Prismatic Synthesis: Gradient-based Data Diversification Boosts Generalization in LLM Reasoning

Jaehun Jung[1,2] **Seungju Han**[1*] **Ximing Lu**[1,2*] **Skyler Hallinan**[3*]
David Acuna[1] **Shrimai Prabhumoye**[1] **Mostofa Patwary**[1]
Mohammad Shoeybi[1] **Bryan Catanzaro**[1] **Yejin Choi**[1]
[1]NVIDIA Research
[2]University of Washington
[3]University of Southern California
hoony123@cs.washington.edu*

Abstract

Effective generalization in language models depends critically on the diversity of their training data. Yet existing diversity metrics often fall short of this goal, relying on surface-level heuristics that are decoupled from model behavior. This motivates us to ask: *What kind of diversity in training data actually drives generalization in language models—and how can we measure and amplify it?* Through large-scale empirical analyses spanning over 300 training runs, carefully controlled for data scale and quality, we show that data diversity can be a strong predictor of generalization in LLM reasoning—as measured by average model performance on unseen out-of-distribution benchmarks. We introduce **G-Vendi**, a metric that quantifies diversity via the entropy of model-induced gradients. Despite using a small off-the-shelf proxy model for gradients, G-Vendi consistently outperforms alternative measures, achieving strong correlation (Spearman's $\rho \approx 0.9$) with out-of-distribution (OOD) performance on both natural language inference (NLI) and math reasoning tasks. Building on this insight, we present **Prismatic Synthesis**, a framework for generating diverse synthetic data by targeting underrepresented regions in gradient space. Experimental results show that Prismatic Synthesis consistently improves model performance as we scale synthetic data—not just on in-distribution test but across unseen, out-of-distribution benchmarks—significantly outperforming state-of-the-art models that rely on 20 times larger data generator than ours. For example, PrismMath-7B, our model distilled from a 32B LLM, outperforms R1-Distill-Qwen-7B—the same base model trained on proprietary data generated by 671B R1—on 6 out of 7 challenging benchmarks.

1 Introduction

The idea that diverse training data leads to better language models is both intuitive and widely acknowledged—a growing body of works support this intuition, linking data diversity to improved robustness, sample efficiency, and generalization [2, 45, 69]. But despite this growing consensus, a fundamental question remains surprisingly underexplored: what kind of diversity actually matters for generalization, and how should we empirically quantify it? While the importance of diversity is broadly recognized, its precise quantification remains elusive—existing measures often rely on task-specific heuristics, or intrinsic textual features such as semantic or lexical variations [58, 67, 18, 19].

*SH, XL and SH are co-second authors. Project Page: Link

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

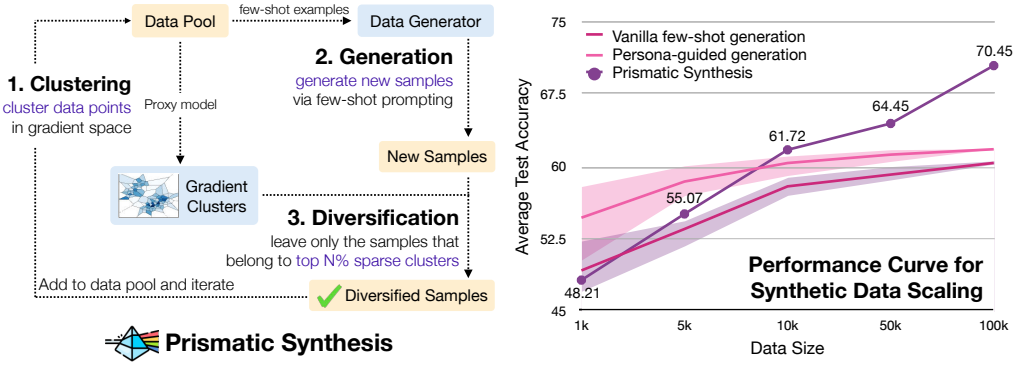

Figure 1: (Left) **Overview of Prismatic Synthesis**. We iteratively (1) cluster samples in a gradient space, (2) generate new samples, and (3) add to the pool only the samples in sparse clusters, consistently improving both the diversity and scale of generated dataset. (Right) Naive scaling of synthetic math data—with no diversification or with a heuristic persona-guided prompting [9]—faces early saturation, when measuring average performance across 7 distinct benchmarks. **Prismatic Synthesis consistently improves model performance beyond 100K, up to million scale synthetic data.** See §3.2.2 for more details.

Yet it remains unclear to what extent these measures actually correlate with downstream model performance—particularly for reasoning tasks, where surface-form diversity offers limited insight.

In this work, we aim to investigate the role of data diversity in LLM reasoning, and how to measure and promote diversity during data curation. To isolate the effect of diversity, we conduct a comprehensive suite of experiments that control for data scale and quality, allowing us to systematically examine how diversity measures relate to model empirical generalization. Our study focuses on post-training scenarios in reasoning tasks, and builds on an extensive empirical framework—generating a synthetic data pool of over 3 million samples and fine-tuning more than 300 models on distinct datasets with varied range of diversity and scale. In both discriminative NLI and generative math reasoning tasks, our analyses reveal three key insights:

- Unless tailored to a specific task, **existing diversity measures often only show moderate to weak correlation with how the model performs on unseen benchmarks**—as they latch onto surface-form features that may not be relevant to the target task.
- We propose G-Vendi, a novel measure that computes entropy of a given dataset in a gradient space. Despite using a small, off-the-shelf proxy model to compute gradients, **G-Vendi strongly correlates (Spearman's $\rho \approx 0.9$) with how the resulting model performs on OOD benchmarks.**
- Diversity often outweighs the scale for out-of-distribution generalization—training on a small dataset with higher diversity can outperform datasets of 10 times larger scale, even when the datasets are drawn from the same data pool. However, **scale is a dominant factor for in-distribution performance, improving models in complementary ways to diversity.**

Leveraging these insights, we seek to further enhance data diversity by strategically generating synthetic data. We find that naive scaling of synthetic data does not improve the model, facing early saturation even with heuristic diversification such as persona-guided prompting [9]. We instead propose **Prismatic Synthesis**, a novel algorithm to scale synthetic data while simultaneously improving the diversity of generated samples. By iteratively (1) clustering existing data in gradient space and (2) rejection-sampling new data that correspond to the sparse clusters, Prismatic Synthesis constantly improves both the diversity and scale of the resulting dataset. Consequently, this simple process empowers consistent improvement in model reasoning beyond million-scale synthetic samples.

We apply our method using off-the-shelf 32B / 72B LLMs to generate two synthetic datasets: PrismMath and PrismNLI. While our datasets are **entirely model-generated without any human annotations**, they produce surprisingly strong models—**often surpassing state-of-the-art models distilled from a 671B teacher-generated trajectories, further verified by ground-truth answers.** For example, our math reasoning model PrismMath-7B yields state-of-the-art results on 6 out of 7 benchmarks (*e.g.,* 57.08% on AIME24 with only SFT), outperforming R1-Distill-Qwen-7B trained on

proprietary data generated by 671B R1 [5]. Overall, these results suggest that strategic diversification may offer greater gains than costly, manual data curation, and highlight the importance of data diversity as a key driver of model generalization.

2 G-Vendi: Diversity Measure that Predicts Generalization

Our first goal in this work is to analyze data diversity—often regarded as an intrinsic property of a dataset—through the lens of its extrinsic, practical impact on model performance. In the following sections, we first introduce our novel diversity measure G-Vendi (§2.1), illustrate how we evaluate diversity measures while controlling for confounders (§2.2), and analyze the results (§2.3).

2.1 G-Vendi Score

We motivate G-Vendi from the formulation of Pruthi et al. [48] that approximates training data influence with first-order gradients. Let $\nabla l(z;\theta)$ denote the back-propagated loss gradient of a data sample z under model parameters θ. The improvement (*i.e.,* loss reduction) on a test sample z' induced by training on a train sample z, is approximately proportional to the dot product between the loss gradients for z and z':

$$l(z';\theta) - l(z';\theta') \propto\!\!\!\sim \nabla l(z;\theta) \cdot \nabla l(z';\theta)$$

In other words, the degree to which training on z helps generalization to z', can be estimated by the similarity between their loss gradients. Extending this insight, we hypothesize and empirically show that when no explicit target distribution of z' is available, promoting the *diversity among the loss gradients of training samples* z may help improve model performance—allowing the training data to cover a broader region of the task distribution, assuming uniform prior over z'.

Collecting Gradient Representation G-Vendi implements this idea by utilizing the gradient representation of each sample in a given dataset. Concretely, let $(x, y) \in \mathcal{D}$ denote a training data point where x is the input and y is the output. G-Vendi represents each sample (x, y) with its normalized loss gradient vector computed under an off-the-shelf proxy model θ:

$$g_\theta(x,y) = \frac{-\nabla \log P(y|x;\theta)}{|| - \nabla \log P(y|x;\theta)||} \in \mathbb{R}^{|\theta|} \tag{1}$$

Note here that each $g_\theta(x, y)$ is a $|\theta|$-dimensional vector, which is prohibitively large even with a small proxy model (*e.g.,* 0.5B). We thus follow prior works [61, 57] to reduce the dimension of $g_\theta(x, y)$ from $|\theta|$ to $d << |\theta|$ via Rademacher random projection [46, 50]:

$$g_\theta^{proj}(x,y) = \Pi^T g_\theta(x,y), \quad \text{where } \Pi \in \mathbb{R}^{|\theta|\times d},\ \Pi_{ij} \sim \mathcal{U}(\{-1,1\}) \tag{2}$$

The projection qualifies as a Johnson-Lindenstrauss transform [17], and thus preserves the inner product between projected gradients with high probability (while significantly reducing their dimension). We set $d = 1024$ in our experiments, and compute the projection for all samples $(x, y) \in \mathcal{D}$; we denote the collected projections as $G \in \mathbb{R}^{|\mathcal{D}|\times d}$.

Importantly, we set the proxy model θ to be a small instruction-tuned model (*e.g.,* Qwen2.5-0.5B-Instruct [49]) without any additional training. This greatly simplifies the gradient collection process compared to prior methods, which often require in-distribution warm-up training, weight decomposition, and gradient aggregation across multiple checkpoints [1, 48, 57, 20]. While the exact formulation may be more accurate in estimating influence to a target instance [61], we find that gradients from an off-the-shelf model θ are still effective in estimating variance *in between* training samples—which does not involve measuring distance to a target distribution. In addition, the efficiency of off-the-shelf gradients makes it suitable for online diversification of large-scale synthetic data, as we discuss in §3.

Measuring Entropy of Gradients We measure the diversity of a dataset $\mathcal{D}$ by computing the entropy of its loss gradients. Specifically, we compute the exponentiated entropy of the normalized covariance matrix of G, *i.e.,* the Vendi Score [8] of G, allowing us to measure the entropy without knowing the support of the underlying gradient distribution. Let K denote the covariance matrix, *i.e.,* $K_{ij} = \left(GG^T\right)_{ij}/|\mathcal{D}| = g_\theta^{proj}(x_i, y_i) \cdot g_\theta^{proj}(x_j, y_j)/|\mathcal{D}|$. Then the G-Vendi score of $\mathcal{D}$ is computed as:

$$\text{G-Vendi}(\mathcal{D}) = \exp\Big(-\sum_i \lambda_i^K \log \lambda_i^K\Big) \tag{3}$$

where λ_i^K is the i-th eigenvalue of K. Intuitively, a low value of G-Vendi implies that most of the variance among gradients is explained by a few directional components, indicating low diversity; a high value of G-Vendi implies no dominant directional component among gradients, indicating high diversity. Compared to other aggregation methods (*e.g.,* average pairwise similarity), Vendi score is significantly more compute-efficient when $|\mathcal{D}| >> d$ and robust when features are correlated with each other [8]. G-Vendi is a positive unbounded scalar score, and inherits several desirable properties of the Vendi score for diversity measurement, such as permutation invariance. We further illustrate the properties of our measure in §C.

2.2 Evaluating Data Diversity Measures

Our next step is to evaluate how well G-Vendi (and existing diversity measures) can predict model performance on unseen benchmarks. Given a diversity measure f along with two datasets $\mathcal{D}_1$ and $\mathcal{D}_2$ of comparable size and quality, our goal is to evaluate whether f satisfies:

$$f(\mathcal{D}_1) > f(\mathcal{D}_2) \Rightarrow \text{Perf}(\mathcal{M}_1) > \text{Perf}(\mathcal{M}_2) \tag{4}$$

where $\mathcal{M}_1$ and $\mathcal{M}_2$ are the same base models trained on $\mathcal{D}_1$ and $\mathcal{D}_2$, and Perf estimates the models' task performance on unseen benchmarks. The most straightforward way of evaluating the desideratum is to (1) prepare many datasets with varied levels of diversity, (2) train a model on each dataset, and (3) compute the correlation between performance and diversity. However, translating this evaluation into practice entails non-trivial experimental design, particularly to control for confounding factors such as data quality. Below, we detail our setup focusing on math reasoning and NLI tasks.

Preparing Datasets To control for the effects of data quality, we generate a large pool of synthetic data using LLMs from which we sample distinct training datasets. Compared to off-the-shelf datasets, synthetic data offers several advantages for our experimental setup: it is more scalable, easier to quality-control (by applying the same data-generating pipeline for all samples), and importantly, has been widely adopted for post-training LLMs on reasoning tasks [16, 44, 63].

We take a few-shot generation approach to generate the data pool, by random-sampling 5 demonstrations from a seed dataset then prompting an LLM to generate novel data points. For seed datasets, we use WANLI [28] for NLI, and a mixture of GSM8k [4] and MATH [15] for math reasoning. We prompt Qwen2.5-72B-Instruct [49] to generate both the new problems and corresponding solutions to create a pool of 1.5M samples for each task. Finally, we prepare training datasets by sampling subsets from this data pool. We create 300 distinct subsets with varied levels of diversity, while controlling for their size (N = 100k, 50k, 10k for math reasoning, N = 50k, 10k for NLI). For more details, we refer the readers to §A.1.

Evaluating Model Generalization Yet another challenge lies in defining Perf, *i.e.,* how to estimate the model's empirical performance on unseen benchmarks. One intuitive approach is to prepare multiple unseen benchmarks $B_1, \cdots B_{|\mathcal{B}|}$ and average the accuracies, i.e., $\frac{1}{|\mathcal{B}|}\sum_i Acc_{B_i}(\mathcal{M})$. However, this metric overlooks the fact that not all benchmarks are equally challenging—for example, a 3% improvement on an elementary arithmetic exam should not be deemed equivalent to 3% improvement in an Olympiad-level benchmark, where even the strongest model performs poorly and thus 3% gap is much more significant. To better reflect the benchmark-specific notion of performance gap, we average the relative accuracy of the current model $\mathcal{M}$ against a strong reference model $\mathcal{M}_{\text{ref}}$:

$$\text{Perf}(\mathcal{M}) \coloneqq \frac{1}{|\mathcal{B}|}\sum_i \frac{\text{Acc}_{\mathcal{B}_i}(\mathcal{M})}{\text{Acc}_{\mathcal{B}_i}(\mathcal{M}_{\text{ref}})} \tag{5}$$

We define $\mathcal{M}_{\text{ref}}$ to be the same base model as $\mathcal{M}$, but trained on the full data pool instead of a subset. Therefore, $\text{Perf}(\mathcal{M}) = 1$ means that the model was able to achieve the same performance as the reference model, despite being trained on a substantially smaller subset.

We initialize $\mathcal{M}$ with Llama-3.2-1B [31] for math reasoning and DeBERTa-v3-large [13] for NLI, and select benchmarks that are considered to be out-of-distribution from our seed datasets. Specifically, we aggregate performance across 7 math benchmarks—SAT-Math, MMLU Elementary, High School,

Table 1: Correlation between model OOD performance and data diversity measures on 100K scale. Compared to baseline measures, **G-Vendi strongly correlates with how the model generalizes to unseen distributions.** The OOD performance for each task is defined as the average relative performance on 7 unseen benchmarks (see §2.2 for more details). For each measure, we fit both log-linear and linear trendlines to the collected data, and report the higher resulting R^2.

Diversity Measure	Math OOD Performance		NLI OOD Performance		Math ID Performance	
	R^2	Spearman's ρ	R^2	Spearman's ρ	R^2	Spearman's ρ
Embedding Vendi	0.659	0.754	0.622	0.841	0.421	0.614
Embedding Dissimilarity	0.583	0.751	0.582	0.806	0.397	0.574
2-gram Entropy	0.549	0.736	0.497	0.664	0.243	0.548
Average Perplexity	0.442	-0.631	0.321	0.597	0.200	-0.341
Skill Set Entropy	0.706	0.812	-	-	0.527	0.695
G-Vendi	**0.823**	**0.899**	**0.791**	**0.893**	**0.697**	**0.780**

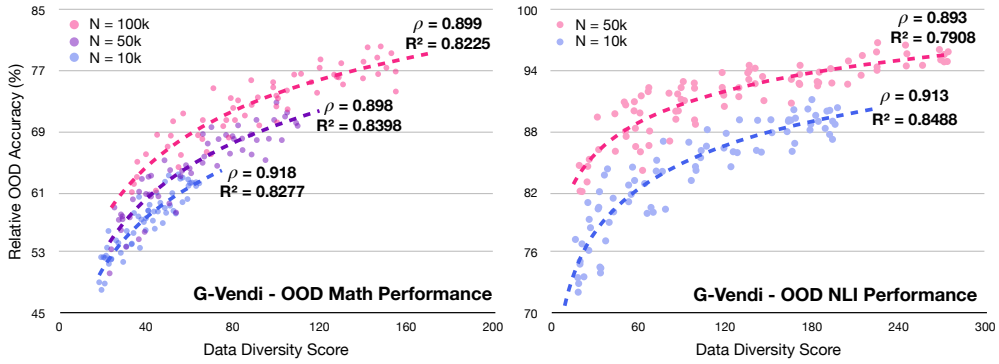

Figure 2: G-Vendi and model OOD performance. **G-Vendi shows a strong log-linear relationship with model performance, when controlling for data scale and quality.** In both tasks, models trained with datasets of high G-Vendi tend to generalize better in OOD benchmarks. Plots for baseline measures are shown in §B.1.

College Math, GSM-IC, Aqua-RAT and Minerva-Math [71, 14, 52, 27, 22]—where the reference model $\mathcal{M}_{\text{ref}}$ achieves meaningful accuracy above 10%—and 7 NLI benchmarks—HANS, WNLI, ANLI, QNLI, NLI Diagnostics, BigBench NLI and ConTRoL [37, 42, 56, 53, 29].

Baselines For baselines, we include *Embedding Vendi*: Vendi score with an off-the-shelf embedding model to represent each sample, and *Embedding DisSim*, which measures the average $1 -$ cosine similarity between all pairs of sample embeddings. We use gte-Qwen-7B-Instruct [25], a state-of-the-art embedding model on MTEB benchmark [39]. We also include two traditional metrics *2-gram Entropy* and *Perplexity*, along with an LLM-based metric Skill-Set Entropy for math reasoning—that prompts LLM to extract "reasoning skills" involved in each sample, then measures the entropy of the extracted skill set distributions [6, 33]. We leave further baseline details in §A.1.

2.3 Evaluation Results

2.3.1 Main Results

G-Vendi strongly predicts empirical generalization. Our main results comparing model generalization against diversity measures are shown in Fig. 2 and Table 1. Overall, we find that G-Vendi strongly predicts model generalization—with $R^2 \approx 0.8$ and Spearman's $\rho \approx 0.9$ in both tasks. Notably, the measure outperforms (1) *Embedding Vendi* that utilizes a state-of-the-art embedding model 14 times larger than the proxy model in G-Vendi, and (2) *Skill Set Entropy* which employs GPT-4 and Qwen2.5-72B-Instruct to taxonomize and extract instance-level skill sets. Overall, the result demonstrates that G-Vendi is surprisingly effective in predicting model generalization—due in part to the gradient representation's stronger alignment with task-relevant cognitive processes, as opposed to the surface form features prioritized in the baselines (§2.3.2).

G-Vendi often outweighs scale for OOD generalization, but scale primarily drives in-distribution performance. As shown in Fig. 2, training on larger scale datasets generally improves OOD

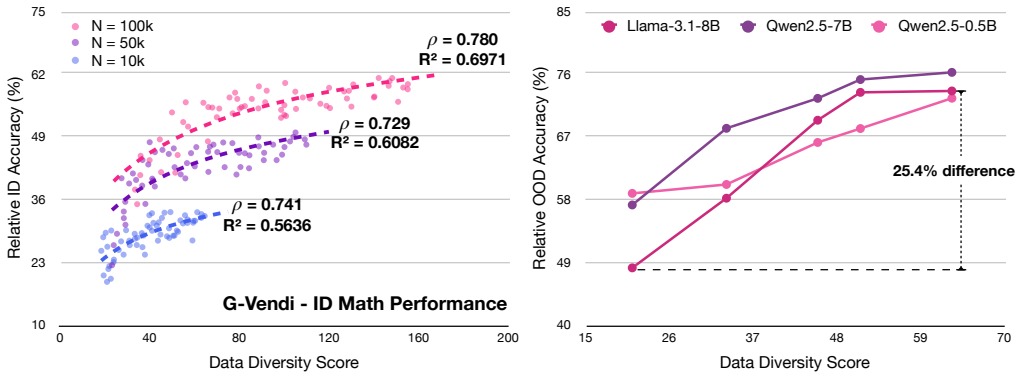

Figure 3: (Left) G-Vendi and in-distribution performance. Compared to OOD, ID performance is more heavily dominated by the scale of the training dataset—*e.g.,* 10K datasets with high diversity are less likely to outperform 50K datasets, compared to OOD results in Fig. 2. But significantly low diversity can still harm in-distribution performance. (Right) Ablation on the student model. **Higher G-Vendi correlates with stronger OOD performance, across model family and scale.**

performance. However, diversity can override the scale, *e.g.,* a 10K dataset with more diversity often outperforms a 100K dataset with less diversity.

We also investigate how data scale and diversity impact in-distribution (ID) performance. In Fig. 3, we evaluate the same model checkpoints as in Fig. 2, but on the in-distribution test set of MATH and GSM8k. Interestingly, while data diversity does correlate positively with ID performance, the strength of correlation is consistently weaker than on OOD across all data scale. While significantly low diversity (*e.g.,* below 40 for $N = 100k$) does harm model's ID performance, above this level, higher data diversity does not necessarily improve performance at each data scale. Furthermore, the overall performance on the ID test sets is generally much lower than on OOD (Fig. 2), where training on 100k diverse samples can recover up to 80% performance of $\mathcal{M}_{\text{ref}}$. These results quantitatively confirm that scale complements diversity by enhancing in-distribution performance, aligning with prior observations in narrower task setups or with task-specific measures [70, 72, 32].

2.3.2 Understanding What G-Vendi Encodes

Despite the surprising effectiveness of G-Vendi, it remains unclear whether the result stems from a fundamental difference in how G-Vendi defines diversity compared to baseline measures, such as *Embedding Vendi*. We analyze this by generating two specialized datasets, *DiverseReason* and *DiversePersona*. Specifically, we adopt the persona-guided generation framework of Ge et al. [9], and prompt Qwen2.5-72B-Instruct to generate math problems conditioned on both the few-shot examples and a random-sampled persona. For *DiverseReason*, we create 100k data points from 10k seed samples, along with only 0.1k personas. Conversely, for *DiversePersona*, we create 100k data points from only 0.1k seed samples, but with 10k personas. This design ensures that we induce a systematic difference between the two datasets, specifically by making *DiverseReason* richer in reasoning patterns but narrower in topical breadth and phrasing than *DiversePersona*.

Table 2 presents the diversity scores of the two datasets and the model performance after training on each dataset. As expected, the model trained on *DiverseReason* shows substantially better performance than with *DiversePersona*, as it has been exposed to a variety of problems with distinct reasoning patterns. Here, we find a notable difference between G-Vendi and *Embedding Vendi*—while G-Vendi consistently assigns higher score to subsets of *DiverseReason*, *Embedding Vendi*

Table 2: **Embedding diversity prioritizes semantic variability, while G-Vendi prioritizes solution template diversity.** We generate *DiverseReason* and *DiversePersona* in a controlled setup, and report their OOD performance and diversity scores along with standard deviation over 5 independent runs. See §F for qualitative examples comparing G-Vendi and *Embedding Vendi*.

Dataset	N = 10k			N = 50k		
	OOD Perf.	G-Vendi	Embedding Vendi	OOD Perf.	G-Vendi	Embedding Vendi
DiverseReason	**54.77 (3.49)**	**63.22 (1.05)**	22.03 (0.89)	**63.48 (2.59)**	**103.63 (1.25)**	30.79 (0.52)
DiversePersona	38.15 (8.02)	51.94 (1.24)	**24.49 (0.45)**	42.14 (6.92)	87.864 (1.19)	**31.16 (0.59)**

gives higher diversity to *DiversePersona*, prioritizing semantic diversity over reasoning patterns. The results suggest that (1) heuristic diversification methods such as using persona may not necessarily improve the type of diversity helpful for the task, and (2) G-Vendi can be a useful tool to capture the task-specific notion of diversity—particularly in reasoning domains—by emphasizing variations in the underlying cognitive process rather than surface-level differences.

2.3.3 Impact of Gradient Proxy Model

Next, we ablate the proxy model used for computing gradient representations on math reasoning. In Table 3, we replace the proxy model with Llama-3.2-1B-Instruct and Qwen2.5-0.5B, and report their rank correlation on 10k subsets with (1) the original G-Vendi measured using Qwen2.5-0.5B-Instruct and (2) model performance. Llama3.2-1B-Instruct serves well as a proxy model, performing slightly better than Qwen2.5-0.5B-Instruct; we posit that this is because the model originates from the same base model as our student model $\mathcal{M}$ (base Llama3.2-1B). In addition, we find that using an instruction-tuned proxy is helpful, as the diversity measured under a base model yields a comparatively weaker correlation with model generalization. Overall, G-Vendi provides a stable estimate of data diversity with consistently high correlation across proxy models, suggesting its effectiveness is not specific to any single configuration.

Table 3: **G-Vendi is stable with proxy models of different sizes and model families.** We report rank correlation with the original proxy model and with model OOD performance on math reasoning.

Proxy Model	ρ w/ Qwen2.5-0.5B-It	ρ w/ OOD Perf.
Qwen2.5-0.5B-It	1	0.898
Llama-3.2-1B-It	0.899	0.909
Qwen2.5-0.5B	0.811	0.772

2.3.4 Impact of Student Model

We also analyze whether G-Vendi is consistent across different student models other than Llama-3.2-1B. In Fig. 3 (Right), we randomly pick 5 distinct subsets from our math data pool, each with 10K samples, train 3 distinct student models on these subsets, then evaluate their OOD performance. Again, datasets with higher diversity consistently lead to better performance across model scales and families. Notably, Llama-3.1-8B [31] trained on the same-sized subsets show up to 25.4% difference in OOD accuracy, depending on the measured diversity of their training sets. The results show that G-Vendi captures a consistent aspect of dataset quality that holds across models—diverse datasets tend to be broadly beneficial, not just tailored to a particular training setup.

3 Prismatic Synthesis

The strong performance of G-Vendi provides us with a compelling prospect—by strategically improving data diversity in gradient space, we can improve our model performance **even without knowledge of the target distribution**. We present Prismatic Synthesis, a simple yet effective framework to improve data diversity by generating novel synthetic data. An overview of Prismatic Synthesis is shown in Fig 1. Starting from a seed dataset, the framework repeats the following 3 steps:

Step 1: Cluster existing samples in gradient space. Following §2.1, we compute the loss gradient of existing samples with an off-the-shelf proxy model, and cluster them using K-means.

Step 2: Generate new samples from existing samples. We prompt an LLM with few-shot examples sampled from the current data pool to generate new data points.

Step 3: Diversify by leaving only the samples in sparse clusters. Among the generated data points, we only add to the pool those samples that belong to sparse clusters (*e.g.,* the top 20% of clusters with the smallest number of members).

Iterating these steps, we collect the samples currently underrepresented in the gradient space, thereby greedily improving the data diversity. In the following section, we describe in detail how we apply Prismatic Synthesis to generate two state-of-the-art datasets, *PrismMath* and *PrismNLI*.

3.1 *PrismMath* and *PrismNLI*

Generation and Diversification The generation process takes the few-shot approach in §2.2. For NLI, we start from WANLI [28] with 103k samples as seed dataset. We use Qwen2.5-72B-Instruct

Table 4: (Top) **PrismMath-7B, distilled from a 32B LLM, outperforms state-of-the-art reasoning models trained on solutions generated by a 671B LLM and further verified by human-written answers.** All scores are `pass@1`. (Bottom) **PrismNLI outperforms the best prior mixture of datasets by 8% across 8 OOD benchmarks.** We present full results with standard deviation and data size comparison in §B.2.

Model	Data Generator	AIME24	AIME25	AMC23	MATH500	MATH^2	Olympiad Bench	GSM8K Platinum	Avg.
Qwen2.5-Math-7B-It	-	14.17	9.91	72.50	83.80	57.62	44.29	96.11	54.06
OpenR1-Qwen-7B	R1-671B	47.91	30.41	87.19	90.60	78.10	67.06	**96.69**	71.14
OpenThinker-7B		27.50	22.50	74.06	84.20	67.62	45.93	93.05	59.27
OpenThinker2-7B		50.00	35.00	88.44	91.40	78.10	**69.63**	93.96	72.36
R1-7B		54.66	33.33	92.50	**92.60**	78.57	68.00	89.91	72.80
PrismMath-7B	R1-32B	**57.08**	**38.33**	**93.75**	92.40	**80.95**	68.30	95.95	**75.25**

Dataset	Data Generator	HANS	WNLI	ANLI R1	ANLI R2	ANLI R3	Diagnostics	BigBench	Control	Avg.
MNLI	ChatGPT, Humans	78.47	63.03	60.10	45.30	42.58	81.88	78.22	42.16	61.47
WANLI		89.25	75.21	61.30	46.50	44.90	83.68	80.81	43.93	65.70
MNLI+FEVER		74.62	66.29	60.20	47.00	41.75	80.89	76.14	48.51	61.93
{WA+M+S}NLI		80.18	69.41	65.00	50.60	45.25	83.77	84.72	50.49	66.18
PrismNLI	Qwen2.5-72B	**92.44**	**78.47**	**73.70**	**61.90**	**57.00**	**86.13**	**86.32**	**58.25**	**74.28**

to generate both the new problem and a corresponding label (either *entailment*, *neutral*, or *label*), given 5-shot examples random-sampled from the existing data pool. For math reasoning, we start from OpenR1-Math [16] with 94k samples. Instead of generating both problem and solution in one pass, we first generate problems by 5-shot prompting Qwen2.5-72B-Instruct. Then we use R1-Distill-Qwen-32B (R1-32B) to annotate solution traces for each generated problem. This two-stage process allows us to distill self-reflecting capabilities of R1-like models, known to be particularly effective in hard reasoning tasks [44, 64]. We dynamically set the number of clusters k to be 1% of the existing data pool size at each iteration, and leave only the samples that correspond to the smallest $k/2$ clusters.

Quality Filtering and Decontamination Our pipeline generates data in an *entirely automated fashion*, i.e., it does not involve any ground-truth answers or manual verification. This is in contrast to dominant approaches for synthetic data generation, which first collect human-written problems and their ground-truth answers, then augment them with model-generated solution traces [23, 40]. To improve data quality, we perform majority-voting based filtering—we sample N solutions for R1-32B for each generated problem, and compare their answers. When the number of majority answers is above a threshold τ, we consider those answers to be "verified", and add them to the pool. We set $N = 3, \tau = 2$ for *PrismMath* and $N = 2, \tau = 2$ for *PrismNLI*. We then run decontamination against all test benchmarks we used (Table 4). We adopt the most conservative methods in literature, by first applying brute-force 10-gram matching and subsequently running LLM-based paraphrase detection [62]. After filtering and decontamination, we are left with 1.0M problem-solution pairs in *PrismMath*, and 515K input-label pairs for *PrismNLI*.

Evaluation Setup We evaluate the quality of our datasets by training models and evaluating them on a suite of benchmarks. For math reasoning, we fine-tune Qwen2.5-Math-7B-Instruct, yielding *PrismMath-7B*. We then compare our model against state-of-the-art distilled reasoning models at 7B scale, such as DeepSeek-R1-Distill-Qwen-7B (R1-7B) and OpenThinker2-7B. Note that these models are trained on heavily curated datasets, whose solution traces are generated by R1-671B and verified based on ground-truth answers. For NLI, we train Deberta-V3-large on PrismNLI, and compare against the same base model trained on (a mixture of) widely used datasets—such as a mixture of WANLI, MNLI and SNLI. We illustrate further experimental details in §A.2.

3.2 Evaluation Results

3.2.1 Main Results

PrismMath-7B outperforms state-of-the-art distilled reasoning models. The results for math reasoning tasks are shown in Table 4 (Top). *PrismMath-7B* yields surprisingly strong performance across benchmarks, outperforming all state-of-the-art baselines distilled for hard reasoning tasks. Notably, despite being distilled from R1-32B without any human verification involved, our model

outperforms OpenThinker2, a model trained on 1.14M samples distilled from R1-671B and further verified using ground-truth answers.

PrismNLI improves by 8% from the best prior data mixture. In Table 4 (Bottom), *PrismNLI* with 515k samples achieves substantially better OOD performance than a mixture of widely-used, large-scale datasets. For example, *PrismNLI* outperforms the mixture of WANLI, MNLI and SNLI by 8% on average OOD accuracy, despite being only the half the size of the baseline and not relying on any human annotation. Overall, these results show that the benefits of strategic diversification may exceed that of expensive curation techniques, such as a stronger data generator or manual verification. In §F, we analyze examples of actual clusters discovered in both math reasoning and NLI tasks.

3.2.2 Impact of Diversification on Synthetic Data Scaling

Is diversification in gradient space truly necessary? Although Prismatic Synthesis shows clear improvement in model performance, it is questionable whether similar gains could have been achieved with heuristic diversification—or even without any diversification, but by just scaling synthetic data. To address this concern, we compare the scaling of synthetic data generated from alternative diversification strategies—vanilla few-shot generation and persona-guided generation [9]—up to 100K samples. Further experimental details are in §A.2.

The results are shown in Fig. 1 (Right). In both vanilla few-shot and persona-guided generation, model performance faces early saturation—around 50K $\sim$ 100K scale. Notably, while persona-guided generation outperforms other methods at low data scale (below 5K), it quickly plateaus, eventually converging to the performance achievable with vanilla few-shot prompting. This result attests to our earlier observation that heuristic diversification—that optimizes for variances in surface form—may ultimately fall short, as it overlooks the task-specific nature of diversity necessary for effective generalization. This contrasts with Prismatic Synthesis, where performance continues to improve even at scale beyond 100K, as demonstrated by PrismMath with 1 million samples (Table 4).

4 Related Work

Data diversity is often considered a key factor in LLM post-training. Motivating from ideas of data selection and active learning [35, 1, 21], prior works have shown that training on diverse instruction-tuning data improves sample efficiency and model robustness [72, 2, 30, 3, 11]. Yet, these approaches often rely on either task-specific heuristics—*e.g.,* variance in instruction metadata [58, 33, 65]—or embedding similarity [60, 66, 41] as a proxy for diversity, which may be insufficient for reasoning-oriented tasks. Gradient-based representation is a promising alternative to these proxies that can approximate training data influence [48], and has been adopted for data selection [68, 61]. Wang et al. [57] proposes a gradient kernel-based approach to instruction-tuning diversity, but requires LoRA adaptation and omits systematic comparison against canonical diversity measures. Our work builds upon these prior works, to (1) provide a large-scale empirical analysis on the impact of data diversity on model generalization, and (2) introduce G-Vendi, a scalable yet task-sensitive measure that requires no model adaptation.

Synthetic data are being increasingly adopted for improving LLM capabilities, particularly for reasoning-heavy domains such as math and code [44, 16, 34]. In these settings, LLMs often play a central role as data generators—augmenting solutions for existing prompts [40, 64], rephrasing human-curated datasets [65, 36], and bootstraping novel problems [54, 51]. Despite the broad applicability of synthetic data, subsequent analyses also report that naive scaling of synthetic data may yield significant duplicates and distributional biases [10, 67]. Several techniques have been proposed to improve data diversity—*e.g.,* conditioning on auxiliary attributes [67, 24, 9] or maximizing pairwise embedding distances [60]—but their efficacy essentially relies on the quality of the heuristic attributes and embeddings in capturing task-specific notion of diversity. Prismatic Synthesis provides a principled alternative to these approaches, allowing for improvements in model generalization through a task-agnostic gradient diversification process.

5 Conclusion

Our work investigates how to quantify and leverage data diversity to improve LLM reasoning. We show that the exponentiated entropy of data samples in a gradient space—as measured by G-Vendi—

strongly correlates with the empirical generalization of the model, significantly outperforming prior metrics that rely on heuristic features. Building on this insight, we introduce Prismatic Synthesis, an algorithm for targeted generation of diverse synthetic data in gradient space. Our resulting datasets, *PrismMath* and *PrismNLI*, yield state-of-the-art models that generalize well not only on in-distribution but also across challenging out-of-distribution benchmarks, highlighting the importance of principled diversification over strong generators and curation pipelines.

References

[1] J. T. Ash, C. Zhang, A. Krishnamurthy, J. Langford, and A. Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds, 2020.

[2] A. Bukharin, S. Li, Z. Wang, J. Yang, B. Yin, X. Li, C. Zhang, T. Zhao, and H. Jiang. Data diversity matters for robust instruction tuning. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3411–3425, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics.

[3] H. Chen, A. Waheed, X. Li, Y. Wang, J. Wang, B. Raj, and M. I. Abdin. On the diversity of synthetic data and its impact on training large language models, 2024.

[4] K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, C. Hesse, and J. Schulman. Training verifiers to solve math word problems, 2021.

[5] DeepSeek-AI, D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi, X. Zhang, X. Yu, Y. Wu, Z. F. Wu, Z. Gou, Z. Shao, Z. Li, Z. Gao, A. Liu, B. Xue, B. Wang, B. Wu, B. Feng, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan, D. Dai, D. Chen, D. Ji, E. Li, F. Lin, F. Dai, F. Luo, G. Hao, G. Chen, G. Li, H. Zhang, H. Bao, H. Xu, H. Wang, H. Ding, H. Xin, H. Gao, H. Qu, H. Li, J. Guo, J. Li, J. Wang, J. Chen, J. Yuan, J. Qiu, J. Li, J. L. Cai, J. Ni, J. Liang, J. Chen, K. Dong, K. Hu, K. Gao, K. Guan, K. Huang, K. Yu, L. Wang, L. Zhang, L. Zhao, L. Wang, L. Zhang, L. Xu, L. Xia, M. Zhang, M. Zhang, M. Tang, M. Li, M. Wang, M. Li, N. Tian, P. Huang, P. Zhang, Q. Wang, Q. Chen, Q. Du, R. Ge, R. Zhang, R. Pan, R. Wang, R. J. Chen, R. L. Jin, R. Chen, S. Lu, S. Zhou, S. Chen, S. Ye, S. Wang, S. Yu, S. Zhou, S. Pan, S. S. Li, S. Zhou, S. Wu, S. Ye, T. Yun, T. Pei, T. Sun, T. Wang, W. Zeng, W. Zhao, W. Liu, W. Liang, W. Gao, W. Yu, W. Zhang, W. L. Xiao, W. An, X. Liu, X. Wang, X. Chen, X. Nie, X. Cheng, X. Liu, X. Xie, X. Liu, X. Yang, X. Li, X. Su, X. Lin, X. Q. Li, X. Jin, X. Shen, X. Chen, X. Sun, X. Wang, X. Song, X. Zhou, X. Wang, X. Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. Zhang, Y. Xu, Y. Li, Y. Zhao, Y. Sun, Y. Wang, Y. Yu, Y. Zhang, Y. Shi, Y. Xiong, Y. He, Y. Piao, Y. Wang, Y. Tan, Y. Ma, Y. Liu, Y. Guo, Y. Ou, Y. Wang, Y. Gong, Y. Zou, Y. He, Y. Xiong, Y. Luo, Y. You, Y. Liu, Y. Zhou, Y. X. Zhu, Y. Xu, Y. Huang, Y. Li, Y. Zheng, Y. Zhu, Y. Ma, Y. Tang, Y. Zha, Y. Yan, Z. Z. Ren, Z. Ren, Z. Sha, Z. Fu, Z. Xu, Z. Xie, Z. Zhang, Z. Hao, Z. Ma, Z. Yan, Z. Wu, Z. Gu, Z. Zhu, Z. Liu, Z. Li, Z. Xie, Z. Song, Z. Pan, Z. Huang, Z. Xu, Z. Zhang, and Z. Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025.

[6] A. Didolkar, A. Goyal, N. R. Ke, S. Guo, M. Valko, T. Lillicrap, D. Rezende, Y. Bengio, M. Mozer, and S. Arora. Metacognitive capabilities of llms: An exploration in mathematical problem solving, 2024.

[7] C. Fourrier, N. Habib, H. Kydlíček, T. Wolf, and L. Tunstall. Lighteval: A lightweight framework for llm evaluation, 2023.

[8] D. Friedman and A. B. Dieng. The vendi score: A diversity evaluation metric for machine learning, 2023.

[9] T. Ge, X. Chan, X. Wang, D. Yu, H. Mi, and D. Yu. Scaling synthetic data creation with 1,000,000,000 personas, 2024.

[10] S. Gunasekar, Y. Zhang, J. Aneja, C. C. T. Mendes, A. D. Giorno, S. Gopi, M. Javaheripi, P. Kauffmann, G. de Rosa, O. Saarikivi, A. Salim, S. Shah, H. S. Behl, X. Wang, S. Bubeck, R. Eldan, A. T. Kalai, Y. T. Lee, and Y. Li. Textbooks are all you need, 2023.

[11] A. Havrilla, A. Dai, L. O'Mahony, K. Oostermeijer, V. Zisler, A. Albalak, F. Milo, S. C. Raparthy, K. Gandhi, B. Abbasi, D. Phung, M. Iyer, D. Mahan, C. Blagden, S. Gureja, M. Hamdy, W.-D. Li, G. Paolini, P. S. Ammanamanchi, and E. Meyerson. Surveying the effects of quality, diversity, and complexity in synthetic data from large language models, 2024.

[12] C. He, R. Luo, Y. Bai, S. Hu, Z. L. Thai, J. Shen, J. Hu, X. Han, Y. Huang, Y. Zhang, J. Liu, L. Qi, Z. Liu, and M. Sun. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems, 2024.

[13] P. He, X. Liu, J. Gao, and W. Chen. Deberta: Decoding-enhanced bert with disentangled attention, 2021.

[14] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[15] D. Hendrycks, C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, and J. Steinhardt. Measuring mathematical problem solving with the math dataset, 2021.

[16] Hugging Face. Open r1: A fully open reproduction of deepseek-r1, January 2025.

[17] W. B. Johnson, J. Lindenstrauss, et al. Extensions of lipschitz mappings into a hilbert space. *Contemporary mathematics*, 26(189-206):1, 1984.

[18] J. Jung, X. Lu, L. Jiang, F. Brahman, P. West, P. W. Koh, and Y. Choi. Information-theoretic distillation for reference-less summarization. In *First Conference on Language Modeling*, 2024.

[19] J. Jung, P. West, L. Jiang, F. Brahman, X. Lu, J. Fisher, T. Sorensen, and Y. Choi. Impossible distillation for paraphrasing and summarization: How to make high-quality lemonade out of small, low-quality model. In K. Duh, H. Gomez, and S. Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4439–4454, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[20] K. Killamsetty, S. Durga, G. Ramakrishnan, A. De, and R. Iyer. Grad-match: Gradient matching based data subset selection for efficient deep model training. In *International Conference on Machine Learning*, pages 5464–5474. PMLR, 2021.

[21] K. Killamsetty, D. Sivasubramanian, G. Ramakrishnan, and R. Iyer. Glister: Generalization based data subset selection for efficient and robust learning, 2021.

[22] A. Lewkowycz, A. Andreassen, D. Dohan, E. Dyer, H. Michalewski, V. Ramasesh, A. Slone, C. Anil, I. Schlag, T. Gutman-Solo, Y. Wu, B. Neyshabur, G. Gur-Ari, and V. Misra. Solving quantitative reasoning problems with language models, 2022.

[23] J. LI, E. Beeching, L. Tunstall, B. Lipkin, R. Soletskyi, S. C. Huang, K. Rasul, L. Yu, A. Jiang, Z. Shen, Z. Qin, B. Dong, L. Zhou, Y. Fleureau, G. Lample, and S. Polu. Numina-math. [https://huggingface.co/AI-MO/NuminaMath-CoT](https://github.com/project-numina/aimo-progress-prize/blob/main/report/numina_dataset.pdf), 2024.

[24] Y. Li, S. Bubeck, R. Eldan, A. D. Giorno, S. Gunasekar, and Y. T. Lee. Textbooks are all you need ii: phi-1.5 technical report, 2023.

[25] Z. Li, X. Zhang, Y. Zhang, D. Long, P. Xie, and M. Zhang. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*, 2023.

[26] H. Lightman, V. Kosaraju, Y. Burda, H. Edwards, B. Baker, T. Lee, J. Leike, J. Schulman, I. Sutskever, and K. Cobbe. Let's verify step by step, 2023.

[27] W. Ling, D. Yogatama, C. Dyer, and P. Blunsom. Program induction by rationale generation: Learning to solve and explain algebraic word problems. *ACL*, 2017.

[28] A. Liu, S. Swayamdipta, N. A. Smith, and Y. Choi. Wanli: Worker and ai collaboration for natural language inference dataset creation, 2022.

[29] H. Liu, L. Cui, J. Liu, and Y. Zhang. Natural language inference in context - investigating contextual reasoning over long texts. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13388–13396, May 2021.

[30] W. Liu, W. Zeng, K. He, Y. Jiang, and J. He. What makes good data for alignment? a comprehensive study of automatic data selection in instruction tuning. In *The Twelfth International Conference on Learning Representations*, 2024.

[31] Llama Team. The llama 3 herd of models, 2024.

[32] S. Longpre, L. Hou, T. Vu, A. Webson, H. W. Chung, Y. Tay, D. Zhou, Q. V. Le, B. Zoph, J. Wei, and A. Roberts. The flan collection: Designing data and methods for effective instruction tuning, 2023.

[33] K. Lu, H. Yuan, Z. Yuan, R. Lin, J. Lin, C. Tan, C. Zhou, and J. Zhou. #instag: Instruction tagging for analyzing supervised fine-tuning of large language models, 2023.

[34] X. Lu, S. Han, D. Acuna, H. Kim, J. Jung, S. Prabhumoye, N. Muennighoff, M. Patwary, M. Shoeybi, B. Catanzaro, and Y. Choi. Retro-search: Exploring untaken paths for deeper and efficient reasoning, 2025.

[35] A. Maharana, P. Yadav, and M. Bansal. D2 pruning: Message passing for balancing diversity and difficulty in data pruning, 2023.

[36] P. Maini, S. Seto, H. Bai, D. Grangier, Y. Zhang, and N. Jaitly. Rephrasing the web: A recipe for compute and data-efficient language modeling, 2024.

[37] R. T. McCoy, E. Pavlick, and T. Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In A. Korhonen, D. Traum, and L. Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy, July 2019. Association for Computational Linguistics.

[38] B. Mirzasoleiman, J. Bilmes, and J. Leskovec. Coresets for data-efficient training of machine learning models, 2020.

[39] N. Muennighoff, N. Tazi, L. Magne, and N. Reimers. Mteb: Massive text embedding benchmark, 2023.

[40] N. Muennighoff, Z. Yang, W. Shi, X. L. Li, L. Fei-Fei, H. Hajishirzi, L. Zettlemoyer, P. Liang, E. Candès, and T. Hashimoto. s1: Simple test-time scaling, 2025.

[41] X. Ni, Y. Gong, Z. Gou, Y. Shen, Y. Yang, N. Duan, and W. Chen. Exploring the mystery of influential data for mathematical reasoning, 2024.

[42] Y. Nie, A. Williams, E. Dinan, M. Bansal, J. Weston, and D. Kiela. Adversarial NLI: A new benchmark for natural language understanding. In D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online, July 2020. Association for Computational Linguistics.

[43] OpenAI. Gpt-4 technical report, 2024.

[44] OpenThoughts Team. Open Thoughts. https://open-thoughts.ai, Jan. 2025.

[45] J. Pang, J. Wei, A. P. Shah, Z. Zhu, Y. Wang, C. Qian, Y. Liu, Y. Bao, and W. Wei. Improving data efficiency via curating llm-driven rating systems, 2025.

[46] S. M. Park, K. Georgiev, A. Ilyas, G. Leclerc, and A. Madry. Trak: Attributing model behavior at scale, 2023.

[47] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical recipes in C (2nd ed.): the art of scientific computing*. Cambridge University Press, USA, 1992.

[48] G. Pruthi, F. Liu, M. Sundararajan, and S. Kale. Estimating training data influence by tracing gradient descent, 2020.

[49] Qwen, :, A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, H. Lin, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Lin, K. Dang, K. Lu, K. Bao, K. Yang, L. Yu, M. Li, M. Xue, P. Zhang, Q. Zhu, R. Men, R. Lin, T. Li, T. Tang, T. Xia, X. Ren, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Wan, Y. Liu, Z. Cui, Z. Zhang, and Z. Qiu. Qwen2.5 technical report, 2025.

[50] B. T. Rakhshan and G. Rabusseau. Tensorized random projections. *ArXiv*, abs/2003.05101, 2020.

[51] V. Shah, D. Yu, K. Lyu, S. Park, J. Yu, Y. He, N. R. Ke, M. Mozer, Y. Bengio, S. Arora, and A. Goyal. Ai-assisted generation of difficult math questions, 2025.

[52] F. Shi, X. Chen, K. Misra, N. Scales, D. Dohan, E. Chi, N. Schärli, and D. Zhou. Large language models can be easily distracted by irrelevant context. *arXiv preprint arXiv:2302.00093*, 2023.

[53] A. Srivastava, A. Rastogi, A. Rao, A. A. M. Shoeb, A. Abid, A. Fisch, A. R. Brown, A. Santoro, A. Gupta, and A. G.-A. et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models, 2023.

[54] S. Toshniwal, W. Du, I. Moshkov, B. Kisacanin, A. Ayrapetyan, and I. Gitman. Openmathinstruct-2: Accelerating ai for math with massive open-source instruction data, 2024.

[55] J. Vendrow, E. Vendrow, S. Beery, and A. Madry. Do large language model benchmarks test reliability?, 2025.

[56] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In T. Linzen, G. Chrupała, and A. Alishahi, editors, *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium, Nov. 2018. Association for Computational Linguistics.

[57] P. Wang, Y. Shen, Z. Guo, M. Stallone, Y. Kim, P. Golland, and R. Panda. Diversity measurement and subset selection for instruction tuning datasets. In *ICLR 2025 Workshop on Navigating and Addressing Data Problems for Foundation Models*, 2025.

[58] Y. Wang, S. Mishra, P. Alipoormolabashi, Y. Kordi, A. Mirzaei, A. Arunkumar, A. Ashok, A. S. Dhanasekaran, A. Naik, D. Stap, E. Pathak, G. Karamanolakis, H. G. Lai, I. Purohit, I. Mondal, J. Anderson, K. Kuznia, K. Doshi, M. Patel, K. K. Pal, M. Moradshahi, M. Parmar, M. Purohit, N. Varshney, P. R. Kaza, P. Verma, R. S. Puri, R. Karia, S. K. Sampat, S. Doshi, S. Mishra, S. Reddy, S. Patro, T. Dixit, X. Shen, C. Baral, Y. Choi, N. A. Smith, H. Hajishirzi, and D. Khashabi. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks, 2022.

[59] K. Wei, R. Iyer, and J. Bilmes. Submodularity in data subset selection and active learning. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 1954–1963. JMLR.org, 2015.

[60] S. Wu, K. Lu, B. Xu, J. Lin, Q. Su, and C. Zhou. Self-evolved diverse data sampling for efficient instruction tuning, 2023.

[61] M. Xia, S. Malladi, S. Gururangan, S. Arora, and D. Chen. Less: Selecting influential data for targeted instruction tuning, 2024.

[62] S. Yang, W.-L. Chiang, L. Zheng, J. E. Gonzalez, and I. Stoica. Rethinking benchmark and contamination for language models with rephrased samples, 2023.

[63] T. Ye, Z. Xu, Y. Li, and Z. Allen-Zhu. Physics of language models: Part 2.1, grade-school math and the hidden reasoning process, 2024.

[64] Y. Ye, Z. Huang, Y. Xiao, E. Chern, S. Xia, and P. Liu. Limo: Less is more for reasoning, 2025.

[65] L. Yu, W. Jiang, H. Shi, J. Yu, Z. Liu, Y. Zhang, J. T. Kwok, Z. Li, A. Weller, and W. Liu. Metamath: Bootstrap your own mathematical questions for large language models, 2024.

[66] S. Yu, L. Chen, S. Ahmadian, and M. Fadaee. Diversify and conquer: Diversity-centric data selection with iterative refinement, 2025.

[67] Y. Yu, Y. Zhuang, J. Zhang, Y. Meng, A. Ratner, R. Krishna, J. Shen, and C. Zhang. Large language model as attributed training data generator: A tale of diversity and bias. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

[68] Z. Yu, S. Das, and C. Xiong. Mates: Model-aware data selection for efficient pretraining with data influence models, 2024.

[69] L. Yuan, Y. Chen, G. Cui, H. Gao, F. Zou, X. Cheng, H. Ji, Z. Liu, and M. Sun. Revisiting out-of-distribution robustness in nlp: Benchmark, analysis, and llms evaluations, 2023.

[70] D. Zhang, J. Wang, and F. Charton. Instruction diversity drives generalization to unseen tasks, 2024.

[71] W. Zhong, R. Cui, Y. Guo, Y. Liang, S. Lu, Y. Wang, A. Saied, W. Chen, and N. Duan. AGIEval: A human-centric benchmark for evaluating foundation models. In K. Duh, H. Gomez, and S. Bethard, editors, *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2299–2314, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[72] C. Zhou, P. Liu, P. Xu, S. Iyer, J. Sun, Y. Mao, X. Ma, A. Efrat, P. Yu, L. Yu, S. Zhang, G. Ghosh, M. Lewis, L. Zettlemoyer, and O. Levy. Lima: Less is more for alignment, 2023.

A Experimental Details

A.1 Evaluating Diversity Measures

Generating Data Pool We further illustrate the experimental setup for evaluating data diversity measures. As described in §2.2, we 5-shot prompt Qwen2.5-72B-Instruct as data generator for both NLI and math reasoning tasks, generating 1.5M samples for both domains. For NLI, we generate both the input (*i.e.,* premise, hypothesis) and the corresponding label in one pass. For math reasoning, we take a two-step process, first generating novel problems given the few-shot examples, then solving each generated problem. Prompts for both tasks are shown in §A.3.

Sampling Subsets We then sample distinct 300 subsets in total from the generated data pool, spanning a varied range of diversity. This process is non-trivial, however, because repeated random sampling of subsets would lead to subsets with similar diversity, thus not covering a wide spectrum of diversity values[2]. We therefore employ 4 distinct sampling strategies to cover the diversity spectrum:

- ***Random Sampling:*** Random-sample a subset of given size from the data pool.
- ***Higher Diversity Sampling:*** Search for a higher diversity subset with cluster-guided balanced sampling. Given representations of samples in the data pool, we first perform K-means clustering over the representations. We then perform balanced sampling across each cluster to up-sample from sparse clusters and down-sample from dense clusters in the given representation space.
- ***Lower Diversity Sampling:*** Reduce diversity by iteratively adding similar samples to the current members of the subset. Given representations of samples in the data pool, we randomly pick a small seed samples to initialize a subset, then iteratively add a batch of new samples whose similarity with at least one of the current subset members is above a predefined threshold.
- ***Mixture Sampling:*** Given a pair of random / high-diversity / low-diversity subsets, randomly mix them to create a new subset of a given size.

We present formal descriptions of *Higher Diversity Sampling* and *Lower Diversity Sampling* in Algorithm 1 and Algorithm 2. Note that both strategies do not search for globally maximal / minimal diversity, which distinguish them from submodular data selection methods [59, 38]; instead, we deliberately introduce adjustable parameters and stochasticity in to the algorithms so that we yield distinct subsets across diversity level, while not overlapping too much with each other. We use both the gradient and embedding representations of data, so that subsets are spread out not only with respect to G-Vendi but also in baseline metrics (*e.g., Embedding Vendi*).

Training Models and Baseline Measures Finally, we train models on the sampled subsets. We train DeBERTa-v3-large and Llama-3.2-1B for our main experiments, which only requires 1 H100 GPU per training run thanks to the small number of parameters. In practice, we parallelize the training of distinct models over up to 8 H100 nodes. Along with G-Vendi, we consider 5 baseline measures widely used in the literature—*Embedding Vendi*, *Embedding Dissimilarity*, *2-gram Entropy*, *Average Perplexity*, *Skill Set Entropy*. *Average Perplexity* is measured with Qwen2.5-0.5B-Instruct, the same model as the gradient proxy model in G-Vendi. *Skill-Set Entropy* is an LLM-based approach where the taxonomy of *skill sets* are first collected by prompting LLMs with each sample in the data pool, then measuring the data diversity via the entropy of *skill sets* in the data points. Since we generate our data pool using MATH and GSM8k as seed set, we borrow the taxonomy of skill sets from Didolkar et al. [6] that extracted skill sets in MATH and GSM8k using GPT-4 [43]. We then prompt Qwen2.5-72B-Instruct to map each problem in our data pool to all corresponding skill sets from the taxonomy. Subsequently, we compute the entropy of skill sets in each subset.

A.2 Prismatic Synthesis

Generating *PrismMath* and *PrismNLI* All datasets used for Prismatic Synthesis can be used for academic settings, licensed by either cc-by-4.0, MIT or Apache 2.0. The prompts shown for generating PrismMath and PrismNLI are shown in §A.3. For math reasoning, we generate solutions with max sequence length of 16K using R1-32B. All experiments and data generation are done in

[2]This is somewhat intuitive because if multiple subsets are sampled from the same population using the same sampling strategy, the resulting test statistics (e.g., the mean) are expected to exhibit low variance.

Algorithm 1 *Higher Diversity Sampling*

Input: Data representation $D \in \mathbb{R}^{|\mathcal{D}| \times d}$, number of clusters k and target subset size N_{target}
Output: Indices of selected subset $S \subseteq \{1, \cdots, |\mathcal{D}|\}$

$S \leftarrow \emptyset$ ▷ Initialize the subset.
$\{c_1, \cdots, c_k\} = \text{K-means}(D)$ ▷ Cluster data. c_i is a set of indices corresponding to cluster i.
while $|S| < N_{\text{target}}$ **do**
  $c \leftarrow \text{random-sample}(\{c_1, \cdots, c_k\}, 1)$ ▷ Randomly pick a sampling cluster.
  $S_{\text{new}} \leftarrow \text{random-sample}\left(c, \left\lceil \frac{N_{\text{target}}}{k} \right\rceil\right)$ ▷ Sample new samples from the chosen cluster.
  $S \leftarrow S_{\text{new}}$ ▷ Add new samples to the subset.
return S ▷ Return the sampled subset.

Algorithm 2 *Lower Diversity Sampling*

Input: Data representation $D \in \mathbb{R}^{|\mathcal{D}| \times d}$, seed set size N_{seed}, batch size N_{batch}, target subset size N_{target}, similarity threshold τ
Output: Indices of selected subset $S \subseteq \{1, \cdots, |\mathcal{D}|\}$

$S \leftarrow \text{random-sample}(\{1, \cdots, |\mathcal{D}|\}$ ▷ Initialize the subset with seed data points.
while $|S| < N_{\text{target}}$ **do**
  $S_{\text{new}} \leftarrow \{i \in \{1, \cdots |\mathcal{D}|\} \setminus S \mid \max_{j \in S} \text{cos-sim}(D_i, D_j) > \tau\}$
  ▷ Find samples that are similar to the current subset members.
  $S_{\text{new}} \leftarrow \text{random-sample}(S_{\text{new}}, N_{\text{batch}})$
  $S \leftarrow S \cup S_{\text{new}}$ ▷ Add new samples to the subset.
return S ▷ Return the sampled subset.

a local cluster without resorting to external APIs—since data generation is mostly parallelizable, we use 16 H100 nodes at maximum to expedite the generation process. We used 4 H100 nodes for fine-tuning models on our data. For decontamination, we first run brute-force 10-gram filtering, then perform LLM-based paraphrase detection following Yang et al. [62]. Specifically, we first match each generated sample with the closest sample in the evaluation benchmarks using an off-the-shelf embedding model, then run Qwen2.5-72B-Instruct to determine whether the generated sample is equivalent to the test sample. The paraphrase detection prompt can be found in §A.3.

Evaluating *PrismMath* and *PrismNLI* As shown in Table 4, we use HANS [37], WNLI, ANLI, Diagnostics [56], BigBench NLI [53] and ConTRoL [29] for NLI evaluation. For math reasoning, we use AIME24/25, AMC23, MATH500 [26], MATH^2 [51], Olympiad Bench [12] and GSM8k-Platinum [55]. Note here that we use a more challenging set of math benchmarks than in §2.2, since we aim to achieve state-of-the-art results on challenging benchmarks through long CoT reasoning, unlike in §2.2 where we primarily investigate the relative performance gap between models trained with distinct datasets. During evaluation, we generate solutions with temp $= 0.6$ andtop-p $= 0.95$, and report `pass@1` for each benchmark. For AIME and AMC, we average `pass@1` over 16 independent runs to compensate for the small benchmark sizes. We use HuggingFace lighteval and math-verify [7] to evaluate answer correctness.

Synthetic Data Scaling We analyze the impact of synthetic data scaling as shown in §3.2.2 and Fig. 1 (Right). We first generate a data pool of 100K samples using vanilla few-shot generation and persona-guided generation, respectively. For persona-guided generation, we random-sample personas from PersonaHub [9] and condition problem generation on each persona, along with 5 random-sampled few-shot examples from data pool (§A.3). After generating the data pool, we sample 5 distinct subsets (at each scale) from each data pool, train Qwen2.5-7B-Math on each dataset, and report the average test accuracy on the same set of benchmarks as above.

A.3 Prompts

Prompt for Math Problem Generation

Given a set of example math problems, create a similar or harder problem inspired by the example problems.

The new problem should be formatted as:
———
[Problem]
your new problem - come up with a non-multiple choice problem, even if the provided examples are multiple choices.
———

Examples:
———
[Problem]
{example_problem_1}
———

`...[few shot examples omitted]`

Prompt for Math Solution Generation

Solve the following problem. Make sure to put your final answer in \boxed{}.

{input_problem}

Prompt for Math Problem Generation with Persona

Example 1:
{example_problem_1}

`...[few shot examples omitted]`

Create a challenging math problem similar to the examples above with the following persona:
{input_persona}

Prompt for NLI Sample Generation

Given examples of Natural Language Inference task, create a novel and more challenging problem with the similar reasoning as the given examples.

Each of the novel example should be formatted as:
—
Premise: premise text
Hypothesis: hypothesis text
Label: label
—

Examples:
—
Premise: {example_premise_1}
Hypothesis: {example_hypothesis_1}
Label: {example_label_1}
—

```
...[few shot examples omitted]
```

Prompt for NLI Sample Verification

Given a pair of premise and hypothesis, determine if the hypothesis is entailed by, or contradicted by, or neutral to the premise.
Your answer should be either ‘Entailment‘, ‘Contradiction‘ or ‘Neutral‘.

Premise: {input_premise}
Hypothesis: {input_hypothesis}
Label (Entailment or Contradiction or Neutral):

Prompt for Contamination Detection

Given two problems, help me determine if the two problems are equivalent.
- Disregard the names and minor changes in word order that appear within.
- If they are equivalent, please answer ’True’, otherwise answer ’False’. Do not respond with anything else.
- If their question prompts are very similar and, without considering the solution process, they produce the same answer, we consider them to be the same question.

Problem 1: {input_problem_1}
Problem 2: {input_problem_2}

B Additional Results

B.1 Results on Baseline Measures

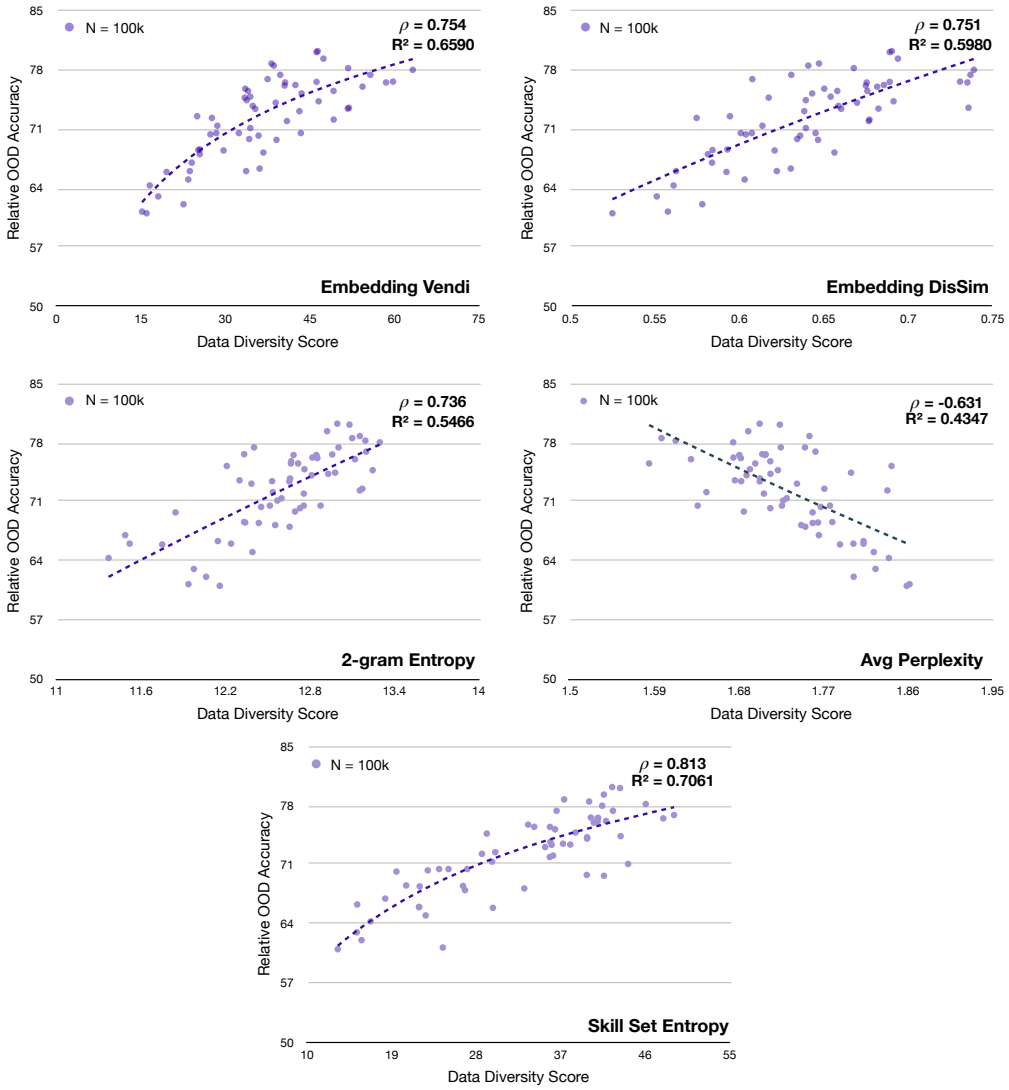

Figure 4: Relationship between baseline diversity measures and model OOD performance, measured in math reasoning tasks. Relative OOD accuracy is averaged across 7 benchmarks, following the same process as in §2.2. Overall, widely-used diversity measures fall behind G-Vendi in their correlation with model performance.

B.2 Extended Results on Prismatic Synthesis

In Table 5 and 6, we present full results on *PrismMath* and *PrismNLI* including the standard error for AIME24 / AIME25 / AMC23 and dataset size comparison.

Table 5: Full evaluation results on math reasoning benchmarks, including dataset size comparison. All scores are `pass@1`, and for AIME and AMC, we report average `pass@1` across 16 runs, and report their standard error.

Model	Data Generator	Data Size	AIME24	AIME25	AMC23	MATH500	MATH^2	Olympiad Bench	GSM8K Platinum	Avg.
Qwen2.5-Math-7B-It	-	-	14.17 (0.83)	9.91 (1.06)	72.50 (1.68)	83.80	57.62	44.29	96.11	54.06
OpenR1-Qwen-7B	R1-671B	114K	47.91 (1.88)	30.41 (1.72)	87.19 (1.92)	90.60	78.10	67.06	**96.69**	71.14
OpenThinker-7B		1.14M	27.50 (1.22)	22.50 (1.97)	74.06 (1.15)	84.20	67.62	45.93	93.05	59.27
OpenThinker2-7B		94K	50.00 (1.67)	35.00 (2.44)	88.44 (1.83)	91.40	78.10	**69.63**	93.96	72.36
R1-Distill-7B		Unknown	54.66 (1.96)	33.33 (1.52)	92.50 (1.24)	**92.60**	78.57	68.00	89.91	72.80
PrismMath-7B	R1-32B	1.0M	**57.08** (1.93)	**38.33** (1.41)	**93.75** (1.06)	92.40	**80.95**	68.30	95.95	**75.25**

Table 6: Full evaluation results on NLI benchmarks, including dataset size comparison.

Dataset	Data Generator	Data Size	HANS	WNLI	ANLI R1	ANLI R2	ANLI R3	Diagnostics	BigBench	Control	Avg.
MNLI	ChatGPT, Humans	393K	78.47	63.03	60.10	45.30	42.58	81.88	78.22	42.16	61.47
WANLI		103K	89.25	75.21	61.30	46.50	44.90	83.68	80.81	43.93	65.70
MNLI+FEVER		601K	74.62	66.29	60.20	47.00	41.75	80.89	76.14	48.51	61.93
{WA+M+S}NLI		943K	80.18	69.41	65.00	50.60	45.25	83.77	84.72	50.49	66.18
PrismNLI	Qwen2.5-72B	515K	**92.44**	**78.47**	**73.70**	**61.90**	**57.00**	**86.13**	**86.32**	**58.25**	**74.28**

C Properties of G-Vendi

As the aggregation process of G-Vendi follows that of Vendi Score [8], it inherits several desirable properties of Vendi score as a diversity metric. More specifically,

- G-Vendi is positive unbounded, and can be interpreted as the effective number of unique samples in a given dataset. That is, if the normalized covariance matrix K of dataset $\mathcal{D}$ satisfies $K_{ij} = 0$ for all $i \neq j \in [1, |\mathcal{D}|]$, then G-Vendi$(\mathcal{D}) = |\mathcal{D}|$. If $K_{ij} = 1$ for all $i \neq j \in [1, |\mathcal{D}|]$, G-Vendi$(\mathcal{D}) = 1$.
- G-Vendi is permutation invariant. That is, permuting the order of data points in $\mathcal{D}$ does not change the value of G-Vendi$(\mathcal{D})$.

In terms of computational complexity, computing eigenvalues would cost $O(n^3)$ for an $n \times n$ matrix [47]. This is prohibitively inefficient to directly compute for our normalized covariance matrix K, since $K = 1/|\mathcal{D}| \cdot GG^T \in \mathbb{R}^{|\mathcal{D}| \times |\mathcal{D}|}$ for the data representation matrix $G \in \mathbb{R}^{|\mathcal{D}| \times d}$. However, we can leverage the fact that eigenvalues of $K = 1/|\mathcal{D}| \cdot GG^T$ equals that of $1/|\mathcal{D}| \cdot G^TG \in \mathbb{R}^{|d| \times |d|}$. Therefore, the time complexity for G-Vendi involving both the computation of G^TG and the eigenvalues of G^TG is $O(d^3 + d^2|\mathcal{D}|) = O(d^2|\mathcal{D}|)$. This is much more efficient than the average pairwise similarity that requires the actual similarity matrix GG^T, which takes $O(d|\mathcal{D}|^2)$.

D Limitations and Future Works

One of our main goals in this work is to investigate the impact of data diversity on the trained model's behavior. While our analyses reveal a strong correlation between G-Vendi and the generalization of the trained model to diverse benchmarks, G-Vendi is not a panacea—the strong correlation can only be discovered via careful control of data quality. Thus, G-Vendi cannot be used as a sole metric for an apple-to-apple comparison between two arbitrary datasets (that originated from completely different data generating processes). Analyzing the interplay between the details of data-generating process (*e.g.,* verification protocol or quality filtering) and data diversity, and their impact on the model performance remains an interesting future direction.

In addition, our measure of model generalization relies on aggregating performance across multiple benchmarks. While aggregation over diverse benchmarks is indeed a standard setup for evaluating LLM capabilities, it may obfuscate the impact of data diversity on a new domain, where not many off-the-shelf benchmarks exist. An interesting future work motivated by this gap, is to compute *target-conditional data diversity*—which not only considers the intrinsic diversity of the training dataset itself, but also the relevance to the target distribution we want to excel on. Incorporating with prior literature on targeted instruction tuning [61], Prismatic Synthesis could evolve to generate diverse synthetic data that are targeted specifically to improve model capabilities in a given benchmark.

E Broader Impacts

Our work contributes to a more principled understanding of data diversity in LLM reasoning, with potential positive implications for building models that generalize better by training on diverse inputs. By proposing scalable methods for measuring and generating cognitively diverse data, our approach could inform better dataset construction practices and reduce overreliance on large, opaque corpora. However, improvements in synthetic data generation may also lower the barrier to creating high-performing models for malicious use, such as for generating more convincing misinformation or deceptive content. Care should be taken to ensure these methods are applied in alignment with ethical guidelines and responsible model deployment.

F Qualitative Examples

F.1 Comparing Closet Sample in Gradient Space and Embedding Space

In Example 1 to 3 below, we qualitatively compare the most similar data point to a given math reasoning sample, measured in the gradient space and the embedding space. We use the same setup as in §2.2, and select GSM8k problems that are easier to understand.

Overall, we find that the closest sample in embedding space often shares the central theme or keywords with the original sample (*e.g.,* a figure collecting Pokemon cards in Example 1). This contrasts with the sample closest in the gradient space, which often does not exhibit any semantic relevance to the original problem, but often shares a very similar solution strategy with the seed sample. These examples qualitatively indicate that similarity in gradient space can better capture the task-specific cognitive process for problem solving.

Example 1

Original Sample

Problem: Jenny has 6 Pokemon cards. Orlando has 2 more cards than Jenny. Richard has three times as many cards as Orlando. How many Pokemon cards do they have in all?

Solution: To solve the problem step by step, we start with the information given:
1. Jenny has 6 cards. This is given directly in the problem statement. `...[omitted]`

Therefore, the total number of Pokemon cards they have all together is $\boxed{38}$ cards.

Closest Sample in Gradient Space

Problem: Levi has 5 lemons. Jayden has 6 more lemons than Levi. Jayden has one-third as many lemons as Eli has while Eli has one-half as many lemons as Ian has. How many lemons do they have in all?

Solution: To solve the problem step by step, we start with the given information and work our way through the relationships between the number of lemons each person has. `...[omitted]`

Therefore, the total number of lemons they all have is $\boxed{115}$ lemons.

Closet Sample in Embedding Space

Problem: Nicole collected 400 Pokemon cards. Cindy collected twice as many, and Rex collected half of Nicole and Cindy's combined total. If Rex divided his card equally among himself and his three younger siblings, how many cards does Rex have left?

Solution: To solve this problem, we follow the steps based on the information provided:
1. Calculate the number of cards Cindy has:`...[omitted]`

So, the number of cards Rex has left is $\boxed{150}$.

Example 2

Original Sample

Problem: The number of math problems that Marvin practiced today is three times as many as the number of problems he solved yesterday. His friend, Arvin, has practiced twice as many math problems on each day. How many math problems have they practiced altogether if Marvin solved 40 math problems yesterday?

Solution: Given that Marvin solved 40 math problems yesterday, we can calculate the number of problems he solved today and the total number of problems solved by both Marvin and Arvin as follows: `...[omitted]`

Therefore, together, Marvin and Arvin have solved $\boxed{480}$ math problems.

Closest Sample in Gradient Space

Problem: Olaf has an aquarium. He has fish in 3 different colors: orange, green, and blue. Blue fish make up half of all the fish in the aquarium. There are 15 fewer orange fish than blue fish. How many green fish are there when the total number of fish in the aquarium is 80?

Solution: Given the total number of fish in the aquarium is 80, and blue fish make up half of all the fish, we can calculate the number of blue fish as follows: `...[omitted]`

Therefore, the number of green fish in the aquarium is $\boxed{15}$.

Closet Sample in Embedding Space

Problem: Angela wants to check her math homework answers with her friends, but some of them aren't done yet. Out of 20 problems, Martha has finished 2, JJenna has finished four times the number Martha did minus 2, and Mark has finished half the number Jenna did. If none of the friends worked on any of the same problems, how many problems have no one but Angela finished?

Solution: To solve the problem, let's break down the solution step by step:

1. **Calculate the number of problems Jenna answered:** `...[omitted]`

Therefore, Angela was able to finish $\boxed{9}$ problems on her own.

Example 3

Original Sample

Problem: The $\$4.55$ in Carol's piggy bank consists of quarters and nickels. There are seven more nickels than quarters. How many nickels does Carol have in her bank?

Solution: To solve the problem, we start by defining variables for the number of nickels and quarters Carol has. Let n represent the number of nickels and q represent the number of quarters. Given that `...[omitted]`

Therefore, Carol has $\boxed{21}$ nickels in her piggy bank.

Closest Sample in Gradient Space

Problem: Billy Goats invested some money in stocks and bonds. The total amount he invested was $\$165{,}000$. If he invested 4.5 times as much in stocks as he did in bonds, what was his total investment in stocks?

Solution: To solve the problem, we start by letting the amount Billy Goats invested in bonds be s. Since he invested 4.5 times as much in stocks, `...[omitted]`

Therefore, the total investment in stocks is $\boxed{\$135{,}000}$.

Closet Sample in Embedding Space

Problem: Maria has 4 dimes, 4 quarters, and 7 nickels in her piggy bank. Her mom gives her 5 quarters. How much money, in dollars, does Maria have now?

Solution: To calculate the total amount of money Maria has in her piggy bank after her mom gives her additional quarters, we proceed as follows:

1. **Calculate the total number of quarters Maria has now:**`...[omitted]`

Therefore, the total amount of money Maria has in her piggy bank is $\boxed{\$3.00}$.

F.2 Example Clusters in Gradient Space

We additionally analyze clusters of data points in gradient space, for both NLI and math reasoning tasks. These examples are specifically selected for interpretability, as many gradient clusters are difficult to analyze. However, we find that when interpretable, the cluster members often exhibit a common, nuanced reasoning strategies rather than superficial topical or semantic overlaps.

NLI Example 1: Membership between concepts

Premise: James Blake `...[omitted]` bus driver, ordered a 42-year old woman to move further back on the bus and to give her seat to a white person. When she did not comply, he called the police and 4 of them came on board. They arrested Rosa Parks who refused to move further back, not because the bus was crowded but because **the front 4 rows** were reserved for white people.

Hypothesis: Rosa Parks was seated in **one of the 4 front rows** of the bus.

Label: Entailment

Premise: The Australian Open, French Open, Wimbledon, and US Open are the 4 most prestigious tennis tournaments in the world. Winning all four in the same calendar year is considered a Grand Slam.

Hypothesis: If a player wins the **French Open and the other three Grand Slam tournaments**, they will have won *all four* Grand Slam tournaments.

Label: Entailment

Premise: American Pharoah won the Triple Crown in 2015, becoming the first horse to do so since Affirmed in 1978. *The Triple Crown consists of the Kentucky Derby, the Preakness Stakes, and the Belmont Stakes.* To win the Triple Crown, a horse must win all 3 races in a single season.

Hypothesis: American Pharoah won the **Kentucky Derby** in 2015.

Label: Entailment

NLI Example 2: Traits of a person

Premise: Henk Fraser is a Dutch former football player and coach. He is the sone of famous Dutch football player, Bert Fraser. Henk coached Sparta Rotterdam, ADO Den Haag and assisted on Ajaex A1. Now he is an assistant coach at Vittesse.

Hypothesis: Henk Fraser **is in** sports industry.

Label: Entailment

Premise: The television series "Smallville" revolves around Clark Kent and his friends as they try to navigate their lives and stop various villains in the fictional town of Smallville, Kansas.

Hypothesis: Clark Kent **lives in** the state of Kansas.

Label: Entailment

Premise: Wandi Rum is a village in southerm Jordan, southeast of the city Aqaba. It is set in an extraordinary landscape of mountains, valleys, dunes and Bedouin camps.

Hypothesis: Wadi Rum **has** Bedouin residents.

Label: Entailment

NLI Example 3: Key information at the beginning of premise

Premise: The American Broadcasting Company **(ABC) is a major American commercial broadcast network**, and the fifth-oldest major network in the US. It broadcasts a wide range of programming, including drama television series, which has become increasingly popular in `[...omitted]` over the past two decades.

Hypothesis: ABC **broadcasts American** drama television series.

Label: Entailment

Premise: **The store's advertisement for a Halloween sale** features a giant spider hanging from the celing and mannequins dressed in elaborate costumes and masks.

Hypothesis: The store sells **Halloween decorations and accessories**.

Label: Entailment

Premise: **Liam believes that several people in a gaming arcade are sitting at computes** while wearing headphones and intensely focused on the screens.

Hypothesis: **Liam believes that people** play computer games in the arcade.

Label: Entailment

NLI Example 4: Aggregating "different" concepts

Premise: The Asian Games were held in Bangkok in **1973**, and in Hiroshima in **1978**.

Hypothesis: The Asian Games were held in **different years**.

Label: Entailment

Premise: The 2024 European Women's Handball Championship will be the 16th edition of European Woman's Handball Championship, the top level woman's handball event organized by the European Handball Federation. The tournament will take place in **Austria, Hungary and Switzerland**.

Hypothesis: The European Women's Handball Championship will be held in **three different countries**.

Label: Entailment

Premise: For the past century the city has been **dominated by the German and French**, and in the past few years by the British.

Hypothesis: The city has been dominated by **different people**.

Label: Entailment

Math Example 1: State tracking by following the information step by step

Problem: Simon, Gerry, and Micky want to have a race with handmade miniature rafts. Simon's raft needs 36 sticks, Gerry's raft needs two-thirds of the number of sticks that Simon needs, and Micky's raft needs 9 sticks more than Simon and Gerry's rafts combined. How many sticks do the three boys need to collect in total?

Solution: To solve this problem, we follow the information given step by step:

1. **Simon's Raft**: Simon needs 36 sticks. `...[omitted]`

Therefore, the total number of sticks the three boys need to collect is $\boxed{129}$ sticks.

Problem: Bella bought stamps at the post office. Some of the stamps had a snowflake design, some had a truck design, and some had a rose design. Bella bought 11 snowflake stamps. She bought 9 more truck stamps than snowflake stamps, and 13 fewer rose stamps than truck stamps. How many stamps did Bella buy in all?

Solution: To solve the problem, we follow the information given step by step:

1. Bella bought 11 snowflake stamps. `...[omitted]`

Thus, the total number of stamps Bella bought is $\boxed{38}$.

Problem: Marcy is the lunch monitor in an elementary school cafeteria. She gives 5 time-outs for running, 1 less than five times that number of time-outs for throwing food, and 1/3 the number of food-throwing time-outs for swearing. If each time-out is 5 minutes, how much time do the students spend in time-out total?

Solution: To solve the problem step by step step, we follow the given instructions closely:

1. **Calculate the number of time-outs for running**: `...[omitted]`

Thus, the students spend a total of $\boxed{185}$ minutes in time-out.

Problem: Rebecca makes her own earrings out of buttons, magnets, and gemstones. For every earring, she uses two magnets, half as many buttons as magnets, and three times as many gemstones as buttons. If Rebecca wants to make 4 sets of earrings, how many gemstones will she need?

Solution: To solve the problem, we follow the information given and calculate step by step:
1. **Calculate the total number of earrings**: `...[omitted]`

Therefore, to make 4 sets of earrings, Rebecca will need $\boxed{24}$ gemstones.

Math Example 2: Computing the remainder

Problem: At the height of cranberry season, there are 60000 cranberries in a bog. 40% are harvested by humans and another 20000 are eaten by elk. How many cranberries are left?

Solution: To solve the problem, we follow these steps:

1. Calculate the total number of cranberries harvested by humans. `...[omitted]`

Therefore, the number of cranberries left in the bog after being harvested by humans and eaten by elk is $\boxed{16000}$ cranberries.

Problem: Barney's grocery store sold out all of its items at the beginning of the pandemic, so they ordered extra items to restock the shelves. However, they ended up ordering far too much and have to keep the leftover items in the storeroom. If they ordered 4458 items, sold another 1561 items that day, and have 575 items in the storeroom, how many items do they have left in the whole store?

Solution: To solve the problem, we follow these steps:

1. Calculate the total number of items Barney's grocery store had after ordering extra items: `...[omitted]`

Therefore, the total number of items left in the whole store is $\boxed{3472}$.

Problem: Out of 804 senior high school students, 75% passed their exams and so got their degree. The rest failed. How many didn't pass their exams?

Solution: To solve the problem, we follow these steps:

1. First, we calculate the number of students who passed their exams and got their degree. `...[omitted]`

Therefore, the number of students who didn't pass their exams is $\boxed{201}$ students.

Problem: Miranda wants to buy a pair of heels she saw online. She saved money for 3 months. Her sister heard that she was sad and gave her $50 for her to buy the heels. If she paid $260 in total for the heels, how much money did she save per month?

Solution: To solve this problem, we follow these steps:

1. First, we need to determine how much money Miranda actually saved before her sister gave her the additional $50. `...[omitted]`

Therefore, Miranda saved $\boxed{\$70}$ per month.

Math Example 3: How old is X now?

Problem: In 10 years, Cornelia will be three times as old as Kilee. If Kilee is currently 20 years old, how old is Cornelia now?

Solution: To solve this problem, we follow these steps:

1. Calculate Kilee's age in 10 years: `...[omitted]`

Therefore, Cornelia is currently $\boxed{80}$ years old.

Problem: In 5 years, Raven will be 4 times as old as Phoebe. If Phoebe is currently 10 years old, how old is Raven?

Solution: To solve this problem, we follow these steps:

1. Calculate Phoebe's age in 5 years:`...[omitted]`

Therefore, Raven is currently $\boxed{55}$ years old.

Problem: After five years, Ron will be four times as old as Maurice. If Ron's age now is 43, how old is Maurice now?

Solution: 1. Calculate Ron's age in five years:
- Ron's current age is 43 years.`...[omitted]`

Therefore, Maurice is currently $\boxed{7}$ years old.

Problem: In four years, Suzy will be twice Mary's age then. If Suzy is 20 now, how old is Mary?

Solution: To solve this problem, we follow these steps:
1. Calculate Suzy's age in four years: `...[omitted]`

Thus, Mary is $\boxed{8}$ years old now.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We propose G-Vendi and demonstrate that it strongly predicts model generalization in reasoning tasks as empirically measured by benchmark performance, and use this knowledge to generate diverse synthetic data which we analyze and evaluate.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See §D.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: Our contribution focuses on empirical evaluation with its motivation theoretically grounded on existing works, which we cite for completeness.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We include detailed experimental design choices and setups for reproducibility in both the main section (§2.2, §3.1) and in the appendix (§A.1, §A.2).

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The code is provided as part of the supplementary material, along with anonymized link for the data samples from *PrismMath* and *PrismNLI*.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We include detailed experimental design choices and setups for reproducibility in both the main section (§2.2, §3.1) and in the appendix (§A.1, §A.2).

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We provide the standard errors for iterated experiments in our extended results in §B.2.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We elaborate on GPU usages for our experiments in §A.1 and §A.2

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: We abide by the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: See §E for the broader impacts of our work.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our domain of analyses focuses on tasks in academic settings, such as inferring the logical relationship between two snippets of text or solving a math problem. We did not scrape any internet data, and our models are narrowly trained on these specific tasks. Thus we did identify a need for specialized safeguards beyond standard ethical considerations.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite all existing assets used by either their URL or the corresponding paper.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We include documentations of our code and data released in the README accompanying the supplementary submission.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The work does not involve crowdsourcing or research conducted with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The work does not involve crowdsourcing or research conducted with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: We did not use an LLM as a development tool in all phases of our research other than for formatting purposes.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.