

Generalizability Theory for Evaluating Fine-Tuned LLMs in Automated Item Generation

Zhifei Li

University of Iowa
zhifei-li@uiowa.edu

Won-Chan Lee

University of Iowa
won-chan-lee@uiowa.edu

Abstract

Automated item generation with large language models (LLMs) is typically evaluated using aggregate accuracy metrics that conflate the quality of the source passage with noise introduced by the generation process itself. We address this gap by embedding a prospective $i : (p \times s \times t)$ Generalizability Theory (G-theory) design into the evaluation of a QLoRA fine-tuned Qwen2.5-7B-Instruct model on the SciQ corpus. Passages (p) serve as the object of measurement; random seeds (s) and prompt templates (t) are fully crossed facets; items are nested within each (p, s, t) cell. Across 9,000 observations and five binary quality metrics, we find that 77–81% of total variance is attributable to the passage, seed and template main effects are negligible ($\leq 0.02\%$), and G-coefficients ($E\hat{\rho}^2$) uniformly exceed 0.97 under the observed design. D-study projections show that a single seed and template already achieves $E\hat{\rho}^2 = 0.87$, while the observed design ($n_s = 3$, $n_t = 3$, $n_i = 2$) reaches 0.98. Code, data, and R analysis scripts are released to support reproducible psychometric evaluation of future item-generation systems.

1 Introduction

Automated item generation (AIG) using large language models (LLMs) has received growing attention as a promising approach for producing educational assessment items at scale and at lower cost (Moore et al., 2022; Kurdi et al., 2020). The operational promise is compelling: given a passage of source text, a fine-tuned model should produce a well-formed, content-valid multiple-choice item—complete with a stem, four labeled options, a correct-answer designation, a rationale, and an evidence sentence—without human authoring effort. Recent work confirms that instruction-tuned models can produce structurally plausible items on scientific passages (Wei et al., 2022; Ouyang et al., 2022; Tian et al., 2026; Dang et al., 2026).

Yet these studies share a common measurement gap: they report aggregate accuracy or human-rated quality scores, leaving unanswered a more fundamental reliability question.

How much of the observed variation in item quality is attributable to the passage being assessed, and how much is irreducible noise introduced by the generation process itself?

This distinction is consequential. A generator whose quality depends primarily on the passage represents genuine comprehension—its errors are informative about passage difficulty. A generator whose quality fluctuates with the random seed, the phrasing of the prompt, or the label assigned to the correct answer is an unreliable instrument: its variance cannot be attributed to anything content-meaningful. Distinguishing these sources of variance requires a measurement framework that isolates the *object of measurement*—the passage—from the *facets of generalization*—the seed, the prompt template, and the item replicate—and that does so simultaneously across multiple quality dimensions.

Research Questions.

1. How well does the fine-tuned language model generate high-quality items, and how effective is the proposed generation pipeline in producing valid, evidence-aligned outputs?
2. How reliable are the generated items across generation conditions (i.e., prompt templates and decoding seeds), and what minimum generation design (e.g., number of prompts, seeds, and items per passage) is required to achieve adequate generalizability?

2 Related Work

Automated item generation has a longer history than is suggested by its recent AI approaches. Early AIG systems relied on hand-crafted item templates

and slot-filling rules applied to structured knowledge bases (Christ et al., 2024; Gierl and Lai, 2012). While controllable, these approaches produced items of limited diversity and required substantial domain-specific engineering. The neural turn in NLP brought sequence-to-sequence models (Dong et al., 2025; Naeiji et al., 2023) that could generate question stems from passages directly, though producing the full multiple-choice structure—stem, four labeled options, a correct designation, and supporting evidence—remained a harder target.

The reliability of evaluation judgments in NLP has received considerable attention, particularly in summarization, where annotator disagreement is high (Liu et al., 2023; Hardy et al., 2019). The standard approach to measuring this disagreement borrows from psychometrics: inter-rater agreement coefficients such as Cohen’s κ (Chen et al., 2001), Krippendorff’s α (Chen et al., 2001), or the intraclass correlation coefficient (ICC) (Shrout and Fleiss, 1979).

G-theory was introduced by Cronbach et al. (1972) as a generalization of classical test theory (CTT) that replaces the single, undifferentiated error term of CTT with a decomposition of variance across the facets of a measurement design. Where CTT asks “how reliable is this test?”, G-theory asks “reliable with respect to which sources of variation, and under which conditions of generalization?” (Shavelson and Webb, 1989). The theory distinguishes two studies: the *G-study*, which estimates variance components from a fully-crossed or nested design, and the *D-study*, which uses those components to project reliability under alternative measurement designs—different numbers of items, raters, or occasions.

Applications of G-theory to NLP and AI systems have been sparse. Chen et al. (2022) applied a univariate G-theory framework to automatic essay scoring, estimating the relative contributions of student, prompt, and scoring occasion to essay-score variance. Zaefarian et al. (2022) discussed variance decomposition for generation evaluation in general terms but did not instantiate a G-theory design. No prior work has applied a G-theory design to the evaluation of an LLM as a generative measurement instrument in terms of prompt templates, decoding seeds and item replicates.

3 Method

Our method consists of two complementary components. First, we fine-tune an LLM to generate structured multiple-choice items from science passages, using a training protocol whose key design choices are motivated by psychometric considerations (Sections 3.1–3.2). Second, we evaluate the fine-tuned model under a prospective univariate Generalizability Theory framework, with passages as the object of measurement. Variance in generated item quality is decomposed across prompt templates and decoding seeds, with items nested within each passage-by-seed-by-prompt cell (Sections 3.3–3.4).

3.1 Dataset and Task Formulation

Corpus. We use SciQ (Welbl et al., 2017), a dataset of 13,679 multiple-choice science questions drawn from biology, chemistry, earth science, and physics textbooks, partitioned into three official splits: 11,679 training, 1,000 validation, and 1,000 testing. From the 1,000 validation passages, we randomly sample 500 to keep the total cell count ($500 \times 2 \times 3 \times 3 = 9,000$) computationally tractable while maintaining a psychometrically adequate passage-level sample size for reliable variance component estimation (Brennan, 2001).

Target schema. We define the generation target as a structured JSON object with five mandatory keys: stem, options, answer, rationale, and evidence. The options field maps labels A, B, C, D to option texts, while answer identifies the correct option label. The evidence field is required to contain one sentence copied verbatim from the support passage.

This schema enforces three item-quality constraints: structural compliance, content grounding, and label consistency. These constraints motivate the multi-metric evaluation design described in Section 3.4.

3.2 Fine-Tuning Protocol

Base model. We fine-tune *Qwen2.5-7B-Instruct* (Qwen et al., 2025), a 7-billion parameter instruction-tuned model with a 128K-token context window. The model’s strong instruction-following capability and support for structured JSON generation make it a suitable base for schema-constrained item generation. All experiments are conducted on a single NVIDIA L40S GPU (48 GB).

Quantization. To reduce GPU memory requirements, we load the base model in 4-bit precision using (Dettmers et al., 2023) with the NF4 quantization type and double quantization enabled. The compute dtype is `bf16`. This configuration reduces the loaded model footprint to approximately 5 GB, leaving sufficient headroom for the LoRA adapter parameters, optimizer states, and training batch activations.

Low-rank adaptation. We attach LoRA adapters (Hu et al., 2022) to the query, key, value, output, gate, up, and down projection matrices of every transformer layer, following the standard practice for instruction-tuned causal language models (Vaswani et al., 2023; Wei et al., 2022). The adapter hyperparameters are rank $r = 16$, scaling factor $\alpha = 32$ (effective scale $\alpha/r = 2.0$), and dropout $p = 0.05$. This introduces approximately 20M trainable parameters on top of the frozen 7B base, a ratio of roughly 0.3%.

Training objective and prompt masking. We use the standard causal language modeling objective (cross-entropy over next-token predictions) applied only to the *assistant turn* of each training example. Concretely, each training instance is formatted as a three-turn chat template: a system prompt establishing the item-writer persona, a user turn containing the support passage and generation instructions, and an assistant turn containing the target JSON object. The input sequence is tokenized in full; the labels for the system and user tokens are set to -100 (ignored by the loss function) so that gradient signal flows exclusively through the assistant JSON output.

Prompt templates. We define three prompt templates (t_1, t_2, t_3) that convey the same generation instructions using different phrasing, ordering, and imperative framing (see Appendix A). All three templates require the same output schema and are designed to convey equivalent task information. During training, all examples are formatted with t_1 (prompt_1), ensuring that the model’s exposure to the task is not confounded with template variance. At evaluation, the trained model is applied to all three templates (t_1, t_2, t_3). Thus, for each passage and decoding seed, outputs are generated under each prompt condition, forming a fully crossed $p \times s \times t$ design. Template t_1 is included at evaluation as the in-distribution condition, providing a baseline against which performance under t_2 and

t_3 can be compared.

Training configuration. We train for 6 epochs with a cosine learning-rate schedule, initial learning rate $\eta = 2 \times 10^{-4}$, warmup ratio 0.03, and weight decay 0. We use paged AdamW with 8-bit optimizer states (Dettmers et al., 2022), per-device batch size 2, and gradient accumulation over 8 steps (effective batch size 16). The maximum sequence length is 2,048 tokens; examples exceeding this are truncated from the right. Gradient checkpointing is enabled to reduce activation memory. The model checkpoint with the lowest validation loss is selected for evaluation. All training hyperparameters are summarized in Table 1.

Hyperparameter	Value
Base model	Qwen2.5-7B-Instruct
Quantization	4-bit NF4, double quant
LoRA rank r	16
LoRA α	32
LoRA dropout	0.05
Target modules	q, k, v, o, gate, up, down
Trainable parameters	$\approx 20\text{M}$ (0.3%)
Training epochs	6
Effective batch size	16
Learning rate	2×10^{-4}
LR schedule	Cosine
Warmup ratio	0.03
Optimizer	Paged AdamW 8-bit
Max sequence length	2,048 tokens
Compute dtype	<code>bf16</code>

Table 1: Fine-tuning hyperparameters.

3.3 G-Theory Measurement Design

Overview. Cronbach et al. (1972) and Brennan (2001) partitions the total variance in a set of observations into components attributable to each source of variation (facet) in the measurement design. We implement a $i : (p \times s \times t)$ design in which the *object of measurement* is the passage and the *facets* are the random seed, the prompt template, and items nested within each passage-by-seed-by-prompt combination. Table 2 summarizes all design facets and their levels.

Facet definitions. *Passages* (p) are the objects of measurement in the G-Theory framework. For each passage, the model generates a set of items under each combination of random seed and prompt template. Since the intended inference concerns the quality of item generation at the passage level, rather than the characteristics of individual generated items, the goal is to estimate each passage’s

Facet	Role	Type	Levels
p	Passage; object of measurement	—	500
s	Seed; crossed random facet	Random	3
t	Prompt; crossed random facet	Random	3
$i : (p \times s \times t)$	Item nested within (p, s, t)	Random	2/cell

Table 2: G-theory measurement design: $i : (p \times s \times t)$. Total cells: $500 \times 2 \times 3 \times 3 = 9,000$.

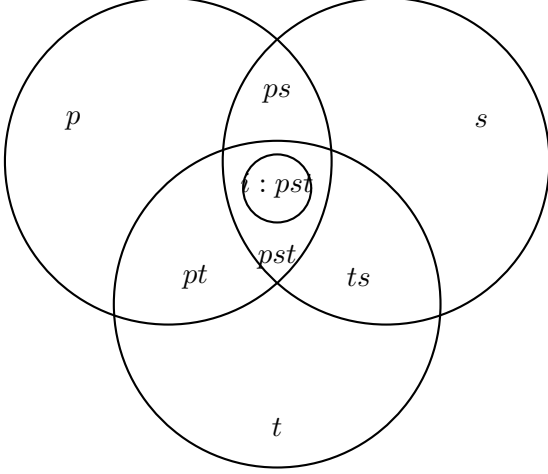


Figure 1: Variance decomposition structure for the G-theory design. The diagram illustrates the main effects of passage (p), seed (s), and prompt template (t), their two-way and three-way interactions, and the nested item effect within the passage-by-seed-by-template condition ($i : pst$).

universe score on each quality metric—that is, the expected quality of items generated from that passage across all admissible generation conditions.

Items nested within passage-by-seed-by-prompt cells ($i : (p \times s \times t)$) represent within-cell replication. For each (p, s, t) combination, we generate $n_i = 2$ items by initializing the random number generator once at the start of the cell and allowing the model’s internal stochastic sampling process to produce distinct outputs. Items are not re-seeded between replicates within a cell, so each pair of items shares an initial RNG state but diverges through token-level sampling. The $i : (p \times s \times t)$ variance component therefore captures item-to-item variability under fixed passage, seed, and prompt conditions.

Seeds ($s \in \{41, 51, 61\}$) control the initialization of the random number generator prior to each (p, s, t) cell. Because temperature sampling ($T = 0.7$, top- $p = 0.9$) introduces stochasticity into generation, different seeds yield different items

from the same passage under the same prompt. The seed facet isolates this source of variance.

Prompt templates ($t \in \{t_1, t_2, t_3\}$) represent three linguistically distinct but informationally equivalent formulations of the item-generation instruction. The prompt facet isolates variance attributable to instruction phrasing, a practically important source of instability in deployed LLM systems.

Fully crossed design. All facets except the item facet are fully crossed: every passage is evaluated under every combination of seed and prompt template. Items are nested within each passage-by-seed-by-prompt combination, denoted $i : (p \times s \times t)$. Thus, for each (p, s, t) cell, multiple item replicates are generated.

This yields a total of

$$n_p \times n_s \times n_t \times n_i = 500 \times 3 \times 3 \times 2 = 9,000$$

observations per metric.

Full crossing of p , s , and t is essential for G-theory, as it permits unconfounded estimation of all main effects and interaction variance components, including the three-way interaction ($p \times s \times t$). For relative decisions about passages, interactions involving p contribute to relative error variance after being scaled by the number of seeds, prompt templates, and nested item replications included in the measurement design.

Variance components. The variance decomposition structure is illustrated in Figure 1. For a single outcome variable Y , the linear decomposition under this design is:

$$Y_{psti} = \mu + \nu_p + \nu_s + \nu_t + \nu_{ps} + \nu_{pt} + \nu_{st} + \nu_{pst} + \nu_{i:pst},$$

where μ is the grand mean; ν_p , ν_s , and ν_t are the main effects of passage, seed, and prompt, respectively; ν_{ps} , ν_{pt} , and ν_{st} are two-way interaction effects; ν_{pst} is the three-way interaction among passage, seed, and prompt; and $\nu_{i:pst}$ represents item-level variability nested within each (p, s, t) cell.

Each component is assumed to be independently distributed with mean zero and variance $\sigma^2 \geq 0$.

$E\hat{\rho}^2$ -coefficient and $\hat{\Phi}$ -coefficient. The estimated *generalizability coefficient* $E\hat{\rho}^2$ and the estimated *dependability coefficient* $\hat{\Phi}$ were computed from the D-study variance components. The D-study variance components are obtained by dividing the corresponding G-study variance compo-

nents by the sample sizes specified for the D-study design.

The estimated generalizability coefficient $\hat{\mathbf{E}}\rho^2$ is the ratio of estimated universe-score variance to estimated observed-score variance under a specified D-study design (Brennan, 2001):

$$\mathbf{E}\hat{\rho}^2 = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \hat{\sigma}_\delta^2} \quad (1)$$

where the relative error variance $\hat{\sigma}_\delta^2$ aggregates all variance components that contribute to rank-order error under the observed design:

$$\hat{\sigma}_\delta^2 = \hat{\sigma}_{ps}^2 + \hat{\sigma}_{pt}^2 + \hat{\sigma}_{pst}^2 + \hat{\sigma}_{i:pst}^2$$

where $\hat{\sigma}_{i:pst}^2$ denotes the variance component of items nested within each (p, s, t) cell.

The Φ -coefficient is the criterion-referenced analogue of the generalizability coefficient, defined as

$$\hat{\Phi} = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \hat{\sigma}_\Delta^2}, \quad (2)$$

where $\hat{\sigma}_\Delta^2$ denotes the *absolute error variance*. Unlike relative error variance, absolute error variance includes all sources of measurement error, except for σ_p^2 . It is computed as

$$\hat{\sigma}_\Delta^2 = \hat{\sigma}_\delta^2 + \hat{\sigma}_s^2 + \hat{\sigma}_t^2 + \hat{\sigma}_{st}^2.$$

D-study projections. The D-study uses the estimated variance components from the G-study to project $\mathbf{E}\hat{\rho}^2$ and $\hat{\Phi}$ under alternative measurement designs, varying the numbers of seeds n'_s , prompt templates n'_t , and item replicates n'_i . This produces a reliability surface over the design space $(n'_i, n'_s, n'_t) \in \{1, 2, 3, 4\} \times \{1, \dots, 5\} \times \{1, \dots, 5\}$.

Because items are nested within each (p, s, t) combination, n'_i indexes replicates within a cell rather than a crossed facet. The D-study therefore evaluates how increasing within-cell replication (items), across-seed variation, or across-prompt variation contributes to reliability.

This surface directly informs practical deployment decisions: given a target reliability and a fixed per-passage generation budget (i.e., total number of model calls), the D-study identifies the optimal allocation of calls across seeds, prompt templates, and within-cell item replicates.

Estimation. All variance components are estimated by restricted maximum likelihood (REML) using the `lmer` function in `lme4` (Bates et al., 2015).

For each metric, we fit the following mixed-effects model:

$$\begin{aligned} \text{score} \sim & (1|\text{passage_id}) \\ & + (1|\text{seed}) \\ & + (1|\text{prompt_id}) \\ & + (1|\text{passage_id}:\text{seed}) \\ & + (1|\text{passage_id}:\text{prompt_id}) \\ & + (1|\text{seed}:\text{prompt_id}) \\ & + (1|\text{passage_id}:\text{seed}:\text{prompt_id}) \\ & + (1|\text{item_uid}) \end{aligned}$$

where $\text{item_uid} = (p, s, t, i)$ uniquely identifies each item replicate nested within a specific passage-by-seed-by-prompt cell, corresponding to the i : $(p \times s \times t)$ design. Variance components that REML drives to the boundary (zero) are retained as zero in all downstream calculations. Each metric is analyzed independently using the same modeling framework, and no cross-metric covariance is estimated.

3.4 Quality Metrics

We collect eight quality metrics per generated item, organized into two measurement tiers. The first tier is used to evaluate overall model performance, whereas the second tier is used for G-Theory analysis.

Tier 1: Scalar text-quality metrics. Three scalar metrics were used to assess continuous aspects of generated text quality.

1. **ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-Lsum** (Lin, 2004) were computed between each generated item and its support passage, with stemming enabled. Because no gold-standard item is available for each evaluation passage, the passage serves as the reference text. In this study, ROUGE is interpreted as an indicator of *lexical grounding*, rather than semantic correctness or overall item quality.
2. **Perplexity** (Chen et al., 2001) was computed under the fine-tuned model as $PPL(y) = \exp\left(-\frac{1}{|y|} \sum_t \log p_\theta(y_t | y_{<t})\right)$. Lower perplexity indicates greater model confidence and more fluent, in-distribution generation.
3. **MAUVE** (Pillutla et al., 2021) was computed once at the corpus level to compare the distribution of generated items with the distribution of support passages. Because MAUVE requires a large sample for stable estimation,

all 9,000 generated texts were pooled for this calculation.

Tier 2: Binary structural metrics. Seven binary metrics were computed from the raw generated output, each taking a value of 0 or 1. These metrics assess whether the item satisfies key structural, grounding, and answer-consistency requirements. Because `json_valid` and `evidence_nonempty` exhibit minimal variability and redundant, we excluded them and retained other five metrics for analysis. These measures capture complementary aspects of item generation, including evidence grounding, alignment between evidence and answer, option-generation quality, and answer selection accuracy, while preserving sufficient variability for reliable variance component estimation.

Summary. Table 3 summarizes all metrics, their type, their computation level, and their role in the G-theory analysis.

Metric	Type	Level	G-study
<code>json_valid</code>	Binary	Per cell	
<code>evidence_nonempty</code>	Binary	Per cell	
<code>evidence_in_passage</code>	Binary	Per cell	
<code>evidence_contains_answer</code>	Binary	Per cell	
<code>answer_correct_soft</code>	Binary	Per cell	
<code>gold_in_options</code>	Binary	Per cell	
<code>correct_label_match</code>	Binary	Per cell	
ROUGE-1	Scalar	Per cell	
ROUGE-2	Scalar	Per cell	
ROUGE-L	Scalar	Per cell	
ROUGE-Lsum	Scalar	Per cell	
Perplexity	Scalar	Per cell	
MAUVE	Scalar	Corpus	—

Table 3: Quality metrics, their type (binary / scalar), computation level (per cell or corpus-wide), and inclusion in the per-metric G-theory decomposition. The checkmark indicates the metric enters the per-cell wide TSV and is modeled in `lme4`.

4 Results

We organize results into two parts that mirror the two measurement tiers defined in Section 3.4. Section 4.1 reports overall model performance across all 9,000 evaluation cells, based on ROUGE scores, perplexity, and MAUVE. Section 4.2 presents the G-theory variance decomposition, reliability coefficients, and D-study projections for the five binary metrics that entered the generalizability analysis.

4.1 Overall Model Performance

Scalar text-quality metrics. Table 4 reports aggregate scalar metric values across all 9,000 cells.

Metric	Value
ROUGE-1	0.3526
ROUGE-2	0.2390
pillutla2021mauvemeasuringgapneural ROUGE-L	0.2735
ROUGE-Lsum	0.2735
Perplexity	3.892
MAUVE (global)	0.0041

Table 4: Aggregate scalar text-quality metrics across all 9,000 evaluation cells. MAUVE is a single corpus-level scalar computed over all generated texts jointly (see Section 3.4).

The scalar metrics suggest that the fine-tuned model produced fluent and passage-grounded items. ROUGE-1 (0.353), ROUGE-2 (0.239), and ROUGE-L/ROUGE-Lsum (0.274) indicate moderate lexical overlap with the source passages, consistent with the objective of generating evidence-aligned items. Perplexity was low (3.89), suggesting fluent, in-distribution outputs. In contrast, the corpus-level MAUVE score was very low (0.004), reflecting the genre mismatch between structured JSON item outputs and unstructured expository source passages rather than necessarily indicating poor generation quality.

4.2 G-theory Analysis

We report the G-theory results independently for each of the five binary metrics. For each metric we present: (1) the estimated G-study variance components and their percentage contributions to total variance, (2) D-study estimates of the G-coefficient $E\hat{\rho}^2$ and $\hat{\Phi}$ -coefficient under the observed design, using the same sample sizes as the G-study ($n_i = 2$, $n_s = 3$, $n_t = 3$), and under alternative sample-size configurations for the generalization facets.

Variance components. Table 5 reports the G-study variance component estimates for each source of variation, with the corresponding percentage of total variance shown in parentheses.

Three findings emerge consistently across the five metrics. First, the passage facet (p) accounts for the largest share of total variance, ranging from 76.54% (ECA) to 80.88% (ACS), indicating that item quality is driven primarily by source-passage characteristics rather than arbitrary generation conditions. Second, the item-within-passage component $i : (p \times s \times t)$ is the largest noise source, con-

Source	EIP	ECA	GIO	ACS	CLM
p (passage)	0.1238 (78.74%)	0.1231 (76.54%)	0.1935 (79.60%)	0.1951 (80.88%)	0.1904 (79.55%)
s (seed)	0.0000 (0.02%)	0.0000 (0.01%)	0.0000 (0.00%)	0.0000 (0.00%)	0.0000 (0.01%)
t (template)	0.0000 (0.01%)	0.0000 (0.00%)	0.0000 (0.00%)	0.0000 (0.00%)	0.0000 (0.00%)
$p \times s$	0.0038 (2.40%)	0.0044 (2.72%)	0.0075 (3.10%)	0.0072 (2.98%)	0.0073 (3.04%)
$p \times t$	0.0000 (0.00%)	0.0000 (0.00%)	0.0000 (0.00%)	0.0000 (0.00%)	0.0000 (0.00%)
$s \times t$	0.0000 (0.00%)	0.0000 (0.00%)	0.0000 (0.00%)	0.0000 (0.01%)	0.0000 (0.01%)
$p \times s \times t$	0.0000 (0.00%)	0.0000 (0.00%)	0.0000 (0.00%)	0.0000 (0.00%)	0.0000 (0.00%)
$i : (p \times s \times t)$	0.0296 (18.83%)	0.0333 (20.73%)	0.0421 (17.30%)	0.0389 (16.13%)	0.0416 (17.40%)

Table 5: G-study variance component estimates, with percentage contributions to total variance shown in parentheses, under the $i : (p \times s \times t)$ design. Values smaller than 1×10^{-4} are displayed as 0.0000. EIP = evidence_in_passage; ECA = evidence_contains_answer; GIO = gold_in_options; ACS = answer_correct_soft; CLM = correct_label_match.

tributing 6.13% (ACS) to 20.73% (ECA) of total variance, suggesting that generated items can still vary meaningfully even under the same passage, seed, and template condition as well as residual unexplained variation. Third, seed (s) and template (t) main effects are negligible, accounting for no more than 0.02% of total variance, while the $p \times s$ interaction contributes a small but nontrivial 2.40%–3.10%. Thus, prompt choice has little practical effect for this fine-tuned model, whereas some passages show modest sensitivity to decoding seed.

G-coefficients $E\hat{\rho}^2$ and $\hat{\Phi}$ -coefficients. Table 6 reports reliability coefficients for each metric under the observed design.

Metric	$E\hat{\rho}^2$	$\hat{\Phi}$
EIP	0.977	0.977
ECA	0.974	0.974
GIO	0.976	0.976
ACS	0.977	0.977
CLM	0.976	0.976
Mean	0.976	0.976

Table 6: G-coefficient ($E\hat{\rho}^2$) and Φ -coefficient ($\hat{\Phi}$) under the observed design ($n_i = 2$, $n_s = 3$, $n_t = 3$). Metric abbreviations as in Table 5.

All five metrics yield G-coefficients in the narrow range $[0.974, 0.977]$, with a mean of 0.976. These values substantially exceed the conventional 0.90 threshold for individual-level decisions (Shavelson and Webb, 1989), indicating that the generator’s quality estimates are highly generalizable across the universe of admissible seeds and prompt templates.

Crucially, $E\hat{\rho}^2$ and $\hat{\Phi}$ are virtually identical for all five metrics (differing by at most 0.0001). This near-equality arises because the seed and template main-effect variance components are negligible

(Table 5), meaning the distinction between relative error (norm-referenced) and absolute error (criterion-referenced) is inconsequential in this application. Whether a practitioner’s goal is to rank passages by generability or to certify that a passage meets an absolute quality standard, the same reliability coefficient applies.

D-study: effect of seeds and templates. Table 7 provides the corresponding D-study variance components, relative and absolute error variances, and reliability coefficients for the observed design ($n_S = 3$, $n_T = 3$, $n_I = 2$). Figure 2 shows the mean G-coefficient across the five metrics as a function of $n'_s \in \{1, \dots, 6\}$ seeds and $n'_t \in \{1, \dots, 5\}$ prompt templates, with $n_i = 2$ items per cell held fixed at the observed design value.

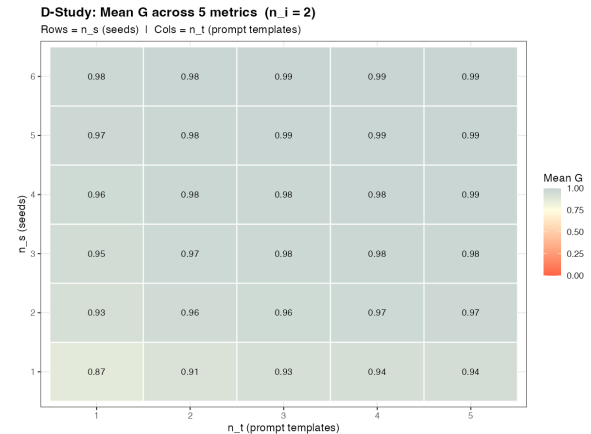


Figure 2: D-study heatmap of mean G-coefficient across five binary metrics as a function of number of seeds (n'_s , rows) and prompt templates (n'_t , columns), with $n_i = 2$ fixed. Even the minimal design ($n'_s = 1$, $n'_t = 1$) yields mean $E\hat{\rho}^2 = 0.87$, already above the conventional 0.80 group-decision threshold (Shavelson and Webb, 1989).

The heatmap reveals a striking pattern: reliability is uniformly high across the entire design

D-study	EIP	ECA	GIO	ACS	CLM
<i>variance components</i>					
p	0.12381	0.12307	0.19354	0.19514	0.19038
S	0.00001	0.00001	0.00000	0.00000	0.00000
T	0.00000	0.00000	0.00000	0.00000	0.00000
$p \times S$	0.00126	0.00146	0.00251	0.00240	0.00242
$p \times T$	0.00000	0.00000	0.00000	0.00000	0.00000
$S \times T$	0.00000	0.00000	0.00000	0.00000	0.00000
$p \times S \times T$	0.00000	0.00000	0.00000	0.00000	0.00000
$I : (p \times S \times T)$	0.00165	0.00185	0.00234	0.00216	0.00231
<i>Error variances and coefficients</i>					
$\hat{\sigma}_\delta^2$	0.00291	0.00331	0.00485	0.00456	0.00473
$\hat{\sigma}_\Delta^2$	0.00292	0.00332	0.00485	0.00456	0.00474
$E\hat{\rho}^2$	0.97709	0.97382	0.97558	0.97716	0.97574
$\hat{\Phi}$	0.97699	0.97377	0.97557	0.97715	0.97570

Table 7: D-study variance components, error variances, and reliability coefficients under the observed design ($n_s = 3$, $n_t = 3$, $n_i = 2$). Passages (p) are the objects of measurement. The generalization facets are random seed (S), prompt template (T), and item replication (I). D-study variance components are obtained by dividing the corresponding G-study variance components by the relevant D-study sample sizes. Metric abbreviations as in Table 5.

space. Even the most minimal possible design—a single seed and a single prompt template ($n'_s = 1$, $n'_t = 1$)—achieves mean $E\hat{\rho}^2 = 0.87$, already above the conventional 0.80 threshold for group-level decisions. Expanding to three seeds and three templates (the observed design) yields 0.98, and further increases in either dimension produce diminishing returns, plateauing at 0.99 for $n'_s \geq 4$ and $n'_t \geq 3$.

The marginal gain from adding a second template ($n'_t : 1 \rightarrow 2$) is smaller than the gain from adding a second seed ($n'_s : 1 \rightarrow 2$) at all seed levels, consistent with the slightly larger $p \times s$ interaction (2.40–3.10%) relative to the effectively zero $p \times t$ interaction. In practice, this means that if a deployment budget permits only two generation conditions per passage, two distinct prompt templates under the same seed yield marginally lower reliability than two seeds under the same template.

D-study: effect of items per cell. Figure 3 shows mean G and Φ as a function of $n'_i \in \{1, \dots, 5\}$, with $n_s = 3$ and $n_t = 3$ fixed.

The n'_i curve is essentially flat: mean $E\hat{\rho}^2$ and $\hat{\Phi}$ are both 0.96 at $n'_i = 1$ and increase by less than 0.02 across the full range to $n'_i = 5$. This result is explained by the structure of the relative error variance (Equation in Section 3.3): the item-within-passage-by-seed-by-template component, $\hat{\sigma}_{i:pst}^2$, is divided by $n_i \times n_s \times n_t$ in the denominator, so with $n_s = 3$ and $n_t = 3$ already averaging out substantial seed and template variance, the marginal

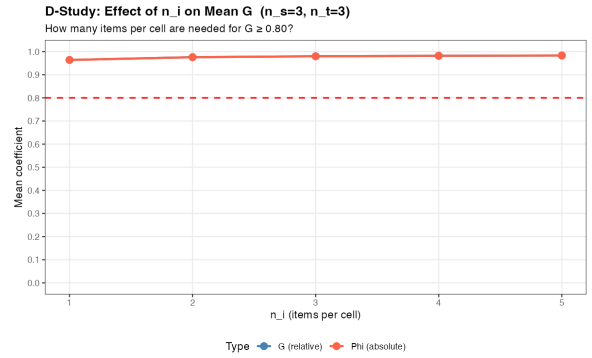


Figure 3: D-study curve showing mean G (relative, blue) and Φ (absolute, red) as a function of items per cell (n'_i), with $n_s = 3$ and $n_t = 3$ fixed. Both coefficients exceed 0.96 even with a single item per cell and are essentially flat across $n'_i \in \{1, \dots, 5\}$, indicating that item replication contributes negligible additional reliability beyond the seed and template design.

contribution of additional item replicates to error reduction is minimal.

Taken together, the D-study results indicate that the dominant design parameter for this generator is n_s : increasing the number of seeds from one to three produces the largest single improvement in reliability (from 0.87 to 0.98 at $n_t = 3$). The observed design ($n_i = 2$, $n_s = 3$, $n_t = 3$) is thus over-specified with respect to items per cell relative to seeds and templates, and future evaluations could reduce n_i to 1 without meaningful loss of reliability, allowing to use the saved generation budget to include more seeds or more templates.

5 Discussion

The results establish three interlocking empirical facts: item-generation quality is primarily a passage-level property (Section 4.2), the fine-tuned generator is highly robust to variation in seed and prompt template, and reliability under even minimal generation conditions exceeds conventional thresholds for group-level decisions.

The G-theory results show that 77–81% of total observed variance across the five binary metrics is attributable to the passage facet p , indicating that item quality is primarily a stable passage-level property rather than a function of arbitrary generation conditions. In classical G-theory terms, large $\hat{\sigma}_p^2$ relative to total variance supports high generalizability because the generator can reliably rank passages by the quality of items produced from them (Brennan, 2001). High-generability passages tend to contain explicit, factually dense statements that map cleanly onto a question–answer structure, whereas lower-generability passages often distribute relevant information across multiple sentences or require implicit inference. This finding suggests that passage selection and pre-screening may be more effective than additional generator tuning for improving AIG quality. The estimated passage effects $\hat{\nu}_p$ could therefore be used in future work to train a passage-suitability classifier, while passages with large $p \times s$ deviations could be flagged for additional seed averaging before deployment.

The near-zero main effects of seed (s) and prompt template (t), which together accounted for no more than 0.02% of total variance across all five metrics, suggest that the fine-tuned model was largely robust to generation conditions. This contrasts with prior findings that LLM outputs in zero-shot or few-shot settings are often sensitive to surface-level prompt changes (Chakraborty et al., 2023; Brown et al., 2020). Within the G-theory framework, the near-zero $\hat{\sigma}_t^2$ indicates that the model generalized beyond the specific training template and internalized the item-generation task rather than the instruction wording. Similarly, the near-zero $\hat{\sigma}_s^2$ suggests that no decoding seed systematically produced higher-quality items across passages. However, the non-zero $p \times s$ interaction indicates that some passages were more sensitive to stochastic sampling than others, explaining why increasing the number of seeds improved reliability in the D-study.

Several limitations should be noted. First, all results are based on a single QLoRA fine-tuned Qwen2.5-7B-Instruct model, and different model families, model sizes, or training regimes may yield different variance structures. Second, the study uses only the SciQ corpus, so the passage-level variance pattern may not generalize to other domains, difficulty levels, or assessment contexts. Third, the G-theory analysis is limited to five binary metrics; future work should extend the framework to scalar and multivariate outcomes such as ROUGE, perplexity, and universe-score covariance. Fourth, the low MAUVE score likely reflects a genre mismatch between structured JSON items and expository passages rather than poor generation quality; a more appropriate comparison would use human-authored multiple-choice items as references. Finally, although high $E\hat{\rho}^2$ provides reliability evidence for stable quality estimates across generation conditions, it does not establish full validity. A stronger validity argument would require expert item review, student response data, think-aloud evidence, and alignment analyses between automated metrics and human judgments (Kane, 2013; Kreiter and Zaidi, 2020).

References

- Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. [Fitting linear mixed-effects models using lme4](#). *Journal of Statistical Software*, 67(1):1–48.
- Robert L. Brennan. 2001. *Generalizability Theory*. Springer, New York.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- Mohna Chakraborty, Adithya Kulkarni, and Qi Li. 2023. [Zero-shot approach to overcome perturbation sensitivity of prompts](#). *Preprint*, arXiv:2305.15689.
- Dandan Chen, Michael Hebert, and Joshua Wilson. 2022. [Examining human and automated ratings of elementary students’ writing quality: A multivariate generalizability theory application](#). *American Educational Research Journal*, 59(6):1122–1156.
- Stanley Chen, Douglas Beeferman, and Ronald Rosenfeld. 2001. Evaluation metrics for language models.

- Paul Christ, Torsten Munkelt, and Jörg M Haake. 2024. Generalized automatic item generation for graphical conceptual modeling tasks. In *CSEDU (1)*, pages 807–818.
- Lee J. Cronbach, Goldine C. Gleser, Harinder Nanda, and Nageswari Rajaratnam. 1972. *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*. John Wiley & Sons, New York.
- Quoc Phu Dang, Phat T. Tran-Truong, Duc-Ly Vu, Long S.T. Nguyen, Quynh T.N. Vo, and Tho Quan. 2026. Enhancing large language model performance for automatic zero-shot multiple-choice question answering via single-token logit prompting. *Computers and Education: Artificial Intelligence*, 10:100578.
- Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. LLM.int8(): 8-bit matrix multiplication for transformers at scale. *Advances in Neural Information Processing Systems*, 35.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Preprint*, arXiv:2305.14314.
- X. Dong, X. Zhang, Z. Li, Q. Hou, J. Xue, and X. Li. 2025. A literature review of research on question generation in education. *PeerJ Computer Science*, 11:e3203.
- Mark J. Gierl and Hollis Lai. 2012. The role of item models in automatic item generation. *International Journal of Testing*, 12(3):273–298.
- Hardy, Shashi Narayan, and Andreas Vlachos. 2019. Highres: Highlight-based reference-less evaluation of summarization. *Preprint*, arXiv:1906.01361.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. *International Conference on Learning Representations*.
- Michael T. Kane. 2013. Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1):1–73.
- Clarence Kreiter and Nikki Bibler Zaidi. 2020. Generalizability theory’s role in validity research: Innovative applications in health science education. *Health Professions Education*, 6(2):282–290.
- Bakir Kurdi, John Leo, Bijan Parsia, Ulrike Sattler, and Saad Al-Emari. 2020. A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, 30(1):121–204.
- Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81.
- Yixin Liu, Alexander R. Fabbri, Pengfei Liu, Yilun Zhao, Linyong Nan, Ruilin Han, Simeng Han, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir Radev. 2023. Revisiting the gold standard: Grounding summarization evaluation with robust human evaluation. *Preprint*, arXiv:2212.07981.
- Steven Moore, Huy A. Nguyen, Norman Bier, Tanvi Domadia, and John Stamper. 2022. Assessing the quality of student-generated short answer questions using GPT-3. In *Educating for a New Future: Making Sense of Technology-Enhanced Learning Adoption*, pages 243–257, Cham. Springer International Publishing.
- Alireza Naeiji, Aijun An, Heidar Davoudi, Marjan Delpisheh, and Muath Alzghool. 2023. Question generation using sequence-to-sequence model with semantic role labels. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2830–2842, Dubrovnik, Croatia. Association for Computational Linguistics.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.
- Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. Mauve: Measuring the gap between neural text and human text using divergence frontiers. *Preprint*, arXiv:2102.01454.
- Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.
- Richard J. Shavelson and Noreen M. Webb. 1989. *Generalizability Theory: A Primer*. Sage Publications, Newbury Park, CA.
- Patrick E. Shrout and Joseph L. Fleiss. 1979. Intra-class correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2):420–428.
- Yu Tian, Linh Huynh, Katerina Christhlf, Shubham Chakraborty, Micah Watanabe, Tracy Arner, and Danielle McNamara. 2026. Cognitively diverse multiple-choice question generation: A hybrid multi-agent framework with large language models. *Electronics*, 15(6).
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need. *Preprint*, arXiv:1706.03762.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. [Finetuned language models are zero-shot learners](#). *Preprint*, arXiv:2109.01652.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017. [Crowdsourcing multiple choice science questions](#). *Preprint*, arXiv:1707.06209.

Ghasem Zaefarian, Viacheslav Iurkov, and Mariia Koval. 2022. [Variance decomposition analysis: What is it and how to perform it – a complete guide for b2b researchers](#). *Industrial Marketing Management*, 107:315–322.

A Prompt Templates for Item Generation

This appendix presents the three prompt templates used to generate multiple-choice science items from passages. The templates were designed to be linguistically distinct while preserving the same task requirements and output structure.

A.1 Prompt Template 1

Create exactly ONE multiple-choice question based ONLY on the Passage.

Rules:

- 4 options labeled A, B, C, D
- Exactly ONE correct answer
- Provide a short rationale
- Provide ONE evidence sentence copied verbatim from the Passage that supports the correct answer
- Output MUST be valid JSON with keys: stem, options, answer, rationale, evidence
- No extra text outside JSON

A.2 Prompt Template 2

Using only the Passage below, write exactly one science multiple-choice question.

Requirements:

- Include 4 answer options labeled A, B, C, and D
- Only one option may be correct
- Include a brief rationale
- Include one sentence copied exactly from the Passage as evidence
- Return valid JSON only with these keys: stem, options, answer, rationale, evidence
- Do not include explanations outside the JSON object

A.3 Prompt Template 3

Read the Passage and generate one multiple-choice science item.

Constraints:

- Base the item only on the Passage
- Provide four choices labeled A, B, C, D
- Mark exactly one correct answer
- Give a short rationale for the correct answer
- Give one verbatim supporting sentence from the Passage
- Output valid JSON with keys exactly: stem, options, answer, rationale, evidence
- Output JSON only