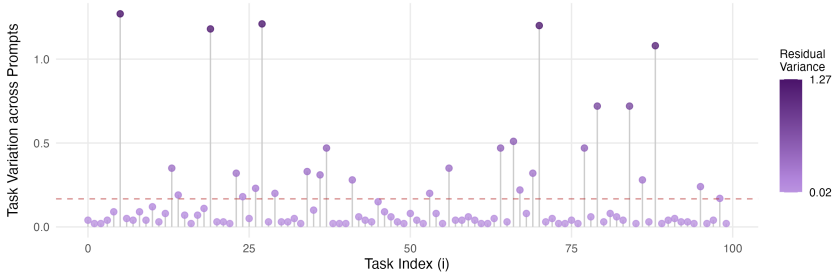


Figure 2: Item-specific variations across prompts.

Table 2: Examples of prompt-sensitive items.

Index	Question and Choices	Answer	Modified z
70	Garden plants on Earth require four resources to stay alive: soil, air, water, and sunlight. How many of these resources are necessary for life to exist on the moon or another planet? Choices: I. 1; II. 2; III. 3; IV. 4	IV. 4	30.44
88	A pitcher throws a 0.15 kg baseball at 43 40 m/s towards the catcher. What is the momentum of the baseball while moving at 40 m/s? Choices: I. 0.025 kg $\times$ m/s; II. 270 kg $\times$ m/s; III. 3.8 kg $\times$ m/s; IV. 6.0 kg $\times$ m/s	IV. 6.0 kg $\times$ m/s	27.44

4.4 MODEL CAPABILITY ESTIMATION

One key output of TRIEVAL is the baseline capability estimate $\hat{\theta}_j$ under the reference prompt. On ARC Challenge, the descending ranking is Llama-8B (1.42), Mistral-7B (0.90), Llama-3B (0.46), OLMoE-7B (0.15), and Llama-1B (-0.93). However, this single capability estimate does not capture how models respond to different prompts. Figure 3 illustrates each model's effective capability $\hat{\theta}_{js}$ across all prompt configurations. Greater color variation within a model indicates larger capability changes across prompts.

These changes differ across models and can alter their relative rankings. On ARC Challenge, $3{,}463$ of $6{,}005$ configurations retain the reference ranking, while approximately 42% configurations yield different results. For example, configuration 2032, with a keyboard enumerator, newline separator, correct-answer-first ordering, a particular phrasing variant, and zero-shot demonstrations, yields the descending ranking Mistral-7B, Llama-8B, OLMoE-7B, Llama-3B, and Llama-1B. Across the other five benchmarks, 55–85% configurations result in rankings different from the reference one.

Note that the evaluation method also affects model rankings. Table 3 compares TRIEVAL's reference ranking with those from average accuracy and a standard IRT model (i.e., Rasch model), counting models assigned different rank positions. Both conventional methods disagree with TRIEVAL on four of six benchmarks. These comparisons establish empirical disagreement, while the controlled simulation in Section 5 compares ranking accuracy against known ground truth.

These results demonstrate that model comparisons depend strongly on prompt choice and evaluation method. The next section examines the model-specific additive prompt effects $\boldsymbol{\alpha}_j$ that explain this variation across prompt dimensions.

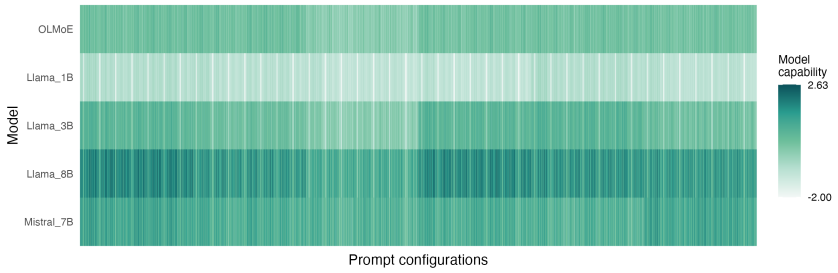


Figure 3: Model capability under each prompt configuration.

Table 3: Number of models (out of five) assigned a different rank position from the TRIEVAL reference.

Benchmark	Average accuracy	Rasch
ARC Challenge	0	0
ARC Easy	2	2
MMLU Abstract Algebra	4	4
MMLU College Computer Science	2	2
MMLU Human Sexuality	3	2
MMLU Moral Disputes	0	0

4.5 MODEL-LEVEL PROMPT SENSITIVITY

Figure 4 presents the estimated model-specific prompt sensitivity $\hat{\boldsymbol{\alpha}}_j$ across all five prompt dimensions and five models. Several patterns are worth noting. First, prompt sensitivity varies substantially across models, indicating meaningful heterogeneity in model-prompt interactions. Second, the five-shot demonstration shows a consistent positive effect across all models, consistent with the established findings that few-shot prompts improve performance. The magnitude of this effect tends to increase with model scale: smaller models benefit less, suggesting limited capacity to exploit in-context examples. Third, within the Llama family, sensitivity patterns are consistent in direction across

models but differ in magnitude. For example, Greek and keyboard in the enumerator dimension (D1) are consistently associated with worse performance, as are polite phrasing variants such as "Could you...", "Please address...", and "Please answer..." in the phrasing dimension (D4). These negative effects are more pronounced for larger models. In contrast, correct-first ordering (D3) degrades performance in Llama-1B but has a diminished effect in larger models, suggesting that positional bias interacts with model scale.

These model-specific prompt effects are also estimated on the other benchmarks (Appendix Figure 7). Pairwise correlations of these effects range from 0.77 to 0.95 among benchmarks, excluding MMLU Abstract Algebra. The correlations between MMLU Abstract Algebra and the others range from 0.25 to 0.49. This suggests that sensitivity patterns are broadly consistent across domains but do not transfer uniformly.

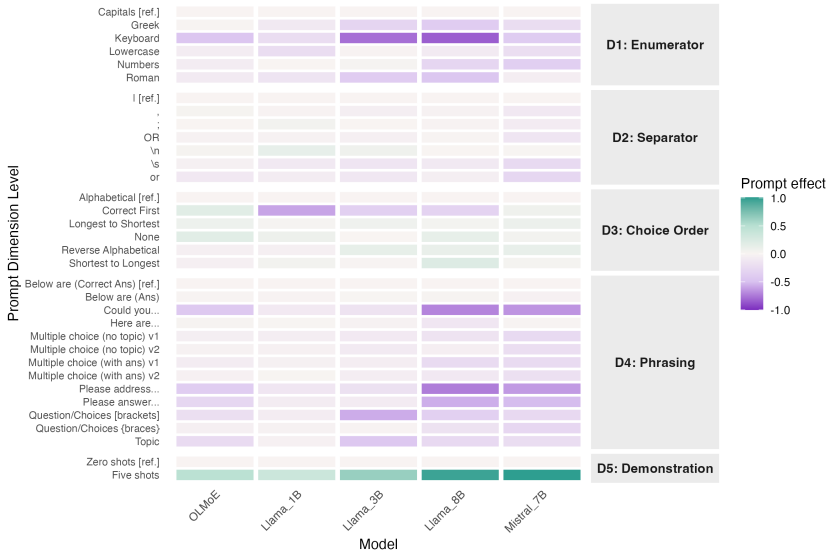


Figure 4: Model-specific prompt sensitivity estimates $\hat{\boldsymbol{\alpha}}_j$ across five prompt dimensions and five models. Each panel corresponds to one prompt dimension.

4.6 Prompt sampling: Stability and evaluation cost

This study assesses whether TriEval can recover full-data estimates from fewer configurations. For each benchmark, we randomly sample $m \in \{200, 600, 1000\}$ of the $6{,}005$ available configurations without replacement, with 50 replications per subset size. We refit TriEval on each subset and compare its estimates with the full-data results.

In general, baseline model ranking agreement improves with subset size (Figure 5). At $m = 1{,}000$, all replications preserve at least 90% of pairwise model rankings, although exact ranking recovery varies across benchmarks. At this subset size, correlations between sampled and full-data estimates exceed 0.90 for both capability and prompt effects, with near-zero bias (Appendix Figure 8). These results support comparable estimation with 83.3% fewer responses, reducing future evaluation costs.

5 Simulation Study

We conduct two simulation studies to evaluate model ranking and parameter recovery, each with 50 replications. Full designs are detailed in Appendix D.

The first study compares TriEval, average accuracy, and a standard IRT model in recovering reference-prompt model rankings of six simulated models. We manipulate the capability gap between adjacent models at two levels, 0.2 and 0.6. Table 4 shows that, at the smaller gap, the mean numbers of incorrectly ranked models are 0.24 for TriEval, 4.48 for average accuracy, and 4.36 for IRT.

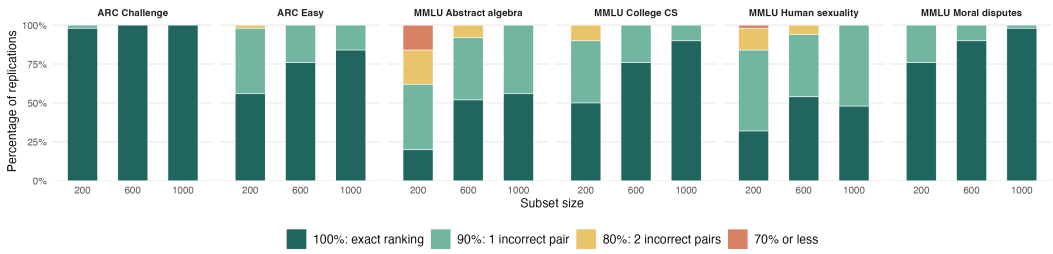


Figure 5: Distribution of pairwise ranking agreement with the full-data results across 50 replications.

Table 4: Number of incorrectly ranked models (out of six) across 50 replications.

Capability gap	Statistic	TRIEVAL	Average accuracy	Rasch IRT
0.2	Min	0	2	2
	Max	2	6	6
	Mean	0.24	4.48	4.36
0.6	Min	0	0	0
	Max	0	6	6
	Mean	0.00	2.66	2.64

The second study evaluates the parameter recovery of TRIEVAL. Three factors are manipulated: model scale $J \in \{6, 20\}$, baseline difficulty variance $\sigma_\beta^2 \in \{1, 2\}$, and the number of prompt configurations $S \in \{200, 2500\}$, resulting in eight conditions. Across all conditions, baseline model capability is well recovered, with the correlation between estimated and true θ_j exceeding 0.98. The recovery of model-specific prompt sensitivity $\boldsymbol{\alpha}_j$ improves as the number of prompt configurations increases, achieving correlations of 0.98 at $S = 2500$ but 0.81 at $S = 200$. These results support the reliability of estimation in the empirical study with 6,005 prompt configurations.

6 LIMITATIONS AND BROADER IMPACT

6.1 LIMITATIONS

TRIEVAL provides a rigorous measurement framework for LLM evaluation across prompt configurations, but several limitations should be noted. The framework models task completion as a binary outcome, while some benchmarks involve partial credit or open-ended generation. Extending TRIEVAL to polytomous or continuous response models is valuable for future work. In addition, the empirical study covers six benchmarks and five open-weight models, intended as an illustration of the TRIEVAL framework rather than an exhaustive evaluation. Broader application with other benchmarks and model families remains an important next step.

6.2 BROADER IMPACT

TRIEVAL addresses a critical reliability problem in LLM evaluation: minor prompt changes can substantially alter observed model performance. TRIEVAL anchors model capability to a reference prompt configuration and quantifies how the capability varies across prompt configurations, separating baseline capability from prompt effects. The prompt sampling results also support a more efficient, lower-cost evaluation by reducing future response collection.

TRIEVAL estimates also have implications beyond evaluation. Model-level sensitivity profiles inform model training by identifying limitations in handling certain prompt structures and guide prompt design by clarifying whether the benefits of prompts generalize across models. Task-level diagnosis identifies tasks with high variability across prompts, informing benchmark curation.

AI USE STATEMENT

In this work, we used generative AI tools to: clean and reformat dataset, and design or provide feedback on research methodology or experiments. We have not used generative AI tools to: generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, implement methods, assist with translation, support qualitative and thematic data analysis, and interpret results. Additionally, we used generative AI tools to: edit a research paper to improve readability, format references, and propose a title or keywords for a research paper. We have reviewed all AI-assisted work. We manually checked the LLM-generated data cleaning code and critically evaluated the LLM-generated feedback on our experiments. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

All information necessary to reproduce the main results is provided. The model specification is fully detailed in Section 3.2, including the model in Equation 1, the hierarchical priors in Equations 2 and 3, and the two-level parameter structure. The RGVEM algorithm is described in Section 3.3 with details provided in Appendix A. The empirical experiment setup is fully described in Section 4.1. The dataset DOVE is publicly available. The five open-weight models used in this study are specified, and all five prompt dimensions with their levels are summarized in Table 1 and Appendix B. The simulation setup is fully described in Appendix D, including all manipulated factors, parameter settings, and the number of replications. The code implementing the framework and reproducing all experimental results is provided at https://anonymous.4open.science/r/TriEval-8756 and will be made publicly available upon publication.

REFERENCES

Stella Biderman, Hailey Schoelkopf, Lintang Sutawika, Leo Gao, Jonathan Tow, Baber Abbasi, Alham Fikri Aji, Pawan Sasanka Ammanamanchi, Sidney Black, Jordan Clive, et al. Lessons from the trenches on reproducible evaluation of language models. *ArXiv preprint*, abs/2405.14782, 2024.

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer, New York, NY, 2006. ISBN 978-0-387-31073-2.

Anwoy Chatterjee, H. S. V. N. S. K. Renduchintala, Sumit Bhatia, and Tanmoy Chakraborty. POSIX: A prompt sensitivity index for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 14550–14565, Miami, Florida, USA, 2024. Association for Computational Linguistics.

April E. Cho, Chun Wang, Xue Zhang, and Gongjun Xu. Gaussian variational estimation for multidimensional item response theory. *British Journal of Mathematical and Statistical Psychology*, 74(S1):52–85, 2021.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. *ArXiv preprint*, abs/1803.05457, 2018.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 herd of models. *ArXiv preprint*, abs/2407.21783, 2024.

Eliya Habba, Ofir Arviv, Itay Itzhak, Yotam Perlitz, Elron Bandel, Leshem Choshen, Michal Shmueli-Scheuer, and Gabriel Stanovsky. DOVE: A large-scale multi-dimensional predictions dataset towards meaningful LLM evaluation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 11744–11763, Vienna, Austria, 2025. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021. URL `https://arxiv.org/abs/2009.03300`.

Valentin Hofmann, David Heineman, Ian Magnusson, Kyle Lo, Jesse Dodge, Maarten Sap, Pang Wei Koh, Chun Wang, Hannaneh Hajishirzi, and Noah A. Smith. Fluid language model benchmarking. In *Second Conference on Language Modeling*, Online, 2025. OpenReview.

Tommi Jaakkola and Michael I. Jordan. Bayesian parameter estimation via variational methods. *Statistics and Computing*, 10:25–37, 2000.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7B. *ArXiv preprint*, abs/2310.06825, 2023.

John P. Lalor and Hong Yu. Dynamic data selection for curriculum learning via ability estimation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 545–555, Online, 2020. Association for Computational Linguistics.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *ArXiv preprint*, abs/2211.09110, 2023.

Chenghao Ma, Jing Ouyang, and Gongjun Xu. Learning latent and hierarchical structures in cognitive diagnosis models. *Psychometrika*, 88(1):175–207, 2023.

Niklas Muennighoff, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Jacob Morrison, Sewon Min, Weijia Shi, Pete Walsh, Oyvind Tafjord, Nathan Lambert, et al. OLMoE: Open mixture-of-experts language models. *ArXiv preprint*, abs/2409.02060, 2024.

Felipe Maia Polo, Ronald Xu, Lucas Weber, Mírian Silva, Onkar Bhardwaj, Leshem Choshen, Allysson Flavio Melo de Oliveira, Yuekai Sun, and Mikhail Yurochkin. Efficient multi-prompt evaluation of LLMs. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=jzkpwcj200.

He Ren, Weicong Lyu, Chun Wang, and Gongjun Xu. A novel method for detecting intersectional DIF: Multilevel random item effects model with regularized gaussian variational estimation. *Psychometrika*, 91(1):70–94, 2026.

Pedro Rodriguez, Joe Barrow, Alexander Miserlis Hoyle, John P. Lalor, Robin Jia, and Jordan Boyd-Graber. Evaluation examples are not equally informative: How should that change NLP leaderboards? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4486–4503, Online, 2021. Association for Computational Linguistics.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations*, Online, 2024. OpenReview.

Charlotte Siska, Katerina Marazopoulou, Melissa Ailem, and James Bono. Examining the robustness of LLM evaluation to the distributional assumptions of benchmarks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10406–10421, Bangkok, Thailand, 2024. Association for Computational Linguistics.

Sonish Sivarajkumar, Mark Kelley, Alyssa Samolyk-Mazzanti, Shyam Visweswaran, and Yanshan Wang. An empirical evaluation of prompting strategies for large language models in zero-shot clinical natural language processing: Algorithm development and validation study. *JMIR Medical Informatics*, 12:e55318, 2024.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *ArXiv preprint*, abs/2206.04615, 2023.

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

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. *ArXiv preprint*, abs/2406.01574, 2024.

Lexin Zhou, Lorenzo Pacchiardi, Fernando Martínez-Plumed, et al. General scales unlock AI evaluation with explanatory and predictive power. *Nature*, 652:58–67, 2026.

Jingming Zhuo, Songyang Zhang, Xinyu Fang, Haodong Duan, Dahua Lin, and Kai Chen. ProSA: Assessing and understanding the prompt sensitivity of LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 1950–1976, Miami, Florida, USA, 2024. Association for Computational Linguistics.

A Derivation of the RGVEM algorithm

A.1 Evidence lower bound

The log marginal likelihood $\ell(\boldsymbol{Y}) = \log p(\boldsymbol{Y})$ is obtained by marginalizing the joint distribution over the latent variables $\mathcal{Z}$,

$$\ell(\boldsymbol{Y}) = \log \int p(\boldsymbol{Y}, \mathcal{Z}) \, \mathrm{d}\mathcal{Z}. \tag{A.1}$$

This integral is intractable. For any density $q(\mathcal{Z})$, multiplying and dividing by $q(\mathcal{Z})$ inside the log function and applying Jensen's inequality yields,

$$\begin{aligned}
\ell(\boldsymbol{Y}) &= \log \int \frac{p(\boldsymbol{Y}, \mathcal{Z})}{q(\mathcal{Z})} q(\mathcal{Z}) \, \mathrm{d}\mathcal{Z} \\
&\geq \int \log \frac{p(\boldsymbol{Y}, \mathcal{Z})}{q(\mathcal{Z})} q(\mathcal{Z}) \, \mathrm{d}\mathcal{Z} \\
&= \mathbb{E}_q \left[\log \frac{p(\boldsymbol{Y}, \mathcal{Z})}{q(\mathcal{Z})} \right] \\
&= \mathbb{E}_q \left[\log p(\boldsymbol{Y}, \mathcal{Z}) \right] - \mathbb{E}_q [\log q(\mathcal{Z})] \\
&= \mathbb{E}_q [\log p(\boldsymbol{Y}, \mathcal{Z})] + H(q) \\
&\equiv \mathcal{L}(q),
\end{aligned} \tag{A.2}$$

where $H(q) = -\mathbb{E}_q[\log q(\mathcal{Z})]$ is the entropy of q, and $\mathcal{L}(q)$ is the ELBO defined in Equation 4.

To see the exact gap between $\ell(\boldsymbol{Y})$ and $\mathcal{L}(q)$, note that $\log p(\boldsymbol{Y})$ is constant with respect to $\mathcal{Z}$ and thus equals $\mathbb{E}_q[\log p(\boldsymbol{Y})]$. Therefore,

$$\begin{aligned}
\ell(\boldsymbol{Y}) &= \mathbb{E}_q[\log p(\boldsymbol{Y})] \\
&= \mathbb{E}_q \left[\log \frac{p(\boldsymbol{Y}, \mathcal{Z})}{p(\mathcal{Z} \mid \boldsymbol{Y})} \right] \\
&= \mathbb{E}_q \left[\log p(\boldsymbol{Y}, \mathcal{Z}) - \log p(\mathcal{Z} \mid \boldsymbol{Y}) \right] \\
&= \mathbb{E}_q \left[\log p(\boldsymbol{Y}, \mathcal{Z}) - \log q(\mathcal{Z}) \right] + \mathbb{E}_q \left[\log q(\mathcal{Z}) - \log p(\mathcal{Z} \mid \boldsymbol{Y}) \right] \\
&= \mathcal{L}(q) + \mathrm{KL}(q(\mathcal{Z}) \parallel p(\mathcal{Z} \mid \boldsymbol{Y})),
\end{aligned} \tag{A.3}$$

where $\mathrm{KL}(q(\mathcal{Z}) \parallel p(\mathcal{Z} \mid \boldsymbol{Y})) \geq 0$ is the KL divergence from the true posterior. Maximizing $\mathcal{L}(q)$ therefore tightens the bound and drives $q(\mathcal{Z})$ toward $p(\mathcal{Z} \mid \boldsymbol{Y})$.

Recall that $\mathcal{L}(q) = \mathbb{E}_q[\log p(\boldsymbol{Y}, \mathcal{Z})] + H(q)$ in Equation A.2, where $H(q)$ does not depend on the model parameters. Thus, maximizing $\mathcal{L}(q)$ is equivalent to maximizing $Q(\boldsymbol{Y}) \equiv \mathbb{E}_q[\log p(\boldsymbol{Y}, \mathcal{Z})]$. The objective function $Q(\boldsymbol{Y})$ can be further expanded as,

$$\begin{aligned}
Q(\boldsymbol{Y}) &= \mathbb{E}_q[\log p(\boldsymbol{Y}, \mathcal{Z})] \\
&= \mathbb{E}_q[\log p(\boldsymbol{Y} \mid \mathcal{Z})] + \mathbb{E}_q[\log p(\mathcal{Z})] \\
&= \sum_{s,j,i} \mathbb{E}_q[\log p(y_{ijs} \mid \theta_{js}, b_{is})] + \sum_{s,j} \mathbb{E}_q[\log p(\theta_{js})] \\
&\quad + \sum_{s,i} \mathbb{E}_q[\log p(b_{is} \mid \beta_{0i})] + \sum_{i} \mathbb{E}_q[\log p(\beta_{0i})],
\end{aligned} \tag{A.4}$$

where $p(y_{ijs} \mid \theta_{js}, b_{is})$ follows the binomial likelihood whose success probability is defined in Equation 1, and $p(\theta_{js})$, $p(b_{is} \mid \beta_{0i})$, and $p(\beta_{0i})$ follow the normal densities specified in Equations 2 and 3.

A.2 Local variational bound

The expectation $\mathbb{E}_q[\log p(y_{ijs} \mid \theta_{js}, b_{is})]$ does not admit a closed-form solution due to the logistic likelihood. Following Jaakkola and Jordan (Jaakkola & Jordan, 2000), auxiliary parameters $\xi_{ijs} > 0$ are introduced to construct a quadratic lower bound on each logistic log-likelihood term,

$$\log p(y_{ijs} \mid \theta_{js}, b_{is}) \geq \log \sigma(\xi_{ijs}) + (y_{ijs} - \tfrac{1}{2})(\theta_{js} - b_{is}) - \frac{1}{2}\xi_{ijs} - \eta(\xi_{ijs}) \left[(\theta_{js} - b_{is})^2 - \xi_{ijs}^2 \right], \tag{A.5}$$

where $\eta(\xi) = \frac{1}{2\xi}\left[\sigma(\xi) - \frac{1}{2}\right]$. Substituting Equation A.5 into Equation A.4 and taking expectations under $q(\mathcal{Z})$ yields the tractable approximation, denoted $Q'(\boldsymbol{Y})$,

$$
\begin{aligned}
Q(\boldsymbol{Y}) \geq \int \Bigg[& \sum_{s,j,i} \Bigg\{ \log \sigma(\xi_{ijs}) + \Big(y_{ijs} - \frac{1}{2}\Big)(\theta_{js} - b_{is}) - \frac{1}{2}\xi_{ijs} - \eta(\xi_{ijs})\Big[(\theta_{js} - b_{is})^2 - \xi_{ijs}^2\Big] \Bigg\} \\
& - \frac{1}{2}\Bigg\{ (SJ + SI + I)\log 2\pi + \sum_{s,j}\left[\log \sigma_{\theta_j}^2 + \frac{(\theta_{js} - \tilde{\boldsymbol{\alpha}}_j^\top \tilde{\mathbf{x}}_s)^2}{\sigma_{\theta_j}^2}\right] \\
& \quad + \sum_{s,i}\left[\log \sigma_{b_i}^2 + \frac{(b_{is} - \beta_{0i})^2}{\sigma_{b_i}^2}\right] + \sum_i \left[\log \sigma_\beta^2 + \frac{\beta_{0i}^2}{\sigma_\beta^2}\right] \Bigg\} \Bigg] q(\mathcal{Z})\, \mathrm{d}\mathcal{Z} \\
\equiv & \int \mathcal{B}(\boldsymbol{Y}, \mathcal{Z}) q(\mathcal{Z})\, \mathrm{d}\mathcal{Z} \\
\equiv & \, Q'(\boldsymbol{Y}).
\end{aligned} \tag{A.6}
$$

The regularized objective is

$$
Q'_{\mathrm{reg}}(\boldsymbol{Y}) = Q'(\boldsymbol{Y}) - \lambda \left[\sum_j \log \sigma_{\theta_j}^2 + \sum_i \log \sigma_{b_i}^2\right]. \tag{A.7}
$$

A.3 E-STEP: VARIATIONAL POSTERIOR UPDATES

Under the mean-field assumption (Bishop, 2006), the optimal variational distribution for each latent variable $z_l \in \mathcal{Z}$ satisfies $q(z_l) \propto \exp \mathbb{E}_{\mathcal{Z}\setminus\{z_l\}}[\mathcal{B}(\boldsymbol{Y}, \mathcal{Z})]$. Thus, all optimal posteriors take Gaussian forms with the following sufficient statistics.

Posterior of θ_{js}.

$$
\sigma_{\theta_{js}}^2 = \frac{\sigma_{\theta_j}^2}{1 + 2\sigma_{\theta_j}^2\left[\sum_i \eta(\xi_{ijs})\right]}, \tag{A.8}
$$

$$
\mu_{\theta_{js}} = \frac{\sigma_{\theta_j}^2 \sum_i \left[y_{ijs} - \frac{1}{2} + 2\eta(\xi_{ijs})\mu_{b_{is}}\right] + \tilde{\boldsymbol{\alpha}}_j^\top \tilde{\mathbf{x}}_s}{1 + 2\sigma_{\theta_j}^2\left[\sum_i \eta(\xi_{ijs})\right]}. \tag{A.9}
$$

Posterior of b_{is}.

$$
\sigma_{b_{is}}^2 = \frac{\sigma_{b_i}^2}{1 + 2\sigma_{b_i}^2\left[\sum_j \eta(\xi_{ijs})\right]}, \tag{A.10}
$$

$$
\mu_{b_{is}} = \frac{-\sigma_{b_i}^2 \sum_j \left[y_{ijs} - \frac{1}{2} - 2\eta(\xi_{ijs})\mu_{\theta_{js}}\right] + \mu_{\beta_{0i}}}{1 + 2\sigma_{b_i}^2\left[\sum_j \eta(\xi_{ijs})\right]}. \tag{A.11}
$$

Posterior of β_{0i}.

$$
\sigma_{\beta_{0i}}^2 = \frac{\sigma_\beta^2 \sigma_{b_i}^2}{S\sigma_\beta^2 + \sigma_{b_i}^2}, \tag{A.12}
$$

$$
\mu_{\beta_{0i}} = \frac{\sigma_\beta^2 (\sum_s \mu_{b_{is}})}{S\sigma_\beta^2 + \sigma_{b_i}^2}. \tag{A.13}
$$

A.4 M-STEP: GLOBAL PARAMETER UPDATES

Given the optimal posterior distributions above, setting the derivative of $Q'_{\mathrm{reg}}(\boldsymbol{Y})$ with respect to each global parameter to zero yields the following closed-form updates.

Augmented prompt sensitivity vector $\tilde{\boldsymbol{\alpha}}_j$.

$$\tilde{\boldsymbol{\alpha}}_j = \left(\sum_s \tilde{\mathbf{x}}_s \tilde{\mathbf{x}}_s^\top \right)^{-1} \left(\sum_s \mu_{\theta_{js}} \tilde{\mathbf{x}}_s \right). \tag{A.14}$$

Residual variance $\sigma^2_{\theta_j}$.

$$\sigma^2_{\theta_j} = \frac{\sum_s \left[(\mu_{\theta_{js}} - \tilde{\boldsymbol{\alpha}}_j^\top \tilde{\mathbf{x}}_s)^2 + \sigma^2_{\theta_{js}} \right]}{S + 2\lambda}. \tag{A.15}$$

Residual variance $\sigma^2_{b_i}$.

$$\sigma^2_{b_i} = \frac{\sum_s \left[(\mu_{b_{is}} - \mu_{\beta_{0i}})^2 + \sigma^2_{b_{is}} + \sigma^2_{\beta_{0i}} \right]}{S + 2\lambda}. \tag{A.16}$$

Baseline difficulty variance σ^2_β.

$$\sigma^2_\beta = \frac{1}{I} \sum_i (\mu^2_{\beta_{0i}} + \sigma^2_{\beta_{0i}}). \tag{A.17}$$

The algorithm iterates between the E-step and M-step until the change in $Q'_{\text{reg}}(\boldsymbol{Y})$ falls below a convergence threshold ϵ.

B Detailed levels of the phrasing dimension

The D4: Phrasing dimension comprises the following 13 distinct prompt templates:

1. Below are multiple-choice questions related to {topic}, each followed by their respective answers. Question: {question} Choices: {choices} Correct Answer:
2. Below are multiple-choice questions related to {topic}. Please provide the correct answer for each question. Question: {question} Choices: {choices} Answer:
3. Could you provide a response to the following question: {question} {choices} Answer:
4. Here are some multiple choice questions along with their answers about {topic}. Question: {question} Choices: {choices} Correct Answer:
5. Please address the following question: {question} {choices} Answer:
6. Please answer the following question: {question} {choices} Answer:
7. Question: [question] Choices: [choices] Answer: [answer] Question: {question} Choices: {choices} Answer:
8. Question: {question} Choices: {choices} Answer:
9. The following are multiple choice questions (with answers) about {topic}. {question} {choices} Answer:
10. The following are multiple choice questions (with answers) about {topic}. Question: {question} {choices} Answer:
11. The following are multiple choice questions (with answers). {question} {choices} Answer:
12. The following are multiple choice questions (with answers). Question: {question} {choices} Answer:
13. Topic: {topic} Question: [question] Choices: [choices] Answer: [answer] Question: {question} Choices: {choices} Answer:

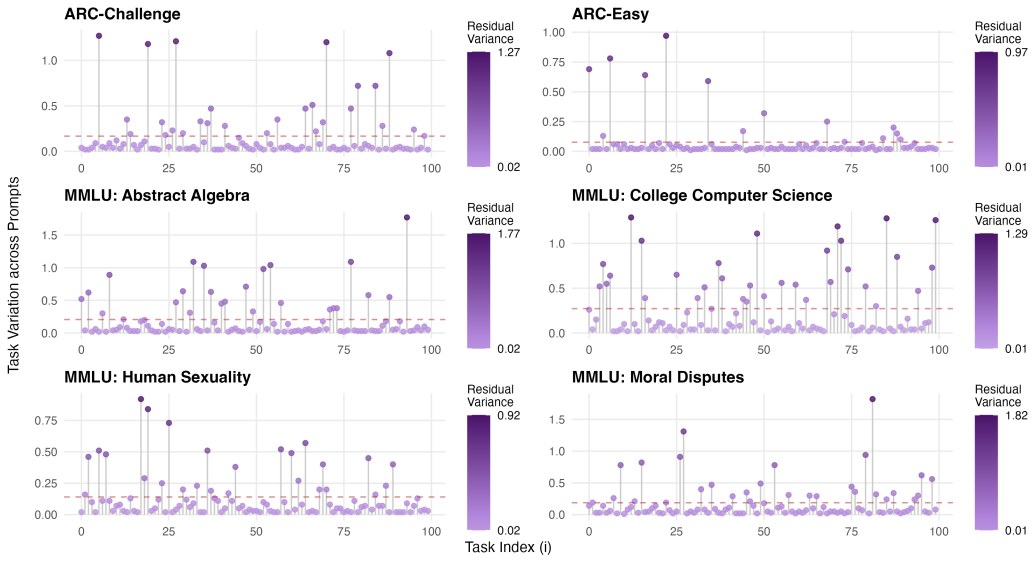


Figure 6: Item-specific variations across prompts on each benchmark.

C Empirical experiment

C.1 Task-level prompt sensitivity

Figure 6 shows task-level prompt sensitivity across benchmarks.

We also examine how filtering prompt-sensitive tasks changes accuracy-based prompt-effect estimates. Here, accuracy-based prompt-effect is defined as the accuracy difference between a target prompt configuration and the reference configuration. Table 5 reports these changes on ARC Challenge. Although the absolute changes are small (-0.03 to 0.05), they can be substantial relative to the unfiltered estimates, as indicated by the percentages in parentheses.

Table 5: Changes in accuracy-based prompt effects after task filtering. Relative changes are shown in parentheses.

Prompt dimension	Min	Max	Mean absolute change
D1: Enumerator	-0.03 (-165%)	0.03 (171%)	0.01 (45%)
D2: Separator	-0.02 (-395%)	0.03 (1236%)	0.01 (113%)
D3: Choice order	-0.03 (-108%)	0.05 (90%)	0.01 (40%)
D4: Phrasing	0.00 (-397%)	0.03 (2705%)	0.01 (156%)
D5: Demonstrations	-0.01 (-2049%)	0.03 (521%)	0.01 (171%)

C.2 Model-level prompt sensitivity

Figure 7 illustrates the model-specific prompt effects on each benchmark.

C.3 Prompt sampling

We use full-data estimates as the comparison baseline. Subsets of 200, 600, and 1,000 configurations require 96.67%, 90.01%, and 83.35% fewer responses, respectively. We repeat sampling 50 times using existing responses to assess variability. In a future evaluation, responses would need to be collected for only one sampled subset, yielding the reductions above.

Reference-prompt model ranking agreement is the proportion of the $\binom{5}{2} = 10$ model pairs whose relative capability rankings match the full-data results. Thus, 100% agreement denotes an exact



Figure 7: Model-specific prompt sensitivity estimates $\hat{\boldsymbol{\alpha}}_j$ across five prompt dimensions and five models on each benchmark.

ranking, 90% denotes one discordant pair, and 80% denotes two discordant pairs. Figure 5 reports the fraction of replications in each agreement category. Exact ranking recovery generally improves with subset size but varies across benchmarks. MMLU Human Sexuality is an exception: exact agreement decreases from approximately 54% at $m = 600$ to 48% at $m = 1{,}000$, despite higher capability correlations.

Recovery of reference-prompt model capability θ_j and model-specific additive prompt effects $\boldsymbol{\alpha}_j$ is evaluated by Pearson correlation, bias, and RMSE relative to their full-data estimates (Figure 8). For both parameters, the bias remains near zero, while RMSE decreases and correlations increase with subset size. At $m = 1{,}000$, all correlations exceed 0.90 and RMSE values are below 0.1.

D Simulation Experiment

We conduct two complementary simulation studies. The first study compares different evaluation methods in recovering the true model ranking under the reference prompt configuration. The second study assesses parameter recovery by the RGVEM estimator. All models are synthetic, the number of tasks is fixed at $I = 100$, and each condition is replicated 50 times.

D.1 Simulation 1: Recovery of Model Rankings

D.1.1 Experiment Setup

We compare rankings from TrIEval, average accuracy, and a standard Rasch model against a known baseline ordering. We fix $I = 100$, $J = 6$, $S = 2500$, and $\sigma_\beta^2 = 1$, and vary the capability gap be-

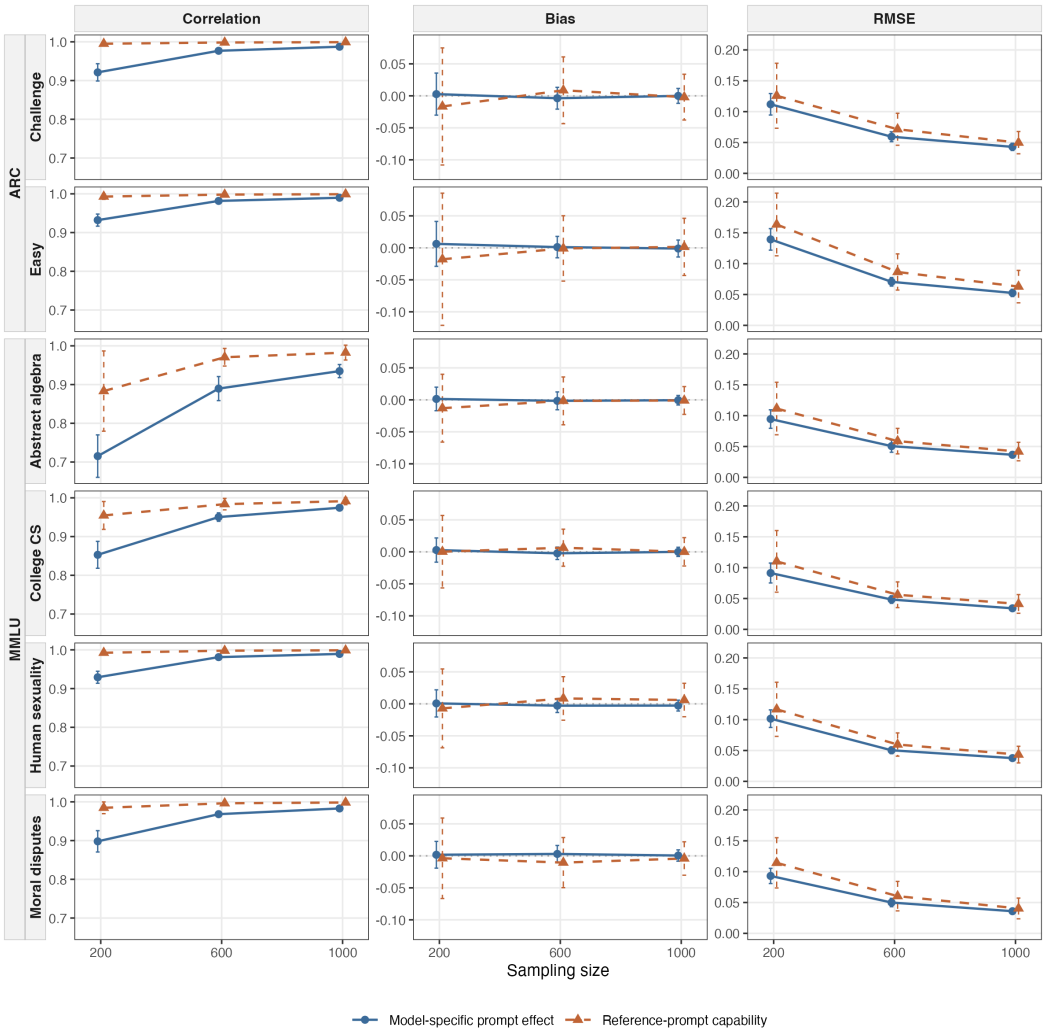


Figure 8: Agreement between sampled-data and full-data estimates across six benchmarks, using 50 replications at each subset size. Columns report Pearson correlation, bias, and RMSE; blue circles represent model-specific additive prompt effects and orange triangles represent reference-prompt capability. Error bars show standard deviations across replications.

tween adjacent models (0.2 or 0.6). For gap 0.2, the capabilities are $(-0.5, -0.3, -0.1, 0.1, 0.3, 0.5)$, and for gap 0.6, the capabilities are $(-1.5, -0.9, -0.3, 0.3, 0.9, 1.5)$. These distinct capabilities define a strict ranking, allowing us to assess how accurately each method recovers the ordering. Ranking error is measured by the number of models assigned an incorrect rank position.

D.1.2 RANKING RECOVERY

As shown in Table 4, at capability gap 0.2, the mean mismatch counts are 0.24 for TRIEVAL, 4.48 for average accuracy, and 4.36 for the Rasch model. At gap 0.6, the corresponding means are 0, 2.66, and 2.64. These results demonstrate a more accurate recovery of the specified baseline ranking by TRIEVAL.

D.2 SIMULATION 2: RECOVERY OF PARAMETERS

D.2.1 EXPERIMENTAL SETUP

To assess the statistical properties of the RGVEM estimator, we conduct a simulation study designed to mirror the empirical evaluation setting. Three factors are manipulated across eight conditions: (i) model scale at $J \in \{6, 20\}$; (ii) baseline difficulty variance at $\sigma_\beta^2 \in \{1, 2\}$; and (iii) prompt structure across five dimensions with levels $(10 \times 5 \times 5 \times 5 \times 2)$ or $(5 \times 5 \times 2 \times 2 \times 2)$, yielding $S \in \{200, 2500\}$ unique configurations.

The number of tasks is fixed at $I = 100$, matching the empirical study. Baseline model capability is set to $\theta_j = -1$ for odd-indexed models and $\theta_j = 1$ for even-indexed models. This alternating design evaluates numerical parameter recovery rather than a strict ranking. Models are divided into two equal groups: the first half have no prompt interactions ($\sigma_{\theta_j}^2 = 0$) and the second half have non-zero prompt interactions ($\sigma_{\theta_j}^2 \neq 0$). On a laptop with an Intel i7-12700H processor, RGVEM converges in under one minute per replication.

Recovery is assessed for the baseline model capability θ_j, model-specific prompt sensitivity $\boldsymbol{\alpha}_j$, and the sparsity structure of the residual interaction variances $\sigma_{\theta_j}^2$ and $\sigma_{b_i}^2$.

D.2.2 RECOVERY OF BASELINE MODEL CAPABILITY

Table 6 reports recovery of θ_j across all conditions. The correlation between estimated and true capability is near-perfect across all conditions, ranging from 0.98 to 1.00, with RMSE between 0.18 and 0.25. Recovery remains strong across different numbers of evaluated models, prompt configurations, and difficulty variances.

Table 6: Recovery of baseline model capability θ_j across simulation conditions. Values are mean (SD) across 50 replications.

J	S	σ_β^2	RMSE	Bias	Correlation
6	200	1	0.21 (0.08)	0.07 (0.12)	0.99 (0.01)
6	200	2	0.22 (0.07)	0.11 (0.10)	0.99 (0.01)
6	2500	1	0.18 (0.06)	0.15 (0.09)	1.00 (0.00)
6	2500	2	0.18 (0.07)	0.12 (0.10)	1.00 (0.00)
20	200	1	0.24 (0.05)	0.15 (0.07)	0.99 (0.01)
20	200	2	0.25 (0.06)	0.17 (0.07)	0.98 (0.01)
20	2500	1	0.25 (0.06)	0.24 (0.06)	1.00 (0.00)
20	2500	2	0.25 (0.07)	0.23 (0.08)	1.00 (0.00)

D.2.3 RECOVERY OF MODEL-SPECIFIC PROMPT SENSITIVITY

Table 7 shows the recovery of $\boldsymbol{\alpha}_j$. Recovery improves substantially with prompt diversity: correlation increases from approximately 0.81 at $S = 200$ to 0.97 at $S = 2500$, accompanied by a marked reduction in RMSE from 0.14 to 0.06. Recovery is stable across model scale and difficulty variance, indicating that prompt diversity is the primary determinant of sensitivity estimation quality. This

confirms that a large number of prompt configurations is necessary to reliably disentangle model-specific prompt sensitivity from the baseline capability under a reference prompt.

Table 7: Recovery of model-specific prompt sensitivity $\boldsymbol{\alpha}_j$ across simulation conditions. Values are mean (SD) across 50 replications.

J	S	σ^2_β	RMSE	Bias	Correlation
6	200	1	0.14 (0.02)	0.01 (0.03)	0.81 (0.05)
6	200	2	0.14 (0.02)	0.02 (0.02)	0.80 (0.05)
6	2500	1	0.05 (0.01)	0.01 (0.01)	0.98 (0.01)
6	2500	2	0.06 (0.01)	0.01 (0.01)	0.97 (0.01)
20	200	1	0.15 (0.01)	0.01 (0.01)	0.81 (0.03)
20	200	2	0.15 (0.01)	0.02 (0.02)	0.81 (0.03)
20	2500	1	0.06 (0.00)	0.01 (0.01)	0.97 (0.00)
20	2500	2	0.06 (0.00)	0.01 (0.00)	0.97 (0.00)

D.2.4 Recovery of sparsity structure

TriEval achieves perfect identification of prompt interactions on model capability ($\sigma^2_{\theta_j}$) across all conditions. Majority models with true non-zero interactions are correctly identified and no spurious interactions are detected, as shown in Figure 9.

Recovery of task-level interactions ($\sigma^2_{b_i}$) is shown in Figure 10, and it is more condition-dependent: true positive rate (TPR) increases substantially when moving from $S = 200$ to $S = 2500$ prompt configurations or from $J = 6$ to $J = 20$ models, while baseline difficulty variance σ^2_β has negligible impact on detection accuracy. The consistently low false positive rate (FPR) across all conditions confirms that the log-penalty successfully induces sparsity, suppressing the detection of non-existent interactions.

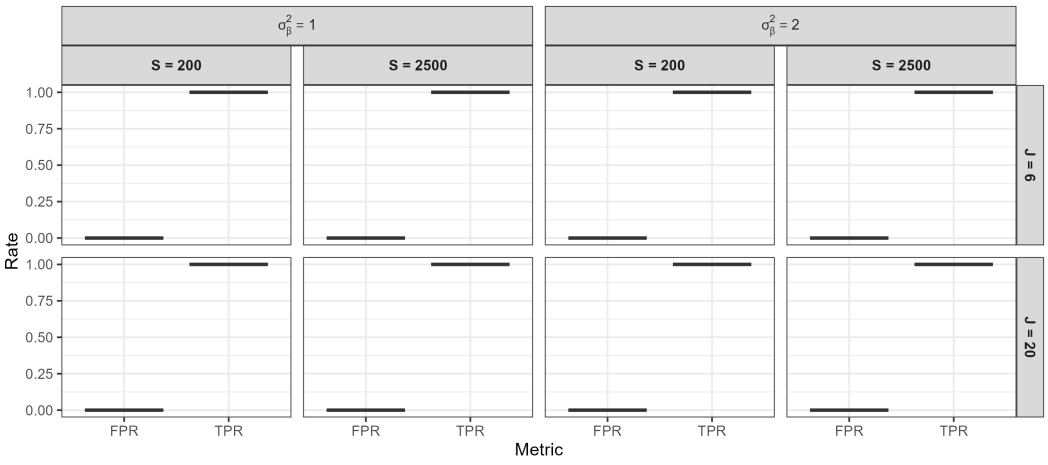


Figure 9: Sparsity recovery of $\sigma^2_{\theta_j}$.

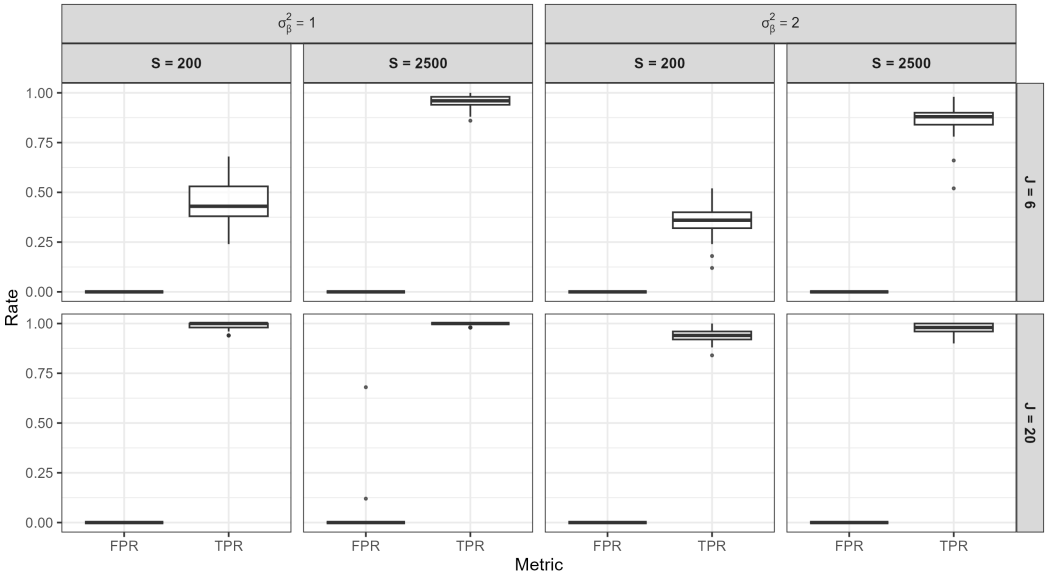


Figure 10: Sparsity recovery of $\sigma^2_{b_i}$.