# Adaptive Sequential Certification for LLM Evaluators: Why Standard Item Selection Hurts and How to Fix It

Anonymous ACL submission

## Abstract

Certifying large language models (LLMs) as automated evaluators is challenging because each additional test item consumes annotation budget. Although sequential testing can reduce this cost, current certification approaches usually process items in arbitrary order and therefore miss opportunities to prioritize informative observations. A natural solution is to borrow item selection from Computerized Adaptive Testing (CAT), which uses Item Response Theory (IRT) to pick maximally informative items. We show that this intuition is wrong: standard CAT item selection, which maximizes Fisher information, is *counterproductive* for Sequential Probability Ratio Testing (SPRT) because the two frameworks optimize opposing objectives. Fisher information peaks at maximum uncertainty ($P \approx 0.5$), whereas SPRT converges fastest under maximum certainty ($P$ near 0 or 1). We derive an SPRT-aware item selection criterion that maximizes expected evidence per observation and demonstrate that IRT-informed ordering achieves up to 86% total annotation savings and improves savings by 5–9 percentage points over random ordering across 29 LLM graders and three benchmarks. Monte Carlo simulations reveal that all ordering strategies, including random, exceed nominal error bounds at the exact decision boundary due to the fundamental non-i.i.d. nature of IRT data, yet empirical decision agreement on real data exceeds 96% because most test cases are far from the boundary.

## 1 Introduction

Automated evaluation with large language models (LLMs), whether for grading, (Sung et al., 2019; Cohn et al., 2024; Gwozdz and Both, 2025), translation quality (Zheng et al., 2023), or open-ended assessment (Kim et al., 2024), requires formal quality assurance before deployment. Sequential Probability Ratio Testing (SPRT) provides a principled framework: under i.i.d. Bernoulli sampling, it certifies or rejects an evaluator by sequentially accumulating evidence with mathematically bounded Type I and Type II error rates (Wald, 1945).

A key limitation of SPRT is that it treats all test items as equally informative. In practice, some items reveal more about evaluator quality than others. Computerized Adaptive Testing (CAT), the standard approach for efficient testing in psychometrics (Wainer et al., 2000), selects items that maximize Fisher information given the current ability estimate. Combining CAT with SPRT seems promising: use IRT to pick informative items, feed them into the sequential test, and reach a decision faster.

We show that this natural combination fails. Fisher information, the standard CAT criterion, is maximized when the probability of a correct response is $P \approx 0.5$, i.e., when the item outcome is maximally uncertain. But SPRT accumulates evidence fastest when item outcomes are maximally *certain*: if the true ability is far from the decision boundary, items whose outcomes are predictable provide the strongest evidence in each direction. The two criteria are diametrically opposed.

We resolve this tension by deriving an SPRT-aware item selection criterion: at each step, select the item that maximizes the magnitude of the expected log-likelihood-ratio drift $|E[\Delta\Lambda]|$ given the current ability estimate. This criterion selects items where the outcome is most predictable, accelerating convergence toward a decision boundary.

Using 29 LLM graders across three public benchmarks, we demonstrate:

- Standard Fisher-CAT item selection *reduces annotation savings by 4–9 percentage points* compared to random ordering, confirming the CAT–SPRT conflict.
- SPRT-aware evidence-maximizing selection and fixed difficulty ordering both *achieve 5–9 percentage point improvements over random*, with up to 86% annotation savings.
- Monte Carlo simulations reveal that boundary error rates *exceed nominal bounds for all strategies* (including random) due to the non-i.i.d. nature of IRT data, yet empirical decision agreement on real data exceeds 96%.
- Cross-dataset ability transfer retains 72–101% of the ordering improvement, demonstrating that *adaptive certification generalizes across benchmarks.*

The paper is structured as follows. Section 2 reviews SPRT, IRT, and CAT. Section 3 derives the evidence-maximizing criterion and presents the CAT-SPRT algorithm. Section 4 describes the experimental setup. Section 5 presents the four-strategy comparison, error rate verification, and transfer experiments. Section 6 positions our contribution in the literature.

## 2 Background

### 2.1 Sequential Probability Ratio Testing (SPRT)

The Sequential Probability Ratio Test (Wald, 1945) evaluates evidence item by item, reaching a decision as soon as sufficient evidence accumulates. Applied to LLM evaluation, SPRT tests whether a model's agreement rate with a reference standard exceeds a quality threshold $\pi_0$.

Formally, SPRT tests $H_0 : p \leq p_0$ against $H_1 : p \geq p_1$, where $p_0 < \pi_0 < p_1$ define an indifference zone. After observing binary agreement outcome $x_i \in \{0, 1\}$, the log-likelihood ratio is updated:

$$\Lambda_n = \sum_{i=1}^{n} x_i \log \frac{p_1}{p_0} + (1 - x_i) \log \frac{1 - p_1}{1 - p_0} \quad (1)$$

Under the i.i.d. Bernoulli model, the test certifies when $\Lambda_n \geq B = \log \frac{1-\beta}{\alpha}$ and rejects when $\Lambda_n \leq A = \log \frac{\beta}{1-\alpha}$, bounding Type I error at $\alpha$ and Type II error at $\beta$.

Throughout the paper, we apply tiered certification at three quality levels (**Gold**, **Silver**, **Bronze**); the per-tier indifference bands $(p_0, p_1)$ are defined in Section 4.3.

### 2.2 Item Response Theory and the Rasch Model

Item Response Theory models the probability of a correct response as a function of **person ability $\boldsymbol{\theta}$** and **item difficulty $\boldsymbol{\beta}$** (Lord and Novick, 1968). The Rasch (1PL) model defines (Rasch, 1960):

$$P(X_{ij} = 1 \mid \theta_j, \beta_i) = \sigma(\theta_j - \beta_i) \quad (2)$$

where **$\boldsymbol{\sigma(\cdot)}$ is the logistic function**. The raw proportion correct is a sufficient statistic for $\theta$ (Fischer and Molenaar, 1995).

We operationalize IRT parameters empirically: a model's ability is its mean agreement rate across items (mapped to the logit scale), and an item's difficulty is the negative logit of the cross-model agreement rate. With 29 models, full parametric estimation produces degenerate parameter estimates (de Ayala, 2009; Cong et al., 2026); the sufficient statistic approach avoids this instability.

### 2.3 Computerized Adaptive Testing (CAT)

In Computerized Adaptive Testing (Wainer et al., 2000), items are selected to maximize the Fisher information about ability:

$$I(\theta; \beta_i) = P_i(1 - P_i) \quad (3)$$

where $P_i = \sigma(\theta - \beta_i)$. Fisher information is maximized when $\theta = \beta_i$ (i.e., $P_i = 0.5$): items are most informative for ability estimation when the examinee has a 50% chance of a correct response.

This criterion is optimal for *ability estimation* but, as we demonstrate, counterproductive for *sequential hypothesis testing.*

## 3 Method: Evidence-Maximizing Item Selection

### 3.1 The CAT–SPRT Conflict

Consider a single observation in the SPRT. The log-likelihood ratio increment is:

$$\Delta\Lambda = x \cdot \underbrace{\log \frac{p_1}{p_0}}_{a>0} + (1-x) \cdot \underbrace{\log \frac{1-p_1}{1-p_0}}_{b<0} \quad (4)$$

For an item with Rasch probability $P = \sigma(\theta - \beta)$, the expected increment is:

$$E[\Delta\Lambda] = P \cdot a + (1-P) \cdot b \quad (5)$$

This is zero at $P^* = \frac{|b|}{a+|b|}$, the response probability at which the expected log-likelihood-ratio increment has zero drift. For the symmetric 0.10-wide bands used here, $P^*$ is close to the midpoint of the indifference interval $(p_0 + p_1)/2$.

The **expected evidence**, the absolute expected increment, is:

$$|E[\Delta\Lambda]| = |P \cdot a + (1-P) \cdot b| \quad (6)$$

This quantity is maximized when $P$ is far from $P^*$, i.e., when the item outcome is most *predictable*.

The conflict is now clear (Figure 1):

- **Fisher information** (Eq. 3) is maximized at $P = 0.5$: maximum uncertainty.
- **Expected evidence** (Eq. 6) is maximized when $|P - P^*|$ is large: maximum certainty.

These criteria are diametrically opposed. Selecting items at $P = 0.5$ yields $|E[\Delta\Lambda]|$ near zero when $P^*$ is near 0.5 (as for Silver and Bronze tiers, cf. Section 4.3), producing the slowest possible SPRT convergence.

### 3.2 SPRT-Aware Adaptive Selection

We propose selecting items to maximize $|E[\Delta\Lambda]|$ at each step, using the current ability estimate $\hat{\theta}$ to compute Rasch probabilities. Algorithm 1 presents the full procedure.

The ability estimate $\hat{\theta}$ is updated as the logit of the running agreement rate (clipped to $[0.01, 0.99]$), with a prior of $\hat{\theta} = 0$ for the first three observations.

Item difficulties $\beta_i$ are computed using leave-one-model-out cross-validation: for each model under test, the difficulty of each item is estimated from the agreement rates of the remaining $N-1$ models, avoiding data leakage.

**Algorithm 1** CAT-SPRT: Evidence-maximizing adaptive certification

**Require:** Items $\{1, \ldots, N\}$ with difficulties $\beta_1, \ldots, \beta_N$
**Require:** SPRT parameters $p_0, p_1, \alpha, \beta$, truncation $n_{\max}$, minimum $n_{\min}$

```
Λ ← 0; θ̂ ← 0; R ← {1, ..., N}
a ← log(p1/p0); b ← log((1 − p1)/(1 − p0))
A ← log(β/(1 − α)); B ← log((1 − β)/α)
for t = 1, ..., n_max do
    i* ← arg max_{i∈R} |σ(θ̂ − β_i)·a + (1 − σ(θ̂ − β_i))·b|
    Observe x_{i*} ∈ {0, 1}
    Λ ← Λ + x_{i*}·a + (1 − x_{i*})·b
    R ← R \ {i*}
    θ̂ ← logit(clip(x̄_t, 0.01, 0.99)) if t ≥ 3, else 0
    if t ≥ n_min and Λ ≥ B then
        return CERTIFY at step t
    else if t ≥ n_min and Λ ≤ A then
        return REJECT at step t
    end if
end for
return CERTIFY if Λ > (A + B)/2, else REJECT
```

### 3.3 Fixed Difficulty Ordering as Simpler Alternative

When difficulty estimates are reliable, a simpler strategy is available: sort all items by difficulty before running SPRT, without adaptive re-selection. Items from harder questions are presented first.

This fixed ordering is equivalent to CAT-SPRT when the ability estimate is stable: both strategies select items where $|P - P^*|$ is maximized. The adaptive approach offers potential advantages when difficulty estimates are noisy (e.g., transferred from another dataset) or when the model's true ability is initially misestimated.

## 4 Experimental Setup

### 4.1 Datasets

We evaluate on three established ASAG benchmarks:

- **ASAP-SAS** (The Hewlett Foundation, 2012): 10 prompts, ordinal labels, within-tolerance agreement ($\pm 1$). 348,261 grading instances.

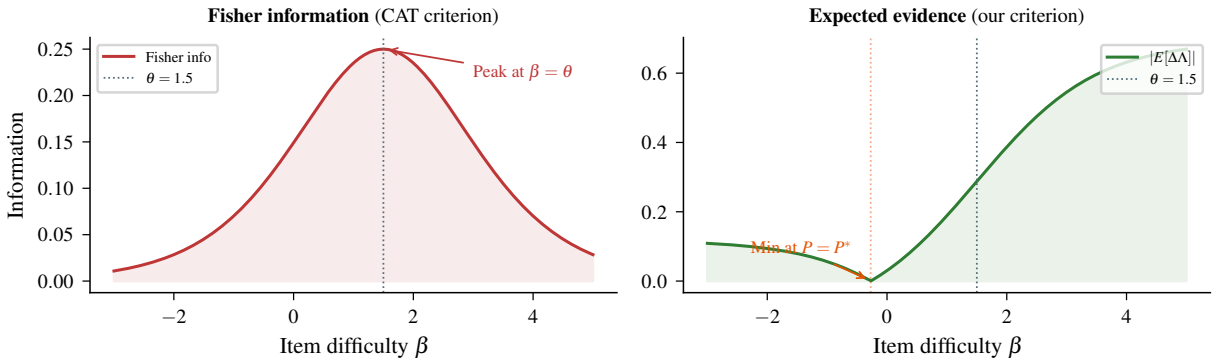


Figure 1: The CAT–SPRT conflict. **Left:** Fisher information (Eq. 3) peaks at $\beta = \theta$ (maximum uncertainty, $P = 0.5$), making it optimal for ability estimation. **Right:** Expected SPRT evidence (Eq. 6) peaks when the item's response probability $P$ is far from the indifference rate $P^*$ (maximum certainty), making it optimal for hypothesis testing. Selecting items by Fisher information *minimizes* SPRT convergence speed.

- **Mohler** (Mohler et al., 2011): 87 questions, ordinal labels (0–5), within-tolerance agreement (±1). 70,818 instances.
- **SciEntsBank** (Dzikovska et al., 2013): 196 questions, 5-way categorical labels, exact-match agreement. 313,316 instances.

Together, the three datasets span 293 questions with 732,395 total grading instances across 29 models.

### 4.2 Models

We evaluate 29 instruction-tuned LLMs spanning 10 model families (Qwen, Gemma, Mistral, Llama, Falcon, DeepSeek, OLMo, IBM Granite, Phi, Yi), with sizes from 4B to 72B parameters. The distribution is skewed toward Qwen variants ($n = 13$); Section 5.5 reports a family-balanced robustness check. All models were prompted with zero-shot rubric-based grading instructions following the protocol of Gwozdz and Both (2025).[1]

### 4.3 Certification Configuration

A certification unit is a (model, question) pair; within each unit, the sequential observations are individual student-response grading instances, which we refer to as items in the SPRT/CAT procedure. All experiments use $\alpha = 0.05$, $\beta = 0.10$, $n_{\max} = 200$, and minimum 15 observations per unit. For random ordering, we average over 20 shuffles and take the majority decision; ties are broken conservatively as rejection. Tier thresholds: **Gold** ($p_0 = 0.80$, $p_1 = 0.90$), **Silver** ($p_0 = 0.65$, $p_1 = 0.75$), and **Bronze** ($p_0 = 0.50$, $p_1 = 0.60$). Each pair is assigned to its highest certifying tier. The tier midpoints (0.85, 0.70, 0.55) reflect established agreement standards in educational measurement, where 80% raw agreement is a common minimum for acceptable inter-rater reliability and 90% approaches the human–human consistency ceiling reported for rubric-based scoring (Stemler, 2004). The uniform indifference zone width of 0.10 ($\delta = 0.05$ half-width) is the minimum for which SPRT maintains nominal error control under Bernoulli data (Wald, 1945); narrower zones ($\delta = 0.025$) cause Type II inflation beyond 10 pp even under ideal conditions, as confirmed by our Monte Carlo analysis (Section 5.3). The three equally spaced tiers map onto natural deployment modes: fully automated (**Gold**), automated with selective human review (**Silver**), and automated with systematic oversight (**Bronze**).

### 4.4 Compared Strategies

We compare four item ordering strategies:

1. Random: Uniform random ordering (baseline).
2. Difficulty-ordered: Fixed hardest-first ordering using LOO-CV difficulty estimates.
3. Fisher-CAT: Adaptive selection maximizing Fisher information $I(\hat{\theta}; \beta_i) = P_i(1 - P_i)$ (standard CAT criterion).

[1] Data and code: `https://anonymous.4open.science/r/irt-sprt-bridge`

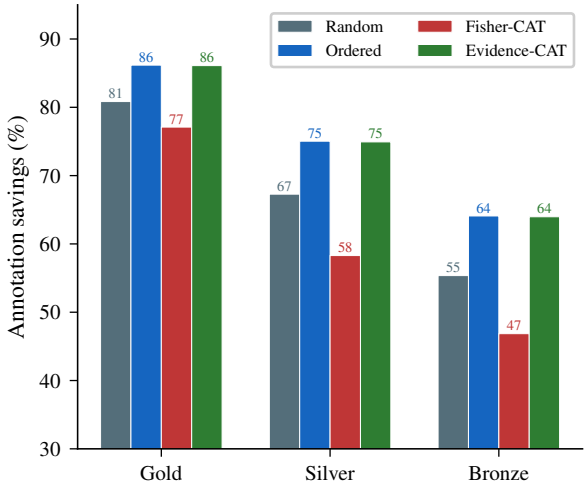


Figure 2: Annotation savings by item selection strategy and certification tier. Fisher-CAT (standard CAT) performs *worse* than random; Evidence-CAT (our criterion) matches fixed difficulty ordering.

| Tier | Strategy | Savings (%) | Agreement (%) |
|---|---|---|---|
| Gold | Random | 80.8 | 99.5 |
| | Ordered | 86.2 | 97.0 |
| | Fisher-CAT | 77.1 | 97.1 |
| | Evidence-CAT | 86.1 | 97.2 |
| Silver | Random | 67.3 | 99.1 |
| | Ordered | 75.0 | 96.7 |
| | Fisher-CAT | 58.3 | 97.3 |
| | Evidence-CAT | 74.9 | 97.0 |
| Bronze | Random | 55.4 | 99.0 |
| | Ordered | 64.1 | 96.6 |
| | Fisher-CAT | 46.9 | 97.3 |
| | Evidence-CAT | 64.0 | 97.4 |

Table 1: Annotation savings and decision agreement with full-sample SPRT across 8,497 (model, question) pairs × 3 tiers. Fisher-CAT is worse than random ordering; Evidence-CAT matches the simpler difficulty-ordered approach.

4. Evidence-CAT: Adaptive selection maximizing $|E[\Delta\Lambda]|$ (CAT-SPRT, our proposed criterion).

For each (model, question, tier) triple, the reference decision is defined by applying the likelihood-ratio rule to all available observations without early stopping. We report decision agreement with this reference and the annotation savings (reduction in test length relative to the number of available observations).

## 5 Results

### 5.1 Four-Strategy Comparison

Figure 2 and Table 1 present the main comparison across 25,491 (model, question, tier) triples.

| Strategy | ASAP-SAS | Mohler | SciEntsBank |
|---|---|---|---|
| Random | 96.6 | 59.1 | 89.7 |
| Ordered | 99.0 | 73.6 | 91.1 |
| Fisher-CAT | 98.8 | 58.8 | 84.1 |
| Evidence-CAT | 99.0 | 73.5 | 91.0 |
| $n$ (pairs) | 290 | 2,523 | 5,684 |

Table 2: Gold-tier annotation savings (%) per dataset. The Fisher penalty is largest on Mohler (−14.7 pp vs. Evidence-CAT) and SciEntsBank (−7.0 pp).

The results confirm the theoretical analysis:

**Fisher-CAT hurts.** Standard CAT item selection reduces savings by 3.7–9.0 percentage points compared to random ordering across all tiers. The effect is strongest at Bronze (−8.5 pp), where $P^* \approx 0.55$ is closest to the Fisher optimum of 0.5.

**Evidence-CAT matches fixed ordering.** The SPRT-aware criterion recovers the full improvement of difficulty-ordered SPRT, achieving 86.1%, 74.9%, and 64.0% savings at Gold, Silver, and Bronze respectively, within 0.1 pp of fixed ordering.

**IRT-informed ordering yields large savings.** Both difficulty-ordered and Evidence-CAT achieve 5–9 pp more savings than random, with up to 86% total savings at Gold.

**Decision agreement.** All non-random strategies show slightly lower agreement with full-sample decisions (96.6–97.4% vs. 99.0–99.5%), reflecting the conservative bias of deterministic ordering versus averaged random shuffles. Table 2 shows per-dataset Gold-tier results.

### 5.2 Per-Dataset Analysis

The effect size varies substantially across datasets. On ASAP-SAS (diverse multi-point rubrics), Evidence-CAT achieves 99.0% savings at Gold, nearly eliminating the annotation cost. On Mohler (moderate diversity, 87 questions), Evidence-CAT achieves 73.5% at Gold. On SciEntsBank (homogeneous 5-way labels, exact match), savings reach 91.1% at Gold. The Fisher-CAT penalty is largest on Mohler: 58.8% vs. 73.5% for Evidence-CAT at Gold (−14.7 pp).

Difficulty-informed ordering is most beneficial when the item pool is heterogeneous,

| Tier | Strategy | Type I | Type II |
|---|---|---|---|
| Gold | Random | 0.008 | 0.440 |
| | Ordered | 0.000 | 0.882 |
| | Evidence-CAT | 0.012 | 0.808 |
| Silver | Random | 0.040 | 0.434 |
| | Ordered | 0.000 | 0.952 |
| | Evidence-CAT | 0.000 | 0.874 |
| Bronze | Random | 0.162 | 0.220 |
| | Ordered | 0.004 | 0.912 |
| | Evidence-CAT | 0.078 | 0.664 |
| Nominal bounds | | $\alpha = 0.05$ | $\beta = 0.10$ |

Table 3: Boundary error rates from Monte Carlo simulation ($N = 500$ per condition, Rasch model with $\sigma_\beta = 1.5$). False-certification rates are less inflated than false-rejection rates but can exceed nominal $\alpha$ for Bronze; Type II rates substantially exceed $\beta$ for *all* strategies, including random: a consequence of the non-i.i.d. violation inherent to IRT data, not of the ordering strategy.

which is typical of multi-rubric deployments with varying student populations.

### 5.3 Error Rate Verification

Since non-random item ordering violates SPRT's i.i.d. assumption, we verify error rates via Monte Carlo simulation. We test at the SPRT boundary rates ($p = p_0$ and $p = p_1$) where error rates are highest. Table 3 reports empirical Type I error (false certification at $p = p_0$) and Type II error (false rejection at $p = p_1$) from $N = 500$ synthetic Rasch datasets per condition, each with 100 items whose difficulties are drawn from a normal distribution with mean 0 and standard deviation $\sigma_\beta = 1.5$, matching the empirical difficulty variance in our datasets. We omit Fisher-CAT from this simulation because Table 1 already shows it is strictly dominated by random ordering; characterizing its boundary error properties would not inform the practical choice among competitive strategies. Appendix B replicates the simulation with 200 items, showing that Type I rates tighten further while Type II inflation at the upper boundary intensifies, ruling out truncation as a confound.

The results reveal a key insight: **false-certification rates are substantially less inflated than false-rejection rates** for Gold and Silver ($\leq 0.08$), though Bronze still exceeds the nominal $\alpha = 0.05$ under some strategies ($\leq 0.16$). The dominant boundary failure mode is Type II inflation. However, **Type II error rates substantially exceed the nominal** $\beta = 0.10$ for all strategies, including random. This is not an artifact of ordering but a fundamental consequence of applying SPRT to non-i.i.d. IRT data: when item responses follow a Rasch model with heterogeneous difficulties, observations are drawn from different Bernoulli distributions, violating the homogeneity assumption that underlies Wald's error guarantees.

Non-random ordering amplifies this effect: difficulty-ordered and Evidence-CAT strategies front-load items with extreme response probabilities, causing faster but more error-prone decisions at the boundary. In practice, this boundary behavior is rarely encountered: our empirical results (Table 1) show 96–99% decision agreement because the vast majority of (model, question) pairs have true agreement rates well above or below the tier thresholds. Deployments concerned about borderline cases, where the true agreement rate is within $\pm 0.02$ of a tier threshold, should fall back to random ordering or full-sample evaluation; IRT-informed ordering is best reserved for the bulk of cases that are far from a decision boundary.

| Transfer | Random | Own | Warm |
|---|---|---|---|
| ASAP → Mohler | 39.8 | 53.4 | 49.6 |
| ASAP → SciEnts. | 78.8 | 83.5 | 82.2 |
| Mohler → ASAP | 97.4 | 98.7 | 98.7 |
| Mohler → SciEnts. | 78.8 | 83.5 | 83.5 |
| SciEnts. → ASAP | 97.4 | 98.7 | 98.7 |
| SciEnts. → Mohler | 39.8 | 53.4 | 53.5 |

Table 4: Cross-dataset transfer: annotation savings (%) with random ordering, own-difficulty Evidence-CAT, and warm-started Evidence-CAT (ability transferred from source dataset). Warm-start retains 72–101% of the improvement over random.

### 5.4 Cross-Dataset Ability Transfer

To assess whether model-level ability estimates transfer across benchmarks, we estimate each model's ability on a source dataset ($A$) and use it to warm-start Evidence-CAT when certifying on a target dataset ($B$). Target-dataset item difficulties are still estimated on $B$ using the same leave-one-model-out procedure. Table 4 reports the annotation savings retained by this warm start relative to the cold-start default ($\hat{\theta} = 0$).

Warm-started Evidence-CAT retains 72–

101% of the improvement over random, with a mean retention of 90% across all transfer pairs. The transfer is most effective when the source dataset provides an accurate model ability estimate (e.g., Mohler → SciEntsBank: 100% retention). Even the weakest transfer (ASAP-SAS → Mohler) retains 72% of savings, indicating that cross-dataset ability transfer recovers more than two thirds of the in-domain ordering benefit when in-domain data is limited.

### 5.5 Practical Pre-Screening Heuristic

As a practical complement, we examine whether a lightweight ability estimate can predict certification outcomes before running SPRT. The Spearman correlation between mean agreement rate (Rasch ability) and mean certification tier is $\rho = 0.93$ ($p < 10^{-13}$).

To verify this is not tautological, we test pilot-based prediction: ability estimated from only 25% of responses per question predicts the full certification tier with $\rho = 0.975$ (bootstrap 95% confidence interval: [0.92, 0.99]), indicating that a quarter of the annotation budget suffices to rank models reliably. To control for model-family clustering (13 of 29 models are Qwen variants), we repeat with one model per family ($k = 10$, 1,000 bootstraps): mean $\rho = 0.912$ (95% confidence interval: [0.83, 0.98]), with 98.9% of iterations exceeding $\rho > 0.8$.

## 6 Related Work

**LLM grading and LLM-as-judge evaluation.** Recent work has evaluated LLMs as graders across benchmarks (Sung et al., 2019; Filighera et al., 2022; Cohn et al., 2024; Gwozdz and Both, 2025), focusing on aggregate metrics with less attention to per-question certification. More broadly, LLM-as-judge frameworks such as MT-Bench (Zheng et al., 2023) and Prometheus (Kim et al., 2024) have established that LLMs can serve as scalable evaluators, yet their reliability analyses remain at the benchmark level rather than certifying individual model–question interactions as we do here.

**IRT for LLM and NLP evaluation.** Lalor et al. (2019) showed that IRT item parameters can be estimated from artificial crowds of neural models without human response data, establishing that machine-generated response patterns carry meaningful psychometric signal, a premise our cross-model profiling shares. Vania et al. (2021) applied IRT to compare 29 NLP benchmarks, revealing that datasets differ substantially in their ability to discriminate among models, which parallels our finding that difficulty-heterogeneous datasets benefit most from IRT-informed ordering. Rodriguez et al. (2021) showed that NLP evaluation examples are not equally informative, and Byrd and Srivastava (2022) demonstrated that IRT difficulty and discrimination of natural language questions can be predicted from textual features. Cong et al. (2026) applied IRT to LLM grading ability estimation, and Martínez-Plumed et al. (2019) used IRT for general AI evaluation. Our work extends this lineage from descriptive analysis to a concrete decision problem: sequential certification of evaluator quality.

**Difficulty-aware data characterization.** Swayamdipta et al. (2020) stratified examples by confidence and variability. Their difficulty signal parallels our cross-model difficulty but targets training data selection rather than evaluation efficiency.

**Efficient evaluation and adaptive testing.** Active learning (Settles, 2012) and Computerized Adaptive Testing (Wainer et al., 2000) are established paradigms for reducing evaluation costs. Zhuang et al. (2023) applied CAT to LLM evaluation, showing that adaptive item selection can recover ability estimates with only 20% of the full benchmark. Polo et al. (2024) used IRT to select small evaluation subsets that approximate full-benchmark scores, and Li et al. (2025) extended this line with learned acquisition policies. These works use CAT for *ability estimation*, where Fisher information is the correct criterion. Our central finding is that for *sequential hypothesis testing*, a fundamentally different objective, the Fisher criterion is counterproductive. Fisher information maximizes at $P = 0.5$ (maximum uncertainty), whereas SPRT evidence maximizes at extreme $P$ values (maximum certainty). This distinction has not been previously noted in the NLP evaluation literature, and any future work combining adaptive item selection with sequential testing must use an evidence-based

criterion rather than Fisher information.

## 7 Discussion

### 7.1 Practical Implications

Our results suggest a two-stage deployment workflow:

**Stage 1: Pre-screening.** Compute the evaluator's mean agreement rate on a small pilot (25% of responses). The pilot-to-full correlation of $\rho = 0.975$ enables early rejection of weak models without running full certification.

**Stage 2: Efficient certification.** For surviving models, run SPRT with difficulty-ordered items. The simpler fixed-ordering approach matches adaptive CAT-SPRT when difficulty estimates are reliable, making it the recommended default. Adaptive selection is preferable when the initial ability estimate is transferred from another dataset or when target-domain difficulty estimates are noisy.

### 7.2 Equivalence of Evidence-CAT and Fixed Ordering under Stable Ability Estimates

Evidence-CAT achieves the same savings as fixed ordering because, with a stable ability estimate, both criteria select the same items: those with $|P - P^*|$ maximized. The adaptive mechanism provides value only when the initial ability estimate is transferred from another dataset or when target-domain difficulty estimates are noisy. This finding suggests that the primary value of IRT for sequential certification lies in the *difficulty estimates themselves*, not in the adaptive selection mechanism.

### 7.3 Implications for Adaptive Item Selection

The CAT–SPRT conflict has direct consequences for downstream system design. Adaptive testing pipelines targeting ability estimation (e.g., Zhuang et al., 2023; Polo et al., 2024) should continue to use Fisher information; pipelines built around sequential decisions (certification gates, online A/B testing, model rollout policies) should substitute the expected-evidence criterion of Section 3.1. A natural next step is to combine the two: use Fisher information to bootstrap a reliable ability estimate during an exploration phase, then switch to evidence-maximizing selection once the estimate stabilizes and a decision becomes actionable.

## 8 Conclusion

We investigated adaptive item selection for sequential certification of LLM-driven graders/evaluators and discovered a fundamental conflict: standard Computerized Adaptive Testing item selection, which maximizes Fisher information, is counterproductive for the Sequential Probability Ratio Test (SPRT) because the two frameworks optimize opposing objectives. We derived an SPRT-aware evidence criterion and demonstrated empirically that IRT-informed item ordering reduces annotation cost by up to 86% across 29 LLM graders and three benchmarks.

The main practical finding is that fixed difficulty ordering, a simpler strategy that does not require adaptive computation, achieves the same savings as the full adaptive algorithm when difficulty estimates are reliable. This reduces deployment to three steps: compute cross-model difficulty estimates, sort items, and run standard SPRT.

The main theoretical finding is the CAT–SPRT conflict itself. As the NLP community increasingly applies psychometric methods to LLM evaluation, distinguishing between the information criteria appropriate for estimation versus testing becomes critical.

## Limitations

**Rasch sufficient statistics vs. parametric IRT.** Our estimates do not separate item discrimination from difficulty. With larger model pools ($N > 50$), parametric models may yield more nuanced profiling.

**Non-random ordering and SPRT error guarantees.** Wald's sequential test assumes i.i.d. observations. Our Monte Carlo simulations empirically characterize the breakdown of nominal SPRT bounds under heterogeneous IRT data, but a formal finite-sample characterization remains future work.

**Model family imbalance.** Our 29-model sample includes 13 Qwen variants. The one-per-family robustness check ($\rho = 0.91$, 98.9%

of bootstraps $> 0.8$) confirms the main finding holds, but future work should evaluate on a more balanced pool.

**Dataset coverage.** Our evaluation covers three English-language ASAG benchmarks from 2011–2013. Generalizability to other languages, tasks, and modern benchmarks remains to be verified, although the theoretical analysis (Section 3.1) is task-independent.

**Benchmark memorization risk.** The three benchmarks used in this study (ASAP-SAS, Mohler, SciEntsBank) date from 2011–2013 and are widely cited in the NLP literature. Modern LLMs may have encountered these datasets, including reference answers and scoring rubrics, during pre-training. If benchmark memorization inflates agreement rates, the *absolute* savings and certification rates reported here may overestimate what would be observed on held-out data. However, the paper's central contribution—the comparison of ordering strategies—is less affected: all four strategies operate on the same data, so any memorization-induced inflation applies uniformly. The theoretical analysis of the CAT–SPRT conflict (Section 3.1) is entirely independent of the benchmark. Evaluating on contemporary or unpublished benchmarks would clarify the extent to which memorization affects absolute certification rates.

## Ethical Considerations

This study uses publicly available benchmark datasets with no personally identifiable information. All experiments are computational reanalyses of existing model predictions; no new data collection involving human participants was conducted. The certification framework is designed to increase human oversight of LLM evaluation, not to replace it.

## References


Matthew Byrd and Shashank Srivastava. 2022. Predicting difficulty and discrimination of natural language questions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL, Short Papers)*, pages 119–130.

Clayton Cohn, Nicole Hutchins, Tuan Le, and Gautam Biswas. 2024. A chain-of-thought prompting approach with LLMs for evaluating students' formative assessment responses in science. In *Artificial Intelligence in Education (AIED 2024)*.

Longwei Cong, Sonja Hahn, Sebastian Gombert, Leon Camus, Hendrik Drachsler, and Ulf Kroehne. 2026. Estimating LLM grading ability and response difficulty in automatic short answer grading via Item Response Theory. *arXiv preprint*.

R. J. de Ayala. 2009. *The Theory and Practice of Item Response Theory*. Guilford Press.

Myroslava O. Dzikovska, Rodney Nielsen, Chris Brew, Claudia Leacock, Danilo Giampiccolo, Luisa Bentivogli, Peter Clark, Ido Dagan, and Hoa Trang Dang. 2013. SemEval-2013 task 7: The joint student response analysis and 8th recognizing textual entailment challenge. In *Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 263–274.

Anna Filighera, Sebastian Ochs, Tim Steuer, and Thomas Tregel. 2022. Your answer is incorrect... Would you like to know why? Introducing a bilingual short answer feedback dataset. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 8577–8591.

Gerhard H. Fischer and Ivo W. Molenaar. 1995. *Rasch Models: Foundations, Recent Developments, and Applications*. Springer.

Jonas Gwozdz and Andreas Both. 2025. Automatic short answer grading using large language models: Comprehensive evaluation of prompt strategies for zero-shot and fine-tuned models. In *Proceedings of the 25th International Conference on Web Engineering (ICWE 2025)*.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. 2024. Prometheus: Inducing fine-grained evaluation capability in language models. In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR)*.

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 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP)*, pages 4249–4259.

Yang Li, Jie Ma, Miguel Ballesteros, Yassine Benajiba, and Graham Horwood. 2025. Active evaluation acquisition for efficient LLM benchmarking. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*.

Frederic M. Lord and Melvin R. Novick. 1968. *Statistical Theories of Mental Test Scores*. Addison-Wesley.

Fernando Martínez-Plumed, Ricardo B. C. Prudêncio, Adolfo Martínez-Usó, and José Hernández-Orallo. 2019. Item response theory in AI: Analysing machine learning classifiers at the instance level. *Artificial Intelligence*, 271:18–42.

Michael Mohler, Razvan Bunescu, and Rada Mihalcea. 2011. Learning to grade short answer questions using semantic similarity measures and dependency graph alignments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 752–762.

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 41st International Conference on Machine Learning (ICML)*.

Georg Rasch. 1960. *Probabilistic Models for Some Intelligence and Attainment Tests*. Danish Institute for Educational Research, Copenhagen.

Pedro Rodriguez, Joe Barrow, Alexander Miserlis 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 59th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Burr Settles. 2012. *Active Learning*. Morgan & Claypool Publishers.

Steven E. Stemler. 2004. A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research, and Evaluation*, 9(4).

Chul Sung, Tejas I. Dhamecha, Swarnadeep Saha, Tengfei Ma, Vinay Reddy, and Rishi Arora. 2019. Pre-training BERT on domain resources for short answer grading. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

The Hewlett Foundation. 2012. Automated student assessment prize: Short answer scoring. Kaggle Competition. `https://www.kaggle.com/c/asap-sas`.

Clara Vania, Phu Mon Htut, William Huang, Dhara Mungra, Richard Yuanzhe Pang, Jason Phang, Haokun Liu, Kyunghyun Cho, and Samuel R. Bowman. 2021. Comparing test sets with Item Response Theory. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1141–1158.

Howard Wainer, Neil J. Dorans, Ronald Flaugher, Bert F. Green, and Robert J. Mislevy. 2000. *Computerized Adaptive Testing: A Primer*, 2nd edition. Lawrence Erlbaum Associates.

Abraham Wald. 1945. Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16(2):117–186.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Yan Zhuang, Qi Liu, Yuting Ning, Weizhe Huang, Rui Lv, Zhenya Huang, Guanhao Zhao, Zheng Zhang, Qingyang Mao, Shijin Wang, and Enhong Chen. 2023. Efficiently measuring the cognitive ability of LLMs: An adaptive testing perspective. *arXiv preprint*. ArXiv:2306.10512.

## A Sensitivity to Agreement Definition

The main analysis uses the conventional agreement definitions for each benchmark: exact-match for SciEntsBank (5-way categorical labels) and within-tolerance $\pm 1$ for the ordinal benchmarks (ASAP-SAS, Mohler). A natural question is whether the results depend on these choices. We replicate the certification analysis with looser tolerances: SciEntsBank with $\pm 1$ ordinal tolerance and Mohler with $\pm 2$ tolerance, leaving ASAP-SAS unchanged (cf. Table 5).

Two observations follow. First, the absolute tier counts depend strongly on the agreement definition, a property of the definition itself, not of our method. Second, the ability–tier correlation $\rho$ (the pre-screening signal from Section 5.5) is preserved across the configurations, with the alternative SciEntsBank definition yielding a stronger correlation (0.92 vs. 0.70) due to better tier discrimination when exact match does not over-compress decisions toward rejection. The Mohler $\pm 2$ value of $\rho = 1.00$ should be read as a floor effect rather

| Config | Gold | Silver | Bronze | Reject | $\rho$ |
|---|---|---|---|---|---|
| SciEnts. exact (used) | 0 | 14 | 63 | 5,621 | 0.70 |
| SciEnts. $\pm 1$ (alt) | 106 | 390 | 1,201 | 4,483 | 0.92 |
| Mohler $\pm 1$ (used) | 257 | 598 | 930 | 1,593 | 0.99 |
| Mohler $\pm 2$ (alt) | 756 | 1,273 | 1,608 | 915 | 1.00 |

Table 5: Tier distribution and ability–tier Spearman correlation $\rho$ under alternative agreement definitions. Looser tolerances shift certification rates (as expected) but $\rho$ remains $\geq 0.70$ throughout, with the SciEntsBank $\pm 1$ variant yielding stronger discrimination than the exact-match baseline.

than as evidence of superior discrimination: under the $\pm 2$ tolerance, 5 of the 6 ordinal cells count as agreement, which saturates the agreement metric and trivializes the model ranking. The principal methodological claim, that IRT-informed item ordering accelerates SPRT certification, does not rely on a particular agreement choice.

## B Error Simulation: Robustness to Item Pool Size

The Monte Carlo simulation in Section 5.3 truncates each synthetic test at $n_{\max} = 100$ items to match typical per-question data availability. To verify the boundary error pattern is not an artifact of this truncation, we replicate the simulation with $n_{\max} = 200$ items per synthetic dataset, holding all other parameters fixed (500 datasets per condition, $\sigma_\beta = 1.5$, $\alpha = 0.05$, $\beta = 0.10$), cf. Table 6.

Two findings emerge. First, Type I error rates tighten further under longer truncation: false certification becomes even less likely with more items, as the test can accumulate additional rejecting evidence before the truncation cap. Second, Type II error rates at the upper decision boundary *worsen* for non-random orderings, approaching 1.0 for difficulty-ordered and Evidence-CAT strategies. The intuition is that difficulty-ordered SPRT front-loads items with low expected agreement, which for a high-ability examinee at the $p = p_1$ boundary lets the log-likelihood drift toward the rejection threshold before easier items are presented; doubling the truncation does not help once the test has already terminated incorrectly. This confirms that the boundary failure is a structural consequence of applying SPRT to non-i.i.d. IRT data, not an artifact of the $n_{\max}$ choice. In practice this boundary regime is rarely encountered: as discussed in Section 5.3, empirical decision agreement exceeds 96% because the vast majority of (model, question) pairs have true rates far from the tier thresholds.

| Tier | Strategy | $n_{\max} = 100$ Type I | $n_{\max} = 100$ Type II | $n_{\max} = 200$ Type I | $n_{\max} = 200$ Type II |
|---|---|---|---|---|---|
| **Gold** | Random | 0.008 | 0.440 | 0.002 | 0.640 |
| | Ordered | 0.000 | 0.882 | 0.000 | 1.000 |
| | Evidence-CAT | 0.012 | 0.808 | 0.002 | 0.988 |
| **Silver** | Random | 0.040 | 0.434 | 0.008 | 0.494 |
| | Ordered | 0.000 | 0.952 | 0.000 | 1.000 |
| | Evidence-CAT | 0.000 | 0.874 | 0.000 | 0.998 |
| **Bronze** | Random | 0.162 | 0.220 | 0.088 | 0.152 |
| | Ordered | 0.004 | 0.912 | 0.000 | 1.000 |
| | Evidence-CAT | 0.078 | 0.664 | 0.002 | 0.998 |
| Nominal bounds | | $\alpha$=0.05 | $\beta$=0.10 | $\alpha$=0.05 | $\beta$=0.10 |

Table 6: Monte Carlo boundary error rates with $n_{\max} = 100$ (main paper, Table 3) versus $n_{\max} = 200$ (this appendix), 500 simulations per condition, $\sigma_\beta = 1.5$. Type I rates tighten under longer truncation, while Type II rates at the upper boundary inflate further for non-random orderings—confirming that the SPRT failure under non-i.i.d. Rasch data is a structural property, not a truncation artifact.