

The Evaluation Scaling Law: Toward More Reliable LLM Assessment via Multi-Prompt Diversity

Hongzhi Luan, Zhaoxin Huan, Linfeng Shi
Longfei Zheng, Xiaolu Zhang, Jun Zhou*

Ant Group

{luanhongzhi.lhz, zhaoxin.hzx, john.slf}@antgroup.com

{zlf206411, yueyin.zxl, jun.zhoujun}@antgroup.com

Abstract

Large Language Model (LLM) evaluation is frequently challenged by prompt sensitivity, which limits the reliability of benchmarks relying on a single template. To address this, we systematically investigate a multi-prompt evaluation framework that aggregates performance across diverse templates, leveraging a Quality-Diversity (QD) prompt selection mechanism to ensure both variety and fidelity. Through extensive empirical analysis, we uncover a novel **Evaluation Scaling Law**: within the range of prompt counts and model families we tested, evaluation reliability tends to improve with template diversity. Specifically, our experiments demonstrate that **evaluation variance** diminishes as the scale of prompts increases, while **ranking monotonicity** exhibits a log-linear improvement relative to the number of templates. By decoupling a model’s intrinsic capability from template-specific fluctuations, our framework facilitates more reliable assessments of LLMs, thereby providing a solid foundation for benchmarking that enables more effective monitoring of training dynamics and rigorous comparisons among state-of-the-art models.

1 Introduction

As Large Language Models (LLMs) rapidly evolve in complexity and capability (OpenAI, 2026; DeepMind, 2026; QwenTeam, 2026), high-fidelity evaluation has become a cornerstone of AI development, guiding both benchmark design and model assessment (Ni et al., 2025; Zhang et al., 2025). It serves as a strategic compass throughout the model life-cycle: informing data curation and architectural decisions during pre-training, as well as safeguarding reliability and alignment during deployment (LingTeam, 2025). Crucially, precise and objective evaluation provides the rigorous metrics needed to distinguish genuine algorithmic progress from

evaluation noise, thereby ensuring a scientifically grounded trajectory for the field.

Despite its pivotal role, current evaluation paradigms rely heavily on single-prompt assessments, making them vulnerable to prompt sensitivity (Shu et al., 2024; Yuan et al., 2025). Research indicates that even minor variations in prompt formatting or phrasing can result in significant fluctuations in model performance (Voronov et al., 2024; Sclar et al., 2024). We identify two fundamental limitations inherent in this "point-estimate" paradigm: (1) **High Measurement Noise**—evaluation results exhibit extreme sensitivity to specific template choices, where perceived performance gains may reflect specific prompt engineering rather than a genuine advancement in general capabilities; and (2) **Weak Ranking Monotonicity**—the noise induced by prompt bias frequently leads to ranking reversals, where superior models or improved checkpoints are erroneously ranked lower than their predecessors. This lack of monotonicity hides true training progress and undermines the reliability of both internal development and cross-model benchmarking.

To address these critical limitations, we systematically investigate a multi-prompt evaluation framework based on a newly discovered **Evaluation Scaling Law**. Through extensive empirical analysis across 3 model families and 10 diverse benchmarks, we show that, within the prompt counts and model families we tested, evaluation reliability tends to improve with template diversity. Specifically, **evaluation noise**—estimated by the fluctuations between adjacent model checkpoints as a proxy, diminishes smoothly as prompt diversity increases. Accordingly, **ranking monotonicity**—measured by Kendall’s τ (Kendall, 1938), improves as a log-linear function of the number of templates. To facilitate the effective implementation of our proposed framework, we further introduce an automated **Quality-Diversity (QD)** prompt

*Corresponding author.

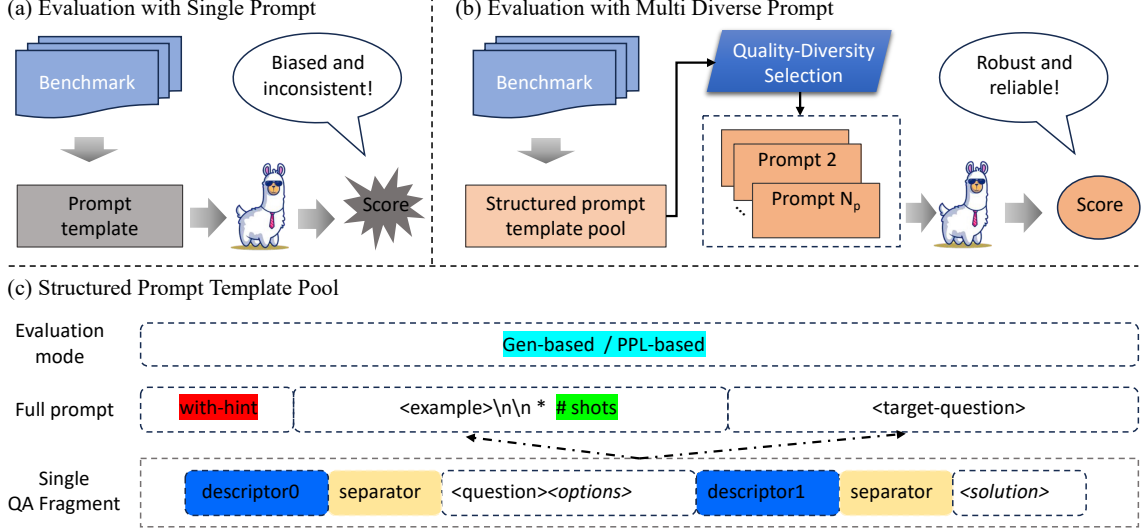


Figure 1: Overview of the proposed framework. (a) **Traditional Evaluation**: A single-prompt paradigm that is often susceptible to template-induced bias and inconsistency. (b) **Multi-Prompt Evaluation (Ours)**: Our proposed protocol that enhances estimation confidence by aggregating scores across diverse templates. (c) **Prompt Candidate Pool**: The modular design space used to generate systematic variations for reliable assessment.

selection strategy that optimizes the prompt set by balancing individual template fidelity with collective diversity. This approach transforms multi-prompt evaluation into a principled optimization problem, providing a solid foundation for monitoring training dynamics and conducting rigorous benchmarking of state-of-the-art models.

Our proposed framework enables researchers to obtain more reliable evaluation trajectories with respect to prompt-induced noise, and to gain a more accurate view of model potential. By introducing the first systematic formulation of evaluation scaling law, this work bridges the gap between theoretical principles and engineering practice, offering a principled and robust benchmark foundation for LLM assessment.

In summary, our core contributions are as follows:

- **The Evaluation Scaling Law.** We establish a theoretical and empirical foundation for multi-prompt evaluation, demonstrating that, in our experiments, evaluation noise diminishes smoothly and ranking monotonicity tends to improve with the diversity and quantity of prompt templates.
- **Quality-Diversity Prompt Selection.** By jointly maximizing individual reliability and inter-prompt diversity, we propose an automated prompt selection strategy that minimizes evaluation noise and reduces the need

for trial-and-error prompt engineering, ensuring maximum reliability gains per prompt.

- **Empirical Analysis and Engineering Guidance.** Through extensive experiments across multiple model families, we provide a rigorous validation of the Evaluation Scaling Law and the efficacy of our QD selection strategy. Our findings identify template diversity as a critical factor of evaluation reliability, offering a high-efficiency heuristic for generating smooth development trajectories and a principled roadmap for large-scale, high-fidelity LLM assessment.

2 Methodology

LLM evaluation is the systematic process of quantifying a model’s capabilities through standardized benchmarks. Formally, let $B = \{(q_i, a_i)\}_{i=1}^N$ denote a benchmark comprising N test samples, where each sample consists of a query q_i and a ground-truth reference a_i . In practice, the query q_i is typically wrapped in a prompt template p . Given an LLM M , the model generates a response $r_{i,p} = M(q_i, p)$. The evaluation seeks to estimate the performance metric E over B as:

$$E(B, M, p) = \frac{1}{N} \sum_{i=1}^N S(a_i, r_{i,p}) \quad (1)$$

where $S(\cdot, \cdot)$ is a task-specific scoring function. For discriminative tasks or multiple-choice bench-

marks, s is often an indicator function, yielding 1 if $r_{i,p}$ aligns with a_i and 0 otherwise.

Under this formulation, the resulting estimate E is inherently susceptible to variance and bias (Chen et al., 2025b), as its outcome is heavily contingent upon the specific choice of p —a phenomenon commonly referred to as prompt sensitivity. To mitigate these effects, we revisit the evaluation methodology and propose a multi-prompt framework as illustrated in Figure 1. Departing from the conventional paradigm that relies on a single prompt (Figure 1 (a)), our approach utilizes a modular candidate pool from which N_p distinct prompts are strategically selected (Figure 1 (b)). We first introduce the multi-prompt evaluation protocol, followed by a Quality-Diversity prompt selection strategy.

2.1 Multi-Prompt Evaluation Protocol

Single prompt evaluation typically suffers from high measurement instability, where model performance fluctuates significantly across different prompt templates for the same query. Following (Miller, 2024), while the sampling variance of a benchmark is often fixed by its size N , the conditional variance—stemming from prompt sensitivity—can be mitigated. To obtain a more reliable estimate of a model’s latent capability μ_i for each question q_i , we propose to marginalize performance over a set of N_p prompt templates.

Let r_{ij} denote the model’s response to question q_i ($i = 1, \dots, N$) when prompted with the j -th template ($j = 1, \dots, N_p$). We define the model’s latent capability on question q_i as μ_i , representing the expected performance across the prompt distribution. The observed score s_{ij} is expressed as $s_{ij} = \mu_i + \epsilon_{ij}$, where ϵ_{ij} denotes the measurement error associated with template j , and we assume $\mathbb{E}[\epsilon_{ij}] = 0$ and $\text{Var}(\epsilon_{ij}) = \sigma_j^2$. Notably, these error terms ϵ_{ij} are typically correlated across different templates.

Following (Miller, 2024), we adopt this zero-mean decomposition purely as a tractable modeling abstraction, not an exact description of prompt effects; the constant-variance and uniform-correlation assumptions are standard simplifications in this line, and ρ captures systematic effects that inflate or depress scores across the whole benchmark (Voronov et al., 2024). We note that the item-wise independence is a favorable idealization whose relaxation is left to future work.

By aggregating scores across N_p templates, the estimated capability for question q_i becomes: $\hat{\mu}_i =$

$\frac{1}{N_p} \sum_{j=1}^{N_p} s_{ij}$. The overall evaluation score $\bar{X} = \frac{1}{N} \sum_{i=1}^N \hat{\mu}_i$ can be decomposed as:

$$\bar{X} = \underbrace{\frac{1}{N} \sum_{i=1}^N \mu_i}_{\text{Latent Mean}} + \underbrace{\frac{1}{NN_p} \sum_{i=1}^N \sum_{j=1}^{N_p} \epsilon_{ij}}_{\text{Aggregate Error}} \quad (2)$$

Under the assumption of item-wise independence, the variance of the aggregate score $\bar{X}$ (conditioned on the fixed set of questions) is:

$$\text{Var}(\bar{X} \mid \{q_i\}_{i=1}^N) = \frac{1}{N^2} \sum_{i=1}^N \text{Var}(\hat{\mu}_i) \quad (3)$$

To characterize the behavior of $\text{Var}(\hat{\mu}_i)$, we assume constant error variance ($\sigma_j^2 = \sigma^2, \forall j$) and a uniform pairwise correlation ρ between prompt-induced errors, such that $\text{Corr}(\epsilon_{ij}, \epsilon_{ik}) = \rho$ for $j \neq k$. The per-item variance is then:

$$\begin{aligned} \text{Var}(\hat{\mu}_i) &= \frac{1}{N_p^2} \left[\sum_{j=1}^{N_p} \text{Var}(\epsilon_{ij}) + \sum_{j \neq k} \text{Cov}(\epsilon_{ij}, \epsilon_{ik}) \right] \\ &= \frac{\sigma^2}{N_p} [1 + (N_p - 1)\rho] \end{aligned} \quad (4)$$

Consequently, the reducible component of the evaluation variance is:

$$\text{Var}(\bar{X}) = \frac{\sigma^2}{NN_p} [1 + (N_p - 1)\rho] \quad (5)$$

This result highlights that the reduction of evaluation variance is fundamentally governed by the inter-prompt correlation ρ . Specifically, there exist three representative regimes of ρ ¹:

- **Case I: Ideal Independence** ($\rho = 0$). When prompt-induced errors are mutually independent, the total variance simplifies to $\text{Var}(\bar{X}) = \frac{\sigma^2}{NN_p}$. In this optimal regime, the variance decays at a rate of $O(1/N_p)$, where using N_p templates provides a precision gain equivalent to increasing the sample size N by a factor of N_p .
- **Case II: Degenerate Redundancy** ($\rho = 1$). If prompt templates are perfectly correlated—implying they trigger identical systematic biases or model idiosyncrasies—the vari-

¹Note that while ρ can theoretically be negative, in the context of LLM prompting, we typically observe $\rho \in [0, 1]$ due to the shared model weights and task knowledge.

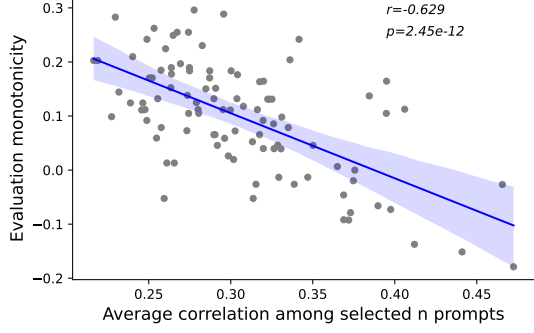


Figure 2: Negative correlation between evaluation monotonicity and inter-prompt correlation.

ance collapses to $\text{Var}(\bar{X}) = \frac{\sigma^2}{N}$, which is identical to the single-prompt baseline. In this scenario, marginalization yields zero statistical benefit.

- **Case III: Practical Partial Correlation** ($0 < \rho < 1$). In practice, LLMs often exhibit shared sensitivities across prompts, leading to a positive correlation $0 < \rho < 1$. Here, the variance reduction follows a law of diminishing returns. Notably, as $N_p \rightarrow \infty$, the variance does not vanish but converges to an irreducible asymptotic lower bound:

$$\lim_{N_p \rightarrow \infty} \text{Var}(\bar{X}) = \frac{\rho\sigma^2}{N} \quad (6)$$

To verify whether reducing inter-prompt correlation effectively enhances evaluation reliability, we first conduct a case study using the OLMo-2-1124-7B² model on the GPQA (Rein et al., 2023) benchmark. Since reliability is not directly observable, we employ evaluation monotonicity as a robust reference proxy (Madaan et al., 2024). This metric quantifies the framework’s ability to consistently recover the true progress of model checkpoints amidst stochastic noise, thereby serving as a high-fidelity indicator of assessment stability. The full sampling and evaluation procedure underlying Figure 2 is provided in appendix A.1.

As illustrated in Figure 2, our findings reveal a significant negative correlation between evaluation monotonicity and inter-prompt correlation. Here, the average correlation $\bar{\rho}$ is calculated as the mean pairwise correlation among N_p templates: $\bar{\rho} = \frac{2}{N_p(N_p-1)} \sum_{1 \leq i < j \leq N_p} \rho_{ij}$. Monotonicity is quantified using Kendall’s τ between evaluation

²<https://huggingface.co/allenai/OLMo-2-1124-7B>

Algorithm 1 Quality-Diversity Prompt Selection

Input: Candidate pool $\mathcal{P} = \{p_1, \dots, p_{|\mathcal{P}|}\}$, Anchor model M_{anc} , Target size N_p
Output: Recommended set $\mathcal{R}$
// Phase I: Quality-aware Filtering
 $\mathbf{S} \leftarrow \text{Evaluate}(M_{anc}, \mathcal{P})$ {Matrix $\mathbf{S} \in \mathbb{R}^{N \times |\mathcal{P}|}$ where columns are $\mathbf{s}_j$ }
 $\bar{\mathbf{s}} \leftarrow \{\bar{s}_1, \dots, \bar{s}_{|\mathcal{P}|}\}$ {Vector of mean scores across N questions}
 $IQR \leftarrow \text{Percentile}(\bar{\mathbf{s}}, 75) - \text{Percentile}(\bar{\mathbf{s}}, 25)$
 $s_{thres} \leftarrow \text{Percentile}(\bar{\mathbf{s}}, 25) - 1.5 \times IQR$
 $\mathcal{P}' \leftarrow \{p_j \in \mathcal{P} \mid \bar{s}_j \geq s_{thres}\}$
if $|\mathcal{P}'| < N_p$ **then**
 $\mathcal{P}' \leftarrow$ Top N_p prompts from $\mathcal{P}$ based on $\bar{s}_j$
end if
// Phase II: Diversity-based Selection
 $\mathbf{S}' \leftarrow$ columns of $\mathbf{S}$ corresponding to $\mathcal{P}'$
 $\mathcal{C} \leftarrow \text{Cluster}(\text{columns of } \mathbf{S}', n_clusters = N_p)$
 $\mathcal{R} \leftarrow \emptyset$
for each cluster $C_k \in \mathcal{C}$ **do**
 $\mathbf{s}^* \leftarrow \arg \min_{\mathbf{s} \in C_k} \sum_{\mathbf{u} \in C_k} \|\mathbf{s} - \mathbf{u}\|_2$ {Find the cluster medoid}
 $p^* \leftarrow \text{MapVectorToPrompt}(\mathbf{s}^*)$
 $\mathcal{R} \leftarrow \mathcal{R} \cup \{p^*\}$
end for
Return $\mathcal{R}$

scores and the chronological sequence of training checkpoints³. These results demonstrate that lower inter-prompt correlation directly translates to higher assessment stability, whereas highly redundant prompts fail to effectively suppress stochastic noise. This encourages us to select a more diverse subset of prompt templates.

2.2 Quality-Diversity Prompt Selection

We propose a two-stage selection strategy to curate a subset of N_p templates that maximize evaluation reliability. The core objective is to jointly minimize inter-prompt redundancy ρ and preserve high-fidelity measurement.

Before selection, we first characterize the candidate pool $\mathcal{P}$ by its empirical behavior. We employ an anchor model⁴ to evaluate all candidates across the benchmark, constructing an evaluation matrix $\mathbf{S} \in \mathbb{R}^{N \times |\mathcal{P}|}$, where s_{ij} denotes the model’s score on the i -th question using the j -th template. Each template $p_j \in \mathcal{P}$ is thus represented by its func-

³See appendix A.4 for more details

⁴Typically, selecting a checkpoint from the early stages of training is sufficient.

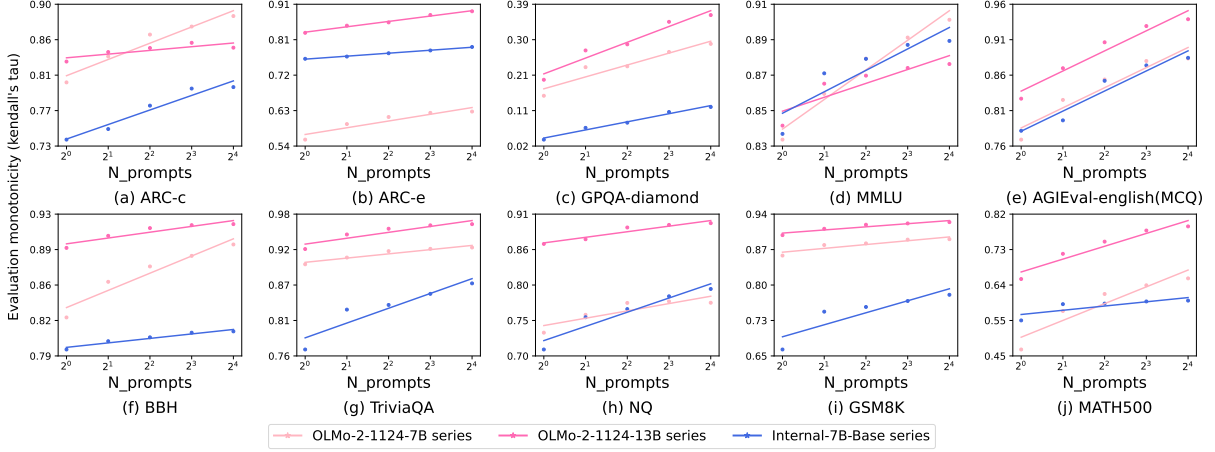


Figure 3: Evaluation scaling law: ranking monotonicity tends to improve with the number of evaluation prompts across all model series and benchmark settings, following a log-linear trend.

tional profile $\mathbf{s}_j = [s_{1,j}, \dots, s_{N,j}]^\top \in \mathbb{R}^N$ (the j -th column of $\mathbf{S}$), while its individual efficacy is captured by the mean score $\bar{s}_j = \frac{1}{N} \sum_i s_{ij}$. These notations form the basis for our selection procedure, which is formalized in Algorithm 1 and detailed below.

I. Quality-aware Filtering. To prevent low-quality prompts from introducing excessive bias, we first prune the candidate pool based on a reliability threshold. Specifically, we apply a boxplot-based outlier detection mechanism to the distribution of mean scores $\{\bar{s}_j\}_{p_j \in \mathcal{P}}$. Using the Interquartile Range (IQR), we identify and remove "pathological" templates that exhibit anomalous performance or fail to provide a stable signal. This stage ensures that the filtered subset $\mathcal{P}'$ consists exclusively of templates that offer high-fidelity measurements of the model's latent capability.

In cases where no outliers are detected, the entire pool is retained ($\mathcal{P}' = \mathcal{P}$). Conversely, to ensure the subsequent diversity-based selection operates on a sufficient set, if pruning results in $|\mathcal{P}'| < N_p$, we retain the top- N_p performing templates as the filtered pool. This heuristic guarantees that the selection process remains robust even when the initial candidate quality is highly variable.

II. Diversity-based Selection. To minimize the inter-prompt correlation ρ , we select N_p templates that exhibit maximally diverse response patterns. We perform clustering on the functional profiles $\{\mathbf{s}_j\}$ of the filtered candidates in $\mathcal{P}'$. These profiles serve as behavioral embeddings that capture the systematic biases and sensitivity signatures of each template across the benchmark distribution. Specifically, we partition the candidate space into N_p

clusters and select the medoid from each cluster to form the final evaluation set $\mathcal{R}$. By partitioning the prompt space based on observed model behavior rather than linguistic surface features, this strategy ensures that $\mathcal{R}$ represents "orthogonal" prompting strategies, thereby maximizing the variance reduction as derived in Eq. (5).

3 Experiments

3.1 Setup

Benchmarks. To provide a comprehensive assessment of our methodology, we select a diverse suite of benchmarks widely used in LLM pre-training, categorized into two primary task types: (1) Multiple-Choice Questions (MCQ): ARC-Challenge/Easy (ARC-c/e) (Clark et al., 2018), GPQA, MMLU (Hendrycks et al., 2021), and the English subset of AGIEval (Zhong et al., 2023); (2) Open-Ended Tasks: Big-Bench Hard (BBH) (Suzgun et al., 2022), TriviaQA (Joshi et al., 2017), Natural Questions (NQ) (Kwiatkowski et al., 2019), GSM8K (Cobbe et al., 2021), and MATH500 (Lightman et al., 2023). These benchmarks span critical domains including General Knowledge, STEM, Symbolic Reasoning, and Mathematics.

Models. Our framework primarily targets evaluation during the pre-training phase. Due to the scarcity of intermediate checkpoints in the open-source community, we validate our framework using a combination of publicly available and our internal models: the OLMo-2-1124 (7B and 13B) (OLMo, 2025) series and our internally pre-trained 7B models. See appendix A.2 and A.3 for more details.

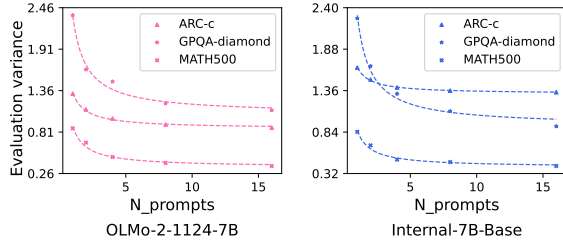


Figure 4: Variance reduction. Evaluation noise steadily decays as the number of prompts increases, demonstrating the framework’s capacity to suppress noise and stabilize performance estimates.

Metrics. Since evaluation reliability is not directly observable, we utilize evaluation monotonicity as our primary metric (Madaan et al., 2024). We employ Kendall’s τ to quantify the alignment between evaluation scores and the sequential checkpoints, where higher monotonicity indicates that the evaluation captures a more stable signal of genuine capability gains. Additionally, we monitor evaluation variance—estimated via performance fluctuations across successive checkpoints (Heineman et al., 2026)—as a proxy to directly quantify the suppression of prompt-induced noise. See appendix A.4 for details.

Prompt Candidate Pool. Inspired by prior works (Sclar et al., 2024; Luan et al., 2025), we construct a modular design space to generate a diverse candidate pool $\mathcal{P}$, as illustrated in Figure 1. Specifically, we implement a few-shot prompting protocol characterized by five fundamental dimensions. Detailed configurations of these factors are provided in appendix A.5, where we provide the corresponding templates and settings for reference. We quantify the diversity of the resulting candidate pool in appendix B.1.

Implementation Details. All experiments are conducted using OpenCompass (OpenCompass, 2023), and we adopt greedy decoding as the default sampling strategy. During evaluation, regex-based pattern matching is used for MCQ tasks, while xVerify (Chen et al., 2025a) is employed for open-ended tasks like mathematics.

3.2 Scaling Reliability Through Multi-Prompt Aggregation

To empirically verify whether marginalizing over multiple prompts enhances evaluation reliability, we conduct a systematic scaling analysis across three pre-trained model families over ten benchmarks. For a target ensemble size N_p , we execute

100 bootstrap trials. In each trial, N_p templates are sampled uniformly from the candidate pool $\mathcal{P}$, with the final evaluation score defined as the expectation across these samples. We then compute the evaluation monotonicity τ for each N_p to characterize the stability of the capability estimation.

As illustrated in Figure 3, the evaluation monotonicity τ tends to improve with the prompt ensemble size within the tested range. This relationship is well captured by a log-linear empirical fit: $\tau(N_p) = \alpha \cdot \log(N_p) + \beta$, detailed fitting parameters are provided in appendix B.2. This empirical scaling law confirms that prompt-induced noise is a systematic impediment that can be predictably suppressed. By increasing N_p , we effectively transition the evaluation metric from "stochastic jitter" to a "stable trajectory", providing a more grounded reflection of model evolution.

3.2.1 Theoretical Intuition

To analyze why employing more prompts enhances evaluation monotonicity, we first examine how multi-prompt evaluation affects the variance of the assessment. Following a similar bootstrap strategy, for each prompt combination N_p , we evaluate $n = 30$ consecutive checkpoints to estimate the evaluation variance⁵. As illustrated in Figure 4, our observations across multiple benchmarks—spanning both OLMo-2 and our internal pre-trained 7B checkpoints—consistently demonstrate that evaluation variance is effectively suppressed as the ensemble size N_p increases, following the decay pattern derived previously.

To explain the monotonicity gain, we model Kendall’s τ through the lens of the Signal-to-Noise Ratio (SNR). For any two checkpoints with a latent capability gap $\Delta\mu$, the probability of a concordant ranking is $P_c = P(\Delta\mu + \Delta\epsilon > 0)$, where $\Delta\epsilon \sim \mathcal{N}(0, 2\sigma_{N_p}^2)$. As σ_{N_p} decays with N_p per Eq. (5), the $\text{SNR} = \frac{\Delta\mu}{\sqrt{2}\sigma_{N_p}}$ increases. Since τ is a linear transformation of P_c (Eq. 7), its growth is governed by the standard normal CDF $\Phi(\text{SNR})$. The inherent diminishing marginal returns of the CDF explain the sub-linear trajectory of $\tau(N_p)$, which is parsimoniously captured by our log-linear scaling law. Here, the coefficient α represents the benchmark-specific efficiency of converting prompt aggregation into evaluation certainty.

A supplementary smaller-model observation,

⁵The OLMo-2-1124-7B checkpoints used here are detailed in appendix A.2

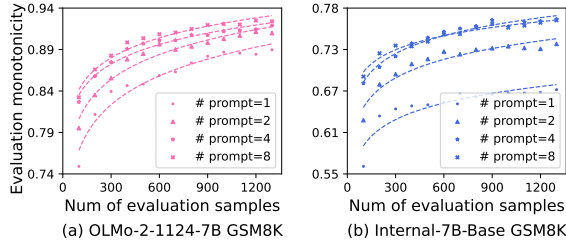


Figure 5: Sample-prompt tradeoff. Given their interchangeability, using more diverse prompts boosts the reliability of smaller benchmarks, allowing them to achieve the same high fidelity as much larger datasets.

corroborating the log-linear trend of Fig. 3, is provided in appendix B.3.

3.3 Sample-Prompt Trade-off

The log-linear scaling observed previously suggests that reliability gains require a proportional increase in computational budget. However, for large-scale benchmarks, evaluating multiple prompts across the entire dataset can be prohibitively expensive. We thus investigate a budget-reallocation strategy: whether comparable reliability can be achieved by reducing the sample size N_s while increasing the prompt ensemble size N_p .

Taking GSM8K ($N = 1319$) as a representative case, Figure 5 illustrates a clear substitutability between sample size and prompt numbers. Specifically, we observe that evaluation monotonicity is not solely determined by the total number of queries ($N_s \times N_p$), but by the strategic balance between them. Notably, for the Internal-7B-Base model, evaluating a subsampled set of 100 queries with 8 prompts yields a higher Kendall’s τ than evaluating the full benchmark with a single prompt, despite a $\sim 40\%$ reduction in total inference cost.

These findings offer two strategic insights for large-scale model assessment: 1) For tasks with limited intrinsic sample sizes, marginalizing over a diverse prompt set serves as a robust mechanism to synthesize evaluation stability that would otherwise require a much larger dataset. 2) For expansive benchmarks, a "subsampled multi-prompt" strategy is often superior to a "full-sample single-prompt" approach. The former not only attains equivalent or superior stability at a lower cost but also enables secondary analyses such as prompt sensitivity and robustness profiling (Chatterjee et al., 2024).

3.4 Effectiveness and Ablation Study of Quality-Diversity Selection

The preceding experiments establish that larger prompt ensembles enhance reliability. However, given a vast candidate pool $\mathcal{P}$, selecting an optimal subset $\mathcal{R}$ of size N_p presents a non-trivial selection dilemma. We evaluate our Quality-Diversity (QD) selection strategy (Algorithm 1) against a stochastic Random Selection baseline. Furthermore, to dissect the individual contributions of our selection components, we conduct an ablation study focusing on two variants: (1) Q-Filter only, which applies quality-aware pruning without diversity considerations, and (2) D-Selection only, which performs functional clustering directly on the raw candidate pool.

As summarized in Table 1, our QD-selection framework consistently achieves superior results, outperforming alternative methods in 15 out of 20 benchmark settings. Notably, while the individual Q-Filter and D-Selection components occasionally yield competitive results, the Random Selection baseline fails to secure a single win. This disparity underscores the necessity of our dual-objective approach: the Q-Filter acts as a reliable signal gatekeeper, while D-Selection serves as a statistical decorrelation mechanism. By balancing individual reliability with collective diversity, our heuristic identifies a more informative and orthogonal subset of the prompt space than stochastic sampling, thereby maximizing the information gain per evaluation trial.

4 Related Work

4.1 Prompt Sensitivity and Evaluation Reliability

The sensitivity of LLMs to prompt templates is well-documented and remains a primary obstacle to objective evaluation. Research in (Voronov et al., 2024; Sclar et al., 2024) demonstrates that even subtle variations in formatting or phrasing can lead to substantial performance discrepancies, occasionally degrading the accuracy of state-of-the-art models to suboptimal levels. This volatility has been further quantified by the Prompt Sensitivity Index (POSIX) in (Chatterjee et al., 2024), which measures model variance across intent-preserving variations, highlighting the fragility of single-prompt assessments.

To mitigate the instability of single-template evaluations, recent research is increasingly shifting

Benchmark	OLMo-2-1124-7B				Internal-7B-Base			
	Random	Q-Filter	D-Selection	QD-Selection	Random	Q-Filter	D-Selection	QD-Selection
ARC-c	0.853	0.830	0.853	0.856	0.798	0.765	0.804	0.826
ARC-e	0.878	0.922	0.843	0.948	0.786	0.817	0.776	0.778
GPQA-D	0.344	0.294	0.359	0.406	0.109	-0.203	0.184	0.294
MMLU	0.871	0.885	0.895	0.895	0.885	0.908	0.830	0.886
AGIEval	0.925	0.869	0.948	0.948	0.870	0.869	0.882	0.908
BBH	0.918	0.922	0.908	0.943	0.810	0.804	0.817	0.843
TriviaQA	0.961	0.961	0.961	0.974	0.854	0.932	0.817	0.934
NQ	0.891	0.908	0.869	0.895	0.786	0.708	0.765	0.856
GSM8K	0.923	0.974	0.935	0.922	0.765	0.725	0.739	0.791
MATH500	0.781	0.778	0.734	0.790	0.595	0.647	0.556	0.608

Table 1: Ablation study of prompt selection strategies. Performance (τ) is reported at $N_p = 8$ to compare QD Selection against Random Selection and its ablated variants (Quality-only and Diversity-only). Bold indicates the best performance.

toward multi-prompt aggregation. (Mizrahi et al., 2024) argues that single-prompt benchmarks inevitably produce inconsistent model rankings and advocates for the use of diverse instruction templates to capture a more comprehensive view of model capabilities. Similarly, PromptEval (Polo et al., 2024) introduces a framework to estimate performance distributions and quantile functions under constrained budgets by leveraging the collective strength of various prompt templates and examples. These methods aim to reduce the “luck of the draw” in prompt selection, fostering more reliable evaluations.

Beyond strategy design, specific efforts have focused on quantifying evaluation reliability. (Madaan et al., 2024) establishes a suite of metrics to measure benchmark variance, particularly ranking monotonicity during pre-training. This line of inquiry has been expanded by studies on the statistical robustness of LLM leaderboards (Luan et al., 2025; Li et al., 2025b).

However, while current literature acknowledges prompt sensitivity and explores basic aggregation, the principled link between multi-prompt diversity and evaluation fidelity remains uncharacterized, and systematic strategies for prompt selection remain relatively under-developed.

4.2 Scaling Laws

Scaling laws increasingly govern the development of LLMs, providing a framework to characterize predictable performance gains as a function of resource investment. These empirical observations have transformed LLM development from a heuristic-based trial-and-error process into a principled engineering discipline across several critical dimensions (Li et al., 2025a). Initially, research

focused on the pre-training phase, where (Kaplan et al., 2020) identified power-law relationships between cross-entropy loss and the scales of compute, parameters, and data. These findings were subsequently refined by (Hoffmann et al., 2022), whose “Chinchilla scaling laws” established the optimal model-data ratio for compute-efficient training. Beyond pre-training, similar patterns have been observed during alignment. Research into RLHF and reward modeling (Rafailov et al., 2024) suggests that the quality of preference alignment scales with the size of the reward model (RM), the volume of RM training data, and the policy size. More recently, this paradigm has been extended to inference, where performance is no longer constrained by the training budget. As demonstrated by OpenAI’s reasoning models (OpenAI, 2024), increasing test-time computation—through strategies such as Chain-of-Thought (CoT) and search-based decoding—yields predictable gains in complex problem-solving. This shift implies that an LLM’s scale is a dynamic property that extends into the execution phase (Snell et al., 2025).

Despite the comprehensive mapping of scaling laws across training, alignment, and inference, the evaluation phase remains a relatively underexplored frontier. Our work pioneers the investigation into Evaluation Scaling Laws, specifically for pre-training. By formalizing how prompt count and diversity govern evaluation fidelity, we provide a principled foundation for more reliable benchmarking.

5 Conclusion

In this paper, we address the critical bottleneck of prompt sensitivity in LLM evaluation, where reliance on a single template compromises relia-

bility due to high stochastic variance and a “narrow filter” effect that obscures latent capabilities. To resolve this, we propose a multi-prompt evaluation framework guided by a Quality-Diversity (QD) selection strategy. By aggregating results across a systematically curated prompt distribution, we shift from a single-prompt point-estimate toward a more reliable estimate with respect to prompt-induced variance. Extensive experiments—using evaluation variance and ranking monotonicity (τ) as metrics—reveal that reliability tends to improve with prompt diversity within the tested range. This log-linear scaling pattern demonstrates that multi-prompt diversity provides the critical foundation for consistent and unbiased model evaluation.

We highly recommend integrating multi-prompt evaluation into standard LLM benchmarking. By mitigating prompt-specific biases, it provides a principled foundation for monitoring training progress, ensuring that model comparisons reflect genuine capabilities rather than stochastic gains from a lucky choice of words.

Limitations

While this study represents a pioneering effort to establish evaluation scaling laws for LLMs, several limitations deserve further investigation. First, our empirical scope centers on dense models during pre-training; extending the framework to Mixture-of-Experts (MoE) and post-training stages remains future work. Our primary validation covers 7B and 13B models (with a supplementary smaller-model observation in the appendix), and whether the trend holds for larger scales (e.g., 32B+) remains open. Second, we do not prescribe a universal, “optimal” N_p : the ideal number of prompts balances budget against reliability, and marginal utility may diminish for highly robust or near-saturated models. Third, evaluation noise is not directly observable, so we estimate it from performance variation across consecutive checkpoints; this approximation may partly conflate prompt-induced noise with genuine training fluctuation, and prompt variation is only one of several noise sources (e.g., dataset sampling, optimization stochasticity). Our conclusions therefore indicate relative, not absolute, noise reduction. Finally, the sample-prompt trade-off assumes a simplified cost model that ignores request-level caching and KV-cache reuse; we nonetheless regard it as reliability-first. Future work will refine the QD strategy for residual edge cases and broader

evaluation scenarios.

Acknowledgements

We sincerely thank the anonymous reviewers and the area chairs for their careful reading and constructive suggestions, which helped improve the clarity and rigor of this paper. During preparation, we used generative AI tools (e.g. Gemini-3.1-pro, Claude Code) for language polishing and formatting assistance; all authors have reviewed the content and take full responsibility for the final version.

References

- Tom B. Brown, Benjamin Mann, Nick Ryder, and et al. 2020. [Language models are few-shot learners](#). *Preprint*, arXiv:2005.14165.
- Anwoy Chatterjee, H S V N S Kowndinya Renduchintala, Sumit Bhatia, and Tanmoy Chakraborty. 2024. [POSIX: A prompt sensitivity index for large language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14550–14565, Miami, Florida, USA. Association for Computational Linguistics.
- Ding Chen, Qingchen Yu, Pengyuan Wang, Mengting Hu, Wentao Zhang, Zhengren Wang, Bo Tang, Feiyu Xiong, Xinchu Li, Chao Wang, Minchuan Yang, and Zhiyu Li. 2025a. [xVerify: Efficient answer verifier for reasoning model evaluations](#). *Preprint*, arXiv:2504.10481.
- Meilin Chen, Jian Tian, Liang Ma, Di Xie, Weijie Chen, and Jiang Zhu. 2025b. [Unbiased evaluation of large language models from a causal perspective](#). *Preprint*, arXiv:2502.06655.
- Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. [Think you have solved question answering? try arc, the ai2 reasoning challenge](#). *Preprint*, arXiv:1803.05457.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. [Training verifiers to solve math word problems](#). *Preprint*, arXiv:2110.14168.
- Google DeepMind. 2026. [Gemini 3.1 pro](#). Technical report, Google DeepMind.
- David Heineman, Valentin Hofmann, Ian Magnusson, Yuling Gu, Noah Smith, Hanna Hajishirzi, Kyle Lo, and Jesse Dodge. 2026. Signal and noise: A framework for reducing uncertainty in language model evaluation. *Advances in Neural Information Processing Systems*, 38:17073–17114.

- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, and 3 others. 2022. [Training compute-optimal large language models](#). *Preprint*, arXiv:2203.15556.
- Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. [TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.
- Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. [Scaling laws for neural language models](#). *Preprint*, arXiv:2001.08361.
- M. Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30:81–93.
- Tom Kwiakowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, and 1 others. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.
- Margaret Li, Sneha Kudugunta, and Luke Zettlemoyer. 2025a. [\(mis\)fitting: A survey of scaling laws](#). *Preprint*, arXiv:2502.18969.
- Yiyang Li, Yonghuang Wu, Ying Luo, Liangtai Sun, Zishu Qin, Lin Qiu, Xuezhi Cao, and Xunliang Cai. 2025b. [Instance-level randomization: Toward more stable LLM evaluations](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 3411–3425, Suzhou, China. Association for Computational Linguistics.
- Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. [Let’s verify step by step](#). *Preprint*, arXiv:2305.20050.
- LingTeam. 2025. [Every activation boosted: Scaling general reasoner to 1 trillion open language foundation](#). *Preprint*, arXiv:2510.22115.
- Hongzhi Luan, Changxin Tian, Zhaoxin Huan, Xiaolu Zhang, Kunlong Chen, Zhiqiang Zhang, and Jun Zhou. 2025. [BOSE: A systematic evaluation method optimized for base models](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16147–16158, Vienna, Austria. Association for Computational Linguistics.
- Lovish Madaan, Aaditya K. Singh, Rylan Schaeffer, Andrew Poulton, Sanmi Koyejo, Pontus Stenetorp, Sharan Narang, and Dieuwke Hupkes. 2024. [Quantifying variance in evaluation benchmarks](#). *Preprint*, arXiv:2406.10229.
- Evan Miller. 2024. [Adding error bars to evals: A statistical approach to language model evaluations](#). *Preprint*, arXiv:2411.00640.
- Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. 2024. [State of what art? a call for multi-prompt llm evaluation](#). *Preprint*, arXiv:2401.00595.
- Shiwen Ni, Guhong Chen, Shuaimin Li, Xuanang Chen, Siyi Li, Bingli Wang, Qiyao Wang, Xingjian Wang, Yifan Zhang, Liyang Fan, Chengming Li, Ruifeng Xu, Le Sun, and Min Yang. 2025. [A survey on large language model benchmarks](#). *Preprint*, arXiv:2508.15361.
- Team OLMo. 2025. [2 OLMo 2 furious](#). *Preprint*, arXiv:2501.00656.
- OpenAI. 2024. Learning to reason with llms. <https://openai.com/index/learning-to-reason-with-llms/>. Accessed: 2024-05-22.
- OpenAI. 2026. [GPT-5.4 Thinking System Card](#). Technical report, OpenAI. Accessed: 2026-03-05.
- OpenCompass. 2023. OpenCompass: A universal evaluation platform for foundation models. <https://github.com/open-compass/opencompass>.
- Felipe M Polo, Ronald Xu, Lucas Weber, Mírian Silva, Onkar Bhardwaj, Leshem Choshen, Allysson F de Oliveira, Yuekai Sun, and Mikhail Yurochkin. 2024. Efficient multi-prompt evaluation of llms. *Advances in Neural Information Processing Systems*, 37:22483–22512.
- QwenTeam. 2026. [Qwen3.5: Accelerating productivity with native multimodal agents](#).
- Rafael Rafailov, Yaswanth Chittepudi, Ryan Park, Harshit Sikchi, Joey Hejna, Bradley Knox, Chelsea Finn, and Scott Niekum. 2024. [Scaling laws for reward model overoptimization in direct alignment algorithms](#). *Preprint*, arXiv:2406.02900.
- David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2023. [GPQA: A graduate-level google-proof qa benchmark](#). *Preprint*, arXiv:2311.12022.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. [Quantifying language models’ sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting](#). *Preprint*, arXiv:2310.11324.

Bangzhao Shu, Lechen Zhang, Minje Choi, Lavinia Dunagan, Lajanugen Logeswaran, Moontae Lee, Dallas Card, and David Jurgens. 2024. [You don’t need a personality test to know these models are unreliable: Assessing the reliability of large language models on psychometric instruments](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5263–5281, Mexico City, Mexico. Association for Computational Linguistics.

Charlie Victor Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2025. [Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning](#). In *The Thirteenth International Conference on Learning Representations*.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. 2022. [Challenging big-bench tasks and whether chain-of-thought can solve them](#). *Preprint*, arXiv:2210.09261.

Anton Voronov, Lena Wolf, and Max Ryabinin. 2024. [Mind your format: Towards consistent evaluation of in-context learning improvements](#). *Preprint*, arXiv:2401.06766.

Jiayi Yuan, Jiamu Zhang, Andrew Wen, and Xia Hu. 2025. [The science of evaluating foundation models](#). *Preprint*, arXiv:2502.09670.

Kun Zhang, Le Wu, Kui Yu, Guangyi Lv, and Dacao Zhang. 2025. [Evaluating and improving robustness in large language models: A survey and future directions](#). *Preprint*, arXiv:2506.11111.

Wanjuan Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. [AGIEval: A human-centric benchmark for evaluating foundation models](#). *Preprint*, arXiv:2304.06364.

A Experiment Setup Details

A.1 Computational Details of Figure 2

Figure 2 reports a case study on the OLMo-2-1124-7B series over GPQA. To obtain each scatter point, we take the first training checkpoint as the anchor model, randomly sample an 8-template ensemble from the 32-template MCQ candidate pool, and evaluate the anchor on this ensemble to obtain its behavioral profile (the vector of scores across templates). The average inter-prompt correlation is then computed as the mean of the

pairwise Pearson correlations of these profiles, $\bar{\rho} = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} \rho_{ij}$ with $n = 8$. Ranking monotonicity is measured by evaluating all 18 selected checkpoints on the same ensemble and computing Kendall’s τ . Repeating this sampling-and-evaluation procedure 100 times yields the 100 points $(\bar{\rho}, \tau)$ shown in Figure 2.

A.2 OLMo-2 Intermediate Checkpoints

To evaluate ranking monotonicity across training stages, we utilize 18 intermediate checkpoints for both OLMo-2-1124-7B and OLMo-2-1124-13B. Each checkpoint represents a distinct phase of pre-training, with specific Hugging Face revisions detailed in Table 2.

A.3 Architecture and Pretraining Setup of Internal 7B Model

We use a GQA (Grouped Query Attention) architecture based on the standard decoder-only Transformer, comprising an embedding layer, alternating layers of attention mechanisms and feed-forward networks. Positional information is handled using RoPE (Rotary Positional Embedding). For training parameters, the pre-trained models are initialized with a standard deviation of 0.006 and optimized using AdamW, with hyperparameters set to $\beta_1 = 0.9$, $\beta_2 = 0.95$, $\epsilon = 1e^{-8}$ and $\text{weight_decay} = 0.1$. The learning rate follows a WSD (Warm-up, Stabilization, Decay) strategy, where the first 1% of training steps involve linear warm-up. Further architectural and pretraining details are available in Table 4.

To ensure consistency with the OLMo-2 evaluation, we select 18 intermediate checkpoints across a total pre-training volume of 5T tokens.

A.4 Metric Details

Ranking Monotonicity: We employ Kendall’s rank correlation coefficient (τ) to quantify the consistency between evaluation scores and the chronological order of training checkpoints. For a sequence of n checkpoints, τ is calculated as:

$$\tau = \frac{4P_c}{n(n-1)} - 1 \quad (7)$$

where P_c is the number of concordant pairs. In this study, we utilize 18 checkpoints to assess evaluation monotonicity. A higher τ signifies that the evaluation more faithfully reflects the steady progression of model capability by minimizing sensitivity to prompt-level noise. Note that τ serves

index	OLMo-2-1124-7B	OLMo-2-1124-13B
0	stage1-step50000-tokens210B	stage1-step30000-tokens252B
1	stage1-step101000-tokens424B	stage1-step60000-tokens504B
2	stage1-step150000-tokens630B	stage1-step90000-tokens755B
3	stage1-step200000-tokens839B	stage1-step120000-tokens1007B
4	stage1-step250000-tokens1049B	stage1-step150000-tokens1259B
5	stage1-step300000-tokens1259B	stage1-step180000-tokens1510B
6	stage1-step350000-tokens1469B	stage1-step210000-tokens1762B
7	stage1-step400000-tokens1678B	stage1-step240000-tokens2014B
8	stage1-step450000-tokens1888B	stage1-step270000-tokens2265B
9	stage1-step500000-tokens2098B	stage1-step300000-tokens2517B
10	stage1-step550000-tokens2307B	stage1-step330000-tokens2769B
11	stage1-step600000-tokens2517B	stage1-step360000-tokens3020B
12	stage1-step650000-tokens2727B	stage1-step390000-tokens3272B
13	stage1-step700000-tokens2937B	stage1-step420000-tokens3524B
14	stage1-step750000-tokens3146B	stage1-step450000-tokens3775B
15	stage1-step800000-tokens3356B	stage1-step480000-tokens4027B
16	stage1-step850000-tokens3566B	stage1-step510000-tokens4279B
17	stage1-step900000-tokens3775B	stage1-step540000-tokens4530B

Table 2: OLMo-2 intermediate checkpoints used for ranking monotonicity evaluation.

here as a proxy for measurement fidelity rather than an assumption of strictly monotonic model improvement.

Evaluation Variance: To quantify the variance of the assessment, we utilize the evaluation noise across a sequence of $n = 30$ consecutive model checkpoints as a proxy. This metric captures the "jitter" in scores that often obscures genuine training progress. Specifically, we define the evaluation variance as the standard deviation (σ) of the scores within this window (Heineman et al., 2026):

$$\sigma = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (s_i - \bar{s})^2} \quad (8)$$

where s_i denotes the evaluation score at checkpoint i , and $\bar{s}$ is the mean score over the window. By calculating σ over a 30-point window, we ensure sufficient statistical power to distinguish stochastic prompt-induced noise from the underlying model improvement. A lower σ indicates a smoother, more reliable evaluation trajectory.

A.5 Prompt Configuration

We construct prompt templates from five dimensions, and ensure that the initial pool $\mathcal{P}$ is sufficiently diverse to represent the broader prompt distribution, providing a robust foundation for the subsequent Quality-Diversity selection, as detailed in Table 5.

- **Evaluation Mode:** We consider both generation-based (decoding the answer

string) and perplexity-based (Brown et al., 2020) modes (Exclusively applied for MCQ tasks). This dimension captures the model’s varying proficiency in discriminative versus generative contexts.

- **Instruction Prefix:** We toggle the presence of task-specific instruction prefixes (e.g., "Answer the following question..."). This factor assesses whether the model requires explicit task grounding to activate its latent knowledge.
- **In-context Exemplars (n -shot):** We vary the number of few-shot examples provided in the context. This dimension modulates the strength of the in-context signal and its efficacy in variance reduction.
- **Task Descriptors:** We utilize a set of semantically equivalent but lexically diverse keywords to describe the task (e.g., using "Question:" vs. "Problem:" vs. "Q:"). These variations are critical for identifying lexical biases in the model’s response patterns.
- **Structural Separator:** We employ different formatting delimiters to separate descriptors and query contents (e.g., newlines, spaces, or specific symbols). This factor probes the model’s sensitivity to the structural layout of the input.

The prompt templates used in our experiments are generated by systematically combining the di-

index	OLMo-2-1124-7B	index	OLMo-2-1124-7B
0	stage1-step900000-tokens3775B	15	stage1-step915000-tokens3838B
1	stage1-step901000-tokens3780B	16	stage1-step916000-tokens3842B
2	stage1-step902000-tokens3784B	17	stage1-step917000-tokens3847B
3	stage1-step903000-tokens3788B	18	stage1-step918000-tokens3851B
4	stage1-step904000-tokens3792B	19	stage1-step919000-tokens3855B
5	stage1-step905000-tokens3796B	20	stage1-step920000-tokens3859B
6	stage1-step906000-tokens3801B	21	stage1-step921000-tokens3863B
7	stage1-step907000-tokens3805B	22	stage1-step922000-tokens3868B
8	stage1-step908000-tokens3809B	23	stage1-step923000-tokens3872B
9	stage1-step909000-tokens3813B	24	stage1-step924000-tokens3876B
10	stage1-step910000-tokens3817B	25	stage1-step925000-tokens3880B
11	stage1-step911000-tokens3822B	26	stage1-step926000-tokens3884B
12	stage1-step912000-tokens3826B	27	stage1-step927000-tokens3889B
13	stage1-step913000-tokens3830B	28	stage1-step928000-tokens3893B
14	stage1-step914000-tokens3834B	29	stage1-step928646-tokens3896B

Table 3: OLMo-2 intermediate checkpoints selected for evaluation variance analysis.

	Internal-7B-Base
vocab size	126464
layer num	28
hidden size	4096
intermediate size	14336
attention heads	32
key value heads	8
# Para. (B)	7.39
# Non-emb. Para. (B)	6.11
sequence length	4096
learning rate	5.23E-04
batch size	1227

Table 4: Architectural specifications and training hyperparameters of the internal 7B model.

Question: <question>
Options:
(A) <textA>
(B) <textB>
(C) <textC>
(D) <textD>
Answer: the answer is <answer>

Question: <question>
Options:
(A) <textA>
(B) <textB>
(C) <textC>
(D) <textD>
Answer: the answer is

mensions described above. To illustrate this, Figures 6 through 8 serve as examples, showcasing three prompt templates with varying configurations.

B Additional Results

B.1 Quantifying Candidate-Pool Diversity

To complement the construction described above, Table 7 quantifies the spread of prompt-wise scores on Internal-7B-Base, evidencing that the templates capture meaningfully different behaviors rather than merely superficial reformatting.

These spreads — up to 11.68 points on GSM8K — indicate that prompt choice materially changes model behavior, reaffirming both the premise of the multi-prompt protocol and the room for diversity-aware selection.

Figure 6: Prompt Template 1 for MCQ tasks. It features a prefix-free, one-shot configuration with descriptors ("Question", "Answer") and a space separator, applicable to both gen-based and ppl-based modes.

B.2 Details of Functional Fitting

We model the relationship using the objective function $\tau = \alpha \cdot \log(N_P) + \beta$, where the parameters were estimated via the *scipy.optimize.curve_fit* function. For each experimental group, we report the fitted values of α and β along with their respective standard errors to account for fitting uncertainty. Additionally, we report the coefficient of determination (R^2) to assess the goodness-of-fit. The consistently high R^2 values across all experimental groups (as shown in Table 6) provide strong

factor	MCQ	Open-ended
mode	[Gen, PPL]	[Gen]
with prefix	[True, False]	[True, False]
# shots	[1, 3]	[1, 3]
descriptors	[("Question", "Answer"), ("Problem", "Solution")]	[("Question", "Answer"), ("Problem", "Solution")], ("Q", "A")
separator	["\n", " "]	["\n", " "]
total	32	24

Table 5: Prompt configurations for different tasks.

Benchmark	OLMo-2-1124-7B			OLMo-2-1124-13B			Internal-7B-Base		
	α	β	R^2	α	β	R^2	α	β	R^2
ARC-c	0.028 $\pm$ 3.9e-03	0.814 $\pm$ 6.6e-03	0.944	0.006 $\pm$ 2.4e-03	0.835 $\pm$ 4.1e-03	0.691	0.025 $\pm$ 3.5e-03	0.739 $\pm$ 5.9e-03	0.944
ARC-e	0.025 $\pm$ 5.9e-03	0.567 $\pm$ 1.0e-02	0.860	0.020 $\pm$ 1.4e-03	0.834 $\pm$ 2.4e-03	0.986	0.011 $\pm$ 4.4e-04	0.763 $\pm$ 7.5e-04	0.995
GPQA-D	0.045 $\pm$ 8.5e-03	0.169 $\pm$ 1.4e-02	0.902	0.059 $\pm$ 8.1e-03	0.208 $\pm$ 1.4e-02	0.948	0.031 $\pm$ 2.4e-03	0.041 $\pm$ 4.1e-03	0.982
MMLU	0.025 $\pm$ 2.9e-03	0.836 $\pm$ 4.9e-03	0.961	0.012 $\pm$ 3.5e-03	0.846 $\pm$ 6.0e-03	0.785	0.018 $\pm$ 5.1e-03	0.845 $\pm$ 8.6e-03	0.807
AGIEval	0.041 $\pm$ 7.7e-03	0.782 $\pm$ 1.3e-02	0.904	0.041 $\pm$ 5.7e-03	0.834 $\pm$ 9.6e-03	0.945	0.041 $\pm$ 6.1e-03	0.778 $\pm$ 1.0e-02	0.937
BBH	0.025 $\pm$ 4.2e-03	0.835 $\pm$ 7.1e-03	0.922	0.008 $\pm$ 1.9e-03	0.900 $\pm$ 3.3e-03	0.866	0.006 $\pm$ 9.8e-04	0.795 $\pm$ 1.7e-03	0.936
TriviaQA	0.010 $\pm$ 1.7e-03	0.904 $\pm$ 2.9e-03	0.915	0.013 $\pm$ 3.3e-03	0.932 $\pm$ 5.7e-03	0.841	0.033 $\pm$ 7.7e-03	0.785 $\pm$ 1.3e-02	0.862
NQ	0.016 $\pm$ 5.1e-03	0.744 $\pm$ 8.6e-03	0.756	0.012 $\pm$ 2.2e-03	0.865 $\pm$ 3.7e-03	0.909	0.030 $\pm$ 5.5e-03	0.721 $\pm$ 9.3e-03	0.909
GSM8K	0.011 $\pm$ 3.0e-03	0.864 $\pm$ 5.0e-03	0.827	0.009 $\pm$ 1.9e-03	0.903 $\pm$ 3.3e-03	0.886	0.035 $\pm$ 1.1e-02	0.692 $\pm$ 1.8e-02	0.781
MATH500	0.063 $\pm$ 1.4e-02	0.501 $\pm$ 2.4e-02	0.875	0.049 $\pm$ 8.2e-03	0.672 $\pm$ 1.4e-02	0.921	0.016 $\pm$ 6.5e-03	0.561 $\pm$ 1.1e-02	0.671

Table 6: Parameter estimation for the logarithmic scaling model. Standard errors and R^2 values are reported to quantify fitting uncertainty and goodness-of-fit.

Answer the following multiple choice questions. Question: <question> Options: (A) <textA> (B) <textB> (C) <textC> (D) <textD> Answer: the answer is <answer> Question: <question> Options: (A) <textA> (B) <textB> (C) <textC> (D) <textD> Answer: the answer is	Following the given example, answer the question last and put your final answer within <code>\boxed{}</code>. Problem: <question> Solution: <answer> Problem: <question> Solution: <answer> Problem: <question> Solution: <answer> Problem: <question> Solution:
---	---

Figure 7: Prompt Template 2 for MCQ tasks. This prefixed, one-shot template utilizes descriptors ("Question", "Answer") and a newline separator, supporting both gen-based and ppl-based evaluation.

Figure 8: Prompt Template 3 for open-ended tasks, utilizing a prefixed, 3-shot setup with ("Problem", "Solution") as descriptors and a space separator. Exclusively compatible with the gen-based mode.

empirical evidence that the log-linear model accurately captures the scaling relationship between prompt count and evaluation reliability. Notably, a higher value of α indicates a more pronounced performance gain derived from multi-prompt evaluation.

Table 7: Score variability across candidate prompts (Internal-7B-Base) evidences non-trivial template diversity.

Benchmark	Min	Max	Mean	Std
ARC-c	75.93	81.36	79.51	1.41
AGIEval	37.76	45.95	41.69	2.60
GSM8K	59.36	71.04	64.54	3.00
MATH500	30.40	35.80	33.13	1.40

Table 8: Ranking monotonicity τ of OLMo-2-0425-1B (supplementary smaller model) as a function of the number of aggregated prompts N_p .

Benchmark	N_p				
	1	2	4	8	16
ARC-c	0.392	0.467	0.535	0.566	0.576
ARC-e	0.069	0.072	0.096	0.109	0.125
AGIEval	0.106	0.194	0.249	0.268	0.343
NQ	0.648	0.706	0.723	0.758	0.771
GSM8K	0.467	0.509	0.601	0.608	0.627
MATH500	0.149	0.179	0.201	0.211	0.234

B.3 Scaling Trends for Smaller Model Sizes

To provide further evidence that the scaling relationship is not specific to a single model size, we additionally evaluate a smaller model, OLMo-2-0425-1B, on the six benchmarks for which its intermediate checkpoints are available. Table 8 reports the resulting ranking monotonicity τ as the number of aggregated prompts N_p grows. The upward trend mirrors the 7B/13B results (Fig. 3), indicating that the log-linear behavior is not an artifact of a particular scale.

B.4 Illustration of the Smoothing Effect

Figure 9 illustrates how marginalization enhances evaluation stability on MATH500. While individual prompt trajectories (gray) exhibit significant point-wise divergence and temporal jitter, the marginalized ensemble (colored) effectively filters out these stochastic oscillations. This smoothing demonstrates that multi-prompt evaluation recovers a more reliable, monotonic signal of model improvement by suppressing the idiosyncratic noise of individual templates.

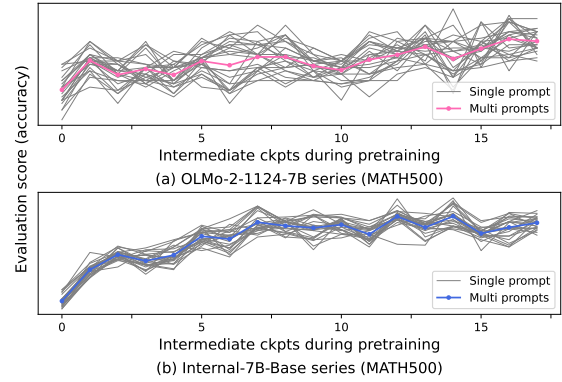


Figure 9: Comparison of performance trajectories. Individual prompts (gray) exhibit significant stochasticity, whereas the marginalized ensemble (colored) yields a substantially smoother and more monotonic progression. This visualization underscores the efficacy of multi-prompt evaluation in mitigating idiosyncratic noise to reveal genuine model improvement.