

Reproducibility Budgets for Open-Weight LLM Evaluation: A Generalizability Theory Study of 7–9B Instruction-Tuned Models

Apex Intelligence

Abstract

LLM benchmarks are the primary tool for comparing language models, guiding deployment, funding, and ranking decisions. Every evaluation involves design choices (prompt template, random seed, numerical precision, answer ordering) typically fixed to single values and assumed inconsequential. Growing evidence challenges this assumption: prompt formatting can shift accuracy by 76 percentage points, and answer ordering can flip model rankings. Yet each study examines one factor in isolation, leaving unanswered which sources dominate when all vary simultaneously and how many items suffice. We address this with Generalizability theory, a psychometric framework for multi-source variance decomposition. In a factorial experiment crossing seven facets across eight models and five benchmarks (2.3M records), we find that item difficulty is the largest variance component (34–38% in multiple-choice), and the model×item interaction is the second largest (30–36% in MC, 15–17% in free-form), exceeding every manipulated design factor by an order of magnitude—models disagree on which items are difficult, not in overall ability. A 3-prompt × 3-seed protocol absorbs this interaction, reducing items from 552 to 178 ($G \geq 0.80$). We observe a format-associated pattern ($n = 5$, exploratory): item-related variance concentrates in multiple-choice (68–72%) but disperses in free-form tasks (31–41%). We release an open-source G-theory toolkit¹.

1 Introduction

LLM benchmarks are the standard mechanism for comparing language models. Evaluation results determine which models get deployed in production, which attract continued funding, and which top public leaderboards. A typical evaluation runs

each model once under a fixed configuration (one prompt template, one random seed, one precision format) and treats the resulting score as a stable measure of model capability. This practice implicitly assumes that these design choices are inconsequential: that a different prompt wording, a different random seed, or a different answer ordering would yield essentially the same result.

A growing body of evidence shows otherwise. Prompt reformatting (rewording instructions without altering their meaning) can shift accuracy by up to 76 percentage points (Sclar et al., 2024; Romanou et al., 2026). Reordering answer options can reverse model rankings (Pezeshkpour and Hruschka, 2024). Numerical precision (Yuan et al., 2025) and random seeds (Dodge et al., 2020; Bui et al., 2025) each contribute additional instability. However, each study examines one source in isolation, holding all others fixed. This piecemeal approach cannot answer the questions that matter most for benchmark design: when all sources vary simultaneously, which ones dominate? Do they interact, e.g., does a prompt change affect some models more than others? Does the answer depend on the benchmark format? Without a unified variance decomposition, benchmark developers cannot determine how many items are needed for reliable measurement, or whether current evaluations are measuring model ability or design noise.

Three lines of prior work each address part of this puzzle (Figure 3): (i) *variance decomposition for ML*—Bouthillier et al. (2021) established ANOVA-based decomposition for hyperparameters and data splits, but not LLM evaluation facets; (ii) *G-theory for LLM evaluation*—Messing (2026) decomposed five methodology facets, but on a single benchmark without infrastructure facets or cross-format comparison; (iii) *Item Response Theory (IRT) for item selection*—Maia Polo et al. (2024); Lalor et al. (2019); Zhou et al. (2026) optimize which items to include, but treat prompt,

¹Code and data are available at https://anonymous.4open.science/r/nlp_variance_gtheory-11E1/.

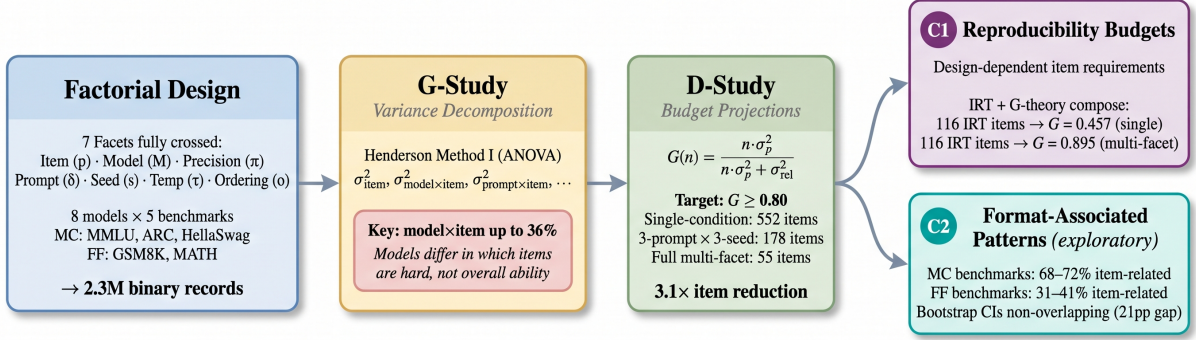


Figure 1: Overview of the G-theory evaluation pipeline. A factorial design crossing seven facets across eight models and five benchmarks produces 2.3M binary records. G-study variance decomposition via Henderson Method I reveals that item difficulty (34–38%) and model × item interaction (30–36%) together dominate in MC benchmarks. D-study projections translate variance components into design-dependent item budgets: a 3-prompt × 3-seed design achieves $G \geq 0.80$ with 178 items (3.1 × reduction). Three contributions emerge: reproducibility budgets (C1), dominant item-level variance (C2), and format-associated variance patterns (C3).

seed, and ordering as fixed, leaving their variance unmodeled.

We provide that unified analysis. Applying Generalizability theory (G-theory) (Shavelson and Webb, 1991; Brennan, 2001), a psychometric framework that decomposes measurement variance into contributions from each design facet and their interactions, to a factorial design crossing seven facets across eight open-weight 7–9B models and five benchmarks (Figure 1; ~2.3M binary records), contrary to the prevailing focus on prompt and seed sensitivity, item difficulty is the largest variance component (34–38% in multiple-choice [MC]), and the model × item interaction is the second largest (30–36% in MC, 15–17% in free-form [FF]), exceeding every manipulated design factor by an order of magnitude and the model main effect by 7–11 ×. Models disagree primarily on *which specific questions* are difficult, rather than differing uniformly in overall ability. This finding has an immediate practical consequence: evaluation designs that average over multiple prompts and seeds absorb this interaction, achieving reliable measurement with a recommended 3-prompt × 3-seed protocol that reduces items by 3.1 × (Figure 2). In an exploratory comparison ($n = 5$ benchmarks), item-related variance concentrates in multiple-choice benchmarks but disperses in free-form tasks, suggesting that reproducibility standards may need per-format calibration.

Our contributions:

- 1. Design-dependent reproducibility budgets.** A recommended 3-prompt × 3-seed protocol achieves $G \geq 0.80$ with 178 MMLU

items versus 552 under single-condition protocols (3.1 × reduction; up to 10 × under the full 1,296-condition design, a theoretical upper bound impractical for routine use). Cross-benchmark budgets range from 93–128 items (multi-facet) versus 1,078–2,768 (single-condition). These budgets complement IRT item selection (Table 7).

- 2. Dominant item-level variance.** Item difficulty is the largest single component in MC benchmarks (34–38%), and the model × item interaction is the second largest (30–36% in MC, 15–17% in FF), exceeding the model main effect by 7–11 × and every manipulated design factor (prompt, seed, precision, ordering) by over an order of magnitude. This reveals item-dependent rather than global ability differences, stable under leave-one-model-out analysis and confirmed by inter-model agreement (Cohen’s $\kappa = 0.33$ –0.46).
- 3. Format-associated variance patterns (exploratory).** Item-related variance (item difficulty + model × item interaction) concentrates in multiple-choice benchmarks (68–72%) but disperses in free-form benchmarks (31–41%), driven by item difficulty heterogeneity ($\rho = 0.82$). This observation ($n = 5$, $p = 0.10$) warrants broader confirmation.

We release an open-source G-theory pipeline and D-study budget calculator for evaluation design optimization.

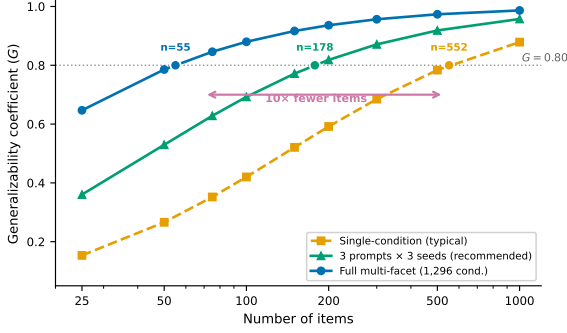


Figure 2: D-study projections for MMLU (single-model baseline). A recommended 3-prompt $\times$ 3-seed design reduces the budget from 552 to 178 items for $G \geq 0.80$; the full multi-facet design (1,296 conditions, theoretical upper bound) reduces further to 55 items.

2 Related Work

Variance Sources in LLM Evaluation. Bouthillier et al. (2021) introduced ANOVA-based variance decomposition for ML reproducibility, establishing that design choices account for a substantial share of benchmark variance. Subsequent work isolated individual sources: Yuan et al. (2025) documented up to 9% accuracy variation from numerical precision; Sclar et al. (2024) and BrittleBench (Romanou et al., 2026) found 50–76% accuracy swings from prompt formatting; Pezeshkpour and Hruschka (2024) and Bui et al. (2025) documented ordering and seed sensitivity. Messina and Scotta (2026) formalized CUDA nondeterminism from floating-point non-associativity. Each study holds other factors fixed, precluding interaction estimation. Madaan et al. (2024) proposed per-factor metrics without crossing factors; holistic evaluation frameworks (Liang et al., 2023; Gao et al., 2023) reduce implementation variance but do not decompose it. Our work crosses seven facets simultaneously, enabling interaction estimation and cross-format comparison (Figure 3).

IRT and Efficient Benchmarking. IRT models item difficulty and discrimination to optimize benchmark efficiency (Lalor et al., 2019; Rodriguez et al., 2021). Maia Polo et al. (2024) and Zhou et al. (2026) developed IRT-based subsampling for efficient evaluation; Perlitz et al. (2023) and Kipnis et al. (2025) explored related compression strategies. These approaches optimize item selection but do not model variance from prompt, seed, or ordering facets. IRT-selected items under a single

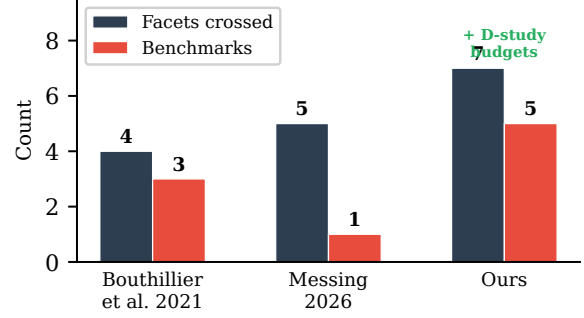


Figure 3: Scope comparison with prior multi-facet variance studies. Our work uniquely combines broad facet crossing, multi-benchmark coverage, and D-study budget projections.

Table 1: Positioning relative to prior multi-facet variance studies. “–” = not addressed.

	Bouthillier et al. (2021)	Messing (2026)	Ours
Facets crossed	4	5	7
Design	Factorial	Factorial	Factorial
Method	ANOVA	G-theory	G-theory
D-study budgets	–	–	✓
Models	ML models	6 LLMs	8 LLMs
Benchmarks	3	1	5
Cross-format	–	–	MC+FF

methodological condition can therefore underestimate required item counts; our results show IRT’s 116 items yield $G = 0.457$ single-condition but $G = 0.895$ multi-facet, demonstrating that item selection and protocol optimization are orthogonal.

G-Theory and Positioning. G-theory (Shavelson and Webb, 1991; Brennan, 2001) provides the canonical framework for multi-facet reliability analysis, with extensions to binary data (Mushquash and O’Connor, 2006; Jiang and Skorupski, 2018). We operationalize differential item functioning (Holland and Wainer, 1993; Clauser and Mazor, 1998) as model $\times$ item interaction. Recent psychometric perspectives on LLM evaluation address construct validity (Kearns, 2026) and measurement properties (Ye et al., 2025). Closest to our work, Messing (2026) applied G-theory to LLM evaluation, decomposing five methodology facets on a single benchmark (judge identity: 43.8%). We extend this in three directions: (i) replacing judge identity with infrastructure facets and additional methodology facets; (ii) crossing facets across five benchmarks for cross-format comparison; and (iii) translating variance components into reproducibility budgets via D-study projections. Table 1 summarizes these distinctions.

3 Method and Experimental Design

3.1 Generalizability Theory

Generalizability theory (G-theory) treats each measurement as one observation from a universe of admissible conditions defined by the crossed facets of an evaluation design (Shavelson and Webb, 1991; Brennan, 2001). The *object of measurement* is the entity whose consistency is of interest. In our setting, test items (p) serve as the object of measurement, and the goal is to determine how reliably item-level scores generalize across the methodological conditions under which they are collected.

A *G-study* decomposes total observed variance into components σ_x^2 for each facet x and their interactions. Each component quantifies how much variability that source contributes to the overall measurement. Large interaction components (e.g., σ_{pM}^2 for item-by-model) indicate that facets do not affect items uniformly but instead create differential difficulty profiles. We estimate variance components via Henderson Method I, an unbiased ANOVA-based estimator for balanced designs (Brennan, 2001, Ch. 3), with 1,000-resample bootstrap confidence intervals (Efron and Tibshirani, 1993) (item-level resampling with replacement, preserving the fully-crossed structure across all other facets) (Li, 2023).

A *D-study* projects the generalizability coefficient under hypothetical designs with different numbers of conditions per facet:

$$G_{\text{item}} = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_{\text{rel}}^2} \quad (1)$$

where $\sigma_{\text{rel}}^2 = \sum_i \sigma_{p \times f_i}^2 / n_{f_i} + \sigma_{\text{residual}}^2 / \prod_i n_{f_i}$ aggregates all item-by-facet interaction variances, each divided by the number of conditions for the corresponding facets (Brennan, 2001, Ch. 4). A D-study projects reliability for a test averaging over n_p items via

$$G(n_p) = \frac{n_p \sigma_p^2}{n_p \sigma_p^2 + \sigma_{\text{rel}}^2} \quad (2)$$

enabling independent verification of all budget projections in Table 7. This prospective capability distinguishes G-theory from classical ANOVA: rather than only attributing variance post hoc, a D-study enables optimization of evaluation designs *before* data collection (Shavelson et al., 1989). Figure 4 illustrates the practical impact: multi-facet designs achieve $G \geq 0.80$ with 55 items versus 552 under

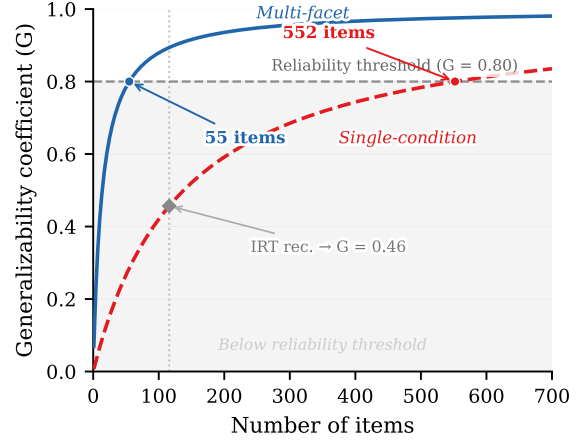


Figure 4: D-study projection: the full multi-facet design (blue, 1,296 conditions; theoretical upper bound, impractical for routine use) reaches $G \geq 0.80$ at 55 items versus 552 under single-condition protocols (red). The recommended 3×3 design achieves 178 items ($3.1 \times$ reduction).

single-condition protocols. We define *item-related variance* as the sum of the item main effect (σ_p^2) and the model $\times$ item interaction (σ_{pM}^2). The latter captures differential item functioning (DIF): items whose difficulty ranking changes across models (Brennan, 2001, Ch. 4). Because DIF is operationally an item-level property (it reflects model-specific item difficulty profiles rather than a global model ability shift); aggregating $\sigma_p^2 + \sigma_{pM}^2$ provides a unified measure of item-driven variance. We report unbundled components throughout to ensure transparency.²

3.2 Experimental Design: MMLU

We design a fully crossed factorial experiment (exp-001) that simultaneously varies seven facets of LLM evaluation on MMLU. Table 2 summarizes the complete design matrix.

Facets. Items (p): 200 MMLU questions via stratified sampling across 10 subjects (20 per subject, four domains: STEM, humanities, social sciences, professional; coefficient of variation [CV] = 3.1%). Models (M): eight open-weight 7–9B instruction-tuned models (Table 3), treated as a quasi-random facet (§3.6). Numerical precision (π): BFloat16 (BF16), Float16 (FP16), Float32 (FP32). Prompt template (δ): six surface perturbations (Sclar et al., 2024) varying layout and wording (Appendix G).

²Notation: p = item, M = model, π = precision, δ = prompt, s = seed, τ = temperature, o = ordering, σ_x^2 = variance component for source x .

Table 2: Fully crossed G-theory design. Exp-001 evaluates MMLU across all facet levels; exp-002 extends to five benchmarks with reduced crossing. [†]BF16 balanced subset (2 temperatures: 0.0, 0.7) for cross-model analysis; the full single-model design uses all 3 levels.

Facet	Levels	exp-001	exp-002
Item (p)	Stratified sample	200	200×5
Model (M)	8 architecture families	8	8
Precision (π)	BF16 / FP16 / FP32	3	1
Prompt (δ)	Paraphrases	6	6
Seed (s)	Fixed set	6	6
Temperature (τ)	0.0 / 0.3 / 0.7	3	2
Ordering (o)	Demo permutations	4	4
Total records		460,800 [†]	2,304,000

Random seed (s): six fixed values (42, 123, 256, 512, 1024, 2048). Sampling temperature (τ): 0.0, 0.3, 0.7. Answer ordering (o): four permutations of few-shot demonstrations (Pezeshkpour and Hruschka, 2024).

Sample sizes. The cross-model BF16 balanced subset yields 460,800 records (8 models $\times$ 200 items $\times$ 6 prompts $\times$ 6 seeds $\times$ 2 temperatures $\times$ 4 orderings). A single-model analysis (Llama) spans all three precisions, yielding 259,200 binary observations across 1,296 fully balanced conditions. The BF16 subset serves as the primary cross-model sample because BF16 is the standard inference precision for these architectures.

3.3 Cross-Benchmark Extension

Exp-002 extends to five benchmarks: MMLU (Hendrycks et al., 2021a), ARC-Challenge (Clark et al., 2018), HellaSwag (Zellers et al., 2019) (MC), GSM8K (Cobbe et al., 2021), and MATH (Hendrycks et al., 2021b) (FF). The design crosses 8 models $\times$ BF16 $\times$ 2 temperatures (0.0, 0.7) $\times$ 6 prompts $\times$ 6 seeds $\times$ 4 orderings $\times$ 200 items, totaling 2,304,000 records. This reduced crossing fixes precision at BF16 to concentrate power on format differences. Ordering operates only on MMLU ($\sigma_{\text{ordering}}^2 = 0$ for other benchmarks, retained for design balance). Subset analysis confirms this preserves the full design’s variance structure ($\Delta G = 0.0008$; Appendix A).

3.4 Evaluation Protocol

All evaluations use generation-based scoring (free-text response with answer extraction) rather than log-likelihood ranking, following Hua et al. (2025). Binary correctness (1/0) is the primary metric; text-exact-match is secondary. Inference uses vLLM 0.9.2 without batch-invariant kernels, pre-

Table 3: Models used in all experiments. RMS = RM-SNorm, LN = LayerNorm.

Short name	Family	Params	Norm
Llama-3.1-8B	Llama-3	8B	RMS
Gemma-2-9B	Gemma-2	9B	RMS
Mistral-7B-v0.3	Mistral	7B	RMS
Qwen3-8B	Qwen-3	8B	RMS
Yi-1.5-9B	Yi-1.5	9B	RMS
OLMo-2-7B	OLMo-2	7B	RMS
InternLM2.5-7B	InternLM-2	7B	RMS
DeepSeek-7B	DeepSeek	7B	RMS

serving CUDA-level nondeterminism as a measurable variance source (Messina and Scotta, 2026; He and Thinking Machines Lab, 2025). Few-shot protocols: 5-shot MMLU/ARC, 4-shot MATH, 8-shot GSM8K, 10-shot HellaSwag.

3.5 Binary Data Assumptions

Standard G-theory assumes continuous data (Shavelson and Webb, 1991); our binary scores violate this. Henderson Method I produces unbiased estimates for balanced designs regardless of distribution (Brennan, 2001, Ch. 12), and restricted maximum likelihood (REML) differs by $<0.5\text{pp}$. A probit simulation, generalized linear mixed model (GLMM) complement ($\rho = 1.0$), and accuracy-range analysis confirm robustness (Appendix A).

All findings in this paper are conditional on binary correctness scoring. Changing the scoring metric can fundamentally reshape variance structure: text-exact-match scoring collapses G_{item} from 0.936 to 0.007 (§4.1), and partial-credit scoring on GSM8K reduces G_{item} from 0.86 to 0.77 (mean $\Delta G = -0.135$; Appendix A). Readers should not assume that the variance decomposition or D-study budgets reported here transfer to non-binary rubrics.

3.6 Model Facet Specification

Our target universe comprises open-weight, instruction-tuned models in the 7–9B parameter range (Table 3); results may not generalize to API-based, base, or $>9\text{B}$ models. The eight models are treated as a quasi-random facet (Meredith, 1993; Putnick and Bornstein, 2016). We report G under random-model assumptions, the more conservative specification: treating models as fixed yields G_{fixed} 0.951–0.990 vs. G_{random} 0.863–0.896 (increase of 0.077–0.113; Table 10), because model $\times$ item interaction (15–36% across benchmarks) enters relative error variance. Consequently, D-study budgets are upper bounds. Leave-one-model-out (LOMO) analysis preserves component rankings across all

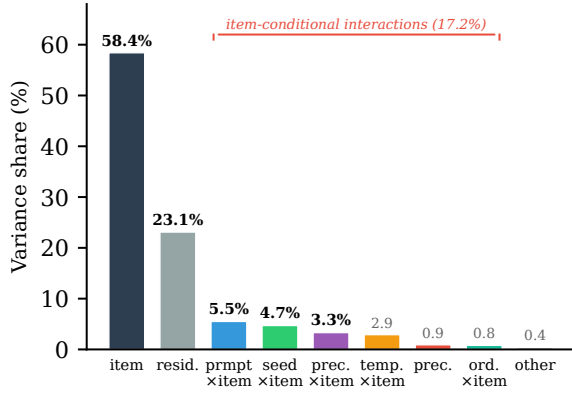


Figure 5: Single-model variance decomposition (Llama-3.1-8B, MMLU). Item difficulty dominates (58.4%); all main facet effects are $<1\%$. Item-conditional interactions total 17.2%.

five benchmarks (Appendix A); four-model subset checks match within $|\Delta G| = 0.035$ (§4.5).

4 Experiments

All experiments use open-weight 7–9B parameter models evaluated with vLLM 0.9.2 on a single NVIDIA RTX PRO 6000 GPU (96GB) with binary correctness scoring.

4.1 Single-Model Baseline

Design. The G-study crosses six facets: 200 items $\times$ 3 numerical precisions (BF16, FP16, FP32) $\times$ 6 prompt templates $\times$ 6 random seeds $\times$ 3 temperature levels (0.0, 0.3, 0.7) $\times$ 4 answer orderings, yielding 259,200 binary observations. Variance components are estimated via Henderson Method I with bootstrap confidence intervals (1,000 resamples).

Results. Item difficulty dominates at 58.4%, with residual at 23.1% (Figure 5). All main facet effects are negligible ($<1\%$ each); the interesting structure is in item-conditional interactions (17.2% total), where prompt $\times$ item (5.5%) is largest. The generalizability coefficient $G_{\text{item}} = 0.936$ (Eq. 1); D-study projection (Eq. 2) gives $G \geq 0.80$ with 55 items.

For text-exact-match scoring, temperature dominates at 79.7% and G_{item} collapses to 0.007, motivating our exclusive use of binary correctness.

4.2 Cross-Model Variance

Extending to eight models (BF16 subset, 460,800 records), Table 5 shows that item difficulty drops from 58.4% to 37.9%, while the model $\times$ item interaction rises to 30.2%, becoming the second-largest

Table 4: LOMO stability across five benchmarks (8 models $\times$ 5 benchmarks = 40 subsets).

Benchmark	G_{full}	CV (%)	Max $ \Delta G $	Most impactful
MMLU	0.896	0.49	0.019	InternLM2.5-7B
ARC	0.892	0.94	0.026	Llama-3.1-8B
HellaSwag	0.872	1.09	0.028	Llama-3.1-8B
GSM8K	0.863	1.25	0.030	Gemma-2-9B
MATH	0.888	1.20	0.029	InternLM2.5-7B

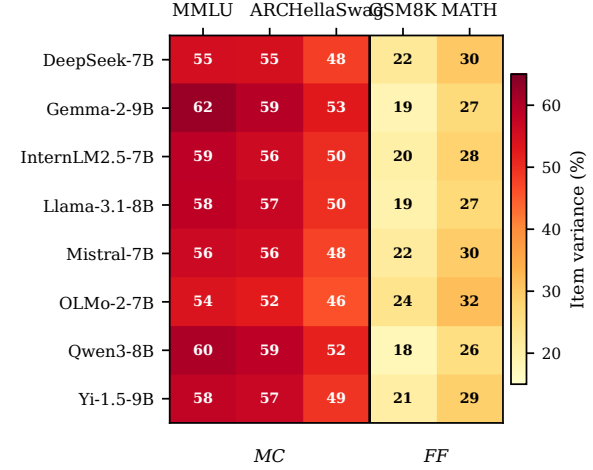


Figure 6: Item variance share (%) by model and benchmark. MC benchmarks (left) show higher item variance (46–62%) than FF benchmarks (right, 18–32%) across all eight models.

component (up to 36% in MC benchmarks, 15–17% in FF; §4.3). The model main effect is 3–5% (MC) and 9–11% (FF)³, suggesting that models differ primarily in *which items* they find difficult rather than in overall ability level. Figure 6 visualizes this pattern across models and benchmarks. Leave-one-model-out (LOMO) analysis (Table 4) confirms this structure is stable: no single model removal changes G by more than 0.030 across all five benchmarks. Inter-model agreement independently quantifies this pattern: pairwise Cohen’s κ ranges from 0.326 (MATH) to 0.461 (MMLU), indicating only fair-to-moderate agreement on individual items despite similar aggregate accuracy, consistent with models differing in *which items* they solve rather than in overall ability (Appendix A).

4.3 Exploratory: Format-Associated Variance Patterns

Table 6 reveals a format-associated pattern: MC benchmarks concentrate item-related variance (item + model $\times$ item) at 68–72%, while FF benchmarks disperse to 31–41% ($n = 5$; permutation

³Ranges rounded to nearest integer; exact values are 2.8–5.4% (MC) and 8.5–10.7% (FF) per Table 6.

Table 5: Variance decomposition for the cross-model MMLU design (8 models, BF16, 460,800 records). Components estimated via Henderson Method I. $G_{\text{item}} = 0.896$.

Component	Var. %	95% CI
item_id	37.9	[33.0, 42.4]
model×item	30.2	[26.9, 33.3]
residual	25.0	[22.9, 27.0]
model	2.8	[1.7, 4.3]
model×prompt	1.6	[1.2, 2.1]
prompt×item	0.9	[0.8, 1.1]
temp.×item	0.6	[0.4, 0.8]
seed×item	0.5	[0.4, 0.7]
ordering×item	0.3	[0.2, 0.4]
all others (each)	<0.1	—

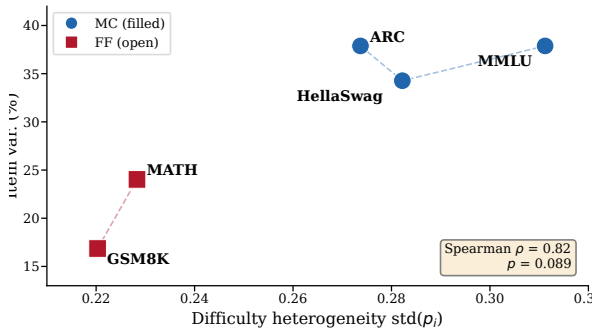


Figure 7: Item difficulty heterogeneity ($\text{std}(p_i)$) versus item_id variance share across five benchmarks. MC benchmarks (filled) exhibit higher difficulty dispersion and item variance than FF benchmarks (open). Spearman $\rho = 0.82$.

$p = 0.10$; item-level $d = 1.73$, benchmark-level CI spans zero); we report this as a descriptive observation. This does not extend to item budgets: HellaSwag (MC, 118 items) exceeds ARC (98) and MATH (101), showing format alone does not predict cost (Figure 8). Figure 7 shows item difficulty heterogeneity ($\text{std}(p_i)$) predicts item variance share ($\rho = 0.82$; Appendix D). Within MMLU, this mechanism is visible at subject granularity: knowledge-intensive subjects (e.g., international law, $G = 0.930$) show high item variance, while reasoning-heavy subjects (e.g., high school mathematics, $G = 0.614$) show $\text{model} \times \text{item}$ exceeding item difficulty, mirroring the MC–FF contrast (Table 12). Per-benchmark details are in Appendix C. All findings are conditional on binary scoring (§3.5).

4.4 Precision Separability

Under binary scoring, BF16 and FP16 are functionally interchangeable (<0.1% interaction

Table 6: Format-associated variance structure across five benchmarks (exp-002, 8 models). Variance components (%) per benchmark; MC = multiple-choice, FF = free-form. Bold = dominant non-residual component. [†]For MC, reports $\text{prompt} \times \text{item}$ interaction (prompt main effect <1%); for FF, reports prompt main effect (item-conditional interactions absorbed into residual).

	MC			FF	
	MMLU	ARC	Hella.	GSM	MATH
item	37.9	37.9	34.3	16.9	24.0
mod.×item	30.2	33.7	36.0	14.6	17.3
model	2.8	3.1	5.4	10.7	8.5
prompt [†]	0.9	1.4	2.1	5.4	0.8
mod.×prmt	1.6	8.0	0.2	8.5	1.7
residual	25.0	15.4	21.4	39.9	43.9
all others	1.6	0.5	0.6	4.0	3.8

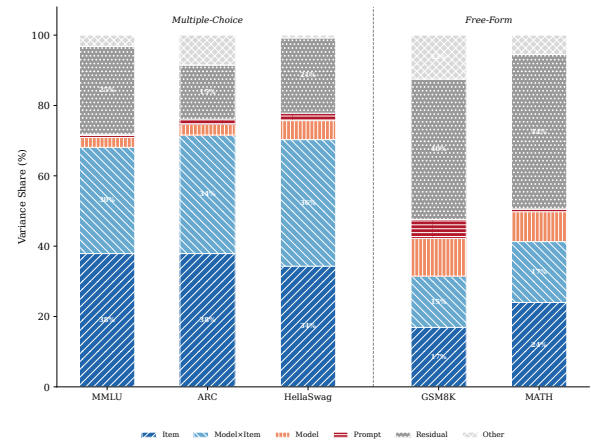


Figure 8: Variance decomposition (Henderson Method I) across five benchmarks and eight models. Multiple-choice benchmarks share an item-dominated structure; HellaSwag is the exception where $\text{model} \times \text{item}$ (36.0%) exceeds item difficulty (34.3%). Free-form benchmarks (GSM8K, MATH) distribute variance across item, model, and interaction components with no single dominant source.

(Yuan et al., 2025)); under text-exact-match, a $\text{precision} \times \text{temperature}$ interaction emerges at 1.3%, consistent with §4.1. Precision need not be controlled for binary evaluation at BF16/FP16 (Shanmugavelu et al., 2024).

4.5 Design-Dependent Reproducibility Budgets

Table 7 compares evaluation designs at the $G \geq 0.80$ threshold using single-model MMLU variance components. The full multi-facet design (1,296 conditions) requires only 55 items for $G \geq 0.80$, a theoretical upper bound impractical for routine evaluation; a single-condition design requires 552 items ($10.0 \times$ ratio, $\text{CV} = 1.9\%$).

The cost–reliability tradeoff across intermedi-

Table 7: D-study reproducibility budgets under different evaluation designs (single-model, MMLU). Total evals = items $\times$ conditions.

Design	Cond.	$n_{G \geq .80}$	Total	G_{200}	G_{116}
Multi-facet	1,296	55	71,280	.936	—
Single-cond.	1	552	552	.592	.457
IRT+multi	1,296	—	150,336	—	.895 [‡]
IRT baseline	—	116	116	—	.457 [†]

[†] PSN-IRT 116 items (Zhou et al., 2026) under single-condition: $G = 0.457$.

[‡] Same 116 items under multi-facet: $G = 0.895$.

Table 8: Intermediate D-study designs ordered by increasing conditions per item. G_{200} and $n_{G \geq 0.80}$ are D-study projections from exp-001 variance components.

Design	Conditions	G_{200}	$n_{G \geq .80}$
Single-condition	1	.592	552
Minimal upgrade (2p)	2	.692	357
Easy upgrade (2p $\times$ 2s)	4	.766	245
Moderate (3p $\times$ 3s)	9	.818	178
Standard (4p $\times$ 4s $\times$ 2t)	32	.869	121
Enhanced (+4 orderings)	72	.871	119
Full multi-facet	1,296	.936	55

ate designs (Table 8) reveals that the steepest gain occurs at the minimal upgrade level (2 prompts): G_{200} jumps from 0.592 to 0.692 at $2\times$ cost. A minimal sufficient design for $G \geq 0.80$ is 3 prompts $\times$ 3 seeds (9 conditions, $n_{G \geq 0.80} = 178$). Beyond 32 conditions, adding orderings provides negligible benefit (ordering $\times$ item $< 1\%$), confirming that prompt and seed diversity are the primary reliability drivers.

Zhou et al. (2026)’s IRT-selected 116 items yield $G = 0.457$ under single-condition but $G = 0.895$ under multi-facet conditions, demonstrating that IRT item selection and G-theory protocol optimization are orthogonal and complementary. Four-model subset checks confirm projection stability ($|\Delta G| \leq 0.029$ for MC, ≤ 0.065 for FF; Table 11; Appendix B).

The recommended 3×3 design balances cost and reliability (178 items, 1,602 evaluations), a $2.9\times$ total-evaluation increase for a $3.1\times$ item reduction. Cross-model projections (Table 9) show multi-facet budgets of 93–128 items, with efficiency ratios of $11\text{--}13\times$ (MC) and $21\text{--}22\times$ (FF).

Projection assumptions. Cross-model budgets increase to 93 items because model $\times$ item interaction enters relative error variance. G_{item} is stable across benchmarks (0.863–0.896, CV = 1.4%). All projections assume the 2024–2025 model generation; §4.6 provides preliminary support.

Table 9: Cross-model D-study reproducibility budgets with 95% bootstrap CIs. $n_{G \geq 0.80}$: items needed. Multi = multi-facet (2,304 cond.); Total = items $\times$ 2,304. Single = single-condition (total evals = item count).

	Multi-facet (2,304 cond.)			Single-cond.		
	Items	95% CI	Total	Items	95% CI	Ratio [95% CI]
<i>Multiple-choice</i>						
MMLU	93	[76, 116]	214K	1,216	[1004, 1491]	13.1 [12.5, 13.7]
ARC	98	[71, 139]	226K	1,078	[793, 1512]	11.0 [10.7, 11.4]
Hella.	118	[93, 151]	272K	1,403	[1120, 1790]	11.9 [11.5, 12.3]
<i>Free-form</i>						
GSM8K	128	[103, 162]	295K	2,768	[2230, 3531]	21.6 [20.7, 23.0]
MATH	101	[86, 122]	233K	2,160	[1853, 2631]	21.4 [20.2, 22.5]

4.6 Scale Validation

To test whether the variance structure generalizes beyond 7–9B models, we conduct within-family and cross-family analyses. Qwen2.5-72B-Instruct on MMLU shows item difficulty at 83.2% ($G_{\text{item}} = 0.979$), with prompt $\times$ item (6.8%) as the dominant interaction, the same ordering observed at 7B. A three-model crossed design (Qwen2.5 3B/7B/72B) preserves the model $\times$ item dominant interaction (38.5%) and item main effect (35.7%). Cross-family, Llama-3.1-70B-Instruct yields item difficulty at 74.5% ($G_{\text{item}} = 0.957$), confirming that item difficulty dominates across a $24\times$ parameter range and two architecture families.

Robustness checks. Three MMLU resamples yield CV = 3.1%; treating models as fixed increases G by 0.077–0.113, confirming conservative budgets (§3.6; Appendix A).

5 Conclusion

Item difficulty is the largest variance component in LLM evaluation (34–38% in MC), and the model $\times$ item interaction is the second largest (30–36% in MC, 15–17% in FF), exceeding all manipulated design factors by an order of magnitude—models disagree on which items are difficult, not in overall ability. Multi-facet designs absorb this interaction: a 3-prompt $\times$ 3-seed protocol achieves $G \geq 0.80$ with 178 items versus 552 single-condition ($3.1\times$ reduction; cross-benchmark: 93–128 items). An exploratory format comparison ($n = 5$) suggests item-related variance concentrates in MC (68–72%) but disperses in FF (31–41%), motivating per-format calibration. We recommend adopting multi-facet designs, reporting item-level accuracy, and calibrating budgets per format.

Limitations

Our design covers eight 7–9B instruction-tuned models; preliminary 72B validation confirms consistent variance structure, but broader scale and temporal coverage is needed. The benchmark sample ($n = 5$) limits statistical power for format-associated observations ($p = 0.10$). All D-study budgets assume binary scoring; partial-credit analysis on GSM8K shows G drops from 0.86 to 0.77, suggesting limited transferability to finer-grained rubrics. Benchmark contamination in pretraining data may systematically bias our variance decomposition: items memorized by all models would be universally answered correctly, artificially deflating model $\times$ item interaction (by reducing inter-model disagreement on those items) and inflating item-level consistency, making D-study budgets optimistic. Our design does not control for contamination; future work should incorporate contamination-aware item sampling.

References

- Xiao Bi, Deli Chen, Guanting Chen, and 1 others. 2024. [DeepSeek LLM: Scaling open-source language models with longtermism](#). *arXiv preprint arXiv:2401.02954*.
- Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Asya Trofimov, Brennan Nichyporuk, Justin Szeto, Nazanin Mohammadi Sepahvand, Edward Raff, KshitiJ Madan, Vikram Voleti, Samira Ebrahimi Kahou, Vincent Michalski, Tal Arbel, Chris Pal, Gaël Varoquaux, and Pascal Vincent. 2021. [Accounting for variance in machine learning benchmarks](#). In *Proceedings of Machine Learning and Systems (MLSys)*.
- Robert L. Brennan. 2001. *Generalizability Theory*. Springer.
- Nghia Tuan Bui, Guergana K. Savova, and Lijing Wang. 2025. [Assessing the macro and micro effects of random seeds on fine-tuning large language models](#). In *Proceedings of IJCNLP-AACL*.
- Zheng Cai, Maosong Cao, Haojiong Chen, and 1 others. 2024. [InternLM2 technical report](#). *arXiv preprint arXiv:2403.17297*.
- Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. [Think you have solved question answering? Try ARC, the AI2 reasoning challenge](#). *arXiv preprint arXiv:1803.05457*.
- Brian E. Clauser and Kathleen M. Mazor. 1998. [Using statistical procedures to identify differentially functioning test items](#). *Educational Measurement: Issues and Practice*, 17(1):31–44.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. [Training verifiers to solve math word problems](#). *arXiv preprint arXiv:2110.14168*.
- Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah A. Smith. 2020. [Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping](#). *arXiv preprint arXiv:2002.06305*.
- Bradley Efron and Robert J. Tibshirani. 1993. *An Introduction to the Bootstrap*. Chapman & Hall, New York.
- Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golber, Jeffrey Hsu, Alain Le Noac’h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ocqueateau, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2023. [A framework for few-shot language model evaluation](#).
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, and 1 others. 2024. [The llama 3 herd of models](#). *arXiv preprint arXiv:2407.21783*.
- Horace He and Thinking Machines Lab. 2025. [Defeating nondeterminism in LLM inference](#). Thinking Machines Lab: Connectionism.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021a. [Measuring massive multitask language understanding](#). In *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021b. [Measuring mathematical problem solving with the MATH dataset](#). In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Paul W. Holland and Howard Wainer, editors. 1993. *Differential Item Functioning*. Lawrence Erlbaum Associates, Hillsdale, NJ.
- Andong Hua, Kenan Tang, Chenhe Gu, Jindong Gu, Eric Wong, and Yao Qin. 2025. [Flaw or artifact? rethinking prompt sensitivity in evaluating LLMs](#). In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, and 1 others. 2023. [Mistral 7b](#). *arXiv preprint arXiv:2310.06825*.
- Zhehan Jiang and William Skorupski. 2018. [A Bayesian approach to estimating variance components within a multivariate generalizability theory framework](#). *Behavior Research Methods*, 50.

- Ryan Othniel Kearns. 2026. [Quantifying construct validity in large language model evaluations](#). *arXiv preprint arXiv:2602.15532*.
- Alex Kipnis, Konstantinos Voudouris, Luca M. Schulze Buschoff, and Eric Schulz. 2025. [metabench – a sparse benchmark of reasoning and knowledge in large language models](#). In *Proceedings of the International Conference on Learning Representations*.
- John P. Lalor, Hao Wu, and Hong Yu. 2019. [Learning latent parameters without human response patterns: Item response theory with artificial crowds](#). In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.
- Guangming Li. 2023. [Which method is optimal for estimating variance components and their variability in generalizability theory? Evidence from a set of unified rules for bootstrap method](#). *PLoS ONE*.
- Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, and 1 others. 2023. Holistic evaluation of language models. *Transactions on Machine Learning Research*.
- Lovish Madaan, Aaditya K. Singh, Rylan Schaeffer, Andrew Poulton, Sanmi Koyejo, Pontus Stenetorp, Sharan Narang, and Dieuwke Hupkes. 2024. [Quantifying variance in evaluation benchmarks](#). *arXiv preprint arXiv:2406.10229*.
- Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. 2024. [tinyBenchmarks: Evaluating LLMs with fewer examples](#). In *Proceedings of the International Conference on Machine Learning (ICML)*.
- William Meredith. 1993. [Measurement invariance, factor analysis and factorial invariance](#). *Psychometrika*, 58(4):525–543.
- Alberto Messina and Stefano Scotta. 2026. [Introducing background temperature to characterise hidden randomness in large language models](#). *arXiv preprint arXiv:2604.22411*.
- Solomon Messing. 2026. [Hidden measurement error in LLM pipelines distorts annotation, evaluation, and benchmarking](#). *arXiv preprint arXiv:2604.11581*.
- Cilia Mushquash and Brian P. O’Connor. 2006. [SPSS and SAS programs for generalizability theory analyses](#). *Behavior Research Methods*.
- Yotam Perlitz, Elron Bandel, Ariel Gera, Ofir Arviv, Liat Ein-Dor, Eyal Shnarch, Noam Slonim, Michal Shmueli-Scheuer, and Leshem Choshen. 2023. [Efficient benchmarking of language models](#). *arXiv preprint arXiv:2308.11696*.
- Pouya Pezeshkpour and Estevam Hruschka. 2024. [Large language models sensitivity to the order of options in multiple-choice questions](#). In *Findings of the North American Chapter of the Association for Computational Linguistics (NAACL)*.
- Diane L. Putnick and Marc H. Bornstein. 2016. [Measurement invariance conventions and reporting: The state of the art and future directions for psychological research](#). *Developmental Review*, 41:71–90.
- Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, and 1 others. 2024. [Gemma 2: Improving open language models at a practical size](#). *arXiv preprint arXiv:2408.00118*.
- Pedro Rodriguez, Joe Barrow, Alexander Hoyle, John P. Lalor, Robin Jia, and Jordan Boyd-Graber. 2021. [Evaluation examples are not equally informative: How should that change NLP leaderboards?](#) In *Proceedings of the Association for Computational Linguistics (ACL)*.
- Angelika Romanou, Mark Ibrahim, Candace Ross, Chantal Shaib, Kerem Oktar, Samuel J. Bell, Anaelia Ovalle, Jesse Dodge, Antoine Bosselut, Koustuv Sinha, and Adina Williams. 2026. [Brittlebench: Quantifying LLM robustness via prompt sensitivity](#). *arXiv preprint arXiv:2603.13285*.
- Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. [Quantifying language models’ sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting](#). In *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Sanjif Shanmugavelu, Mathieu Taillefumier, Christopher Culver, Oscar Hernandez, Mark Coletti, and Ada Sedova. 2024. [Impacts of floating-point non-associativity on reproducibility for HPC and deep learning applications](#). In *SC’24 Workshops (IEEE)*.
- Richard J. Shavelson and Noreen M. Webb. 1991. *Generalizability Theory: A Primer*. SAGE Publications.
- Richard J. Shavelson, Noreen M. Webb, and Glenn L. Rowley. 1989. [Generalizability theory](#). *American Psychologist*.
- Team OLMo, Pete Walsh, Luca Soldaini, and 1 others. 2025. [2 OLMo 2 furious](#). *arXiv preprint arXiv:2501.00656*.
- Francis Tuerlinckx, Frank Rijmen, Geert Verbeke, and Paul De Boeck. 2006. [Statistical inference in generalized linear mixed models: A review](#). *British Journal of Mathematical and Statistical Psychology*.
- An Yang, Anfeng Li, Baosong Yang, and 1 others. 2025. [Qwen3 technical report](#). *arXiv preprint arXiv:2505.09388*.
- Haoran Ye, Jing Jin, Yuhang Xie, Xin Zhang, and Guojie Song. 2025. [Large language model psychometrics: A systematic review of evaluation, validation, and enhancement](#). *arXiv preprint arXiv:2505.08245*.

Alex Young, Bei Chen, Chao Li, and 1 others. 2024. [Yi: Open foundation models by 01.AI](#). *arXiv preprint arXiv:2403.04652*.

Jiayi Yuan, Hao Li, Xinheng Ding, Wenya Xie, Yu-Jhe Li, Wentian Zhao, Kun Wan, Jing Shi, Xia Hu, and Zirui Liu. 2025. [Understanding and mitigating numerical sources of nondeterminism in LLM inference](#). *arXiv preprint arXiv:2506.09501*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. [HellaSwag: Can a machine really finish your sentence?](#) In *Proceedings of the Association for Computational Linguistics (ACL)*.

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. 2026. [Lost in benchmarks? rethinking large language model benchmarking with item response theory](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*. Oral presentation.

A Additional Robustness Analyses

We report nine robustness analyses that verify the stability of the variance decomposition under alternative analytical choices.

GLMM complement. A generalized linear mixed model (GLMM) with marginal logit link produces variance component rankings that perfectly preserve all seven grouped component orderings ($\rho = 1.0$) relative to Henderson Method I (Jiang and Skorupski, 2018; Tuerlinckx et al., 2006). This confirms that the linear variance decomposition is robust to the binary outcome distribution.

Accuracy-range stability. The `item_id` variance share remains stable across subsets with mean accuracy from 0.62 to 0.72 ($r = -0.048$). This rules out distortion from binary floor or ceiling effects at observed accuracy levels.

Estimator robustness. We compare Henderson Method I with restricted maximum likelihood (REML-EM, 500 iterations, practical convergence $\Delta < 0.003$ percentage points). All variance components differ by < 0.5 percentage points between estimators. The largest discrepancy occurs on the residual component (0.47pp); all substantive components (item, model $\times$ item, model) differ by < 0.3 pp. Henderson Method I is preferred for our balanced designs because it produces unbiased estimates without distributional assumptions (Brennan, 2001, Ch. 12).

Probit simulation. A probit simulation matched to our 1,296-condition design confirms that Henderson Method I recovers variance component rankings from dichotomized data (Spearman $\rho = 0.732$ between binary and continuous rankings, excluding residual; `item_id` dominant in all 100 replications; top-5 ordering correct in 79% of replications). Bias analysis by difficulty stratum shows $|\text{bias}| = 0.28\text{--}0.37$ for items with $p \in [0.5, 0.8]$, where the Bernoulli variance ceiling $p(1-p)$ is largest. This analysis covers the observed accuracy range ($p \in [0.3, 0.9]$); extreme-difficulty items ($p < 0.3$ or $p > 0.9$) are rare in our sample ($< 8\%$ of items) and may exhibit larger deviations. Henderson Method I absorbs this Bernoulli-inherent variance into the residual, explaining inflated bias near medium difficulty; systematic component ordering is robustly recovered across all replications. Binary scoring inflates residual variance by absorbing the Bernoulli $p(1-p)$ ceiling, which enters σ_{rel}^2 and increases D-study item requirements. Within a matched probit simulation, binarization increases D-study budgets by approximately 36% ($99 \rightarrow 135$ items for $G \geq 0.80$), driven by residual inflation ($16\% \rightarrow 41\%$) and item variance compression ($50\% \rightarrow 32\%$). Facet $\times$ item interactions (the components entering σ_{rel}^2) are best preserved under binarization (MAPE = 12.8%, range 8.7–17.5%), while residual variance inflates as expected from Bernoulli $p(1-p)$ noise (MAPE = 156%). D-study budgets derived from binary data are therefore conservative: continuous-score designs would require equal or fewer items for the same G threshold. Null calibration (ratio = 1.0008) confirms that this conservatism does not distort relative component magnitudes.

CI method comparison. Bootstrap (1,000 item-level resamples with replacement) and delete-1 jackknife confidence intervals agree closely: width ratios range from 0.96 to 1.04 across all variance components. Bootstrap CIs capture item-sampling variability conditional on these 8 specific models; model-level stability is assessed separately via LOMO (below). The agreement indicates that the dependence structure introduced by item-level bootstrap resampling (which creates ~ 127 unique items per 200-item resample) does not materially distort interval estimates. Stratified resampling by subject (10 subjects $\times$ 20 items) produces CIs within 4% of simple item-level bootstrap, confirming that the within-subject dependence structure

does not materially affect interval estimates.

Facet specification sensitivity. Treating temperature and ordering as fixed (rather than random) facets changes G_{item} by only 0.37 percentage points (all-random: 0.820 vs. fixed temperature+ordering: 0.824). The small difference reflects the negligible temperature $\times$ item (0.65%) and ordering $\times$ item (0.54%) variance shares. The random-facet assumption is therefore robust to facet classification.

Leave-one-model-out (LOMO). Across all 40 leave-one-model-out subsets (8 models $\times$ 5 benchmarks), G_{item} ranges from 0.833 to 0.919, with CV $\leq 1.25\%$ per benchmark. No single model removal changes G by more than 0.030 (GSM8K, removing Gemma-2-9B). Item-related variance rankings are preserved in all 40 subsets. Table 4 (main text) reports per-benchmark stability. This supports treating models as exchangeable within the target universe of 7–9B instruction-tuned models (§3.6).

Inter-model agreement. We computed pairwise Cohen’s κ and Fleiss’ κ across all model pairs on each benchmark. Mean pairwise κ ranges from 0.326 (MATH) to 0.461 (MMLU); Fleiss’ κ ranges from 0.309 (GSM8K) to 0.456 (MMLU). HellaSwag uses six models ($\kappa = 0.332$) due to missing OLMo and Yi data. The fair-to-moderate agreement independently confirms that the model $\times$ item interaction (15–36% of variance across benchmarks) reflects genuine differential item functioning: models disagree substantially on which items are difficult, consistent with prior DIF findings in educational measurement (Holland and Wainer, 1993; Clauser and Mazor, 1998).

Scoring metric sensitivity. To probe metric conditionality, we compared binary correctness with step-level partial credit on GSM8K using the full 8-model crossed design (200 items, 2,304 conditions per item). Under partial credit, G_{item} drops from 0.863 to 0.768, with residual variance rising from 39.9% to 64.9% and item variance falling from 16.9% to 10.3%. All eight models individually show $G_{\text{partial}} < G_{\text{binary}}$ (mean $\Delta G = -0.135$, range -0.025 to -0.309), with ΔG magnitude moderated by model accuracy: higher-accuracy models have fewer incorrect responses whose scores shift under partial credit (e.g., OLMo at 70.1% accuracy shows $\Delta G = -0.025$, while Llama at 49.0% shows $\Delta G = -0.309$). The mechanism is clear: 48.1% of incorrect responses have numerically correct arithmetic steps but incorrect

problem setup, receiving partial score = 1.0 despite binary score = 0; this compresses item discrimination and inflates residual variance. Binary scoring amplifies prompt sensitivity by collapsing reasoning quality into pass/fail, confirming that all findings in this paper are metric-conditional.

Facet specification sensitivity: random vs. fixed model. Table 10 reports per-benchmark generalizability coefficients under random and fixed model specifications. The consistent ΔG of 0.077–0.113 confirms that random-model budgets are strictly more conservative across all benchmarks.

Table 10: Generalizability coefficients under random vs. fixed model facet specification across five benchmarks.

Benchmark	G_{random}	G_{fixed}	ΔG
MMLU	0.896	0.984	+0.088
ARC	0.892	0.990	+0.098
HellaSwag	0.872	0.984	+0.113
GSM8K	0.863	0.951	+0.089
MATH	0.888	0.965	+0.077
Mean	0.882	0.975	+0.093

Null calibration. We generate 100 Bernoulli-distributed datasets with item difficulties matched to the empirical distribution but no systematic facet effects. The observed-to-null variance ratio is 1.0008, confirming that Henderson Method I does not introduce systematic bias when applied to binary outcomes. The null simulation also verifies that the item variance share (58.4%) is not an artifact of the Bernoulli $p(1-p)$ ceiling: the null produces item variance of 0.08%, three orders of magnitude below the observed value. Across all 100 replications, the same three components (item, model $\times$ item, residual) emerge as the top three (set consistency 100%), with Kendall’s $\tau = 0.81$ between binary and continuous rankings.

Reduced crossing validation. To verify that the reduced crossing in exp-002 (BF16 only $\times$ 2 temperatures) does not introduce confounding, we extracted the corresponding subset from the fully crossed exp-001 design. The generalizability coefficient differs by only 0.0008 ($G = 0.9361$ vs. 0.9369), and all non-precision variance components differ by less than 2 percentage points. The 10pp increase in item variance share (58.4% $\rightarrow$ 68.5%) is fully explained by the removal of precision-related components ($\sim 4.5\%$ total), which redistributes across remaining sources. This confirms that the exp-002 design preserves the variance structure of the full crossing.

B Projection Validation Details

Internal consistency check. Table 11 reports consistency checks from four exp-001 models (DeepSeek, Llama, Mistral, Qwen3) to the full eight-model design across all five benchmarks. MC benchmarks show $|\Delta G| \leq 0.029$ ($<3.5\%$ error), while FF benchmarks show larger deviations ($|\Delta G| \leq 0.065$, $\sim 7\%$) due to limited model diversity in the four-model subset. FF predictions consistently underestimate actual G (conservative); MC predictions show mixed directions with small deviations ($|\Delta G| \leq 0.029$).

Table 11: Internal consistency: four-model subset vs. eight-model G_{item} . FF subsets underestimate actual G (conservative); MC shows mixed directions with small deviations.

Benchmark	Predicted G	Actual G	$ \Delta G $	Error%
<i>Multiple-choice</i>				
MMLU	0.888	0.896	0.009	0.96
ARC	0.904	0.892	0.012	1.38
HellaSwag	0.901	0.872	0.029	3.33
<i>Free-form</i>				
GSM8K	0.803	0.863	0.059	6.88
MATH	0.823	0.888	0.065	7.30
Mean	—	—	0.035	3.97

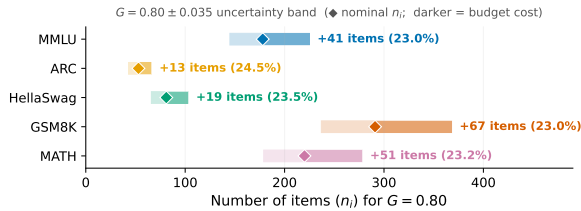


Figure 9: Effect of $|\Delta G| = 0.035$ estimation error on item budgets. The $\sim 23\%$ increase is consistent across benchmarks.

Budget sensitivity to estimation error. The effect of $|\Delta G| = 0.035$ estimation error on item budgets is shown in Figure 9. The percentage increase is remarkably stable (23.0–24.5%, mean 23.2%), because the relative budget adjustment $\Delta n/n$ depends on the G -formula ratio rather than on absolute variance components. Absolute Δn ranges from 13 (ARC) to 67 (GSM8K), scaling with the baseline budget.

Cross-benchmark G stability. G_{item} values cluster in a narrow band across all five benchmarks (0.863–0.896, $\text{CV} = 1.4\%$), despite qualitatively

different variance structures: MC benchmarks concentrate 68–72% in item-related components while GSM8K disperses variance across item (17%), model (11%), and interaction (15%) sources (Table 6). This stability implies that D-study budget recommendations calibrated on one benchmark provide reasonable order-of-magnitude guidance for others. However, the 0.033 range in G_{item} translates to a 71–128 item range for multi-facet $G \geq 0.80$ (Table 9), and the variance compositions differ substantially; benchmark-specific G-studies remain necessary to identify dominant variance sources and optimize condition selection.

C Format-Associated Variance: Per-Benchmark Details

This section supplements the compressed format-associated analysis in §4.3 with per-benchmark details.

MC benchmark exceptions. HellaSwag is the sole MC case where model×item slightly exceeds item (36.0% vs. 34.3%), though their sum (70.3%) remains in the MC range (68–72%). Among MC benchmarks, ARC-Challenge shows an anomalously high model×prompt interaction (8.0% vs. $<2\%$ for MMLU and HellaSwag), suggesting that science-domain items are differentially sensitive to option formatting across model families.

FF benchmark patterns. Free-form benchmarks differ markedly from MC. On GSM8K, variance distributes across item (16.9%), model×item (14.6%), model (10.7%), and model×prompt (8.5%); prompt main effect is 5.4%. MATH shows an intermediate pattern: item leads at 24.0% with model×item at 17.3%, but residual variance is substantially larger (43.9%). Item-level regression reveals that FF residual variance is predominantly explained by the Bernoulli variance ceiling $p(1-p)$ ($R^2 = 0.47\text{--}0.86$; Appendix F): medium-difficulty items maximize binary outcome variance regardless of facet structure, so the elevated FF residuals reflect measurement resolution rather than unmodeled substantive sources.

Difficulty dispersion as variance predictor. Difficulty dispersion ($\text{std}(p_i)$) strongly predicts item variance share ($\rho = 0.82$, $n = 5$ benchmarks; Appendix D; reasoning-heavy subjects show G_{item} as low as 0.614, Appendix E). Format constrains the noise floor, but difficulty heterogeneity determines variance magnitude within it.

Item-level bootstrap stability. Bootstrap resampling (1,000 item resamples per benchmark) confirms that the MC–FF separation is robust to item selection: 95% CIs for item-related variance (item + model $\times$ item) are [65.2, 75.0]% (MC) and [28.0, 44.2]% (FF), with no overlap (minimum gap 21.0 percentage points).

D Cross-Benchmark Difficulty Heterogeneity

Item difficulty heterogeneity is shown in Figure 7 (main text). MC benchmarks cluster at higher dispersion (0.27–0.31) and higher item variance (34–38%), while FF benchmarks occupy the lower-left region (0.22–0.23, 17–24%). The cross-benchmark Spearman $\rho = 0.82$ ($p = 0.089$, $n = 5$) is consistent in direction with the within-MMLU correlation ($\rho = 0.806$, $p = 0.005$, $n = 10$ subjects); these are independent analyses at different granularities (benchmarks vs. subjects within MMLU), providing multi-scale support for the item difficulty heterogeneity mechanism discussed in §4.3.

E MMLU Subject-Level Variance Decomposition

Table 12 reports per-subject variance decomposition for 10 MMLU subjects (20 items each, 8 models, 2,304 conditions per item). G_{item} ranges from 0.614 (high_school_mathematics) to 0.930 (international_law), with mean 0.843. Knowledge-intensive subjects (international_law, clinical_knowledge, world_religions) show $\text{item_id} > 40\%$, while reasoning-heavy subjects (high_school_mathematics, abstract_algebra) show $\text{model} \times \text{item} > \text{item_id}$, consistent with models differing more in reasoning strategies than in factual recall. The high_school_mathematics item_id of 10.4% referenced in §4.3 is the lowest across all subjects, below even GSM8K (16.9%).

F FF Residual Variance Analysis

To investigate whether the elevated residual variance in free-form benchmarks (GSM8K 39.9%, MATH 43.9%) reflects unmodeled substantive sources or measurement properties of binary scoring, we regress item-level residual variance on item characteristics across all five benchmarks (Table 13).

Table 12: Subject-level variance decomposition within MMLU (20 items per subject, 8 models, 2,304 conditions per item). Sorted by G_{item} descending.

Subject	Acc.	item%	mod. $\times$ it.%	G_{item}
international_law	.717	44.4	24.1	.930
clinical_knowledge	.581	47.5	30.0	.923
world_religions	.826	42.2	27.1	.910
computer_security	.740	37.0	24.9	.908
college_physics	.436	31.6	30.1	.876
philosophy	.782	33.0	36.1	.868
abstract_algebra	.457	24.1	32.8	.822
us_foreign_policy	.864	20.4	36.9	.802
professional_medicine	.671	19.9	41.7	.778
high_school_math	.385	10.4	41.9	.614
Mean	—	31.1	32.6	.843

The dominant predictor is the quadratic difficulty effect ($p < 10^{-5}$ on all benchmarks): items near $p = 0.5$ show maximal residual variance, while items near the floor or ceiling show minimal residual. This inverted-U pattern directly mirrors the Bernoulli variance ceiling $p(1-p)$. Item content features (number of reasoning steps, answer length) are not significant predictors for GSM8K ($p > 0.38$), confirming that the difficulty–residual relationship is statistical rather than content-driven.

FF benchmarks show substantially higher R^2 (mean 0.66) than MC benchmarks (mean 0.25), consistent with FF items spanning a wider difficulty range. The MC–FF contrast in residual magnitude is therefore a mathematical consequence of the difficulty distribution under binary scoring, not evidence of additional unmodeled variance sources in free-form evaluation.

Table 13: Item-level residual variance regression. Difficulty² = quadratic (inverted-U) term reflecting the Bernoulli $p(1-p)$ ceiling. GSM8K additionally tests n_{steps} and answer_length (both $p > 0.38$).

Benchmark	R^2 (full)	R^2 (diff. only)	Diff. ² p
<i>Multiple-choice</i>			
MMLU	0.377	0.347	<0.001
ARC	0.221	0.204	<0.001
HellaSwag	0.156	0.140	<0.001
<i>Free-form</i>			
GSM8K	0.468	0.449	<0.001
MATH	0.856	0.845	<0.001

G Model and Prompt Details

Models. See Table 3 (main text) for the eight models. Full HuggingFace identifiers: meta-llama/Llama-3.1-8B-Instruct (Grattafiori et al., 2024), google/gemma-2-9b-it (Riviere et al., 2024), mistralai/Mistral-7B-Instruct-v0.3 (Jiang

et al., 2023), Qwen/Qwen3-8B (Yang et al., 2025), 01-ai/Yi-1.5-9B-Chat (Young et al., 2024), allenai/OLMo-2-1124-7B-Instruct (Team OLMo et al., 2025), internlm/internlm2_5-7b-chat (Cai et al., 2024), deepseek-ai/deepseek-llm-7b-chat (Bi et al., 2024).

Prompt templates. Our six prompt templates follow a surface perturbation design inspired by BrittleBench (Romanou et al., 2026). For multiple-choice benchmarks, templates vary along three dimensions: option layout (lettered vs. numbered), instruction wording (direct vs. contextual), and answer prompt cue. For free-form benchmarks, templates vary chain-of-thought elicitation strategy. All six templates are semantically equivalent: they request the same task with different surface forms, to isolate formatting effects from content effects. ARC-Challenge and HellaSwag use the same six surface patterns as MMLU with domain-appropriate system instructions; MATH mirrors the GSM8K pattern with `\boxed{}` answer formatting.

MMLU templates (representative MC format):

1. The following are multiple choice questions (with answers) about {subject}.\n\n{question}\nA. ... D. ... \nAnswer:
2. Answer the following {subject} questions by selecting the correct option.\nQuestion: {question}\n(A) ... (B) ... (C) ... (D) ... \nCorrect answer:
3. Below are {subject} multiple-choice problems. Pick the right letter.\nQ: {question}\n A) ... B) ... C) ... D) ... \nA:
4. Answer the following {subject} question by selecting A, B, C, or D.\n{question}\n- A. ... - B. ... - C. ... - D. ... \nAnswer:
5. Below is a {subject} exam question with four possible answers. Choose the correct one.\n{question}\nOption A: ... Option B: ... Option C: ... Option D: ... \nCorrect:
6. Question ({subject}): \n{question}\nChoices: (A) ... (B) ... (C) ... (D) ... \nYour answer:

GSM8K templates (representative FF format):

1. Solve the following math problem step by step.\nQ: {question}\nA: Let’s think step by step.
2. Answer the following math word problem. Show your work.\nQuestion: {question}\nSolution:
3. Solve each problem below. Write your reasoning, then give the final answer after ####.\nProblem: {question}\nReasoning:
4. You are a math tutor. Solve the problem and show your steps.\n{question}\nSteps:
5. Work through this math question carefully.\nQ: {question}\nWork:
6. Math problem: \n{question}\nLet’s solve this step by step.

H Software Environment

All experiments were conducted with Python 3.10, PyTorch 2.1, and CUDA 12.8 on a single NVIDIA RTX PRO 6000 GPU (96 GB). Inference used vLLM 0.9.2 with HuggingFace Transformers 4.40. Variance component estimation used StatsModels 0.14 and SciPy 1.12. A full list of pinned dependencies is available in the released code repository.

I Use of AI Assistance

We used Claude for coding, code drafting, debugging, and language polishing.