Confident Rankings with Fewer Items: Adaptive LLM Evaluation with Continuous Scores

Anonymous ACL submission

Abstract

Computerized Adaptive Testing (CAT) has proven effective for efficient LLM evaluation on multiple-choice benchmarks, but modern LLM evaluation increasingly relies on generation tasks where outputs are scored continuously rather than marked correct/incorrect. We present a principled extension of IRT-based adaptive testing to continuous bounded scores (ROUGE, BLEU, LLM-as-a-Judge) by replacing the Bernoulli response distribution with a heteroskedastic normal distribution. Building on this, we introduce an uncertainty aware ranker with adaptive stopping criteria that achieves reliable model ranking while testing as few items and as cheaply as possible. We validate our method on five benchmarks spanning n-gram-based, embedding-based, and LLM-as-judge metrics. Our method uses 2% of the items while improving ranking correlation by 0.12 τ over random sampling, with 95% accuracy on confident predictions.

1 Introduction

Rigorous evaluation of large language models is essential, but current practice faces two challenges. The first is cost: exhaustive evaluation becomes expensive as the number of models, test items, and metrics grows. The second is methodological: score differences are often reported without significance testing, leading many studies to mistake statistical noise with improved performance and produce non-replicable results (Dror et al., 2018). Efficient evaluation methods that maintain statistical validity remain underexplored.

Computerized Adaptive Testing (CAT) dynamically selects informative test items to estimate model capabilities with far fewer evaluations than exhaustive benchmarking, and has emerged as a promising approach for efficient evaluation of large language models (Liu et al., 2024). These methods use adaptive use Item Response Theory (IRT) as their theoretical foundation (Martínez-Plumed et al., 2016; Maia Polo et al., 2024; Rodriguez et al., 2021), with recent work extending this to fully adaptive item selection (Hofmann et al., 2025).

Existing approaches for LLM evaluation focus exclusively on multiple-choice datasets where responses can only be correct or incorrect. However, many of the tasks that reflect real use cases of LLMs such as summarization, dialogue, instruction following and machine translation must be scored on a continuous scale. These scoring strategies cover traditional metrics such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004), and embedding similarity-based scores such as BERTScore (Zhang et al., 2020) and COMET (Rei et al., 2020). LLM-as-judge evaluation (Zheng et al., 2024; Liu et al., 2023), now widely adopted for assessing instruction following and open-ended generation, also produces ordinal ratings that can be normalized and treated as continuous.

We present a continuous extension of IRT-based adaptive testing that accommodates real-valued scores while preserving the mathematical structure of binary CAT. Our key insight is to replace the Bernoulli response distribution with a heteroskedastic normal distribution that maintains the same logistic mean function while introducing variance that mimics the Bernoulli structure. This preserves the natural property that variance is highest when outcomes are most uncertain (μ=0.5) and shrinks at the boundaries where scores are constrained.

Continuous CAT already reduces evaluation cost substantially, but further savings are possible when we consider how evaluation results are typically used. Model evaluation is fundamentally comparative: the primary goal is often to determine whether one model is better than others. Building on continuous CAT, we introduce an adaptive multi-model ranking method that efficiently obtains statistically significant rankings by monitoring uncertainty estimates around model scores. Rather than test-

ing each model to a fixed precision threshold, we stop as soon as models are well-differentiated from each other based on pairwise confidence at a user-specified level. This focuses the testing effort on close competitors while reducing items for clearly separated models. We further observe that when distinguishing two models, testing either one yields information about their relative performance, allowing cost-aware allocation that preferentially tests cheaper models to maximize uncertainty reduction per dollar spent.

We validate our approach on five generation benchmarks that span summarization (GovReport, BioLaySumm2025), named entity recognition (Nemotron PII), question answering (TruthfulQA) and translation (FLORES). For each dataset, we evaluate 5 disjoint sets of 4 hold-out models and multiple metric types (ROUGE, BLEU, BERTScore, COMET, readability indices, LLM-as-judge). Our adaptive ranker achieves 0.73 Kendall's τ correlation with ground-truth rankings while using only 2% of the full evaluation budget, outperforming random sampling by 0.12 τ. Compared to fixed-length CAT, adaptive stopping provides an additional 32% item reduction and 42% cost savings.

Our contributions are:

1. A principled extension of IRT-based CAT from binary to continuous bounded outcomes via heteroskedastic normal distribution

2. Adaptive multi-model ranking algorithm with pairwise stopping and cost-aware allocation

3. Empirical validation across diverse generation tasks, metrics, and model scales

The continuous CAT framework opens adaptive testing to the full spectrum of modern LLM evaluation, and the adaptive multi-model ranker enables efficient comparisons of generation quality across model candidates. We make our code available on GitHub [link]

2 Related Work

Early studies on efficient model evaluation used IRT to analyze benchmark properties such as item difficulty and discrimination (Martínez-Plumed et al., 2016; Lalor et al., 2016; Vania et al., 2021; Rodriguez et al., 2021). Maia Polo et al. (2024) and Kipnis et al. (2025) used IRT to identify representative static subsets, while Hofmann et al. (2025) developed a fully adaptive framework that dynamically selects items during evaluation. However, all existing approaches assume binary outcomes and cannot accommodate the continuous scores. Lalor et al. (2016) applied IRT to construct evaluation scales for NLP systems, and Prudêncio et al. (2015) and Lalor et al. (2019) showed that IRT models can be fit using response patterns from model ensembles rather than human annotations. Chen et al. (2019) proposed a Beta-distributed IRT model applied to modelling classifier confidence scores.

This follows the tradition in psychometrics, with Samejima (1973) introducing the continuous response model, and Noel and Dauvier (2007) proposing Beta-distributed models that naturally respect $[0, 1]$ bounds. Both introduce additional complexity to Fisher Information calculations, complicating item selection. Beta models also cannot accommodate exact boundary values, requiring either ad hoc transformations or zero-and-one inflated extensions (Molenaar et al., 2022). Related continuous-response formulations include linear factor-analytic indices for difficulty and information, and Rasch-type models for continuous responses in large-scale learning systems (Ferrando, 2009; Deonovic et al., 2020). These typically assume homoskedastic residuals or prioritize scalable scoring over closed-form CAT information. Mellenbergh (1994) developed generalized linear IRT with simpler closed-form solutions, but assumes constant variance across the score range. Our heteroskedastic normal formulation combines simplicity with appropriate variance structure and maintains direct compatibility with standard CAT algorithms.

3 Background

Item Response Theory (IRT) (Hambleton et al., 1991) models the probability of a correct response as a function of the test-taker's latent ability θ and item parameters. The simplest model, the 1-Parameter Logistic (1PL) model, defines this probability as:

$$P(X = 1|\theta, a, b) = \frac{1}{1 + \exp(-a(\theta - b))} \quad (1)$$

where b is item difficulty and a is a discrimination parameter shared across all items. When ability matches difficulty ($\theta = b$), the probability of success is exactly 0.5. Higher a yields steeper

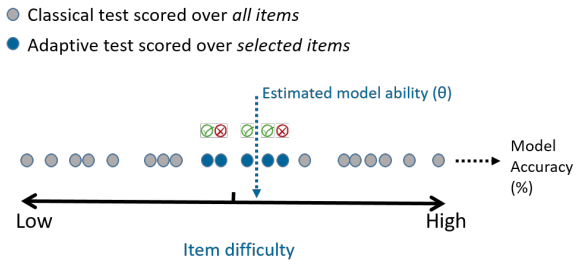

Figure 1: Adaptive testing focuses on items around the model ability, skipping thoses for which it would most certainly (i.e., uninformatively) get high or low scores.

response curves that better differentiate between ability levels.

Computerized adaptive testing (CAT) (Wainer, 2000) dynamically selects items to efficiently estimate a test-taker's ability. The algorithm initializes with a prior distribution over θ, typically $\mathcal{N}(0, \sigma_0^2)$ where the prior variance is a hyperparameter controlling initial uncertainty. At each iteration, it selects the item that maximizes Fisher Information at the current ability estimate $i^* = \arg\max_i I(\hat{\theta}|b_i)$.

After observing the response, the ability estimate and posterior are updated via Bayesian updating. As items accumulate, the standard error decreases:

$$\text{SE}(\hat{\theta}) = \frac{1}{\sqrt{\sum_{i=1}^{n} I_i(\hat{\theta})}} \tag{2}$$

Testing terminates when this standard error falls below a predetermined threshold, yielding a precise ability estimate with fewer items than fixed-length testing.

Fisher Information quantifies how much an observation tells us about θ. For the 1PL model with binary outcomes, this is derived directly from Bernoulli variance:

$$I(\theta|a, b) = a^2 \cdot P(\theta, b) \cdot (1 - P(\theta, b)) \tag{3}$$

Information is highest when $P = 0.5$ (maximum uncertainty about the outcome) and approaches zero as P approaches 0 or 1 (outcome nearly certain). This creates heteroskedastic variance as a function of the predicted probability $P(\theta, b)$.

4 Continuous CAT

To extend CAT to continuous outcomes, we require a response distribution for scores in $[0, 1]$ rather than binary $\{0, 1\}$. We retain the logistic mean function from standard IRT, which captures the relationship between ability, difficulty, and expected performance, and replace the Bernoulli distribution with a normal distribution:

$$X|\theta, b, k \sim \mathcal{N}(\mu(\theta, b), \sigma^2(\theta, b)) \tag{4}$$

where the mean follows the logistic function:

$$\mu(\theta, b) = \frac{1}{1 + \exp(-(\theta - b))} \tag{5}$$

We preserve the Bernoulli variance structure $\text{Var}(X) = P(1 - P)$ by defining:

$$\sigma^2(\theta, b) = k \cdot \mu(\theta, b) \cdot (1 - \mu(\theta, b)) \tag{6}$$

where $k = 1/a^2$ is a noise parameter capturing measurement precision. We estimate a single k per dataset-metric combination from calibration data.

The Fisher Information for this model is:

$$I(\theta|b, k) = \frac{(d\mu/d\theta)^2}{\sigma^2} = \frac{\mu(1 - \mu)}{k} \tag{7}$$

Comparing with the 1PL Fisher Information $I = a^2 \cdot P(1 - P)$, we see that $k = 1/a^2$: high discrimination ($a > 1$) corresponds to low noise ($k < 1$), and vice versa. Both parameterizations capture the same underlying question: how reliably does an observed response reflect true ability? This equivalence indicates that our continuous extension constitutes a natural generalization that unifies the treatment of measurement precision across binary and continuous settings.

A limitation of the normal distribution is its unbounded support: it can technically assign probability to values outside $[0, 1]$. However, the heteroskedastic variance structure mitigates this concern: as μ approaches the boundaries, $\sigma^2 = k \cdot \mu(1 - \mu)$ shrinks, concentrating the distribution within the valid range. Alternative distributions such as the Beta are naturally bounded but do not preserve the connection to binary IRT and complicate the Fisher Information derivation. In practice, we find that the normal approximation performs well across the metrics we consider.

The CAT algorithm itself remains unchanged. Item selection still maximizes Fisher Information at the current ability estimate, with the information formula now incorporating k. Posterior updates follow the same Bayesian logic with normal likelihoods. The standard error formula $\text{SE}(\hat{\theta}) = 1/\sqrt{\sum_i I_i}$ applies as before.

5 Parameter Estimation

We estimate both b_i and k from historical evaluation data by fitting the model to observed score distributions across a population of models.

Item Difficulty. We adopt the 1PL model and estimate item difficulties from observed scores across calibration models. Following standard IRT convention, we center the ability scale such that the mean model ability is zero. Under this parameterization, the maximum likelihood estimate for item difficulty simplifies to:

$$b_i = -\text{logit}(\hat{p}_i) = \log\left(\frac{1-\hat{p}_i}{\hat{p}_i}\right) \tag{8}$$

where $\hat{p}_i$ is the average score on item i across all calibration model-temperature configurations. When scores cluster in a narrow range (common for metrics like ROUGE where most models achieve 0.3–0.5), the resulting difficulties may not span the full ability range. We therefore apply min-max normalization to $[\epsilon, 1-\epsilon]$ before the logit transformationn.

Noise Parameter. We estimate a global k from the calibration data using method-of-moments. Given item difficulties $\{b_i\}$ and model ability estimates $\{\theta_j\}$ (computed as $\theta_j = \text{logit}(\bar{y}_j)$ where $\bar{y}_j$ is the model's average scaled score), we obtain:

$$k = \frac{\sum_{i,j}(y_{ij} - \mu_{ij})^2}{\sum_{i,j}\mu_{ij}(1-\mu_{ij})} \tag{9}$$

where $\mu_{ij} = \text{logit}^{-1}(\theta_j - b_i)$ is the predicted mean score for model j on item i. This formula directly estimates the variance inflation factor: when $k = 1$, observed variance matches the Bernoulli structure exactly; $k > 1$ indicates noisier measurements than binary responses; $k < 1$ indicates more precise measurements.

Filtering Negative Discrimination Items. Some items may anti-correlate with model ability. On these items, weaker models score higher than stronger models, rendering them effectively uninformative. Because we use a 1PL model with a global discrimination parameter, we filter such items by computing the Pearson correlation between item scores and model abilities across the calibration set, and excluding items with negative correlation from the adaptive item pool.

6 Adaptive Ranker

Standard CAT approaches evaluate each model independently to a fixed precision threshold, ignoring the comparative nature of model evaluation. This can be inefficient, since a model that clearly dominates or trails the field may receive the same number of test items as one locked in a close competition with a neighbor. We propose an adaptive multi-model ranking algorithm that directly optimizes for the comparative goal. Given M models to evaluate on a shared item bank, the algorithm runs independent CAT runs for each model, but monitors pairwise confidence in the ranking and terminates when all adjacent model pairs are statistically well-differentiated at a user-specified confidence level.

Pairwise confidence. Given ability estimates $\hat{\theta}_i$ and $\hat{\theta}_j$ with standard errors SE_i and SE_j, we compute the probability that model i outperforms model j under a normal approximation:

$$P(\theta_i > \theta_j) = \Phi\left(\frac{\hat{\theta}_i - \hat{\theta}_j}{\sqrt{\text{SE}_i^2 + \text{SE}_j^2}}\right) \tag{10}$$

where Φ is the standard normal CDF. This quantifies our confidence in the pairwise ordering given current uncertainty in both estimates.

Adaptive stopping. Traditional CAT terminates when each model's standard error falls below a threshold ϵ, regardless of how the models compare. We instead terminate when all adjacent pairs in the current ranking satisfy $P(\theta_i > \theta_j) > 1 - \frac{1-\gamma}{2}$ or $P(\theta_i > \theta_j) < \frac{1-\gamma}{2}$, corresponding to a two-sided confidence interval at level γ, or reach a maximum item budget n_{max}. This focuses testing effort where it matters: models that are clearly separated require fewer items to establish their relative ordering, while close competitors receive additional testing until their difference reaches statistical significance or the budget is exhausted. If two models remain indistinguishable at the budget limit, they can be reported as tied rather than forced into an arbitrary ordering.

Cost-aware allocation. When a pair (i, j) requires additional testing to reach confidence threshold γ, we must choose which model to evaluate next. Since testing either model reduces uncertainty in the pairwise comparison, we select the

Algorithm 1 Adaptive Multi-Model Ranking

Require: Models $\{1, \ldots, M\}$, item bank $\{(b_i, k_i)\}$, costs $\{c_m\}$, confidence γ, max items n_{max}

Initialize $\hat{\theta}_m \sim \mathcal{N}(0, 1)$, $\text{SE}_m = 5$ for all m
Warm-up: Administer n_{init} items to each model using MFI selection
while not all adjacent pairs confident or at max items **do**
  Rank models by $\hat{\theta}_m$
  Identify uncertain pairs: $\mathcal{U} = \{(i, j) : \gamma > P(\theta_i > \theta_j) > 1 - \gamma \text{ and } n_i < n_{\text{max}} \text{ and } n_j < n_{\text{max}}\}$
  if $\mathcal{U} = \emptyset$ **then**
    break
  end if
  Collect candidate models: $\mathcal{C} = \{m : m \in (i, j) \text{ for some } (i, j) \in \mathcal{U}\}$
  Select model: $m^* = \arg\max_{m \in \mathcal{C}} \frac{\text{SE}_m^2}{(n_m+1) \cdot c_m}$
  Select item: $i^* = \arg\max_i I(\hat{\theta}_{m^*} | b_i, k_i)$
  Administer item i^* to model m^*, observe score y
  Update $\hat{\theta}_{m^*}$ and SE_{m^*}
end while
return Ranking with pairwise confidence scores

model that provides the greatest expected uncertainty reduction per unit cost. We quantify this as:

$$\text{value}_m = \frac{\text{SE}_m^2}{(n_m + 1) \cdot c_m} \tag{11}$$

where n_m is the number of items already administered to model m and c_m is its per-item evaluation cost. The numerator reflects current uncertainty; the denominator captures diminishing returns (each additional item contributes less) weighted by cost. This criterion naturally allocates more items to cheaper models when resolving uncertain comparisons, yielding additional cost savings beyond item reduction alone.

Algorithm. The complete procedure combines continuous CAT estimation with pairwise stopping and cost-aware allocation under a total budget constraint. Rather than allocating a fixed maximum number of items per model, we specify a total cost budget B that can be flexibly distributed across models. After a warm-up phase that administers initial items to all models via MFI selection, the algorithm iteratively identifies uncertain pairs, selects the most cost-effective model to test, chooses the maximally informative item for that model, and updates posteriors. Testing continues until either all adjacent pairs reach the confidence threshold, or the total budget B is exhausted.

7 Experiments

We evaluate our continuous CAT extension and adaptive ranker across five generation tasks spanning summarization, machine translation, question answering, and named entity recognition. Our goal is to approximate ground-truth rankings from full dataset evaluation using as few items as possible. We compare our adaptive ranker against random sampling, and examine whether the confidence estimates can reliably tell apart genuine performance differences from statistical ties.

7.1 Experimental Protocol

We select tasks representing diverse generation settings and metric types (Table 1). This selection covers n-gram overlap (ROUGE, BLEU), learned embeddings (BERTScore, COMET), readability indices (FKGL), LLM-as-judge, and span-level F1. We exclude BERTScore from GovReport as all models achieved near-identical scores (range $<$0.02), yielding no meaningful ground-truth ranking. Full prompts and scoring details are provided in Appendix A.

We evaluate 21 models from 6 families: OpenAI GPT, Google Gemini, Amazon Nova, Meta Llama, Mistral, and Qwen (see Appendix A for full list). For each dataset-metric pair, we collect exhaustive evaluation scores by running all models on all items, establishing ground-truth rankings. Each model is evaluated at four temperature settings ($T \in \{0.0, 0.4, 0.7, 1.0\}$), which yields 84 model-temperature configurations per item. Our

Table 1: Evaluation datasets. FKGL: Flesch-Kincaid Grade Level (Kincaid et al., 1975).

Dataset	Size	Split	Task	Metrics
BioLaySumm (Goldsack et al., 2024)	1,376	val	Lay summarization	ROUGE-L BERTScore FKGL
GovReport (Huang et al., 2021)	973	val	Document summarization	ROUGE-L
TruthfulQA (Lin et al., 2022)	817	val	Open-ended QA	LLM-Judge BERTScore (diff.)
FLORES (NLLB Team et al., 2022)	1,012	devtest	Translation (TR→EN)	COMET BLEU
Nemotron-PII (Steier et al., 2025)	2,000	test[1]	NER/PII detection	F1 (span-level)

main experiment uses five-fold cross-validation over models. In each fold, we randomly select 4 models as the hold-out set to be ranked adaptively; all temperature variants of hold-out models are excluded from calibration. The remaining 17 models (68 configurations) provide calibration data for estimating item parameters (b_i, k). We fit parameters via maximum likelihood on the calibration set, then simulate adaptive evaluation on the hold-out models at $T = 0$. We set the confidence threshold $\gamma = 0.95$ for pairwise decisions and require a minimum of 10 items per model before stopping.

For each hold-out set, we compare the adaptive ranker to a random sampling baseline which allocates the total budget B across models by iteratively selecting a random affordable model and a random unseen item to administer. We assess ranking quality using Kendall's τ correlation between predicted and ground-truth rankings computed from point estimates (mean ability $\hat{\theta}$). For efficiency, we report the mean number of items administered per model.

7.2 Main Results

Table 2 presents results across all dataset-metric combinations. On 7 of 9 dataset-metric pairs, the adaptive ranker achieves higher Kendall's τ than the random sampling baseline. The largest gains occur on discriminative metrics: +12% on BioLaySumm ROUGE-L (0.957 vs 0.853), +26% on TruthfulQA BERTScore (0.450 vs 0.190), and +22% on FLORES BLEU (0.803 vs 0.580). These results demonstrate that our adaptive ranker provides genuine performance gains over random sampling. The adaptive ranker uses 52–101 items per hold-out set, and achieves 0.73 τ with the full dataset ranking while using only 2% of the items.

The discrimination parameter a predicts where adaptive methods provide the most benefit. On high-discrimination metrics ($a > 3$), the adaptive ranker consistently outperforms baseline. On low-discrimination metrics like Nemotron F1 ($a = 1.42$), individual items provide less information about ability, limiting the advantage of intelligent selection.

Distributional Conformance. Our continuous IRT extension assumes scores follow a heteroskedastic normal distribution with variance $\sigma^2 = k \cdot \mu(1 - \mu)$. We examine how well each metric conforms to this assumption by measuring the R^2 between observed and predicted variance across score bins (see Appendix B for details). Conformance varies substantially: BERTScore and BLEU exhibit moderate fit ($R^2 = 0.24$–0.36), while ROUGE-L and COMET show poor conformance despite high ranking accuracy. We find no positive correlation between conformance and ranking quality ($r = -0.12$); instead, discrimination a is the stronger predictor ($r = 0.68$). This suggests that violations of the heteroskedastic variance assumption do not degrade ranking performance when items are sufficiently discriminative.

7.3 Tie Detection

The τ correlation in Table 2 measures ranking accuracy based on point estimates, but point estimates alone do not indicate when a ranking is reliable. Reporting that model A outperforms model B when the difference is due to noise leads to non-reproducible results. Our method uses the IRT model's posterior to detect statistical ties, where a pair is reported as a tie when confidence falls below 95%

Table 2: Ranking performance across benchmarks. Higher discrimination (a) indicates more informative items. % Used is items administered relative to full evaluation.

Dataset	Metric	Size	a	Adaptive $\tau \uparrow$	Baseline $\tau \uparrow$	Items $\downarrow$	% Used $\downarrow$
BioLaySumm	ROUGE-L	1376	4.13	**0.957**±0.13	0.853±0.20	82±17	1.5±0.3%
BioLaySumm	BERTScore	1376	3.40	**0.903**±0.17	0.743±0.36	88±9	1.6±0.2%
BioLaySumm	FKGL	1376	2.34	**0.800**±0.25	0.713±0.40	88±15	1.6±0.3%
GovReport	ROUGE-L	973	4.79	0.800±0.20	**0.823**±0.22	98±30	2.5±0.8%
TruthfulQA	LLM-Judge	817	2.59	**0.490**±0.38	0.400±0.49	93±14	2.9±0.4%
TruthfulQA	BERTScore	817	2.65	**0.450**±0.43	0.190±0.46	92±7	2.8±0.2%
FLORES	BLEU	1012	3.12	**0.803**±0.23	0.580±0.31	92±8	2.3±0.2%
FLORES	COMET	1012	4.07	**0.677**±0.33	0.503±0.41	101±11	2.5±0.3%
Nemotron	F1	2000	1.36	0.673±0.19	**0.707**±0.29	52±5	0.7±0.1%
Overall (mean)			3.16	**0.73**	0.61	88	2.0%

Table 3: Tie detection against ground-truth ties from bootstrap CIs on full evaluation. Adapt/GT Tie% are the fraction of pairs reported as ties for adaptive ranker and ground truth. Tie P/R/F1 are the precision, recall and F1 on tie detection quality of the adaptive ranker compared to the full evaluation. Conf Acc is the accuracy of the adaptive ranker on its non-tie predictions.

Dataset	Metric	Adapt Tie%	GT Tie%	Tie P ↑	Tie R ↑	Tie F1 ↑	Conf Acc ↑
BioLaySumm	ROUGE-L	14%	3%	0.60	1.00	0.62	1.00
BioLaySumm	BERTScore	42%	13%	0.36	0.99	0.43	1.00
BioLaySumm	FKGL	37%	17%	0.46	0.95	0.54	0.98
GovReport	ROUGE-L	18%	10%	0.39	0.73	0.43	0.94
TruthfulQA	LLM-Judge	76%	47%	0.50	0.89	0.60	0.77
TruthfulQA	BERTScore	100%	73%	0.73	1.00	0.82	–
FLORES	BLEU	58%	30%	0.48	0.99	0.58	1.00
FLORES	COMET	57%	37%	0.52	0.94	0.56	0.89
Nemotron	F1	65%	7%	0.08	1.00	0.14	1.00
Overall (mean)		52%	26%	0.46	0.94	0.52	0.95

Table 3 presents tie detection performance of the adaptive ranker against ground-truth ties derived from bootstrap confidence intervals on full evaluation. The adaptive ranker is consistently conservative: Adapt Tie% exceeds GT Tie% for all metrics. Tie detection recall is high (0.77–1.00), which means that our method catches most ground-truth ties. For pairs where one model is ranked strictly higher than the other, overall accuracy is 95%. TruthfulQA BERTScore reports 100% ties because all models score within a 0.7% range, and the differences are too small to resolve within the budget constraints.

Two metric-dataset combinations show lower performance: GovReport ROUGE-L (recall 0.73) and TruthfulQA LLM-Judge (confident accuracy 0.77). In both cases, this can be traced to specific model comparisons within single holdout sets that have consistently biased estimates across seeds. For TruthfulQA LLM-Judge, one holdout contains three models that are tied in the ground-truth ranking, but the adaptive ranker consistently ranks one model below the others, detecting the ties in only 1–3 of 20 seeds. For GovReport ROUGE-L, a single ground-truth tie in one holdout is never detected (0/20 seeds). Even with these outliers, the adaptive ranker achieves 95% overall accuracy on confidently ranked pairs and 94% recall in detecting ground truth ties, perfectly aligning with the 95% confidence cutoff of the method.

7.4 Ablation: Adaptive vs Fixed-Length CAT

The adaptive ranker terminates based on pairwise confidence, allocating more items to close competitors and fewer to well-separated models, and further manages cost by preferentially choosing

Table 4: Comparison with fixed-length CAT that administers the same number of items to all models.

Dataset	Metric	Adapt $\tau \uparrow$	Fixed $\tau \uparrow$	$\Delta\tau$	Item Savings	Cost Savings
BioLaySumm	BERTScore	0.90	0.82	−0.08	36%	46%
BioLaySumm	FKGL	0.80	0.78	−0.02	31%	38%
BioLaySumm	ROUGE-L	0.96	0.97	+0.02	39%	47%
FLORES	BLEU	0.80	0.78	−0.02	33%	43%
FLORES	COMET	0.68	0.80	+0.12	34%	44%
GovReport	ROUGE-L	0.80	0.85	+0.05	35%	46%
Nemotron	F1	0.68	0.76	+0.08	23%	31%
TruthfulQA	BERTScore	0.45	0.31	−0.14	28%	41%
TruthfulQA	LLM-Judge	0.49	0.52	+0.03	29%	41%
Overall		0.73	0.73	0.00	32%	42%

cheaper models. To isolate the value of this strategy, we compare against fixed-length CAT. For each holdout set, we administer the same number of items n to all models, where n is set to the maximum items used by the adaptive ranker for any model in that holdout.

Table 4 shows that the two methods achieve the same average τ, but adaptive uses 32% less items and 42% less cost. This shows that the additional items used by the fixed-length CAT are superfluous, and the stopping mechanism leads to significant efficiency gains with no performance degradation.

7.5 Generalization to Unseen Model Families

Our main experiments use cross-validation over individual models, but calibration and hold-out models may share architectural similarities within families. To test whether item parameters generalize to genuinely novel architectures, we conduct a family-stratified evaluation: we hold out Mixtral-8x7b-instruct, Qwen3-32b, and Nova-pro-v1 and calibrating only on OpenAI, Meta/Llama, and Google/Gemini models (14 total).

Table 5 shows results. The adaptive ranker matches or outperforms baseline on 7 of 9 dataset-metric pairs, with substantial gains on TruthfulQA BERTScore (+60%) and GovReport ROUGE-L (+30%). These results suggest that item difficulty parameters transfer across model families: an item that is difficult for Llama models tends to also be difficult for Qwen or Nova models.

Table 5: Family holdout evaluation. Calibration on OpenAI, Llama, Gemini (14 models); evaluation on hold-out Mistral, Qwen, Nova (3 models).

Dataset	Metric	Base $\tau \uparrow$	Adapt $\tau \uparrow$
BioLaySumm	BERTScore	0.53±0.52	**1.00**±0.00
BioLaySumm	FKGL	1.00±0.00	1.00±0.00
BioLaySumm	ROUGE-L	**1.00**±0.00	0.87±0.27
FLORES	BLEU	0.67±0.33	**1.00**±0.00
FLORES	COMET	0.73±0.33	0.73±0.33
GovReport	ROUGE-L	0.56±0.31	**0.85**±0.28
Nemotron	F1	**0.93**±0.20	0.67±0.33
TruthfulQA	BERTScore	0.27±0.55	**0.87**±0.27
TruthfulQA	LLM-Judge	0.40±0.55	**0.60**±0.33
Overall		0.68	**0.84**

8 Conclusion

We present a principled extension of IRT-based adaptive testing from binary to continuous bounded scores. Building on this foundation, we introduce an adaptive multi-model ranking algorithm with pairwise stopping and cost-aware allocation. Across five generation benchmarks spanning different metrics, our method achieves 0.73 Kendall's τ correlation with ground-truth rankings and 95% accuracy on confident predictions while using only 2% of items. Notably, our method can reliably rank models from entirely unseen families, and remains robust to empirical deviations from the assumed variance structure.

For future work, extending our to model item-specific discrimination parameters can improve item selection. Incorporating variable item costs can enable cost-optimal item selection at a more granular level. Extending to multi-metric evaluation can allow ranking models on several metrics simultaneously in a single adaptive run, further increasing evaluation efficiency. Developing methods for more efficient item parameter estimation such as transfer learning from related benchmarks or few-shot calibration can reduce the cold-start cost for new datasets and enable faster deployment.

9 Limitations

Our method requires calibration data from existing model evaluations to estimate item parameters, creating a cold-start problem for new benchmarks. This upfront cost of exhaustive evaluation on calibration models is amortized over subsequent adaptive evaluations, but represents a barrier to immediate deployment on new datasets. Furthermore, item parameters must be re-estimated for each metric separately, which might add overhead when comprehensive multi-metric evaluation is needed.

Our stopping criterion evaluates each adjacent pair independently at significance level α, which controls the per-comparison error rate but not the family-wise error rate (FWER). With M models and $M-1$ adjacent comparisons, the probability of at least one incorrectly ordered pair under the null hypothesis is approximately $1-(1-\alpha)^{M-1}$, reaching 18% for five models at $\alpha = 0.05$. Corrections such as Holm-Bonferroni could provide FWER control by requiring stricter thresholds for the most confident pairs at the cost of additional test items, potentially negating much of the efficiency gain from adaptive selection.

10 Ethical Considerations

Large language model inference carries substantial environmental costs, with energy consumption and carbon emissions scaling with the number of API calls. By reducing evaluation from exhaustive testing to approximately 2% of items, our method directly decreases the environmental footprint of model comparison. As the number of candidate models, benchmarks, and evaluation metrics continues to grow, efficient evaluation methods become increasingly important for sustainable AI development.

Beyond environmental benefits, efficient evaluation has implications for equity in AI research. Exhaustive evaluation across large benchmarks favors well-resourced organizations. By dramatically reducing the number of required items, our approach can lower barriers for smaller laboratories, students and independent researchers to perform meaningful model comparisons. Additionally, our method's explicit uncertainty quantification and tie detection encourages more honest reporting practices.

One potential concern is benchmark gaming: if certain items are predictably selected as most informative, model developers could optimize specifically for those items rather than general capability. Periodic re-calibration of data and item banks can mitigate this risk.

References

Yu Chen, Telmo Silva Filho, Ricardo B Prudencio, Tom Diethe, and Peter Flach. 2019. β^3-irt: A new item response model and its applications. In *The 22nd international conference on artificial intelligence and statistics*, pages 1013–1021. PMLR.

Benjamin Deonovic, Maria Bolsinova, Timo Bechger, and Gunter Maris. 2020. A Rasch model and rating system for continuous responses collected in large-scale learning systems. *Frontiers in Psychology*, 11:500039.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392. Association for Computational Linguistics.

Pere J Ferrando. 2009. Difficulty, discrimination, and information indices in the linear factor analysis model for continuous responses. *Applied Psychological Measurement*, 33(1):9–24.

Tomas Goldsack, Carolina Scarton, Matthew Shardlow, and Chenghua Lin. 2024. Overview of the BioLaySumm 2024 shared task on the lay summarization of biomedical research articles. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 122–131, Bangkok, Thailand. Association for Computational Linguistics.

Ronald K Hambleton, Hariharan Swaminathan, and H Jane Rogers. 1991. *Fundamentals of Item Response Theory*. Sage Publications.

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

Luyang Huang, Shuyang Cao, Nikolaus Parulian, Heng Ji, and Lu Wang. 2021. Efficient attentions for long document summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1419–1436, Online. Association for Computational Linguistics.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas for Navy enlisted personnel. *Research Branch Report*, 8(75).

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. 2016. Building an evaluation scale using item response theory. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 648–657. Association for Computational Linguistics.

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 and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4240–4250. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81. Association for Computational Linguistics.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Qi Liu, Yan Zhuang, Haoyang Bi, Zhenya Huang, Weizhe Huang, Jiatong Li, Junhao Yu, Zirui Liu, Zirui Hu, Yuting Hong, Zachary A Pardos, Haiping Ma, Mengxiao Zhu, Shijin Wang, and Enhong Chen. 2024. Survey of computerized adaptive testing: A machine learning perspective. *arXiv preprint arXiv:2404.00712*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using GPT-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

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*, volume 235, pages 34303–34326. PMLR.

Fernando Martínez-Plumed, Ricardo BC Prudêncio, Adolfo Martínez-Usó, and José Hernández-Orallo. 2016. Making sense of item response theory in machine learning. In *Proceedings of the Twenty-second European Conference on Artificial Intelligence*, pages 1140–1148.

Gideon J Mellenbergh. 1994. Generalized linear item response theory. *Psychological Bulletin*, 115(2):300–307.

Dylan Molenaar, Mariana Curi, and Jorge L Bazán. 2022. Zero and one inflated item response theory models for bounded continuous data. *Journal of Educational and Behavioral Statistics*, 47(6):693–735.

NLLB Team, Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Yvonnick Noel and Bruno Dauvier. 2007. A beta item response model for continuous bounded responses. *Applied Psychological Measurement*, 31(1):47–73.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics.

RB Prudêncio, José Hernández-Orallo, and Adolfo Martınez-Usó. 2015. Analysis of instance hardness in machine learning using item response theory. In *Second international workshop on learning over multiple contexts in ECML*, volume 3.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 2685–2702. Association for Computational Linguistics.

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 and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4486–4503. Association for Computational Linguistics.

Fumiko Samejima. 1973. Homogeneous case of the continuous response model. *Psychometrika*, 38(2):203–219.

Amy Steier, Andre Manoel, Alexa Haushalter, and Maarten Van Segbroeck. 2025. Nemotron-PII: Synthesized data for privacy-preserving AI.

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

Howard Wainer. 2000. *Computerized Adaptive Testing: A Primer*, 2nd edition. Lawrence Erlbaum Associates.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. BERTscore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

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. 2024. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36.

A Experimental Setup

All datasets were evaluated using the same set of 21 models spanning 6 families: OpenAI GPT (GPT-5-mini, GPT-5-nano, GPT-4.1-mini, GPT-4.1-nano, GPT-4o-mini), Google Gemini (Gemini-2.5-flash, Gemini-2.5-flash-lite, Gemini-2.0-flash, Gemini-2.0-flash-lite), Amazon Nova (Nova-pro-v1, Nova-lite-v1, Nova-micro-v1), Meta Llama (Llama-4-maverick-17b, Llama-4-scout-17b, Llama-3.3-70b, Llama-3.2-3b, Llama-3.1-8b), Mistral (Mistral-7b-instruct-v0.2, Mixtral-8x7b-instruct-v0.1), and Qwen (Qwen3-32b, Qwen3-coder-30b-a3b). Models were accessed via their native APIs (OpenAI, Google) or Amazon Bedrock (Nova, Llama, Mistral, Qwen). Each model was queried at 4 temperature settings: $T \in \{0.0, 0.4, 0.7, 1.0\}$, resulting in 84 model-temperature configurations per item. Note that GPT-5 models do not expose a temperature parameter; for these models, the temperature setting was ignored, effectively treating all 4 configurations as identical.

BioLaySumm2025-PLOS. We used the BioLaySumm2025-PLOS dataset (Goldsack et al., 2024), a biomedical lay summarization benchmark containing scientific articles paired with plain-language summaries. We evaluated on the `validation` split, which contains 1,376 items. Each model was prompted with the instruction: *"Summarize the following scientific article in plain language for a general audience:"* followed by the full article text. This yielded 115,584 total inference responses (1,376 items $\times$ 21 models $\times$ 4 temperatures). After scoring, all metrics were linearly scaled to $[0, 1]$ based on the observed minimum and maximum values across all model-temperature configurations for each item.

FLORES-Turkish-English. We used the FLORES-200 dataset (NLLB Team et al., 2022), a massively multilingual translation benchmark. We evaluated on the Turkish-to-English language pair using the `devtest` split, which contains 1,012 sentence pairs. Each model was prompted with the instruction: *"Translate the following Turkish text to English. Output only the translation, nothing else."* followed by the source Turkish sentence. This yielded 85,008 total inference responses (1,012 items $\times$ 21 models $\times$ 4 temperatures).

GovReport-Summarization. We used the GovReport dataset (Huang et al., 2021), a long-document summarization benchmark containing U.S. government reports paired with expert-written summaries. We evaluated on the `validation` split, which contains 973 documents. Each model was prompted with the instruction: *"Summarize the following government report:"* followed by the full report text. This yielded 81,732 total inference responses (973 items $\times$ 21 models $\times$ 4 temperatures).

Nemotron-PII. We used the Nemotron-PII dataset (Steier et al., 2025), a named entity recognition benchmark for detecting personally identifiable information (PII) and protected health information (PHI) across 54 entity types. We evaluated on a random sample of 2,000 items from the `test` split. Each model received a system prompt containing the complete list of valid entity labels and 4 few-shot examples demonstrating the expected output format (`EXTRACTED_TEXT | LABEL`), followed by the document text. This yielded 168,000 total inference responses (2,000 items $\times$ 21 models $\times$ 4 temperatures). Responses were scored using standard span-level F1, where a predicted entity is considered correct only if both the extracted text and entity type exactly match a gold annotation.

TruthfulQA. We used the TruthfulQA dataset (Lin et al., 2022), a benchmark designed to evaluate model truthfulness on questions that commonly elicit false or misleading answers. We evaluated on the `validation` split of the generation configuration, which contains 817 questions. Each model was given a system message *"You are a helpful assistant."* and prompted directly with the question. This yielded 68,628 total inference responses (817 items $\times$ 21 models $\times$ 4 temperatures).

Responses were scored using two metrics. First, we computed a differential BERTScore using DeBERTa-xlarge-mnli embeddings: for each response, we calculate the maximum BERTScore F1 against all correct answers and the maximum F1

against all incorrect answers, then compute the differential as $(\max_{\text{correct}} - \max_{\text{incorrect}} + 1)/2$, normalized to $[0, 1]$. This rewards responses semantically closer to correct answers while penalizing those closer to known misconceptions. Second, we used LLM-as-judge with GPT-4.1-nano, employing a 10-point rubric that penalizes answers matching known incorrect responses and rewards alignment with verified correct answers. Although common practice favors strong judge models, preliminary experiments showed high correlation between GPT-4.1-nano, GPT-4.1-mini, and GPT-4.1 scores when ground-truth correct and incorrect answer sets are available, justifying the use of the more efficient model.

B Conformance Analysis Details

Our continuous IRT extension assumes scores follow a heteroskedastic normal distribution with variance $\sigma^2 = k \cdot \mu(1 - \mu)$, where μ is the predicted mean and k is the estimated noise parameter. To examine how well each metric conforms to this assumption, we bin items by their predicted mean μ and compute the observed variance within each bin. We then measure conformance as the R^2 between observed and predicted variance across bins.

Table 6 reports conformance alongside discrimination and ranking quality. Notably, BioLaySumm ROUGE-L achieves the highest τ (0.957) with strongly negative R^2 (-8.2), whereas BioLaySumm BERTScore has the best conformance ($R^2 = 0.36$) but lower τ (0.903).

Table 6: Distributional conformance. R^2 measures fit between observed and predicted variance across score bins.

Dataset	Metric	a	R^2	τ
BioLaySumm	ROUGE-L	4.13	-8.2	0.957
BioLaySumm	BERTScore	3.40	0.36	0.903
BioLaySumm	FKGL	2.34	-666	0.800
GovReport	ROUGE-L	4.79	0.24	0.800
TruthfulQA	LLM-Judge	2.59	-1.2	0.490
TruthfulQA	BERTScore	2.65	-1524	0.450
FLORES	BLEU	3.12	0.34	0.803
FLORES	COMET	4.07	-20.3	0.677
Nemotron	F1	1.36	-20.5	0.673

C Use of AI Assistants

AI Assistants have been used in this paper to paraphrase text written by the authors, and to generate experimental details from code for Appendix A. All AI generated text was reviewed and corrected by the authors.