OPTION-ORDER EFFECTS IN LLM EVALUATION: SAMPLING VARIABILITY AND MEASUREMENT PRECISION

Anonymous authors
Paper under double-blind review

ABSTRACT

Option-order sensitivity complicates multiple-choice evaluation of large language models (LLMs), but its magnitude must be interpreted relative to repeated-generation variability and measurement precision. We evaluate six open-weight LLMs on 25 text-based items from the International Cognitive Ability Resource under eight cyclic option rotations, with two greedy and five stochastic administrations per rotation, yielding 8,400 item responses. We estimate domain-specific abilities using a between-item multidimensional item response theory model with fixed parameters reconstructed from human response summaries and standardized factor loadings. A mixed-effects analysis separates shared rotation effects, model-specific rotation effects, and residual variation across draws. Under stochastic decoding ($T = 0.7$), the estimated order component has an SD of 0.08–0.10 ability units, compared with 0.30–0.33 across draws. Under greedy decoding, 99.7% of scoreable repeated item-response pairs are identical, yet the order component has an SD of 0.28–0.30. Thus, repeatability under a fixed presentation does not imply robustness to option order. Across both regimes, these pooled components remain below half the mean information-based conditional standard error of measurement. These comparisons are conditional on the reconstructed response model and do not establish equivalence between models. Our findings show how reporting order sensitivity alongside repeated-generation variability and test information can clarify what short evaluations support.

1 INTRODUCTION

Benchmark results are central to claims about the capabilities of large language models (LLMs), but a reported score depends on both the model and the evaluation procedure. Changes in the order of the answer-options, prompt formatting, and inference configuration can alter the measured accuracy and, in some cases, the model rankings (Alzahrani et al., 2024; Pezeshkpour & Hruschka, 2024; Sclar et al., 2024; Yuan et al., 2025). Such sensitivity complicates comparisons: an observed score difference may reflect the capability of interest, a systematic response to the evaluation procedure, or variation across repeated runs. Distinguishing these sources requires both repeated evaluations and an explicit account of measurement uncertainty.

We study the sensitivity of option-order in multiple-choice evaluation. Reordering the same answer options preserves their content, yet can change an LLM's answer. Previous work documents preferences for particular option labels or positions and proposes methods to reduce their influence (Zheng et al., 2024; Pezeshkpour & Hruschka, 2024). However, the size of an observed order effect needs a reference point. Under stochastic decoding, repeating an otherwise identical evaluation can also change the score (Miller, 2024; Madaan et al., 2024). A difference between two option orders therefore does not, by itself, establish how much variation is attributable to order. Nor does it establish whether that variation is large relative to the precision with which the test measures performance. Our aim is to place order sensitivity in both contexts: repeated-run variability and measurement precision.

Item response theory (IRT) provides a framework for the second comparison. Accuracy remains a useful outcome and can be accompanied by uncertainty estimates (Miller, 2024). IRT additionally

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.

models each item's difficulty and discrimination, so an ability estimate $\hat{\theta}$ can be accompanied by the precision the items themselves provide: the conditional standard error of measurement, $\mathrm{CSEM}(\theta) = I(\theta)^{-1/2}$, where $I(\theta)$ is the test information (Lord, 1980; Embretson & Reise, 2000). We use the CSEM as a reference for what the instrument can resolve, not as a significance threshold for effects estimated from repeated administrations (Section 2.5). We ask whether variation in $\hat{\theta}$ across option orders exceeds variation across repeated stochastic generations, and how both compare with the CSEM of a single administration.

To keep the scale independent of the six models under study, we fix item parameters at values derived from human response data for the International Cognitive Ability Resource (ICAR; Condon & Revelle, 2014), a public-domain cognitive ability test. Calibrating items on LLM responses is the main alternative, but it can be unreliable when few models are available (Jiang et al., 2026), and it would fit the scale to the very responses whose order sensitivity we aim to measure. External calibration rests on a different assumption: that the human-calibrated response model adequately describes LLM responses. Fixing the parameters gives every model the same scoring rule, but it does not by itself establish measurement equivalence between humans and LLMs, so we interpret all scores conditionally on this assumption.

We administer 25 ICAR items (16 verbal reasoning, 9 letter–number series) to six open-weight models of up to 14B parameters under fixed inference settings. We exclude matrix-reasoning and three-dimensional-rotation items because converting them to text would itself change the assessment. Each model receives eight cyclic rotations of the answer options, each administered twice under greedy decoding ($T{=}0$) and five times under sampling ($T{=}0.7$), yielding 56 administrations per model and 8,400 item-level responses (Section 2). The design targets cyclic rotations in two domains; it does not characterize all option permutations or all LLM benchmarks.

We address three research questions:

- **RQ1.** How does variation in estimated ability across option rotations compare with variation across repeated draws, and how does this comparison depend on the decoding regime?
- **RQ2.** How large are these sources of variation relative to the mean information-based CSEM of each subscale?
- **RQ3.** Which differences between models can the instrument resolve, given its own measurement error?

Our main findings are:

- **Under sampling, order variation is smaller than sampling variation.** At $T{=}0.7$, the order SD of a model's ability estimate is 0.08–0.10 on the calibrated scale (0.11–0.14 of the mean CSEM), whereas the SD across repeated draws is 0.30–0.33 (0.41–0.44 of the mean CSEM), about 3.5 times larger.
- **Repeatability does not imply robustness to option order.** Under greedy decoding, 99.7% of repeated item responses are identical, yet the order SD is 0.28–0.30 (0.39–0.42 of the mean CSEM). Most of this variation is model-specific: rotations shift each model's estimate by different amounts and in different directions, while the effect shared by all models is small (0.07–0.08). The estimated order component is smaller under stochastic decoding, but this comparison does not establish whether averaging repeated runs accounts for the reduction.
- **The instrument, not option order, limits comparisons between models.** Neither source of variation exceeds half a CSEM, but the subscales' own measurement error is large. Although 11–14 of the 15 model pairs differ significantly in each subscale and temperature, only the highest- versus lowest-scoring pair exceeds the minimum detectable difference (≈ 2.0 on the θ scale) in both subscales at both temperatures; one further pair does so only for letter–number series under greedy decoding.

These findings qualify the interpretation of option-order effects. Under greedy decoding, order variation is largely model-specific and exceeds variation across repeated draws. Under stochastic decoding, the estimated order component is smaller than the draw component. In both regimes, these pooled SDs remain below half the mean CSEM. This does not establish that order effects are

absent or that debiasing is unnecessary. It shows why their magnitude should be reported alongside repeated-run variability and measurement uncertainty when benchmark results are used to compare models.

1.1 Related Work

Option-order sensitivity and its mitigation. LLMs can exhibit preferences for particular option labels or positions when answering multiple-choice questions (Zheng et al., 2024; Pezeshkpour & Hruschka, 2024). Their observed choice distributions and sensitivity to prompt perturbations also depend on whether evaluation uses first-token probabilities or answers extracted from generated text (Wang et al., 2024). Inference-time mitigation strategies include averaging predictions across option permutations (Zheng et al., 2024; Säuberli et al., 2025) and estimating and correcting the model's prior preference over option labels (PriDe; Zheng et al., 2024). Related work addresses position bias more broadly: PINE modifies attention and positional treatment to make inference invariant to document order (Wang et al., 2025). Our study complements these approaches by examining the magnitude of option-order variation relative to repeated-generation variability and the test information available for estimating ability.

IRT for LLM evaluation. IRT has been applied to benchmark compression (Maia Polo et al., 2024; Kipnis et al., 2025), adaptive evaluation (Hofmann et al., 2025; Truong et al., 2025), and diagnosis of benchmark quality (Rodriguez et al., 2021; Zhou et al., 2026). These approaches draw on LLM response data to estimate item characteristics or learn models of item difficulty. The reliability of such estimates depends on the calibration setting: Jiang et al. (2026) identify conditions under which small or non-normally distributed model samples lead to unreliable item-level and ranking inferences. Related studies use psychometric models to compare human and LLM performance, assess human–AI complementarity, and evaluate LLMs as simulated respondents (Gor et al., 2024; Liu et al., 2025; Srivatsa et al., 2025; Säuberli et al., 2025). For example, Liu et al. (2025) find that repeated responses from individual LLMs yield narrower estimated proficiency distributions than those observed among human respondents in their college-algebra assessment. Our focus is the interpretation of evaluation variability. We reconstruct fixed item parameters from ICAR's human response data, independently of the LLM responses under study. Within each domain, we compare option-order variation, sampling variation, and shifts induced by scoring decisions on the same ability scale. Test information supplies a conditional precision reference for these comparisons. Its interpretation depends on the reconstructed response model adequately describing LLM performance.

2 Method

2.1 Instrument

The International Cognitive Ability Resource (ICAR; Condon & Revelle, 2014) is a public-domain cognitive ability measure of 60 items in four types: 9 Letter and Number Series (LN), 11 Matrix Reasoning, 16 Verbal Reasoning (VR), and 24 Three-dimensional Rotation items. It was validated on 96,958 online participants aged 14–90 from 199 countries, each of whom completed a random subset of items. We use the 25 text-based items (16 VR, 9 LN) and exclude the 35 image-based items.

To define the scale independently of the evaluated models, we derive two-parameter logistic (2PL) parameters for each item from the human data: its proportion correct p_j and its standardized loading λ_j on its domain factor in a correlated two-factor solution. Using the equivalence between the normal-ogive model and factor analysis of dichotomous variables (Takane & de Leeuw, 1987),

$$\tau_j = \Phi^{-1}(1 - p_j), \qquad a_j = D\,\frac{\lambda_j}{\sqrt{1-\lambda_j^2}}, \qquad b_j = \frac{\tau_j}{\lambda_j}, \qquad D = 1.702, \tag{1}$$

where Φ is the standard normal CDF and D rescales normal-ogive slopes to the logistic metric. The two factors correlate at $\rho = 0.904$. All item parameters and ρ are held fixed when scoring model responses.

2.2 MODELS AND INFERENCE

We evaluate six open-weight instruction-tuned models from three families: Phi-4 (Abdin et al., 2024) and Phi-4-mini (Microsoft et al., 2025); Qwen3 (Yang et al., 2025) and Qwen3.5 (Qwen Team, 2026); and Gemma 4 (Gemma Team, 2026). By Ollama tag, they are `phi4-mini:3.8b`, `phi4:latest` (14B), `qwen3:4b`, `qwen3.5:9b`, `gemma4:e2b`, and `gemma4:12b`. Models are served locally through Ollama v0.34.1 (Ollama, 2026) with Q4_K_M GGUF quantization (Gerganov & ggml contributors, 2026). Each model runs on one of two Apple-silicon laptops with identical software. In every condition we set `top_p`$= 1$ and `top_k`$= 0$, which disables nucleus and top-k truncation, so temperature is the only source of sampling randomness; `num_ctx`$= 8192$ and `num_predict`$= 2048$. Each administration uses a fixed seed, shared across items and models (Appendix A.1).

Four models (Qwen3, Qwen3.5, and both Gemma 4 models) support a thinking mode and use it by default. A pilot showed that, with thinking enabled, most deliberation moved into a separate reasoning field in model-specific proportions (7.2× the explanation length for `qwen3:4b`, 2.3× for `qwen3.5:9b`), and `gemma4:12b` exhausted its token budget before producing an answer. We therefore disable thinking for all models (`think=False`); the reasoning field was empty in all 8,400 responses. Each item is sent in its own stateless request.

Output is constrained by a JSON schema with an `explanation` field and a `final_answer` field whose value must be one of the eight presented option texts, so the task is forced-choice recognition under grammar-constrained decoding. In every response, the explanation is generated before the answer. Because each call is stateless, an administration is a logical grouping of 25 calls that share a rotation, temperature, and seed, not a continuous session; calls are executed in a shuffled order within each model and domain.

2.3 EXPERIMENTAL DESIGN

We manipulate two factors:

Option order. Each item has eight options, including the non-substantive options "None of these" and "I don't know." We apply eight cyclic rotations of all eight options, so that each keyed answer appears in every presented position exactly once. Rotation is applied to all items in an administration. Options are displayed with position numbers (1)–(8), but the output schema only accepts option text, so a model cannot answer with a position number.

Decoding. Each rotation is administered under greedy decoding (T=0, two draws, which lets us measure determinism directly) and under sampling (T=0.7, five draws).

An *administration* is one pass through all 25 items under a fixed rotation, temperature, and draw. Each model therefore completes $8 \times (2 + 5) = 56$ administrations, or 1,400 item responses, for 8,400 responses in total.

2.4 IRT SCORING MODEL

Let $x_{ij} \in \{0, 1\}$ be the scored response to item j in administration i, and let $\boldsymbol{\theta}_i = (\theta_{i,\mathrm{VR}}, \theta_{i,\mathrm{LN}})^\top$ be the latent abilities for that administration. Each item loads on exactly one domain $d(j)$ (a between-item two-dimensional IRT model; Reckase, 2009):

$$P_j(\boldsymbol{\theta}_i) \equiv \Pr(x_{ij} = 1 \mid \boldsymbol{\theta}_i) = \sigma\big(a_j\big(\theta_{i,d(j)} - b_j\big)\big)\,, \tag{2}$$

where σ is the logistic function. Because a_j and b_j are fixed (Eq. 1), only $\boldsymbol{\theta}_i$ is estimated. We place a bivariate normal prior on it, $\boldsymbol{\theta}_i \sim \mathcal{N}(\mathbf{0}, \Sigma)$ with unit variances and correlation ρ, and estimate each domain's ability by its expected a posteriori (EAP) value (Bock & Mislevy, 1982):

$$\hat{\theta}_{id} = \frac{\int \theta_d\, L_i(\boldsymbol{\theta})\, \phi(\boldsymbol{\theta}; \mathbf{0}, \Sigma)\, d\boldsymbol{\theta}}{\int L_i(\boldsymbol{\theta})\, \phi(\boldsymbol{\theta}; \mathbf{0}, \Sigma)\, d\boldsymbol{\theta}} \approx \frac{\sum_q \theta_d^{(q)}\, L_i(\boldsymbol{\theta}^{(q)})\, w_q}{\sum_q L_i(\boldsymbol{\theta}^{(q)})\, w_q}, \qquad L_i(\boldsymbol{\theta}) = \prod_{j=1}^{25} P_j(\boldsymbol{\theta})^{x_{ij}} \big[1 - P_j(\boldsymbol{\theta})\big]^{1 - x_{ij}}, \tag{3}$$

where $\{\boldsymbol{\theta}^{(q)}, w_q\}$ are the nodes and normalized prior weights of a 61×61 quadrature grid with range -6 to 6. EAP yields finite estimates for all-correct response patterns, which matters here because several models are at or near ceiling.

2.5 MEASUREMENT PRECISION

The conditional standard error of measurement (CSEM) is the precision supplied by the items alone. Because each item loads on a single domain and responses are locally independent, the information matrix is diagonal, and the CSEM is computed separately for each domain:

$$I_d(\theta) = \sum_{j:\, d(j)=d} a_j^2\, P_j(\theta)\big[1 - P_j(\theta)\big], \qquad \mathrm{CSEM}_d(\theta) = I_d(\theta)^{-1/2}. \tag{4}$$

We evaluate it at each model's mean estimate $\bar{\theta}_{md}$ within each temperature, and write $\overline{\mathrm{CSEM}}_d$ for its average over the six models. The CSEM is our yardstick for what the instrument can resolve.

2.6 MISSING RESPONSES

Of the 8,400 responses, 86 (1.0%) could not be scored, almost all because generation reached the 2,048-token limit before producing valid JSON. These failures come from model behavior rather than the harness. Most (68) come from `qwen3:4b`, which cycles through candidate answers on the same hard items until the limit, reproducibly across administrations. The rest are spread across `qwen3.5:9b` (7), `phi4-mini` (7), and the two Gemma models (2 each). Because the failures concentrate on the hardest items, they are not missing at random. We therefore score them as incorrect in all main analyses, so that all 25 items enter the likelihood function.

2.7 ANALYSIS

All analyses are run separately for each domain (LN, VR) and temperature.

Variance decomposition. Our primary analysis is a linear mixed-effects model fitted to the administration-level ability estimates,

$$\hat{\theta}_{mrk} = \mu + \alpha_m + \beta_r + v_{mr} + \varepsilon_{mrk}, \qquad v_{mr} \sim \mathcal{N}(0, \sigma^2_{m\times r}), \quad \varepsilon_{mrk} \sim \mathcal{N}(0, \sigma^2_{\mathrm{draw}}), \tag{5}$$

where α_m and β_r are fixed effects of model m and rotation r, v_{mr} is a random intercept for each model $\times$ rotation cell, and ε_{mrk} is variation across repeated draws k within a cell (two draws at $T{=}0$, five at $T{=}0.7$). Three quantities summarize option-order effects. The shared rotation effect is $\hat{\sigma}_{\mathrm{rot}}$, the SD of the eight estimated rotation effects $\hat{\beta}_r$, tested with a Wald $\chi^2(7)$ test. The model-specific rotation effect is $\hat{\sigma}_{m\times r}$. Their combination, $\hat{\sigma}_{\mathrm{order}} = (\hat{\sigma}^2_{\mathrm{rot}} + \hat{\sigma}^2_{m\times r})^{1/2}$, is the order variation experienced by a given model. We report all quantities as SDs on the θ scale and compare them with $\overline{\mathrm{CSEM}}_d$.

Distinguishability. We test all 15 pairwise model contrasts from the model's estimated marginal means with Holm correction (Holm, 1979), and call a pair distinguishable only if its difference also exceeds the minimum detectable difference $\mathrm{MDD}_d = 1.96\sqrt{2}\,\overline{\mathrm{CSEM}}_d$, the 95% threshold for separating two single-administration estimates whose standard error equals the mean CSEM. Because every administration reuses the same 25 items, Eq. 5 treats correlated replicates as independent. Its tests and contrast standard errors therefore reflect how reproducible the runs are, not how precisely the instrument measures ability; we interpret them descriptively and judge magnitudes against the CSEM.

Item-level dispersion. As an item-level check, we fit a logistic regression, that is, a binomial generalized linear model (GLM) with a logit link, to the binary item responses: $\mathrm{correct} \sim \mathrm{rotation} + \mathrm{model} + \mathrm{draw} + \mathrm{item}$, with standard errors clustered on administration (model $\times$ rotation $\times$ draw). The dispersion ratio is the Pearson χ^2 statistic divided by its residual degrees of freedom, computed per model at $T{=}0.7$; values above 1 indicate more variation across draws than binomial sampling alone predicts.

3 RESULTS

Figure 1 shows each model's mean $\hat{\theta}$ by rotation under greedy decoding, Figures 2 and 3 show the mean and SD across draws at $T{=}0.7$, and Table 1 summarizes the variance decomposition.

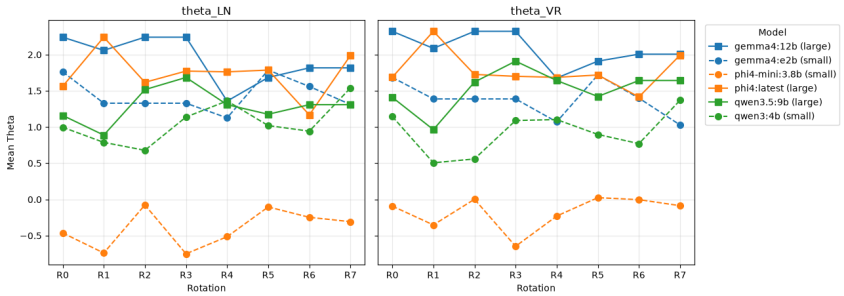


Figure 1: Mean $\hat{\theta}$ by option rotation for each model under greedy decoding (T=0). Differences between rotations within a model show the model-specific order effect.

For temperature 0.7, both mean and the standard deviation of the theta scores of each LLM and each option rotation is presented in Figure 2 and Figure 3.

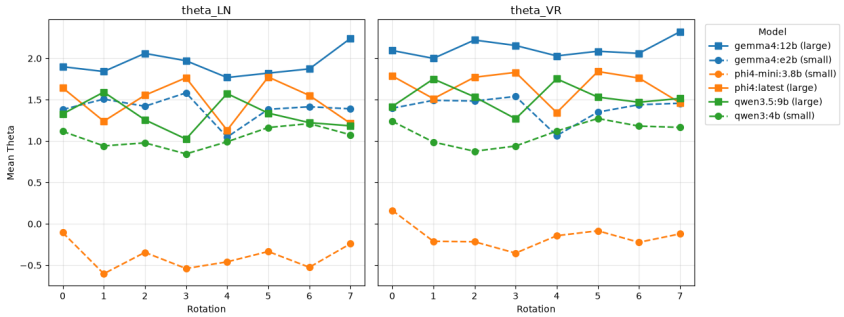


Figure 2: Mean $\hat{\theta}$ across the five draws at each option rotation, by model (T=0.7).

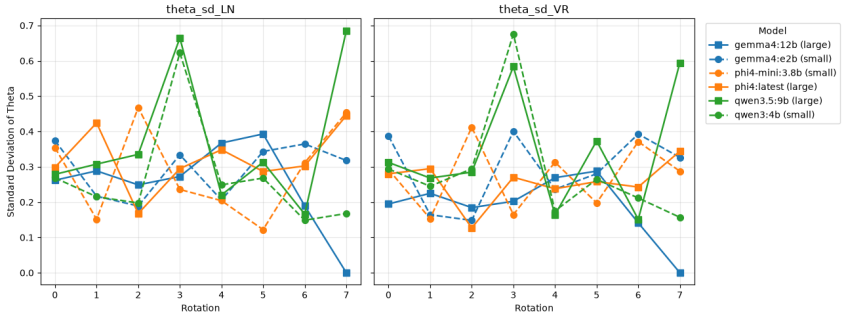


Figure 3: SD of $\hat{\theta}$ across the five draws at each option rotation, by model and domain (T=0.7).

The dominant source of instability depends on the decoding regime (RQ1). Under greedy decoding (T=0), repeating a rotation barely moved $\hat{\theta}$. Of the 1,192 repeated greedy item pairs with two scoreable responses, 1,189 (99.7%) were identical, and all three differences occurred in `phi4-mini:3.8b`. The draw SD was 0.01 (1–2% of the mean CSEM), and the median absolute difference between the two draws was 0 in all 48 model $\times$ rotation cells in both domains. Option order, however, did move $\hat{\theta}$. The rotation effect shared by all models was small ($\hat{\sigma}_{\text{rot}} = 0.07$–$0.08$; LN $\chi^2(7) = 3.15$, $p = .87$; VR $\chi^2(7) = 3.28$, $p = .86$), but model-specific rotation effects were large ($\hat{\sigma}_{m\times r} = 0.27$–$0.29$): each rotation shifted each model's estimate by a different amount and in a different direction. The resulting order SD for a given model, $\hat{\sigma}_{\text{order}}$, was 0.28–0.30 (Table 1; Figure 1). Under sampling (T=0.7) the pattern reversed. The draw SD rose to 0.30–0.33, while $\hat{\sigma}_{m\times r}$ fell roughly fourfold to 0.07, giving $\hat{\sigma}_{\text{order}} = 0.08$–$0.10$ (LN $\chi^2(7) = 8.00$, $p = .33$; VR

Table 1: Variance components of $\hat{\theta}$ by decoding regime and domain.

	$T = 0$		$T = 0.7$	
	LN	VR	LN	VR
Standard deviations (θ units)				
$\hat{\sigma}_{\text{rot}}$	0.075	0.071	0.067	0.051
$\hat{\sigma}_{m\times r}$	0.292	0.273	0.074	0.067
$\hat{\sigma}_{\text{order}}$	0.302	0.282	0.100	0.084
$\hat{\sigma}_{\text{draw}}$	0.007	0.013	0.326	0.299
Mean CSEM	0.724	0.721	0.737	0.735
Relative to mean CSEM				
$\hat{\sigma}_{\text{rot}}$ / CSEM	0.10	0.10	0.09	0.07
$\hat{\sigma}_{\text{order}}$ / CSEM	0.42	0.39	0.14	0.11
$\hat{\sigma}_{\text{draw}}$ / CSEM	0.01	0.02	0.44	0.41
Shared rotation effect, $\chi^2(7)$ (p)	3.15 (.87)	3.28 (.86)	8.00 (.33)	5.67 (.58)
MDD (95%, θ)	2.01	2.00	2.04	2.04
Distinguishable pairs (of 15)	2	1	1	1

Note. Components are from Eq. 5. $\hat{\sigma}_{\text{rot}}$: rotation effect shared by all models; $\hat{\sigma}_{m\times r}$: model-specific rotation effect; $\hat{\sigma}_{\text{order}} = (\hat{\sigma}^2_{\text{rot}} + \hat{\sigma}^2_{m\times r})^{1/2}$: order variation experienced by a given model; $\hat{\sigma}_{\text{draw}}$: variation across repeated draws. Ratios divide by the mean CSEM across models. $\text{MDD} = 1.96\sqrt{2}\,\overline{\text{CSEM}}$.

$\chi^2(7) = 5.67$, $p = .58$). Sampling variation then exceeded order variation by a factor of 3.3–3.6. The dominant source of variation therefore differs by decoding regime: option order under greedy decoding and repeated draws under stochastic decoding. The smaller order component at $T = 0.7$ should not be attributed to averaging alone, because temperature also changes the response distribution.

Both sources of variation are smaller than the instrument's resolution (RQ2). The largest source of variation in each regime stayed well below one CSEM: order variation under greedy decoding reached 0.39–0.42 of the mean CSEM, and sampling variation at T=0.7 reached 0.41–0.44. Under sampling, order variation was 0.11–0.14 of the mean CSEM. An auxiliary dispersion analysis at T=0.7 showed that draw-level instability was uneven across models: `qwen3:4b` (dispersion ratio 2.69) and `qwen3.5:9b` (1.57) varied more than binomial sampling alone predicts, while the other four models were at or below that expectation (0.66–1.14).

Statistical significance overstates which models can be distinguished (RQ3). Nearly all pairwise contrasts were statistically significant after Holm correction (11–14 of 15 pairs per subscale and temperature). These tests treat each model's administrations as independent replicates, but every administration reuses the same 25 items, so their standard errors reflect how reproducible the runs are, not how precisely the instrument measures ability. Against the CSEM-based minimum detectable difference (≈ 2.0 at both temperatures), only one or two pairs per subscale were distinguishable. The two models at the extremes, `gemma4:12b` and `phi4-mini:3.8b`, were distinguishable in both subscales at both temperatures (T=0: LN $\Delta = -2.33$, VR $\Delta = -2.25$; T=0.7: LN $\Delta = -2.42$, VR $\Delta = -2.28$). At T=0, `phi4-mini:3.8b` versus `phi4:latest` also cleared the threshold in LN ($\Delta = 2.14$) but not in VR ($\Delta = 1.95$ versus MDD $= 2.00$). All remaining pairs, including differences approaching 2.0, were statistically significant but below the instrument's resolving power at both temperatures.

4 Discussion

For the six models and two short subscales studied here, the size of option-order effects depends on the decoding regime. Under greedy decoding, the estimated order SD is about 0.3 on the θ scale and is driven mainly by model-specific rotation effects. Under stochastic decoding, the estimated order SD is about 0.1, compared with a draw SD of about 0.3. In both regimes, every source of variation stays below half a CSEM, and the instrument's measurement error, not option order, limits which models can be told apart.

Implications for order debiasing. Prior work treats option-order sensitivity as a robustness failure (Alzahrani et al., 2024; Pezeshkpour & Hruschka, 2024) and targets it with inference-time debiasing (Zheng et al., 2024; Wang et al., 2025). Our decomposition gives a more specific picture. The component shared across models, the kind of systematic position preference that prior-based methods such as PriDe estimate and remove, was small in both regimes and not statistically significant. Model-specific rotation effects were larger under greedy decoding. Averaging independent draws reduces Monte Carlo variability in the estimated score for a fixed rotation, but does not necessarily reduce differences between the expected scores of different rotations. Averaging predictions across rotations, as in permutation-based debiasing, is a separate intervention that we do not directly evaluate here. Our protocol also differs from much of the prior literature in how answers are elicited and scored: models reason in text before selecting an option under a schema that accepts only option text, whereas many reported effects rely on first-token probabilities over option labels (Wang et al., 2024). Order sensitivity may therefore depend strongly on the elicitation and scoring protocol (Tam et al., 2024).

Sampling variability dominates under stochastic decoding. Under sampling, draw-to-draw variation reached 0.41–0.44 of the mean CSEM, 3.3–3.6 times the order variation in the same regime. This agrees with work that treats run-to-run variability as a source of evaluation error (Miller, 2024; Madaan et al., 2024; Mizrahi et al., 2024) and with evidence that even nominally deterministic inference varies with system configuration (Yuan et al., 2025). Our contribution is to express this variability on a psychometric scale, relative to what the test can resolve. The dispersion analysis shows that it is not uniform across models: `qwen3:4b` and `qwen3.5:9b` varied more than binomial sampling predicts. For `qwen3:4b`, this may partly reflect its truncated responses, which recur on the same items and are scored as incorrect. If the pattern holds more broadly, the number of draws needed for a stable estimate may need to be set per model.

Measurement precision reframes what "distinguishable" means. Nearly all pairwise contrasts were significant after Holm correction, yet only one or two of 15 pairs per subscale exceeded the MDD. The significance tests treat each model's 16 (at $T{=}0$) or 40 (at $T{=}0.7$) administrations per domain as independent replicates, although all reuse the same 25 items, so their standard errors describe how reproducible the runs are rather than how precisely ability is measured. This echoes the finding that small benchmark subsets cannot reliably rank closely matched models (Yauney et al., 2026). Unlike most applications of IRT to LLM evaluation (Kipnis et al., 2025; Truong et al., 2025; Zhou et al., 2026), which use it to estimate or compress rankings, we use it to supply a measurement-error benchmark against which both model differences and nuisance effects can be judged. When runs are cheap and repeatable, significance is nearly guaranteed for any nonzero difference, and a precision-based threshold is a more conservative criterion for claiming that two models differ in ability.

Limitations. First, we study six open-weight models of up to 14B parameters on 25 items from two ICAR subscales (Condon & Revelle, 2014); other item types, languages, and larger commercial models are untested. Second, the subscales are short and poorly targeted at the strongest models, so their CSEM is large and ceiling-level estimates depend on the prior. Third, our protocol departs from both standard deployment and the human calibration: thinking mode is disabled, answers are grammar-constrained to the option texts, the 2,048-token limit truncated 1% of responses, and rotating "None of these" and "I don't know" creates option orders that the calibration sample never saw. Fourth, the MDD ($1.96\sqrt{2}\times$ mean CSEM) is one reasonable convention rather than an established standard, and a generalizability-theory decomposition (Brennan, 2001) could treat items, rotations, and draws as crossed facets more formally. Finally, the human-calibrated parameters assume measurement invariance between humans and LLMs (Meredith, 1993), which remains an open question (Jiang et al., 2026; Säuberli et al., 2025; Srivatsa et al., 2025); a violation would affect both the ability estimates and the CSEM benchmark.

Future work. Natural extensions include other item formats, elicitation and scoring protocols, and decoding regimes; harder or adaptively targeted items that place more models mid-scale, where the CSEM is smallest; and formal tests of measurement invariance between human and LLM respondents on this item bank (Liu et al., 2025; Srivatsa et al., 2025; Säuberli et al., 2025).

AI USE STATEMENT

In this work, we used generative AI tools to refine research ideas and hypotheses, to provide feedback on the research methodology and experimental design, to help implement the administration and analysis code, and to help draft the interpretation of results. We have not used generative AI tools to generate synthetic data sets, to assist with translation, or to support qualitative or thematic analysis. The remaining tasks with required disclosure (developing theoretical models, formulating mathematical claims, and writing proofs) are not applicable to this work.

We have reviewed all AI-assisted work. Every citation was checked against its original source, and AI-suggested descriptions of prior work were corrected where they misstated a paper's claims or venue. AI-assisted code was reviewed, tested with built-in audits of option rotation, model configuration, and greedy-decoding determinism, and run by the authors. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Section 2 and Appendix A.1 give every setting needed to reproduce the administrations: model tags, quantization, decoding parameters, the seed formula, the exact prompt, the output schema, and the scoring rule. Item parameters follow from the ICAR human data through Eq. 1, and the scoring model, measurement-precision quantities, and analyses are specified in Sections 2.4–2.7. Our administration code records, for every call, the model digest, quantization, hardware, rotation map, seed, and raw output. ICAR items are not reproduced here because the ICAR project distributes them to registered researchers; they are available from the project.

REFERENCES

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.

Norah Alzahrani, Hisham Alyahya, Yazeed Alnumay, Sultan AlRashed, Shaykhah Alsubaie, Yousef Almushayqih, Faisal Mirza, Nouf Alotaibi, Nora Al-Twairesh, Areeb Alowisheq, M Saiful Bari, and Haidar Khan. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13787–13805. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.744. URL https://aclanthology.org/2024.acl-long.744/.

R. Darrell Bock and Robert J. Mislevy. Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, 6(4):431–444, 1982.

Robert L. Brennan. *Generalizability Theory*. Springer, 2001. doi: 10.1007/978-1-4757-3456-0. URL https://link.springer.com/book/10.1007/978-1-4757-3456-0.

David M. Condon and William Revelle. The international cognitive ability resource: Development and initial validation of a public-domain measure. *Intelligence*, 43:52–64, 2014. URL https://personality-project.org/revelle/publications/condon.icar.14.pdf.

Susan E. Embretson and Steven P. Reise. *Item Response Theory for Psychologists*. Lawrence Erlbaum Associates, 2000. URL https://www.routledge.com/Item-Response-Theory-for-Psychologists/Embretson-Reise/p/book/9781410605269.

Gemma Team. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

Georgi Gerganov and ggml contributors. llama.cpp: LLM inference in C/C++, 2026. URL https://github.com/ggml-org/llama.cpp.

Maharshi Gor, Hal Daumé III, Tianyi Zhou, and Jordan Boyd-Graber. Do great minds think alike? investigating human-AI complementarity in question answering with CAIMIRA. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 21533–21564. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.1201. URL https://aclanthology.org/2024.emnlp-main.1201/.

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*, 2025. URL https://openreview.net/forum?id=mxcCg9YRqj.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Han Jiang, Sunbeom Kwon, Jinwen Luo, Ziang Xiao, and Susu Zhang. Can we trust item response theory for AI evaluation? *arXiv preprint arXiv:2607.15190*, 2026. URL https://arxiv.org/abs/2607.15190.

Alex Kipnis, Konstantinos Voudouris, Luca M. Schulze Buschoff, and Eric Schulz. metabench: A sparse benchmark of reasoning and knowledge in large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/4ebc26584810a189ef1e4f173aba4319-Paper-Conference.pdf.

Yunting Liu, Shreya Bhandari, and Zachary A. Pardos. Leveraging LLM respondents for item evaluation: A psychometric analysis. *British Journal of Educational Technology*, 56(3):1028–1052, 2025. doi: 10.1111/bjet.13570. URL https://bera-journals.onlinelibrary.wiley.com/doi/10.1111/bjet.13570.

Frederic M. Lord. *Applications of Item Response Theory to Practical Testing Problems*. Lawrence Erlbaum Associates, 1980. URL https://www.routledge.com/Applications-of-Item-Response-Theory-To-Practical-Testing-Problems/Lord/p/book/9780898590067.

Lovish Madaan, Aaditya K. Singh, Rylan Schaeffer, Andrew Poulton, Sanmi Koyejo, Pontus Stenetorp, Sharan Narang, and Dieuwke Hupkes. Quantifying variance in evaluation benchmarks. *arXiv preprint arXiv:2406.10229*, 2024. URL https://arxiv.org/abs/2406.10229.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinyBenchmarks: evaluating LLMs with fewer examples. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 34303–34326. PMLR, 2024. URL https://proceedings.mlr.press/v235/maia-polo24a.html.

William Meredith. Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, 58(4):525–543, 1993. doi: 10.1007/BF02294825. URL https://doi.org/10.1007/BF02294825.

Microsoft, Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, et al. Phi-4-Mini technical report: Compact yet powerful multimodal language models via mixture-of-LoRAs. *arXiv preprint arXiv:2503.01743*, 2025.

Evan Miller. Adding error bars to evals: A statistical approach to language model evaluations. *arXiv preprint arXiv:2411.00640*, 2024. URL https://arxiv.org/abs/2411.00640.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. State of what art? a call for multi-prompt LLM evaluation. *Transactions of the Association for Computational Linguistics*, 12:933–949, 2024. doi: 10.1162/tacl_a_00681. URL https://aclanthology.org/2024.tacl-1.52/.

Ollama. Ollama, 2026. URL https://github.com/ollama/ollama. Version 0.34.1.

Pouya Pezeshkpour and Estevam Hruschka. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 2006–2017. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-naacl.130. URL https://aclanthology.org/2024.findings-naacl.130/.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Mark D. Reckase. *Multidimensional Item Response Theory*. Springer, New York, 2009.

Pedro Rodriguez, Joe Barrow, Alexander 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. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.346. URL https://aclanthology.org/2021.acl-long.346/.

Andreas Säuberli, Diego Frassinelli, and Barbara Plank. Do LLMs give psychometrically plausible responses in educational assessments? In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pp. 266–278. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.bea-1.21. URL https://aclanthology.org/2025.bea-1.21/.

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*, 2024. URL https://arxiv.org/abs/2310.11324.

KV Aditya Srivatsa, Kaushal Kumar Maurya, and Ekaterina Kochmar. Can LLMs reliably simulate real students' abilities in mathematics and reading comprehension? In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pp. 988–1001. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.bea-1.75. URL https://aclanthology.org/2025.bea-1.75/.

Yoshio Takane and Jan de Leeuw. On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, 52(3):393–408, 1987.

Zhi Rui Tam, Cheng-Kuang Wu, Yi-Lin Tsai, Chieh-Yen Lin, Hung-yi Lee, and Yun-Nung Chen. Let me speak freely? a study on the impact of format restrictions on large language model performance. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 1218–1236. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-industry.91. URL https://aclanthology.org/2024.emnlp-industry.91/.

Sang T. Truong, Yuheng Tu, Percy Liang, Bo Li, and Sanmi Koyejo. Reliable and efficient amortized model-based evaluation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 60238–60265. PMLR, 2025. URL https://proceedings.mlr.press/v267/truong25c.html.

Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. "my answer is C": First-token probabilities do not match text answers in instruction-tuned language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 7407–7416. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.441. URL https://aclanthology.org/2024.findings-acl.441/.

Ziqi Wang, Hanlin Zhang, Xiner Li, Kuan-Hao Huang, Chi Han, Shuiwang Ji, Sham Kakade, Hao Peng, and Heng Ji. Eliminating position bias of language models: A mechanistic approach. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/e389b15166cf98966ba058965a8c17e3-Abstract-Conference.html.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Gregory Yauney, Shahzaib Saqib Warraich, and Swabha Swayamdipta. How reliable is language model micro-benchmarking? In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2510.08730.

Jiayi Yuan, Hao Li, Xinheng Ding, Wenya Xie, Yu-Jhe Li, Wentian Zhao, Kun Wan, Jing Shi, Xia Hu, and Zirui Liu. Understanding and mitigating numerical sources of nondeterminism in LLM inference. In *Advances in Neural Information Processing Systems*, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/f80094a824ba5912d4a2de169c404a40-Paper-Conference.pdf.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2309.03882.

Hongli Zhou, Hui Huang, Ziqing Zhao, Lvyuan Han, Huicheng Wang, Kehai Chen, Muyun Yang, Wei Bao, Jian Dong, Bing Xu, Conghui Zhu, Hailong Cao, and Tiejun Zhao. Lost in benchmarks? rethinking large language model benchmarking with item response theory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 35085–35093, 2026. doi: 10.1609/aaai.v40i41.40814. URL https://ojs.aaai.org/index.php/AAAI/article/view/40814.

A Appendix

A.1 Prompt and inference settings

Each item is sent as a single user message in its own stateless request to the Ollama chat API. The message is identical for every model, item, rotation, and temperature; only the question text and the order of the answer options change. {QUESTION} is the item stem and {OPTIONS} is the list of answer choices in the order given by the current rotation, formatted as shown below the template.

Prompt template

```
You are completing an assessment item.

Read the question carefully and solve it.

Question:
{QUESTION}

Answer options:
{OPTIONS}

First, explain your reasoning.

Explanation requirements:
- State the main rule, relationship, or calculation you used.
- State the result you derived.
- Use only one solution approach.
- Do not evaluate every answer choice individually.
- Do not restart the analysis after reaching a conclusion.
- Do not repeat the question.

Then provide your final answer itself, not the option number.

Important rules:
- Your final answer must be exactly one of the answer choices as
↪  written.
- Your final answer must be supported by your explanation.
- Do not provide an option number.
```

```
- Do not choose an answer merely because it is numerically,
↪  alphabetically,
  or semantically closest to the result you derived.
- Before giving the final answer, verify that it exactly matches
↪  one of the
  listed answer choices.

Stop after providing the explanation and final answer.
```

Output format. Responses are constrained to JSON with two fields, `explanation` followed by `answer`, via Ollama's structured-output. Because `explanation` precedes `answer`, the model writes its reasoning before committing to a choice.

Option display. Options are listed one per line as `(1) <text>` through `(8) <text>`, in the order given by the current rotation.

Output schema.

```
{"type": "object",
 "properties": {
   "explanation":  {"type": "string"},
   "final_answer": {"type": "string", "enum": [<8 presented option texts>]}},
 "required": ["explanation", "final_answer"],
 "additionalProperties": false}
```

Request parameters.

```
{"model": "<tag>", "think": false, "format": <schema>,
 "messages": [{"role": "user", "content": <prompt>}],
 "options": {"temperature": 0.0 | 0.7, "top_p": 1.0, "top_k": 0,
             "num_ctx": 8192, "num_predict": 2048, "seed": <seed>}}
```

Seed for rotation r, draw k, temperature T: $20260915 + 10000\,r + 100\,k + 10\,T$.

Scoring. The returned `final_answer` is matched to the presented options after normalization (lowercasing, whitespace collapsing, and trimming of quotes and trailing punctuation) and mapped back to the item's canonical option order. A response is correct if it matches the keyed option. Responses whose JSON output is incomplete are recorded as missing and scored as incorrect in the IRT analysis (Section 2.6).

Effort. Explanation length is counted with a single reference tokenizer (`cl100k_base`) for all models, so that lengths are comparable across tokenizer families.