# BEYOND TOKEN ENTROPY: MEASURING DISTINCT FORMS OF OUTPUT VARIATION IN LLM REASONING

**Anonymous authors**

Paper under double-blind review

## ABSTRACT

Reinforcement-learning-based post-training has made output diversity an increasingly important concept in LLM reasoning, but "diversity" can refer to several behaviorally distinct forms of variation. We study three observable proxies: lexical richness, semantic dispersion, and equation-pattern variation. In an original three-model pilot across StrategyQA, MMLU College Mathematics, and ARC-Challenge, the proxies exhibit markedly different pooled associations with reasoning accuracy. A benchmark-centered reanalysis shows that the lexical-richness association remains comparatively stable ($r \approx 0.784$), semantic dispersion is near zero ($r \approx 0.019$), and the initially strong negative equation-pattern association is substantially attenuated ($r \approx -0.405$ versus pooled $r = -0.834$). We therefore interpret the study as evidence that output diversity is measurement- and benchmark-dependent rather than as evidence that any diversity axis causally improves or harms reasoning. We additionally report an expanded five-model evaluation spanning 1.5B and 7B checkpoints on GSM8K, MMLU College Mathematics, and ARC-Challenge. The expanded study exposes substantial variation in accuracy, generation truncation, and answer-extraction reliability, reinforcing the need to separate behavioral measurements from inference-pipeline artifacts. Together, these results motivate more careful, axis-specific measurement and controlled training interventions before diversity-based prescriptions are made for RL algorithms.

## 1 INTRODUCTION

**Diversity in RL-based reasoning.** Reinforcement learning with verifiable rewards (RLVR) and related post-training methods have driven substantial progress in LLM reasoning (Guo et al., 2025; Yang et al., 2024b; Yu et al., 2025). Recent work also suggests that diversity can matter during reasoning optimization. Yao et al. (2025) report a positive relationship between solution diversity and reasoning potential and introduce an explicit diversity-aware objective, while Wang et al. (2025) localize much of the RL signal to a minority of high-entropy "forking" tokens. These findings motivate studying diversity, but they do not imply that every form of variation has the same relationship with reasoning performance.

**The gap: diversity is not a single measurable property.** Prior work uses "diversity" to describe solution-level variation, token-level entropy, semantic variation, and behavioral heterogeneity (Yao et al., 2025; Kirk et al., 2024; Wang et al., 2025). These quantities are not interchangeable. Two outputs may differ lexically while following nearly identical reasoning, while two structurally different solution paths may converge to the same answer. Conversely, apparent structural variation may also arise from unstable or unsuccessful reasoning rather than productive exploration. We therefore argue that diversity analyses should state which form of variation is actually measured instead of treating diversity as a single scalar.

**This work.** We study three complementary operational views of sampled model outputs: lexical richness, semantic dispersion, and equation-pattern variation. Our framework explicitly separates conceptual constructs from the text-based proxies used to estimate them. In the original pilot, lexical richness is operationalized with type–token ratio (TTR), semantic dispersion with pairwise sentence-embedding distance, and equation-pattern variation with the number of distinct equation-like expressions across samples. These are imperfect behavioral proxies, not direct measurements of token entropy or general reasoning-tree structure.

The study is observational. We ask whether the proxies exhibit distinguishable associations with reasoning accuracy and whether those associations remain stable after benchmark-level mean differ-



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

ences are removed. We additionally report a five-model expanded evaluation, spanning 1.5B and 7B checkpoints, to characterize model- and benchmark-level performance heterogeneity and inference-pipeline failure modes. Because complete raw generations were not retained for every expanded run, that experiment is not presented as an independent replication of the diversity correlations.

**Findings.** The original pilot shows different pooled associations across proxies. Benchmark centering leaves the lexical-richness association almost unchanged, drives the semantic association close to zero, and substantially weakens the negative equation-pattern association. The strongest defensible conclusion is therefore not that one form of diversity is beneficial and another harmful, but that the observed relationships depend on what is measured and on benchmark context. The expanded five-model evaluation further shows that generation truncation and answer-extraction failures can vary dramatically across checkpoints, creating an additional source of confounding.

**Contributions.**

- We formulate output diversity as a collection of distinct behavioral properties rather than a single scalar and explicitly distinguish conceptual axes from operational proxies.
- We reanalyze the original three-model pilot with a benchmark-centered association that separates within-benchmark model variation from between-benchmark mean differences.
- We show that lexical richness, semantic dispersion, and equation-pattern variation have different robustness profiles: the lexical association is comparatively stable, while the equation-pattern association is strongly benchmark-sensitive.
- We supplement the pilot with a five-model evaluation spanning 1.5B and 7B checkpoints, revealing substantial heterogeneity in accuracy, truncation, and answer extraction.
- We identify concrete measurement and evaluation failure modes that future diversity-aware training studies should control before drawing causal conclusions.

## 2 Related Work

**Reinforcement learning for LLM reasoning.** DeepSeek-R1 (Guo et al., 2025), Kimi k1.5 (Kimi Team et al., 2025), and related systems demonstrate the effectiveness of rule-based rewards and policy-optimization methods for reasoning. DAPO (Yu et al., 2025) introduced a collection of techniques for stabilizing large-scale RL, including Clip-Higher and Dynamic Sampling. Separately, Yao et al. (2025) introduced an explicit token-level diversity objective selectively applied to positive samples. Other work addresses value estimation, length bias, and training efficiency (Yuan et al., 2025; Liu et al., 2025; Zhang et al., 2025). Open zero-RL reproductions include SimpleRL-Zoo (Zeng et al., 2025) and Open-Reasoner-Zero (Hu et al., 2025).

**Diversity in RL and language models.** Diversity-driven exploration and quality-diversity methods have a long history in reinforcement learning (Hong et al., 2018; Eysenbach et al., 2019; Wu et al., 2023). In language models, Kirk et al. (2024) study how RLHF changes generalization and diversity, and Li et al. (2025) study diversity preservation during supervised fine-tuning. Most closely related, Yao et al. (2025) measure solution diversity using equation-level variation and report a positive relationship with reasoning potential. Our initial equation-pattern result appears to contrast with that finding, motivating a central question of this paper: whether apparently similar notions of diversity remain comparable once their operational definitions and benchmark context are separated.

**Token-level reasoning signals.** Wang et al. (2025) show that high-entropy minority tokens can act as important decision points during reasoning, while Lin et al. (2025) identify tokens that disproportionately influence correctness. Gandhi et al. (2025) show that particular reasoning behaviors can drive self-improvement. These results motivate token-level analysis but do not establish that inference-time lexical richness is equivalent to policy entropy. Our analysis keeps these quantities conceptually separate.

# 3 PROBLEM FORMALIZATION

## 3.1 PROBLEM SETTING AND NOTATION

We study observable variation in the outputs of pretrained and post-trained language-model checkpoints. Let $\mathcal{M}$ denote the set of model checkpoints and $\mathcal{B}$ the set of reasoning benchmarks. Benchmark $b \in \mathcal{B}$ contains questions $Q_b$. For model $m \in \mathcal{M}$ and question $q \in Q_b$, we draw $G$ stochastic generations

$$O_{mbq} = \left\{ o_{mbq}^{(1)}, o_{mbq}^{(2)}, \ldots, o_{mbq}^{(G)} \right\}. \tag{1}$$

The sampled responses are used to estimate output variation. Separately, a greedy-decoded response $\hat{o}_{mbq}$ is used for primary accuracy measurement. With gold answer $y_q$,

$$A_{mbq} = \mathbb{I}[\mathrm{Ans}(\hat{o}_{mbq}) = y_q], \tag{2}$$

and benchmark-level accuracy is

$$A_{mb} = \frac{1}{|Q_b|} \sum_{q \in Q_b} A_{mbq}. \tag{3}$$

These quantities characterize final model behavior at inference time and are not interpreted as causal effects of the training procedures that produced the checkpoints.

## 3.2 OPERATIONALIZING OUTPUT VARIATION

We distinguish conceptual properties from the statistics used to approximate them. Given $O_{mbq}$, we write

$$D_k(O_{mbq}), \qquad k \in \{L, M, E\}, \tag{4}$$

where $L$, $M$, and $E$ denote lexical richness, semantic dispersion, and equation-pattern variation, respectively.

**Lexical-richness proxy ($D_L$).** For response $o$, let $V(o)$ be the set of unique whitespace-tokenized word types and let $|o|$ be its total word-token count. We define

$$D_L(m, b, q) = \frac{1}{G} \sum_{i=1}^{G} \frac{|V(o_{mbq}^{(i)})|}{|o_{mbq}^{(i)}|}. \tag{5}$$

This quantity measures average lexical richness within responses, not variation between responses and not token-level policy entropy. Standard TTR is also sensitive to text length, a known issue in lexical-diversity measurement (McCarthy and Jarvis, 2010).

**Semantic-dispersion proxy ($D_M$).** Each sampled response is embedded as $e_i = \mathrm{Enc}(o_{mbq}^{(i)})$ using a frozen Sentence-BERT encoder (Reimers and Gurevych, 2019). We define

$$D_M(m, b, q) = \frac{2}{G(G-1)} \sum_{i<j} [1 - \cos(e_i, e_j)]. \tag{6}$$

Larger values indicate greater separation in the selected embedding space. Full-response embeddings can compress differences between individual reasoning steps, so $D_M$ is treated as semantic dispersion rather than a complete measure of semantic reasoning diversity.

**Equation-pattern variation proxy ($D_E$).** For each response, let $E(o)$ be the set of equation-like substrings extracted by the exact regular expression reported in Appendix A. We define

$$D_E(m, b, q) = \left| \bigcup_{i=1}^{G} E(o_{mbq}^{(i)}) \right|. \tag{7}$$

This captures the number of distinct equation-like patterns across samples. It does not measure general reasoning-tree structure and cannot distinguish valid alternative strategies from incorrect expressions.

For all three proxies, the benchmark-level value is

$$D_{mb}^{(k)} = \frac{1}{|Q_b|} \sum_{q \in Q_b} D_k(O_{mbq}), \qquad k \in \{L, M, E\}. \tag{8}$$

### 3.3 Research Question

**RQ:** Do lexical richness, semantic dispersion, and equation-pattern variation in sampled LLM outputs exhibit distinct relationships with reasoning accuracy, and how sensitive are these relationships to benchmark and model context?

Because the evaluated checkpoints were not produced under controlled, matched training interventions, all relationships are interpreted observationally rather than causally.

### 3.4 Benchmark-Adjusted Association Analysis

A pooled correlation can conflate between-benchmark differences with within-benchmark model variation. We therefore supplement the pooled analysis with benchmark centering. For proxy $k$,

$$\widetilde{D}_{mb}^{(k)} = D_{mb}^{(k)} - \frac{1}{|\mathcal{M}|} \sum_{m' \in \mathcal{M}} D_{m'b}^{(k)}, \tag{9}$$

and

$$\widetilde{A}_{mb} = A_{mb} - \frac{1}{|\mathcal{M}|} \sum_{m' \in \mathcal{M}} A_{m'b}. \tag{10}$$

We then compute

$$r_k^{\text{within}} = \operatorname{corr}\left(\widetilde{D}^{(k)}, \widetilde{A}\right). \tag{11}$$

This removes benchmark-level mean differences before estimating the remaining association across models. Because the pilot contains only nine non-independent model–benchmark observations, all such coefficients are treated as descriptive sensitivity analyses.

## 4 Experimental Setup

### 4.1 Study Overview

We distinguish two experimental components. **Original diversity pilot:** three 1.5B checkpoints evaluated on StrategyQA, MMLU College Mathematics, and ARC-Challenge, with four stochastic samples per question and one greedy response. This experiment is the source of the diversity-proxy measurements. **Expanded model-suite evaluation:** five successfully completed checkpoints spanning 1.5B and 7B scales evaluated on GSM8K, MMLU College Mathematics, and ARC-Challenge, with $G = 8$ stochastic generations plus one greedy generation per question. The expanded experiment is used to characterize performance heterogeneity and inference reliability; it is not an independent replication of the diversity correlations because complete raw generations were not retained for every expanded run.

### 4.2 Models

The original diversity pilot evaluates Qwen2.5-Math-1.5B-Instruct (Yang et al., 2024b), SimpleRL-Zoo-1.5B (Zeng et al., 2025), and Open-Reasoner-Zero-1.5B (Hu et al., 2025). These public checkpoints have related Qwen-family architectures but differ in post-training histories; model identity is therefore a confound rather than a controlled treatment.

The expanded evaluation uses Qwen2.5-1.5B-Instruct, Qwen2.5-Math-1.5B-Instruct, Open-Reasoner-Zero-1.5B, DeepSeek-R1-Distill-Qwen-1.5B, and Qwen2.5-7B-Instruct. The Qwen2.5 instruction checkpoints are part of the public Qwen2.5 family (Yang et al., 2024a); Qwen2.5-Math uses specialized mathematical post-training (Yang et al., 2024b). We report checkpoint identity directly and avoid attributing performance differences to any single training stage.

### 4.3 BENCHMARKS

The pilot uses 30 questions each from StrategyQA (Geva et al., 2021), MMLU College Mathematics (Hendrycks et al., 2021), and ARC-Challenge (Clark et al., 2018). The StrategyQA subset comes from the available training split. Because public language models may have encountered these examples during pretraining or post-training, we do not treat StrategyQA as a strict held-out generalization benchmark.

The expanded evaluation uses 25 fixed questions per benchmark from GSM8K (Cobbe et al., 2021), MMLU College Mathematics, and ARC-Challenge. These benchmarks span different reasoning domains; we use the terms *reasoning accuracy* and *cross-benchmark performance* rather than claiming strict out-of-distribution status.

### 4.4 INFERENCE PROTOCOL

The original pilot used one greedy response for accuracy and $G = 4$ sampled responses at temperature $0.7$ and top-$p = 0.9$, with a 150-token generation limit. We retain these details because they define the measurements reported in the original diversity tables, while treating the low token limit as a limitation.

The expanded evaluation uses one greedy response and $G = 8$ stochastic generations per question with temperature $0.7$, top-$p = 0.9$, and a maximum of 1024 new tokens. All models receive a reasoning prompt ending in an explicit `FINAL:` answer field. The same fixed questions are used for every model within a benchmark.

All evaluations use 4-bit NormalFloat (NF4) quantization with double quantization and half-precision computation, following the QLoRA quantization scheme (Dettmers et al., 2023). Models are loaded sequentially on a single NVIDIA T4 GPU.

### 4.5 ANSWER EXTRACTION AND RELIABILITY METRICS

Predictions are extracted with a cascaded parser supporting explicit `FINAL:` fields, common "final answer" variants, boxed answers, and numeric fallbacks. We report parser failure rather than silently mapping an unparseable response to a substantive model error.

For model $m$ and benchmark $b$, truncation rate is

$$T_{mb} = \frac{1}{N_{mb}} \sum_{i=1}^{N_{mb}} \mathbb{I}[\ell_i = L_{\max}], \tag{12}$$

where $\ell_i$ is generated length and $L_{\max}$ is the configured token limit. Answer-extraction failure is

$$F_{mb} = \frac{1}{N_{mb}} \sum_{i=1}^{N_{mb}} \mathbb{I}[\mathrm{Ans}(o_i) = \varnothing]. \tag{13}$$

We report $T_{mb}$ and $F_{mb}$ alongside accuracy because severe truncation or parsing failure can make an apparent model difference an evaluation artifact.

## 5 RESULTS

### 5.1 ORIGINAL DIVERSITY PILOT

Table 1 reproduces the original pilot measurements under the revised proxy terminology. The descriptive pattern is heterogeneous: lexical richness is higher for SimpleRL and ORZ on StrategyQA and ARC, while equation-pattern variation is much larger on MMLU College Mathematics. This benchmark dependence motivates the adjusted analysis below.

### 5.2 POOLED VERSUS BENCHMARK-CENTRED ASSOCIATIONS

Table 2 contrasts the original pooled coefficients with benchmark-centered coefficients recomputed from the reported Table 1 values. The lexical association is nearly unchanged. Semantic dispersion

**Table 1:** Original pilot measurements. $D_L$: lexical-richness proxy; $D_M$: semantic-dispersion proxy; $D_E$: equation-pattern variation proxy.

| Model | Benchmark | Acc. | $D_L$ | $D_M$ | $D_E$ |
|---|---|---|---|---|---|
| Qwen2.5-Math-1.5B | StrategyQA | 30.0 | .566 | .381 | 1.7 |
| | MMLU-Math | 30.0 | .570 | .116 | 5.1 |
| | ARC-Challenge | 53.3 | .572 | .213 | 1.5 |
| SimpleRL-Zoo-1.5B | StrategyQA | 63.3 | .933 | .375 | .3 |
| | MMLU-Math | 20.0 | .648 | .233 | 2.6 |
| | ARC-Challenge | 63.3 | .933 | .282 | .3 |
| Open-Reasoner-Zero | StrategyQA | 53.3 | .878 | .616 | .6 |
| | MMLU-Math | 16.7 | .656 | .364 | 4.4 |
| | ARC-Challenge | 70.0 | .947 | .309 | .2 |

**Table 2:** Pooled and benchmark-centered associations with reasoning accuracy. Coefficients are descriptive because the nine model–benchmark observations are non-independent.

| Proxy | Pooled $r$ | Centred $r$ |
|---|---|---|
| Lexical richness ($D_L$) | +0.777 | +0.784 |
| Semantic dispersion ($D_M$) | +0.205 | +0.019 |
| Equation-pattern variation ($D_E$) | -0.834 | -0.405 |

becomes essentially zero. Most importantly, the magnitude of the equation-pattern association falls from $-0.834$ to approximately $-0.405$, indicating that a substantial portion of the original pooled effect is associated with benchmark-level differences.

Figure 1 makes the central sensitivity result visually explicit: benchmark adjustment affects the three proxies very differently.

We therefore do not treat the original Pearson $p$-values as confirmatory evidence. The appropriate interpretation is sensitivity of descriptive effect direction and magnitude to benchmark adjustment.

### 5.3 Expanded Five-Model Evaluation

Table 3 reports the five successfully completed expanded runs. Qwen2.5-7B-Instruct achieves the highest observed greedy accuracy on all three benchmarks, but the 25-item evaluation yields wide uncertainty intervals (Appendix C). More importantly for methodology, reliability varies strongly by checkpoint. DeepSeek-R1-Distill-Qwen-1.5B exhibits severe truncation and parser failure on MMLU, while Qwen2.5-7B-Instruct remains below 6% on both reliability measures.

These results do not replicate the diversity analysis; they instead demonstrate that checkpoint scale, prompt-response format, generation length, and extraction reliability can materially affect the quantities from which diversity and accuracy are measured.

## 6 Analysis and Discussion

### 6.1 Interpreting the Lexical-Richness Association

The benchmark-centered lexical association ($r \approx 0.784$) is the most stable pattern in the pilot. However, $D_L$ is TTR-based and therefore measures lexical richness, not inter-response variation or token-level uncertainty. TTR is length-sensitive (McCarthy and Jarvis, 2010); a model could also generate the same lexically rich response repeatedly and still obtain a high $D_L$. We therefore do not interpret the result as direct evidence that increasing token entropy during RL would improve reasoning. Instead, it motivates future work using genuine inter-response lexical distances and logit-based measures when model internals are available.

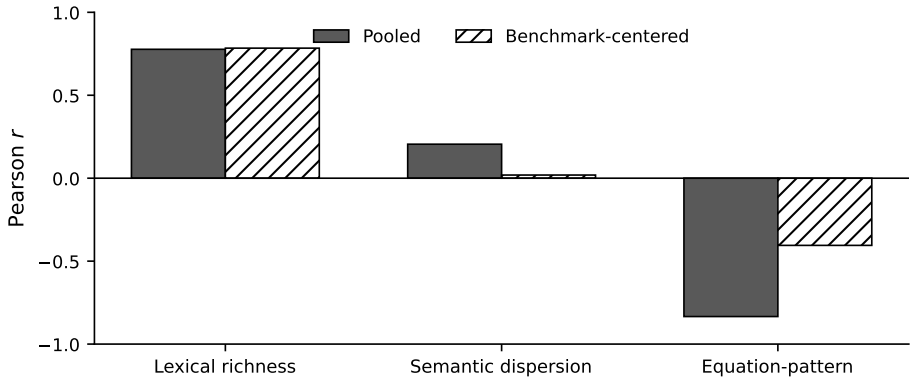


**Figure 1:** Sensitivity of the three proxy–accuracy associations to benchmark centering. The lexical-richness coefficient is nearly unchanged, semantic dispersion moves close to zero, and the magnitude of the equation-pattern coefficient is substantially reduced.

**Table 3:** Expanded evaluation. Accuracy is greedy accuracy on 25 questions per benchmark. Max truncation and parser failure are the largest rates observed across benchmark/generation-type cells for each checkpoint.

| Model | ARC | GSM8K | MMLU | Max trunc. | Max parser fail. |
|---|---|---|---|---|---|
| Qwen2.5-1.5B-Instruct | 68% | 52% | 40% | 6.0% | 22.0% |
| Qwen2.5-Math-1.5B-Instruct | 56% | 72% | 56% | 8.5% | 19.5% |
| Open-Reasoner-Zero-1.5B | 44% | 52% | 40% | 9.0% | 32.0% |
| DeepSeek-R1-Distill-Qwen-1.5B | 48% | 64% | 20%† | 78.5% | 77.0% |
| Qwen2.5-7B-Instruct | 88% | 88% | 64% | 6.0% | 6.0% |

†Not interpreted as a reliable performance estimate because most MMLU generations reached the token limit and failed answer extraction.

### 6.2 Why Is Equation-Pattern Variation Benchmark-Sensitive?

The pooled equation-pattern coefficient is strongly negative, but its magnitude falls substantially after benchmark centering. This directly addresses the risk of cross-benchmark confounding: mathematical tasks naturally elicit more equation-like strings, and benchmark difficulty also differs. In Table 1, Qwen2.5-Math has both the highest MMLU equation-pattern value ($D_E = 5.1$) and the highest MMLU accuracy among the three pilot models (30.0%), contradicting a simple interpretation that more equation patterns necessarily indicate confused guessing.

We therefore do not interpret the negative pooled coefficient as evidence that genuine structural diversity is harmful. At most, it shows that this particular equation-pattern proxy covaries negatively with pooled accuracy in the original pilot. Distinguishing valid alternative reasoning strategies from unstable reasoning requires a correctness-conditioned, domain-general structural representation.

### 6.3 Semantic Dispersion and Encoder Sensitivity

The original Sentence-BERT association is small and positive, while the previously reported BGE-based sensitivity check changes sign. Neither result provides evidence of a stable semantic-dispersion relationship in this small pilot. We interpret the sign change as insufficient evidence for a consistent association, not as proof that the true effect is exactly zero. Future work should evaluate step-level or reasoning-aware encoders rather than treating a single full-response embedding as definitive.

### 6.4 What the Expanded Evaluation Adds

The five-model evaluation partially addresses the original scale concern by including a 7B checkpoint and nonidentical post-training families. It also exposes why larger generation budgets and explicit reliability accounting matter: the DeepSeek MMLU cell has truncation of 72–78.5% and parser

failure of 72–77%, whereas Qwen2.5-7B remains at 4–6% on MMLU. Thus a low observed accuracy can reflect model behavior, evaluation-pipeline failure, or both. Diversity statistics derived from truncated strings can be distorted in the same way.

### 6.5 Hypotheses for Future Diversity-Aware Training

Our results do not establish that manipulating any proxy during training would improve performance. Instead, they motivate controlled intervention studies in which initialization, data, reward, and optimization settings are held fixed while a diversity objective is varied. The comparatively stable lexical association motivates testing whether token-level exploration produces corresponding inference-time behavioral changes. The benchmark sensitivity of $D_E$ motivates correctness-conditioned structural metrics that distinguish valid alternative solutions from failed exploration. Finally, the unstable semantic result motivates step-aware semantic representations before semantic diversity is accepted or rejected as a training target. Diversity-aware RL objectives such as Yao et al. (2025) provide a natural intervention setting for these questions.

### 6.6 Limitations

**Primary sample size.** The diversity-proxy analysis still contains only three models, three benchmarks, and nine aggregate cells. The expanded five-model evaluation broadens scale and reliability analysis but cannot independently reproduce the diversity correlations because the complete raw generations were not retained for every run.

**Proxy validity.** $D_L$ measures intra-response lexical richness, $D_M$ depends on the chosen encoder, and $D_E$ captures equation-like surface patterns rather than general reasoning structure. More faithful proxies may change the observed relationships.

**Benchmark dependence.** The attenuation of $D_E$ after benchmark centering demonstrates that pooled cross-domain correlations can be misleading when tasks differ in both expected response form and difficulty. The StrategyQA pilot additionally uses a training split and is therefore not interpreted as strict held-out generalization.

**Quantization and inference.** All evaluations use 4-bit NF4 quantization (Dettmers et al., 2023); no matched full-precision comparison is available. The pilot's 150-token limit is especially restrictive for reasoning models. The expanded study improves the token budget to 1024 but still contains checkpoint-specific truncation.

**Training-history confounding.** The evaluated checkpoints differ in scale, specialization, training data, distillation, and post-training procedures. The design cannot isolate the causal contribution of RL or any single training choice.

**Observational design.** Establishing whether changing a diversity objective causes performance improvements requires controlled training interventions rather than post hoc correlations across public checkpoints.

## 7 Conclusion

We studied output diversity in LLM reasoning as a collection of distinct observable properties rather than a single scalar. In the original pilot, lexical richness, semantic dispersion, and equation-pattern variation exhibit markedly different pooled associations with reasoning accuracy. Benchmark-centered analysis shows that the lexical-richness association is comparatively stable, semantic dispersion provides little consistent signal, and the initially strong negative equation-pattern association is substantially attenuated after benchmark-level differences are removed.

An expanded five-model evaluation further demonstrates that model scale, generation truncation, and answer-extraction reliability can materially affect reasoning evaluation. Together, these findings support a measurement-focused conclusion: claims about "diversity" should specify the form of variation being measured and test robustness to benchmark and inference conditions before drawing conclusions about reinforcement-learning objectives. Our results remain observational and do not establish causal prescriptions for training.

## AI USE STATEMENT

Generative AI tools were used for language editing, organization, LaTeX formatting, and consistency checking. The authors are responsible for all scientific claims, analyses, experimental results, and citations.

## REFERENCES


Peter Clark, Isaac Cowhey, Oren Etzioni, et al. Think You Have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. *arXiv:1803.05457*, 2018.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, et al. Training Verifiers to Solve Math Word Problems. *arXiv:2110.14168*, 2021.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient Finetuning of Quantized LLMs. *Advances in Neural Information Processing Systems*, 36, 2023.

Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. Diversity Is All You Need: Learning Skills without a Reward Function. *International Conference on Learning Representations*, 2019.

Kanishk Gandhi, Ayush Chakravarthy, Anikait Singh, Nathan Lile, and Noah D. Goodman. Cognitive Behaviors that Enable Self-Improving Reasoners, or, Four Habits of Highly Effective STaRs. *Conference on Language Modeling*, 2025.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. Did Aristotle Use a Laptop? A Question Answering Benchmark with Implicit Reasoning Strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361, 2021.

Daya Guo, Dejian Yang, Haowei Zhang, et al. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv:2501.12948*, 2025.

Dan Hendrycks, Collin Burns, Steven Basart, et al. Measuring Massive Multitask Language Understanding. *International Conference on Learning Representations*, 2021.

Zhang-Wei Hong, Tzu-Yun Shann, Shih-Yang Su, et al. Diversity-Driven Exploration Strategy for Deep Reinforcement Learning. *Advances in Neural Information Processing Systems*, 31, 2018.

Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, Xiangyu Zhang, and Heung-Yeung Shum. Open-Reasoner-Zero: An Open Source Approach to Scaling Up Reinforcement Learning on the Base Model. *Advances in Neural Information Processing Systems*, 38, 2025.

Kimi Team, Angang Du, Bofei Gao, et al. Kimi k1.5: Scaling Reinforcement Learning with LLMs. *arXiv:2501.12599*, 2025.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, et al. Understanding the Effects of RLHF on LLM Generalization and Diversity. *International Conference on Learning Representations*, 2024.

Ziniu Li, Congliang Chen, Tian Xu, et al. Preserving Diversity in Supervised Fine-Tuning of Large Language Models. *International Conference on Learning Representations*, 2025.

Zicheng Lin, Tian Liang, Jiahao Xu, Qiuzhi Liu, et al. Critical Tokens Matter: Token-Level Contrastive Estimation Enhances LLM's Reasoning Capability. *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

Zichen Liu, Changyu Chen, Wenjun Li, et al. Understanding R1-Zero-Like Training: A Critical Perspective. *arXiv:2503.20783*, 2025.

Philip M. McCarthy and Scott Jarvis. MTLD, vocd-D, and HD-D: A Validation Study of Sophisticated Approaches to Lexical Diversity Assessment. *Behavior Research Methods*, 42(2):381–392, 2010.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks. *Proceedings of EMNLP-IJCNLP*, pages 3982–3992, 2019.

Shenzhi Wang, Le Yu, Chang Gao, et al. Beyond the 80/20 Rule: High-Entropy Minority Tokens Drive Effective Reinforcement Learning for LLM Reasoning. *Advances in Neural Information Processing Systems*, 38, 2025.

Shuang Wu, Jian Yao, Haobo Fu, et al. Quality-Similar Diversity via Population Based Reinforcement Learning. *International Conference on Learning Representations*, 2023.

An Yang, Baosong Yang, Beichen Zhang, et al. Qwen2.5 Technical Report. *arXiv:2412.15115*, 2024a.

An Yang, Beichen Zhang, Binyuan Hui, et al. Qwen2.5-Math Technical Report: Toward Mathematical Expert Model via Self-Improvement. *arXiv:2409.12122*, 2024b.

Jian Yao, Ran Cheng, Xingyu Wu, Jibin Wu, and Kay Chen Tan. Diversity-Aware Policy Optimization for Large Language Model Reasoning. *Advances in Neural Information Processing Systems*, 38, 2025.

Qiying Yu, Zheng Zhang, Ruofei Zhu, et al. DAPO: An Open-Source LLM Reinforcement Learning System at Scale. *arXiv:2503.14476*, 2025.

Yufeng Yuan, Qiying Yu, Xiaochen Zuo, et al. VAPO: Efficient and Reliable Reinforcement Learning for Advanced Reasoning Tasks. *arXiv:2504.05118*, 2025.

Weihao Zeng, Yuzhen Huang, Qian Liu, et al. SimpleRL-Zoo: Investigating and Taming Zero Reinforcement Learning for Open Base Models in the Wild. *arXiv:2503.18892*, 2025.

Xiaojiang Zhang, Jinghui Wang, Zifei Cheng, et al. SRPO: A Cross-Domain Implementation of Large-Scale Reinforcement Learning on LLM. *arXiv:2504.14286*, 2025.

# A Diversity Metric Implementation Details

## A.1 Lexical-Richness Proxy

For each sampled response, whitespace tokenization is used to compute TTR. Responses shorter than three tokens are excluded. This choice is retained to reproduce the original pilot, but TTR is explicitly interpreted as lexical richness and not as inter-response diversity or policy entropy.

## A.2 Semantic-Dispersion Proxy

The primary encoder is `all-MiniLM-L6-v2`, producing a single embedding for each full response. Pairwise cosine distance is averaged over all response pairs. A previously reported sensitivity analysis used `BAAI/bge-base-en-v1.5`; the estimated association changed sign, motivating cautious interpretation.

## A.3 Equation-Pattern Variation Proxy

The exact regular expression is

```
[\d\w]+\s*[+\-*/=<>]+\s*[\d\w]+
```

Matches are lowercased and stripped of whitespace before the union is formed. The regex can match some non-mathematical surface patterns and does not validate equations; these are known construct-validity limitations.

# B Proxy Comparison

**Table 4:** Operational proxies and known failure modes.

| Proxy | Measures | Known limitations |
|---|---|---|
| $D_L$ | Mean within-response TTR | Length-sensitive; not inter-response diversity; not token entropy |
| $D_M$ | Pairwise cosine distance of response embeddings | Encoder-dependent; full-response embeddings can obscure step-level differences |
| $D_E$ | Number of distinct regex-matched equation-like strings | Math-centric; can match surface punctuation; cannot distinguish valid from incorrect equations |

# C Expanded Evaluation Details

## C.1 Accuracy Uncertainty

With only $n = 25$ questions per model–benchmark cell, one question changes accuracy by four percentage points. We therefore use the Wilson interval

$$\mathrm{CI}_{\mathrm{Wilson}} = \frac{\hat{p} + \frac{z^2}{2n} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2}{4n^2}}}{1 + \frac{z^2}{n}}, \qquad z = 1.96. \tag{14}$$

Selected 95% intervals are: Qwen2.5-7B on ARC, 88% [70.0, 95.8]; Qwen2.5-Math-1.5B on GSM8K, 72% [52.4, 85.7]; and Qwen2.5-1.5B on MMLU, 40% [23.4, 59.3]. The intervals emphasize that small absolute differences between 25-item cells should not be overinterpreted.

## C.2 Full Reliability Summary

**Table 5:** Expanded experiment reliability rates. Sample rows aggregate 200 stochastic generations; greedy rows contain 25 generations.

| Model | Benchmark/type | Accuracy | Trunc. | Parser fail. |
|---|---|---|---|---|
| Qwen2.5-1.5B | ARC greedy | .680 | .000 | .000 |
| | ARC sample | .635 | .000 | .080 |
| | GSM8K greedy | .520 | .000 | .000 |
| | GSM8K sample | .490 | .000 | .005 |
| | MMLU greedy | .400 | .040 | .120 |
| | MMLU sample | .315 | .060 | .220 |
| Qwen2.5-Math-1.5B | ARC greedy | .560 | .000 | .000 |
| | ARC sample | .555 | .015 | .020 |
| | GSM8K greedy | .720 | .040 | .040 |
| | GSM8K sample | .800 | .005 | .005 |
| | MMLU greedy | .560 | .080 | .120 |
| | MMLU sample | .515 | .085 | .195 |
| ORZ-1.5B | ARC greedy | .440 | .040 | .320 |
| | ARC sample | .495 | .080 | .265 |
| | GSM8K greedy | .520 | .040 | .000 |
| | GSM8K sample | .600 | .025 | .000 |
| | MMLU greedy | .400 | .040 | .320 |
| | MMLU sample | .325 | .090 | .315 |
| DeepSeek-R1-Distill | ARC greedy | .480 | .120 | .160 |
| | ARC sample | .385 | .180 | .330 |
| | GSM8K greedy | .640 | .000 | .000 |
| | GSM8K sample | .710 | .000 | .000 |
| | MMLU greedy | .200 | .720 | .720 |
| | MMLU sample | .195 | .785 | .770 |
| Qwen2.5-7B | ARC greedy | .880 | .000 | .000 |
| | ARC sample | .845 | .000 | .000 |
| | GSM8K greedy | .880 | .000 | .000 |
| | GSM8K sample | .885 | .000 | .000 |
| | MMLU greedy | .640 | .040 | .040 |
| | MMLU sample | .680 | .060 | .060 |