FIDELITY-DIVERSITY METRICS FOR TEXT DATASETS

Anonymous authors
Paper under double-blind review

ABSTRACT

As language modeling technology matures, there is an increasing research focus on the composition and curation of datasets used to train these models. For instance, practitioners commonly seek to augment high-quality datasets with additional text to enhance the performance of models trained on that data. However, informed decisions about data augmentation require more nuanced assessments about data quality. We build on work measuring the precision and recall of generative models to develop a pair of metrics that quantify (1) fidelity, capturing how closely candidate text resembles reference data, and (2) diversity, capturing how well it covers the modes of the reference dataset. Our metrics are based on optimal transport divergence functionals between discrete text summaries. In experiments on M2D2 text datasets, we show that these metrics are able to disentangle a lack of fidelity from a lack of diversity in deficient candidate text. In further experiments, our metrics detect diversity deficits in synthetic GSM8K-style math datasets, which correlate with degradations in downstream accuracy of language models finetuned on this synthetic data.

1 INTRODUCTION

In tandem with the development of large language models (LMs), the research community has developed quantitative evaluation metrics to assess the quality of text generated by these models. In this paper we take a broader perspective on the evaluation of text, reimagining this style of quantitative evaluation as a tool to improve our understanding of datasets used for training and finetuning LMs. We are particularly interested in synthetic training data, which bridges model evaluation and dataset analysis: in this setting, generated text is both the training data input and sampled output of LMs (Gunasekar et al., 2023; Wang et al., 2023).

The predominant metrics for evaluating LM-generated text compare model outputs to reference text using a scalar numerical score. These metrics can be computed either on the instance level, comparing one model output to one reference (Zhang et al., 2020), or between i.i.d. samples from distributions of generated text and reference text (Pillutla et al., 2021). We study the latter scenario, comparing an *evaluation dataset* collected from one distribution of text to a *reference dataset*. We seek to develop more nuanced metrics for insight into data quality, disaggregating our measure of data quality along two axes: *fidelity*, how similar evaluation text is to the reference, and *diversity*, how broadly evaluation text covers the modes of the reference text distribution.

We quantify fidelity and diversity through two separate notions of transport between discrete measures. Our metrics operate on pretrained text embeddings; each dataset is summarized by a discrete measure consisting of point masses on a finite set of embeddings that represent modes of the data distribution. Fidelity is defined through the cost of a greedy nearest neighbor assignment between evaluation modes and reference modes. We arrive at this notion through an optimal transport argument, providing an alternative motivation for nearest-neighbor metrics that have previously been proposed as a measure of data fidelity (Sajjadi et al., 2018). Diversity is defined as the *gap* between this greedy assignment and the global cost of optimal transport between the two measures. Taken together, these metrics decompose the Wasserstein transport cost between dataset representations into two components: costs arising from lack of fidelity (geometric mismatch in support) and costs arising from lack of diversity (discrepancy of mass distribution).

We study the behavior of these metrics in two settings. First, we seek to validate their efficacy in a setting with ground-truth labels that reflect a colloquial understanding of fidelity and diversity. Using text from the

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

M2D2 dataset (Reid et al., 2022) annotated with topic labels, we design an experiment that controls the topics contained in the evaluation dataset and measures the sensitivity of our metrics as we vary the overlap and coverage of topics relative to the reference dataset. Second, we demonstrate that our metrics are sensitive to more subtle variations in fidelity and diversity, in a setting where no clear topic structure exists. We design a protocol for generating synthetic variants of the GSM8K dataset (Cobbe et al., 2021) that allows us to control for the diversity of problems contained in each generated variant. We show that our diversity metric tracks diversity in these datasets and in turn, is correlated with downstream GSM8K accuracy when we finetune LMs on these synthetic datasets.

Our main contributions are summarized as follows. First, we propose principled fidelity-diversity metrics based on a decomposition of the Wasserstein transport distance between distributions. Second, we devise a controlled experiment to validate that these metrics behave intuitively based on data with ground-truth topic labels. Third, we devise a protocol for controlling diversity when constructing synthetic datasets. Using this protocol, we construct synthetic variants of the GSM8K dataset with varying degrees of diversity; we show that our metric is sensitive to the diversity in these datasets, and that diversity in these synthetic training datasets directly correlates with downstream accuracy of models trained on this data.

2 Related work

Evaluating generative models. There is a large body of prior work that has sought to evaluate generative models through distributional divergence measures. Widely used examples include the Inception Score (Salimans et al., 2016) or the Fréchet Inception Distance (FID; (Heusel et al., 2017)), which approximates both distributions in a learned feature space by multivariate Gaussian distributions, comparing their means and covariances using the Wasserstein distance. Another popular metric for evaluating generative models is the MAUVE score (Pillutla et al., 2023; 2021; Yang et al., 2026), which reports a summary scalar of a divergence frontier. However, inspecting the divergence frontier alone does not readily lend itself to disaggregated analyses of fidelity and diversity. Indeed, both FID and MAUVE are unchanged when the evaluation and reference sets are swapped, so they cannot distinguish the two effects our metrics separate. Finally, we note that it is also possible to measure diversity without a reference corpus, but to instead quantify intrinsic properties of generated text via metrics like the Vendi score put forth in (Friedman & Dieng, 2023). We refer to (Celikyilmaz et al., 2020; Ito et al., 2025) for surveys of such methods, and to Appendix C.1 for a detailed comparison of our fidelity-diversity metrics with FID, MAUVE, and Vendi.

A separate line of work has sought to disentangle different aspects of a generative model's quality through the lens of *precision* and *recall* metrics. This perspective was introduced through the work of (Sajjadi et al., 2018), among others, and was subsequently developed into a variety of practical methodologies for evaluating generative models (Kynkäänniemi et al., 2019; Naeem et al., 2020; Alaa et al., 2022; Jiralerspong et al., 2023; Djolonga et al., 2020). While these methods have been successful for applications involving image generation, they have seen more limited use in the text domain. Closest to our work is (Le Bronnec et al., 2024), which develops precision-recall metrics specifically for evaluating language models. Their framework adapts the metrics of (Kynkäänniemi et al., 2019) to the text domain, and empirically demonstrates their efficacy in large-scale open-ended text generation tasks. We provide further discussion of methodological and empirical differences between our fidelity-diversity metrics and (Le Bronnec et al., 2024)'s precision-recall metrics in Appendix C.2.

Optimal transport for text comparison. The Wasserstein distance between mixing measures has previously been adopted for comparing text corpora in (Bing et al., 2025; 2022), and has more broadly been advocated as a loss function in latent variable models at least since the works (Nguyen, 2013; 2015). In particular, (Bing et al., 2025, Theorem 1) shows that the Wasserstein distance is the most discriminative metric between mixing measures that is jointly convex and consistent with the ground metric over mixture components. We view this result as strong motivation for adopting the Wasserstein distance in our work. When one relaxes the metric requirement, stronger divergences have also recently been considered (Manole & Ho, 2022; Nguyen et al., 2024). From a different perspective, optimal transport has been used to compare text documents without quantization (Kusner et al., 2015; Yurochkin et al., 2019; Alvarez-Melis & Fusi, 2020). In either case, we are

not aware of past attempts at using it to disentangle notions of fidelity and diversity. Finally, we mention a separate line of work, in which optimal transport has been used as a means of aggregating quality metrics for language models using tests for stochastic dominance (Rioux et al., 2024).

3 METHODOLOGY

3.1 PROBLEM SETTING

Let $\mathcal{T}$ be a finite vocabulary of tokens, and let $\mathcal{T}^*$ denote the set of all sequences of tokens of arbitrary finite length. Assume that we are given two text datasets of the form $\overline{\mathcal{U}} = \{\overline{U}_1, \ldots, \overline{U}_n\} \subseteq \mathcal{T}^*, \overline{\mathcal{V}} = \{\overline{V}_1, \ldots, \overline{V}_m\} \subseteq \mathcal{T}^*$. We interpret $\overline{\mathcal{U}}$ as an *evaluation dataset*—a collection of text documents whose fidelity and diversity we wish to evaluate against the *reference dataset* $\overline{\mathcal{V}}$. Our aim is to define metrics for quantifying the quality of the evaluation corpus that explicitly distinguish between its fidelity and diversity.

Irrespective of the downstream application, our metric construction first embeds the two datasets into a common space $\mathbb{R}^L$, where the dimension $L \geq 1$ does not depend on the lengths of the various documents. Concretely, assume that we have access to an embedding model $\mathcal{E} : \mathcal{T}^* \to \mathbb{R}^L$, such as a Sentence Transformer (Reimers & Gurevych, 2019), and define the embedded documents $U_i = \mathcal{E}(\overline{U}_i)$ and $V_j = \mathcal{E}(\overline{V}_j)$, to obtain the embedded datasets $\mathcal{U} = \{U_1, \ldots, U_n\} \subseteq \mathbb{R}^L, \mathcal{V} = \{V_1, \ldots, V_m\} \subseteq \mathbb{R}^L$.

In order to motivate the definition of our metrics, we will treat each corpus as a collection of i.i.d. draws from probability distributions over $\mathbb{R}^L$,

$$U_1, \ldots, U_n \sim P, \quad V_1, \ldots, V_m \sim Q. \tag{1}$$

Our fidelity and diversity metrics will be based on discrete representations of these probability distributions, described next, which will be estimated on the basis of the two datasets.

3.2 DISCRETE MEASURE REPRESENTATIONS

Quantization. We seek discrete summaries of the distributions P and Q in the form of *mixing measures*: discrete probability distributions which respectively take the following form:

$$G = \sum_{i=1}^{k} \alpha_i \delta_{\theta_i}, \quad H = \sum_{i=1}^{k} \beta_i \delta_{\eta_i},$$

where $\alpha = (\alpha_1, \ldots, \alpha_k)$ and $\beta = (\beta_1, \ldots, \beta_k)$ are elements of the $(k-1)$-dimensional simplex Δ_k, and $\theta_1, \eta_1, \ldots, \theta_k, \eta_k$ lie in a parameter space $\Theta \subseteq \mathbb{R}^L$ for some $k \geq 1$. Here, δ_θ denotes the Dirac mass (degenerate probability distribution) at θ. By potentially setting some of the mixing proportions α_i and β_i to zero, we can assume without loss of generality that both mixing measures have the same number of components k. We will interpret the atoms θ_i and η_i as k latent semantic topics that are captured by the evaluation and reference datasets, respectively. In turn, the mixing probabilities α_i and β_i denote the relative weight assigned to topic i under each corpus.

There are many ways of constructing such quantizations. Perhaps the most natural approach is to approximate P and Q with mixture models. We assume, for simplicity of exposition, that P and Q admit respective densities p and q. Concretely, given a parametric family of densities $F = \{f_\theta : \theta \in \Theta\}$ over $\mathbb{R}^L$, we say that P is a mixture model with mixing measure G if its density takes the form

$$p(x) = \int_\Theta f_\theta(x) dG(\theta) = \sum_{i=1}^{k} \alpha_i f(x; \theta_i), \quad x \in \mathbb{R}^L. \tag{2}$$

For instance, the class F could be the set of Gaussian densities $N(\mu, \Sigma)$, with respect to the location-scale parameter $\theta = (\mu, \Sigma)$, or a kernelized Gaussian mixture family with a learned kernel.

In practice, the datasets (1) are used to learn estimators $\widehat{G} = \sum_{i=1}^{k} \widehat{\alpha}_i \delta_{\widehat{\theta}_i}$, $\widehat{H} = \sum_{i=1}^{k} \widehat{\beta}_i \delta_{\widehat{\eta}_i}$ of the mixing measures G and H, giving rise to the data-dependent quantizations of a corpus. For instance, if quantization arises via a mixture approximation, one can use the Expectation-Maximization algorithm to approximate the maximum likelihood estimator under model (2), which we shall pursue in Section 4.1. More generally, many clustering procedures can be understood as producing discrete mixing measure summaries such as $\widehat{G}$ and $\widehat{H}$, and therefore fit naturally within this framework.

Empirical measures. There may be situations in which the data does not naturally contain interpretable semantic topics, in which case quantization becomes much less informative. In that case, we can use the empirical measures $\frac{1}{n}\sum_{i=1}^{n} \delta_{U_i}$ and $\frac{1}{m}\sum_{i=1}^{m} \delta_{V_i}$ as canonical data-dependent discrete representations of P and Q. We illustrate this in Section 4.2. We emphasize that the metrics we propose in the next section are independent of the way in which we arrive at discrete representations G and H of the distributions P and Q.

3.3 Optimal transport fidelity and diversity metrics

We now introduce fidelity and diversity metrics for comparing the evaluation distribution P to the reference distribution Q, via their mixing measures G and H. Our metrics are based on the *optimal transport* problem (Villani, 2003; Peyré & Cuturi, 2019; Chewi et al., 2025) between G and H, which we recall next.

Definition 1. *Given two mixing measures $G = \sum_{i=1}^{k} \alpha_i \delta_{\theta_i}$ and $H = \sum_{i=1}^{k} \beta_i \delta_{\eta_i}$, we denote by $\Pi(G, H)$ the set of doubly stochastic matrices which are couplings of G and H:*

$$\Pi(G, H) = \left\{ \gamma \in \mathbb{R}_{+}^{k \times k} : \sum_{i'=1}^{k} \gamma_{i'j} = \beta_j, \sum_{j'=1}^{k} \gamma_{ij'} = \alpha_i, \text{for all } i, j \in \{1, \dots, k\} \right\}.$$

Given a metric $d(\cdot, \cdot)$ on Θ, the 1-Wasserstein distance between G and H is defined by

$$W_1(G, H) = \inf_{\gamma \in \Pi(G,H)} \sum_{i,j=1}^{k} \gamma_{ij} d(\theta_i, \eta_j),$$

and any solution to the above optimization problem is called an optimal transport coupling.

The Wasserstein distance measures the most efficient way of redistributing the probability mass of G onto that of H, as measured by the metric d on the underlying space. In doing so, it induces a global notion of distance between the evaluation distribution P and the reference distribution Q. Our goal will now be to decompose the distance $W_1(G, H)$ into two complementary metrics which capture two distinct aspects of the evaluation distribution: the fidelity and the diversity of G relative to H.

Fidelity metric. To isolate how well the support of G aligns with that of H, we will define a third mixing measure H_G, defined as the best approximation of G using *only* atoms from H. That is, letting $\mathcal{P}(H)$ denote the set of mixing measures whose support is contained in that of H, we define

$$H_G = \operatorname*{argmin}_{\mu \in \mathcal{P}(H)} W_1(G, \mu).$$

Intuitively, H_G is obtained by reassigning all atoms of G onto the atoms of H, without requiring that all atoms of H receive mass. We then define the *fidelity* of G relative to H as follows:

$$\mathcal{F}(G, H) = \exp\left\{ -\frac{W_1(G, H_G)}{\kappa(H)} \right\}, \quad \text{with } \kappa(H) := \mathbb{E}_{\eta, \eta' \sim H}[d(\eta, \eta')] = \sum_{i,j=1}^{k} \beta_i \beta_j d(\eta_i, \eta_j).$$

A simple derivation (see App. E) shows that the Wasserstein distance between G and H_G is available in closed-form, and arises from a nearest-neighbor search: One has the equivalent expression

$$\mathcal{F}(G,H) = \exp\left\{-\frac{1}{\kappa(H)}\sum_{i=1}^{k}\alpha_i \cdot \min_{1\le j\le k} d(\theta_i,\eta_j)\right\}. \tag{3}$$

The fidelity thus measures how close the atoms of G are to *some* atom of H. It is large when each atom of G lies near an atom of H, indicating heuristically that each semantic topic captured by the evaluation corpus comes close to some topic in the reference corpus; see Figure 1(a) for an illustration. However, the fidelity can still be large if *some* of the atoms of H are far from those of G, which reflects a lack of diversity rather than fidelity; see Figure 1(b). In this case, all topics in the evaluation corpus G still come close to the reference corpus H, but this is achieved by being faithful to only one semantic topic in H.

We include the normalization $\kappa(H)$ so that the fidelity can be interpreted on an absolute scale. In particular, we expect a mixing measure G that has high fidelity to have its atoms located closer to those of H than the typical distance between the atoms of H themselves. One can thus loosely think of e^{-1} as a lower bound on typical values of $\mathcal{F}(G,H)$ for high-fidelity text.

Diversity metric. We further define the *diversity* of G relative to H as

$$\mathcal{D}(G,H) = \exp\left\{-\frac{1}{\kappa(H)}\Big(W_1(G,H) - W_1(G,H_G)\Big)\right\}.$$

This quantity captures the extent to which G spreads its mass around the support of H. To make this interpretation clear, we consider the special case where the mixing measures are uniform: $\alpha_i = \beta_i = 1/k$ for all $i = 1,\dots,k$. In this case, the optimal transport coupling is always achieved by a permutation: Denoting by $\mathcal{S}_k$ the permutation group on k elements, one has

$$\mathcal{D}(G,H) = \exp\left\{-\frac{\Delta}{\kappa(H)}\right\}, \quad \text{with } \Delta = \frac{1}{k}\min_{\sigma\in\mathcal{S}_k}\sum_{i=1}^{k}\Big\{d(\theta_i,\eta_{\sigma(i)}) - \min_{1\le j\le k} d(\theta_i,\eta_j)\Big\}.$$

The quantity Δ measures the difference between the cost of the greedy nearest neighbor matching from the atoms of G to those of H, and the optimal transport cost, which must respect the constraint of mapping mass from G onto all atoms of H. Intuitively, if G has high diversity, it should spread its mass evenly around the atoms of H, so these two costs should be similar, and Δ should be small.

Fidelity-diversity decomposition. Together, our definitions of fidelity and diversity induce a decomposition

$$\exp\left(-\frac{1}{\kappa(H)}W_1(G,H)\right) = \mathcal{F}(G,H)\cdot\mathcal{D}(G,H).$$

of the total Wasserstein distance between G and H, suitably scaled and exponentiated, into its contributions arising from lack of fidelity (geometric mismatch in support), and diversity (discrepancy in mass distribution). We refer to Figure 1 for an illustration of the behavior of our metrics on several examples.

4 Experiments

We conduct two sets of experiments. The first seeks to validate our method against ground-truth categorizations of text data from the M2D2 corpus (Reid et al., 2022). Our method produces measurements which correlate well with the amount of diversity implied by these categorizations. The second experiment seeks to demonstrate the utility of our method for assessing diversity in synthetic training data. To that end, we devise a controlled procedure to generate multiple synthetic GSM8K-style math datasets of similar fidelity and increasing diversity. We find that our measurements of fidelity and diversity align well with the synthesis procedure: datasets that

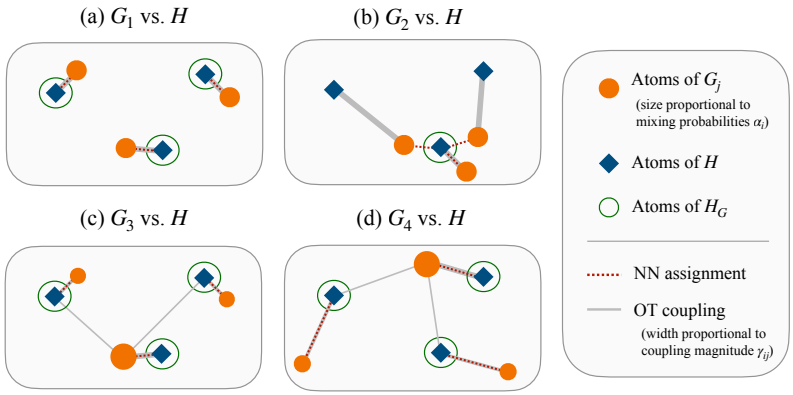


Figure 1: Illustration of four evaluation mixing measures $G_1, \ldots, G_4$ (with atoms in orange), and one reference mixing measure H (with atoms in blue), lying in dimension $L = 2$. The dotted lines denote the nearest blue point to each orange point, whereas the grey lines denote the magnitude of the optimal coupling between the measures G_i and H. In each figure, the fidelity $\mathcal{F}_i = \mathcal{F}(G_i, H)$ is determined by a weighted average of the length of the dotted lines, whereas the diversity $\mathcal{D}_i = \mathcal{D}(G_i, H)$ can be read as the difference between weighted averages of the grey and dotted lines. The top row illustrates settings where all mixing probabilities are equal, and the optimal coupling is a perfect matching, whereas the bottom row illustrates settings where the mixing probabilities are not uniform. Figure (a) illustrates high fidelity and perfect diversity. Figure (b) illustrates high fidelity and low diversity. Figure (c) illustrates high fidelity and fair diversity. Figure (d) illustrates low fidelity and fair diversity. In summary, we have $\mathcal{F}_1 = \mathcal{F}_2 = \mathcal{F}_3 \geq \mathcal{F}_4$, and $1 = \mathcal{D}_1 \geq \mathcal{D}_3 = \mathcal{D}_4 \geq \mathcal{D}_2$.

are constructed to be more diverse receive higher diversity scores, while all achieve similar fidelity scores. Finetuning a language model on each synthetic dataset further demonstrates that our diversity measurements correlate well with downstream model performance when evaluated on a held-out test set.

4.1 M2D2 EXPERIMENTS

We perform several controlled experiments using human-annotated text data to empirically validate whether our metrics successfully disaggregate fidelity and diversity. At a high level, we construct multiple datasets of similar fidelity but varying levels of diversity, and then verify that our metrics computed on these datasets agree with the expected trends. We also perform these experiments on different kinds of text data, demonstrating the general applicability of our method.

We begin by collecting data from the M2D2 S2ORC and M2D2 WIKIPEDIA test splits in the Paloma benchmark (Magnusson et al., 2024; Reid et al., 2022). M2D2 S2ORC contains Semantic Scholar papers and is organized by academic field, while M2D2 WIKIPEDIA contains text from Wikipedia articles and is categorized according to the Wikipedia ontology. We construct a total of ten different collections of data: each of the first five collections contains ten randomly sampled categories from M2D2 S2ORC, while each of the remaining five consists of ten randomly sampled categories from M2D2 WIKIPEDIA. All ten collections are then embedded using E5_{small} (Wang et al., 2024), with mean-pooling to produce sentence embeddings of the data. We also validate the robustness of our metrics on several other embeddings in App. D.

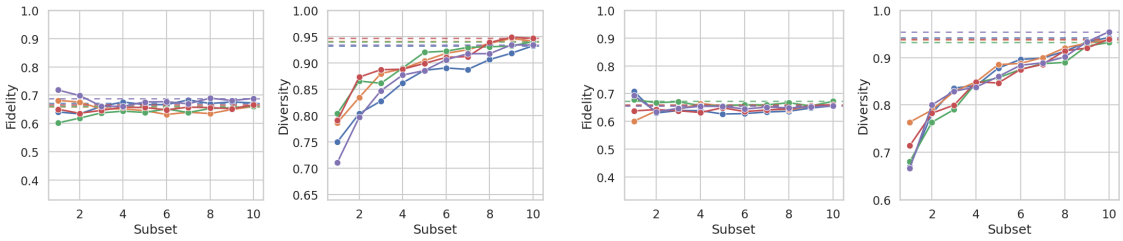
We now construct multiple datasets from each collection, controlling the diversity in each dataset by varying the number of categories included, and fixing fidelity across datasets by selecting an appropriate reference dataset for each collection. In particular, we would like the reference dataset to encompass all topics in that collection, so any other dataset composed of categories from that collection should have uniformly high fidelity with respect to the reference. A natural choice of reference is to pool together a subset of every category in the collection. We collect only the first half of each category for the reference and construct all other datasets from the remaining samples, to avoid duplicate samples across evaluation and reference datasets. As a numerical

baseline, we also compute our metrics between all remaining samples and the reference, since the remaining samples can be understood as being drawn from the same underlying distribution as the reference. These baseline values thus capture ideal fidelity and diversity for the given collection.

We construct ten datasets of varying diversity from those remaining samples. In particular, for every $i \in \{1, \ldots, 10\}$, we construct a dataset composed of the second half of samples in categories 1 through i, and refer to it as subset i. For instance, subset 1 contains the second half of samples from category 1, while subset 10 contains the second halves of all ten categories. We highlight that these subsets are nested, where subset $i+1$ contains all of subset i, and so on. By design, smaller subsets are strictly less diverse than larger ones, as they contain fewer topics. However, all subsets should be of comparable fidelity with respect to the reference.

To apply our method to the constructed subsets, we first quantize each dataset using Gaussian mixture models (GMMs). Since the reference and subsets span a wide range of sizes, we independently fit a GMM for each and select the number of mixture components using the Bayesian Information Criterion (BIC). The fidelity and diversity scores are then computed directly on the mixing measure underlying the selected mixture model. For further details about this quantization step, see App. A.4.

Examining Fig. 2 and Table 1, observe that the fidelity scores do not typically exhibit significant association with the number of categories in the data, as expected. We address the two exceptions (S2ORC Collection 2 and WIKIPEDIA Collection 3) in App. A.2. On the other hand, in all ten collections, the diversity score improves steadily and has a statistically significant OLS regression coefficient with respect to the number of categories. This is again consistent with our expectations and serves as evidence that **our metrics consistently disaggregate fidelity and diversity across different types of text data**. We also provide a more detailed discussion of these experiments with bootstrapped confidence intervals (App. A.4), as well as additional experiments verifying that our fidelity score responds to changes in fidelity (App. A.1).



(a) M2D2 S2ORC (b) M2D2 WIKIPEDIA

Figure 2: Fidelity and diversity scores for collections 0, 1, 2, 3, and 4 from M2D2 S2ORC and M2D2 WIKIPEDIA. The x-axes indicate the subset being evaluated against the reference. Dashed lines indicate baseline fidelity-diversity values. See App. A.3 for a list of the ten categories in each collection and App. A.4 for bootstrapped confidence bands for all collections.

	Fidelity		Diversity	
Coll.	β	p	β	p
0	+0.0011	0.586	**+0.0203**	0.001
1	-0.0031	0.083	**+0.0157**	< 0.001
2	**+0.0054**	0.001	**+0.0130**	< 0.001
3	+0.0016	0.064	**+0.0140**	< 0.001
4	-0.0014	0.523	**+0.0209**	0.001

(a) M2D2 S2ORC

	Fidelity		Diversity	
Coll.	β	p	β	p
0	-0.0021	0.456	**+0.0246**	< 0.001
1	+0.0037	0.064	**+0.0191**	< 0.001
2	-0.0009	0.329	**+0.0244**	< 0.001
3	**+0.0023**	0.019	**+0.0221**	< 0.001
4	-0.0014	0.445	**+0.0244**	< 0.001

(b) M2D2 WIKIPEDIA

Table 1: Ordinary least squares (OLS) regression coefficients (β) of fidelity-diversity on subset index across collections. We also report t-test p-values, with null hypothesis $\beta = 0$. Bold entries indicate statistical significance ($p < 0.05$).

4.2 SYNTHETIC GSM8K EXPERIMENTS

A popular approach for training models given limited data is to augment available training data with *synthetic data* from a strong existing model. Although many works explore the potential for significant boosts in performance from using high-quality synthetic data (Liu et al., 2023; Yang et al., 2024; Su et al., 2025), efficiently diagnosing synthetic data quality ahead of model training remains a key practical challenge. Our metric enables direct quantitative assessments of diversity in synthetic data, which we find to be correlated with the performances of models trained on that data.

4.2.1 SETUP

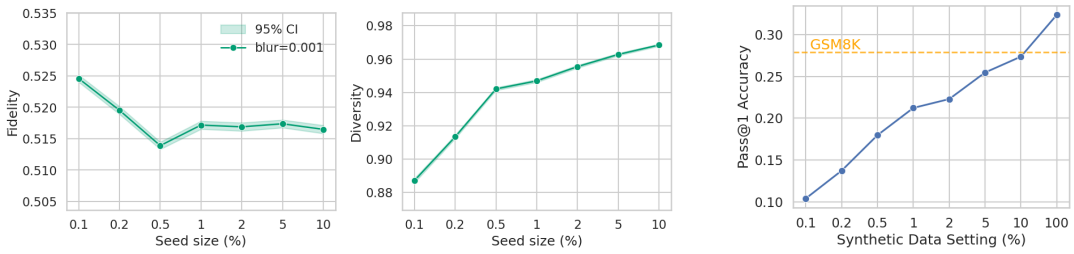
We generate synthetic GSM8K-style datasets of varying diversities using a combination of GPT-4.1, -5.2, and -5.4 (OpenAI, 2024; 2026). To generate synthetic problems, we provide GPT-4.1 with a problem from GSM8K, the "seed" problem, and prompt it to generate a structurally similar rewrite of the seed problem. Using k seed problems from the original n problems in GSM8K, we generate a synthetic dataset by prompting GPT-4.1 $(n-k)/k$ times for each seed problem, generating multiple synthetic examples based on each seed. By varying the number of seed problems k, we control the amount of diversity in the resulting synthetic dataset: if $k = 1$, then every synthetic problem is a variant of just one problem from the original GSM8K; if $k = n$, the resulting dataset should have comparable diversity to the original GSM8K. We emphasize that k serves only as a controlled ground-truth proxy for diversity that practitioners do not typically have. The question our experiment addresses is therefore whether our diversity score, computed from the data alone, recovers the diversity ordering induced by the construction, and whether the property it recovers is relevant to downstream training.

We generate synthetic datasets for each $k \in \{0.1\%, 0.2\%, 0.5\%, 1\%, 2\%, 5\%, 10\%\} \times n$, and refer to each synthetic dataset by the percentage of GSM8K training samples in its seed data. For each generated problem, we then generate step-by-step solutions using GPT-5, adopting a similar prompting strategy as (Toshniwal et al., 2024). For more details about the synthetic data generation pipeline, see App. B.1.

Applying our metric requires an appropriate reference. The GSM8K training split is inadequate as a reference for comparing synthetic datasets, since it confounds the distributional shift in linguistic style arising from the synthetic data generation process with differences in underlying structure of the math problems that we wish to measure. Instead, we generate a fully synthetic reference dataset using the entire GSM8K training split as the seed ($k = n$): for each GSM8K training example, we generate a single synthetic problem-solution pair. Critically, we only include the synthetic generations in this reference dataset. Using a fully synthetic reference allows us to better isolate differences in structural diversity when comparing to other synthetic datasets.

We then embed all datasets using E5_{small} with mean-pooling to produce separate embeddings for each problem and each solution, and collect all problem and solution embeddings. We omit the quantization step, as we found that the synthetic data embeddings are not amenable to common clustering methods, suggesting that the embeddings themselves do not contain well-separated clusters (for further discussion of this point, see App. B.2). We then compute fidelity-diversity scores on the embeddings themselves, with uniform weights and Euclidean distances between embeddings serving as the ground metric. While we can efficiently compute exact fidelity scores, we approximate the diversity scores via debiased Sinkhorn divergences (Feydy et al., 2019) (Fig. 3a).

To investigate the relationship between training data diversity and downstream performance, we perform supervised finetuning (SFT) on a Llama-3.2-1B base model (Llama Team, 2024) with synthetic data. See Fig. 3b for model performance on the GSM8K test set after finetuning on different synthetic datasets, including the reference dataset seeded with the full GSM8K training split (labeled as 100%). For comparison, we also finetune a base model on the original GSM8K training split and evaluate it on the test split. We provide further details on finetuning in App. B.4.



(a) Fidelity + diversity of synthetic GSM8K-style datasets (b) Llama-3.2-1B

Figure 3: The x-coordinates correspond to the synthetic dataset being evaluated or finetuned on, identified by the percentage of GSM8K training samples present in the seed data. **(a)** Fidelity-diversity scores with OT approximated by debiased Sinkhorn divergences, with entropic regularization hyperparameter $= 0.001$. The shaded regions in both plots correspond to 95% confidence intervals computed from bootstrapped estimates of sample variance. **(b)** Exact-match accuracies of Llama-3.2-1B base models on the GSM8K test split after finetuning on synthetic datasets. The dashed orange line indicates test accuracy achieved by a base model finetuned on the GSM8K training split.

4.2.2 Discussion

We find that our diversity metric effectively recovers ground-truth differences in structural diversity, and that diversity correlates strongly with downstream model performance.

Observe that the base model finetuned on the fully synthetic reference dataset (100%) actually achieves stronger GSM8K performance than the one finetuned on the original GSM8K training set. We reproduce this finding on a larger Llama-3.2-3B model (Fig. 8), suggesting that our synthetic data generation process produces high-quality data given sufficiently diverse seed data. This points to improvements in model performance being the result of increasing data diversity, rather than merely the inclusion of more GSM8K training examples.

Meanwhile, diversity scores across the seven datasets exhibit strong correlation with downstream test accuracies of models finetuned on the data ($r = +\mathbf{0.973}$, p-value < 0.001). In contrast, fidelity scores do not exhibit statistically significant correlation ($r = -0.670$, $p = 0.099$), further supporting the premise that our metrics are capable of disaggregating fidelity and diversity. We additionally verify that the fidelity metrics are responsive to failures of fidelity in this setting (App. B.3).

The sensitivity of our metrics on these synthetic datasets is particularly notable, given that the embeddings capture information about problem structure in a manner so subtle that conventional clustering methods fail.

5 Conclusion

The aim of our work has been to develop fidelity-diversity metrics for comparing text datasets. On the one hand, we introduce new constructions for this problem by leveraging optimal transport between mixing measures. On the other hand, we show that modern embedding models are sufficiently expressive for our metrics to detect human-interpretable variations in fidelity and diversity within large-scale datasets. While we expect our metrics to be applicable for a variety of tasks such as language model evaluation, our primary experiments focused on leveraging them to quantify the quality of synthetically generated datasets.

Finally, we note that there is a growing literature emphasizing the fragility of evaluation metrics for text data arising from open-ended language models (Stein et al., 2023; Räisä et al., 2025). Though we do not propose solutions to these challenges, we hope our work nevertheless demonstrates the utility of such metrics in current generative modeling applications.

AI USE STATEMENT

In this work, we used generative AI tools for generating synthetic data sets, cleaning and reformatting datasets, implementing methods, and providing feedback on experiments.

We have not used generative AI tools for developing theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, supporting qualitative and thematic data analysis, or interpreting results. Assisting with translation was not applicable to this work.

Additionally, we used generative AI tools for creating or modifying scientific figures or images, as well as suggesting experimental parameters. We have reviewed all AI-assisted work. We validate the quality of our synthetic datasets by comparing the performance of models finetuned on synthetic data with that of models finetuned on human-written training data. LLM-generated code was verified and tested for correctness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We include our synthetic datasets in supplementary material and also provide a complete description of the data generation procedure, including the specific prompts used, in App. B.1. We will also open-source our code implementing our method and all experiments if accepted. We include detailed descriptions of our experimental setups in Appendices A.3 and B.1, as well as a description of the finetuning procedure used in the GSM8K experiments in App. B.4.

REFERENCES

Ahmed Alaa, Boris Van Breugel, Evgeny S Saveliev, and Mihaela van der Schaar. How faithful is your synthetic data? sample-level metrics for evaluating and auditing generative models. In *International Conference on Machine Learning*. PMLR, 2022. 2

David Alvarez-Melis and Nicolo Fusi. Geometric dataset distances via optimal transport. *Advances in Neural Information Processing Systems*, 33:21428–21439, 2020. 2

Xin Bing, Florentina Bunea, Seth Strimas-Mackey, and Marten Wegkamp. Likelihood estimation of sparse topic distributions in topic models and its applications to wasserstein document distance calculations. *The Annals of Statistics*, 50(6):3307–3333, 12 2022. doi: 10.1214/22-AOS2229. URL https://doi.org/10.1214/22-AOS2229. 2

Xin Bing, Florentina Bunea, and Jonathan Niles-Weed. Estimation and inference for the wasserstein distance between mixing measures in topic models. *Bernoulli, forthcoming*, 2025. 2

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. Evaluation of text generation: A survey. *arXiv preprint arXiv:2006.14799*, 2020. 2

Sinho Chewi, Jonathan Niles-Weed, and Philippe Rigollet. *Statistical optimal transport*. Springer, 2025. 4

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. 2, 16

Josip Djolonga, Mario Lucic, Marco Cuturi, Olivier Bachem, Olivier Bousquet, and Sylvain Gelly. Precision-recall curves using information divergence frontiers. In *International Conference on Artificial Intelligence and Statistics*, pp. 2550–2559. PMLR, 2020. 2

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouve, and Gabriel Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 2681–2690, 2019. 8

Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. *Transactions on Machine Learning Research*, 2023. 2

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, et al. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*, 2023. 1

Moritz Hardt and Eric Price. Tight bounds for learning a mixture of two gaussians. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pp. 753–760, 2015. 16

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 2

Nhat Ho and XuanLong Nguyen. Convergence rates of parameter estimation for some weakly identifiable finite mixtures. *The Annals of Statistics*, 44(6):2726, 2016. 16

Takumi Ito, Kees van Deemter, and Jun Suzuki. Reference-free evaluation metrics for text generation: A survey. *arXiv preprint arXiv:2501.12011*, 2025. 2

Marco Jiralerspong, Joey Bose, Ian Gemp, Chongli Qin, Yoram Bachrach, and Gauthier Gidel. Feature likelihood divergence: evaluating the generalization of generative models using samples. In *Advances in Neural Information Processing Systems*, 2023. 2

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances. In *International conference on machine learning*, pp. 957–966. PMLR, 2015. 2

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *Advances in neural information processing systems*, 32, 2019. 2

Florian Le Bronnec, Alexandre Verine, Benjamin Negrevergne, Yann Chevaleyre, and Alexandre Allauzen. Exploring precision and recall to assess the quality and diversity of llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024. 2, 23

Bingbin Liu, Sebastien Bubeck, Ronen Eldan, Janardhan Kulkarni, Yuanzhi Li, Anh Nguyen, Rachel Ward, and Yi Zhang. TinyGSM: achieving >80% on GSM8k with small language models, 2023. URL https://arxiv.org/abs/2312.09241. 8

Llama Team. The llama 3 herd of models. *arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783. 8, 20

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Dan S. Weld. S2orc: The semantic scholar open research corpus, 2020. URL https://arxiv.org/abs/1911.02782. 14

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. URL https://arxiv.org/abs/1711.05101. 20

Ian Magnusson, Akshita Bhagia, Valentin Hofmann, Luca Soldaini, Ananya Harsh Jha, Oyvind Tafjord, Dustin Schwenk, Evan Pete Walsh, Yanai Elazar, Kyle Lo, Dirk Groeneveld, Iz Beltagy, Hannaneh Hajishirzi, Noah A. Smith, Kyle Richardson, and Jesse Dodge. Paloma: A benchmark for evaluating language model fit, 2024. URL https://arxiv.org/abs/2312.10523. 6

Tudor Manole and Nhat Ho. Uniform convergence rates for maximum likelihood estimation under two-component gaussian mixture models. *arXiv preprint arXiv:2006.00704*, 2020. 16

Tudor Manole and Nhat Ho. Refined convergence rates for maximum likelihood estimation under finite mixture models. In *International Conference on Machine Learning*, pp. 14979–15006. PMLR, 2022. 2

Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In *International Conference on Machine Learning*. PMLR, 2020. 2

Huy Nguyen, Nhat Ho, and Alessandro Rinaldo. Sigmoid gating is more sample efficient than softmax gating in mixture of experts. *Advances in Neural Information Processing Systems*, 37:118357–118388, 2024. 2

XuanLong Nguyen. Convergence of latent mixing measures in finite and infinite mixture models. *The Annals of Statistics*, pp. 370–400, 2013. 2

XuanLong Nguyen. Posterior contraction of the population polytope in finite admixture models. *Bernoulli*, pp. 618–646, 2015. 2

OpenAI. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774. 8

OpenAI. Openai gpt-5 system card, 2026. URL https://arxiv.org/abs/2601.03267. 8

Gabriel Peyré and Marco Cuturi. *Computational optimal transport: With applications to data science*. Now Foundations and Trends, 2019. 4

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. Mauve: Measuring the gap between neural text and human text using divergence frontiers. *Advances in Neural Information Processing Systems*, 34:4816–4828, 2021. 1, 2

Krishna Pillutla, Lang Liu, John Thickstun, Sean Welleck, Swabha Swayamdipta, Rowan Zellers, Sewoong Oh, Yejin Choi, and Zaid Harchaoui. Mauve scores for generative models: Theory and practice. *arXiv:2212.14578*, 2023. URL https://arxiv.org/abs/2212.14578. 2, 25

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 25

Ossi Räisä, Boris Van Breugel, and Mihaela Van Der Schaar. Position: All current generative fidelity and diversity metrics are flawed. *arXiv preprint arXiv:2505.22450*, 2025. 9

Machel Reid, Victor Zhong, Suchin Gururangan, and Luke Zettlemoyer. M2d2: A massively multi-domain language modeling dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 964–975, 2022. 2, 5, 6

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019. 3

Gabriel Rioux, Apoorva Nitsure, Mattia Rigotti, Kristjan Greenewald, and Youssef Mroueh. Multivariate stochastic dominance via optimal transport and applications to models benchmarking. *Advances in Neural Information Processing Systems*, 37:39190–39223, 2024. 3

Subham Sekhar Sahoo, Zhihan Yang, Yash Akhauri, Johnna Liu, Deepansha Singh, Zhoujun Cheng, Zhengzhong Liu, Eric Xing, John Thickstun, and Arash Vahdat. Esoteric language models: A family of any-order diffusion llms, 2026. URL https://arxiv.org/abs/2506.01928. 25

Mehdi SM Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. *Advances in neural information processing systems*, 31, 2018. 1, 2

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016. 2

George Stein, Jesse Cresswell, Rasa Hosseinzadeh, Yi Sui, Brendan Ross, Valentin Villecroze, Zhaoyan Liu, Anthony L Caterini, Eric Taylor, and Gabriel Loaiza-Ganem. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. *Advances in Neural Information Processing Systems*, 2023. 9

Dan Su, Kezhi Kong, Ying Lin, Joseph Jennings, Brandon Norick, Markus Kliegl, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. Nemotron-cc: Transforming common crawl into a refined long-horizon pretraining dataset, 2025. URL https://arxiv.org/abs/2412.02595. 8

Shubham Toshniwal, Wei Du, Ivan Moshkov, Branislav Kisacanin, Alexan Ayrapetyan, and Igor Gitman. OpenMathInstruct-2: Accelerating ai for math with massive open-source instruction data, 2024. URL https://arxiv.org/abs/2410.01560. 8

Cédric Villani. *Topics in Optimal Transportation*, volume 58. American Mathematical Soc., 2003. 4

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training, 2024. URL https://arxiv.org/abs/2212.03533. 6

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 13484–13508, 2023. 1

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference, 2024. URL https://arxiv.org/abs/2412.13663. 25

Larry Wasserman. *All of statistics: a concise course in statistical inference*, volume 26. Springer, 2004. 16

Zhihan Yang, Wei Guo, Shuibai Zhang, Subham Sekhar Sahoo, Yongxin Chen, Arash Vahdat, Morteza Mardani, and John Thickstun. Continuous diffusion scales competitively with discrete diffusion for language, 2026. URL https://arxiv.org/abs/2605.18530. 2

Zitong Yang, Neil Band, Shuangping Li, Emmanuel Candès, and Tatsunori Hashimoto. Synthetic continued pretraining, 2024. URL https://arxiv.org/abs/2409.07431. 8

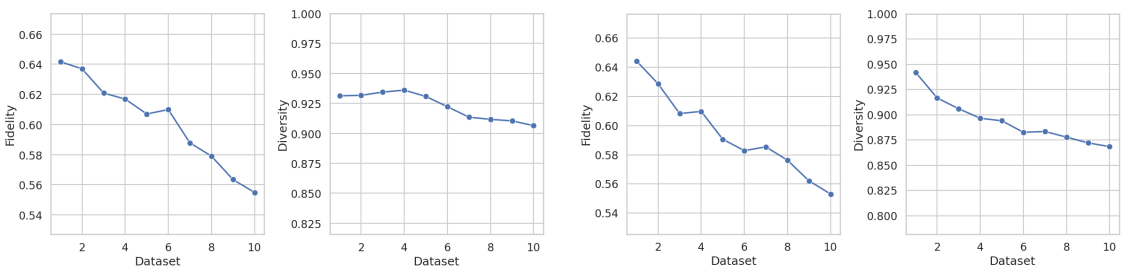
Mikhail Yurochkin, Sebastian Claici, Edward Chien, Farzaneh Mirzazadeh, and Justin M Solomon. Hierarchical optimal transport for document representation. *Advances in neural information processing systems*, 32, 2019. 2

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020. 1

A M2D2 EXPERIMENTS

A.1 VALIDATING FIDELITY

To verify whether our fidelity scores are actually sensitive to changes in fidelity, we also construct two experiments in which we incrementally shift the evaluation dataset away from the reference. More concretely, we select ten categories $C_1, \ldots, C_{10}$ for the reference, as well as ten additional categories $D_1, \ldots, D_{10}$ disjoint from the reference. Each dataset i consists of data from subdomains $D_1, \ldots, D_i, C_{i+1}, \ldots, C_{10}$. Thus, lower values of i correspond to datasets closer to the reference, while higher values correspond to being further from the reference. We again embed the data using E5_{small} and fit spherical GMMs for all datasets. As shown in Fig. 4, our fidelity scores are capable of detecting gradual degradations in fidelity. We list the exact categories used for this experiment in App. A.3.



(a) M2D2 S2ORC (b) M2D2 WIKIPEDIA

Figure 4: Fidelity & diversity scores for controlled distribution-shift experiments. Fidelity scores for both collections exhibit significant negative association with the number of categories disjoint from the reference; computing the OLS regression coefficient β and t-test p-value for the null hypothesis $\beta = 0$ gives $\beta = \mathbf{-0.0096}$ ($p < 0.001$) for M2D2 S2ORC and $\beta = \mathbf{-0.0093}$ ($p < 0.001$) for M2D2 WIKIPEDIA.

A.2 ANALYSIS OF FIDELITY

For WIKIPEDIA Collection 3, the observed trend is fragile in the sense that dropping the last data point (subset 10) gives $\beta = +0.0017$ ($p = 0.097$), which is no longer significant. Alternatively, applying a conservative Bonferroni correction with the number of tests $m = 10$ to our significance threshold of $\alpha = 0.05$ makes this insignificant. We note that the diversity trends do not depend on this correction: every diversity coefficient in Table 1 has $p \leq 0.001$ and remains significant under the same Bonferroni correction.

We believe the trend in S2ORC Collection 2 can be attributed to noise in the GMM fits on these small datasets. In particular, the first subset in S2ORC Collection 2 only contains 1,386 points, with BIC selecting 20 components for the GMM. Over all subsets in all 10 collections, this constitutes the fewest points per component (ppc) $= 69.3$. In fact, if we rank the 100 subsets by ppc, the first three subsets in S2ORC Collection 2 are among the eight lowest, all with substantially fewer points per component than the reference set (ppc $= 137.4$). This suggests that the GMMs for these subsets may have been noisier fits, explaining the lower fidelity scores for these early subsets. This sensitivity to k in smaller datasets is a practical limitation of our implementation of the quantization step.

A.3 DATA

M2D2 S2ORC consists of text sampled from academic papers organized into a two-layer categorization by academic field. Top-level domains were first provided in S2ORC (Lo et al., 2020), while subdomains follow the paper's arXiv categorization. M2D2 WIKIPEDIA contains text from Wikipedia articles categorized by the top two levels of the Wikipedia ontology.

Each of the following collections consists of ten subdomains sampled uniformly at random (without replacement) from all available subdomains in M2D2 S2ORC or M2D2 WIKIPEDIA.

A.3.1 M2D2 S2ORC COLLECTIONS

Collection 0: math.PR, cs.PL, q-fin.CP, cond-mat.quant-gas, eess.SP, stat.ML, q-fin.PR, q-bio.NC, nucl-th, physics.atm-clus

Collection 1: physics.optics, math.RA, physics.class-ph, econ.TH, q-bio.MN, hep-th, Philosophy, physics_l1, nucl-ex, math.NA

Collection 2: math.CO, cond-mat_l1, math.RA, stat.ML, cs.CE, plasm-ph, cs.GR, atom-ph, math.CA, math.SP

Collection 3: cs.CV, cs.NA, q-fin.PM, stat.ML, physics.app-ph, q-bio.CB, eess.AS, math.FA, math.QA, math.CA

Collection 4: math.CA, cond-mat.str-el, Art, astro-ph.IM, q-fin.CP, math_l1, math.SP, astro-ph.EP, q-fin.GN, math.OC

For the fidelity validation experiment from App. A.1, the ten M2D2 S2ORC reference subdomains were all selected from the top-level computer science domain (*cs.AI, cs.AR, cs.CC, cs.CE, cs.CG, cs.CL, cs.CR, cs.CV, cs.CY, and cs.DB*). The ten disjoint subdomains were selected to minimize topical overlap with the reference, and included subdomains from physics, philosophy, and art (*astro-ph.CO, atom-ph, chem-ph, Philosophy, cond-mat.soft, astro-ph.EP, Art, cond-mat.quant-gas, astro-ph.HE, and cond-mat.stat-mech*).

A.3.2 M2D2 WIKIPEDIA COLLECTIONS

Collection 0: Culture_and_the_arts__Games_and_Toys, Health_and_fitness__Exercise, History_and_events__By_continent, Natural_and_physical_sciences__Nature, Religion_and_belief_systems__Belief_systems, Culture_and_the_arts__Culture_and_Humanities, Human_activites__Human_activities, Society_and_social_sciences__Society, Culture_and_the_arts__Mass_media, Culture_and_the_arts__The_arts_and_Entertainment

Collection 1: Natural_and_physical_sciences__Biology, Health_and_fitness__Health_science, Culture_and_the_arts__Culture_and_Humanities, Health_and_fitness__Self_care, Religion_and_belief_systems__Major_beliefs_of_the_world, Health_and_fitness__Human_medicine, Religion_and_belief_systems__Allah, Mathematics_and_logic__Logic, Health_and_fitness__Exercise, Technology_and_applied_sciences__Engineering

Collection 2: Culture_and_the_arts__Performing_arts, Health_and_fitness__Self_care, Natural_and_physical_sciences__Physical_sciences, Health_and_fitness__Exercise, Human_activites__Human_activities, Philosophy_and_thinking__Thinking, Society_and_social_sciences__Social_sciences, Religion_and_belief_systems__Major_beliefs_of_the_world, Culture_and_the_arts__Sports_and_Recreation, Health_and_fitness__Health_science

Collection 3: Religion_and_belief_systems__Major_beliefs_of_the_world, Natural_and_physical_sciences__Physical_sciences, Culture_and_the_arts__Sports_and_Recreation, Technology_and_applied_sciences__Engineering, History_and_events__By_region, History_and_events__By_continent, History_and_events__By_period, Technology_and_applied_sciences__Computing, Natural_and_physical_sciences__Biology, Culture_and_the_arts__Games_and_Toys

Collection 4: Culture_and_the_arts__The_arts_and_Entertainment, History_and_events__By_continent, Health_and_fitness__Self_care, Religion_and_belief_systems__Belief_systems, Culture_and_the_arts__Games_and_Toys, Mathematics_and_logic__Logic, History_and_events__By_period, Culture_and_the_arts__Culture_and_Humanities, Culture_and_the_arts__Sports_and_Recreation, Culture_and_the_arts__Mass_media

For the fidelity validation experiment from App. A.1, the ten M2D2 WIKIPEDIA reference subdomains were selected from the natural and applied sciences (*Biology, Earth_sciences, Nature, Physical_sciences, Mathematics, Logic, Agriculture, Computing, Engineering, Transport*), while the disjoint subdomains were focused on the arts, history, and culture (*Culture_and_Humanities, Games_and_Toys, Mass_media, Performing_arts, Culture_and_the_arts__Sports_and_Recreation, The_arts_and_Entertainment, Visual_arts, History_and_events_By_continent, History_and_events__By_period, History_and_events_By_region*).

A.4 STATISTICAL METHODS

All M2D2 datasets were quantized using Gaussian Mixture Models (GMMs) with spherical covariances. We fit the GMMs using maximum likelihood estimation, approximated via the Expectation-Maximization algorithm. The number of mixture components k was independently selected for each dataset by minimizing the Bayesian Information Criteria (BIC) across models fitted with

$$k \in \{5, 10, 15, 20, 25, 30, 40, 50, 60, 70, 80, 90, 100\}.$$

We adopt a ground metric for atoms $\theta_1 = (\mu_1, \sigma_1^2 \mathrm{Id}_z)$, $\theta_2 = (\mu_2, \sigma_2^2 \mathrm{Id}_z)$ which was previously proposed in the theoretical literature on location-scale Gaussian mixture models (Hardt & Price, 2015; Ho & Nguyen, 2016; Manole & Ho, 2020):

$$d(\theta_1, \theta_2) = \sqrt{\|\mu_1 - \mu_2\|_2^2 + z|\sigma_1^2 - \sigma_2^2|}.$$

We report estimates of our metrics together with 95% Gaussian variability intervals of the form $\widehat{\theta} \pm 1.96\widehat{\sigma}$, where $\widehat{\theta}$ denotes a fidelity or diversity metric, and $\widehat{\sigma}^2$ denotes a nonparametric bootstrap estimate of its variance (Wasserman, 2004) (Fig. 5). We use 500 bootstrap replications for each variance estimate.

B SYNTHETIC GSM8K EXPERIMENTS

B.1 SYNTHETIC GSM8K DATA GENERATION

GSM8K (Cobbe et al., 2021) is a curated collection of 8.8k grade-school math word problems created by human problem writers, divided into a 7.4k training split and a 1.3k test split. Each example consists of a problem-solution pair, where the problem can be solved using some sequence of the four elementary operations (addition, subtraction, multiplication, division) and has a well-defined positive integer answer. The solutions contain the key reasoning steps expressed in natural language, followed by the final answer.

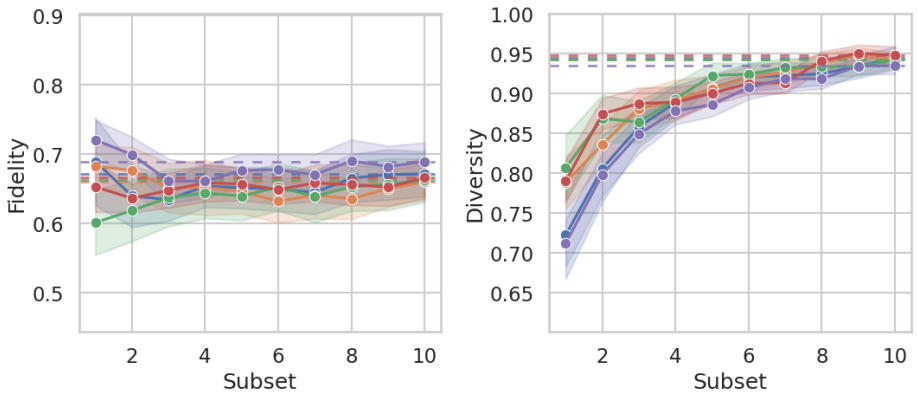
To elaborate on the synthetic dataset construction, for each of the {0.1%, 0.2%, 0.5%, 1%, 2%, 5%, 10%} synthetic datasets, we used the first {0.1%, 0.2%, 0.5%, 1%, 2%, 5%, 10%} problems in the GSM8K training set as the seed. Within each dataset, we generated the same number of synthetic rewrites for every seed problem; to ensure that the final datasets were all of similar sizes, this meant generating {999, 499, 199, 99, 49, 19, 9} synthetic rewrites for each seed problem. Then, for each dataset, we appended the seed problems to the generated synthetic problems and shuffled the problems before storing them.

The synthetic data generation pipeline consisted of three steps. We first generated questions using GPT-4.1 with temperature=1.5 to encourage variety in the generations. We then generated solutions for each synthetic question using GPT-5.2 with reasoning effort="high" and verbosity="medium" to encourage correct solutions with clear reasoning steps. Finally, we checked for proper answer formatting and whether the final answer was a positive integer. Occasionally, a generated question would yield a nonpositive or noninteger answer. For all invalid problem-solution pairs, we used GPT-5.4 (again with reasoning effort="high" and verbosity="medium") to adjust the numbers in the problem and solution so as to yield positive integer answers, repeating this procedure as needed. See the following subsections for the exact prompt templates used in each step.

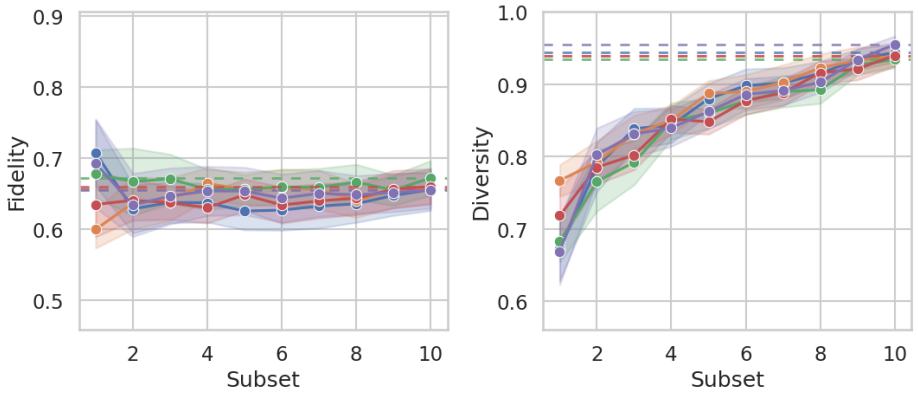
B.1.1 QUESTION GENERATION PROMPT

```
Help the user create a new math dataset by generating problems one at a time.
    You will be given a reference problem from an existing dataset, from which
    you should generate a similar problem.
Be creative with the names and contexts. Each problem you generate should be
    different. The problem must have a positive integer answer.

Here are some examples of how to complete this task.
```



(a) M2D2 S2ORC



(b) M2D2 WIKIPEDIA

Figure 5: Bootstrapped estimates of 95% confidence intervals of fidelity and diversity scores, indicated by the shaded regions. The solid and dashed lines are identical to those in Fig. 2.

```
Problem:
Natalia sold clips to 48 of her friends in April, and then she sold half as many
     clips in May. How many clips did Natalia sell altogether in April and May?

Write another problem similar to this one.
Ryan sold 8 packs of gum in September, and then he sold twice as many packs in
    October. How many packs did Ryan sell in total during September and October?

Problem:
Weng earns $12 an hour for babysitting. Yesterday, she just did 50 minutes of
    babysitting. How much did she earn?

Write another problem similar to this one.
Ali makes $48 an hour as a tutor. On Mondays, she tutors for 45 minutes. How
    much does she make on Mondays?

Problem:
```

```
Betty is saving money for a new wallet which costs $100. Betty has only half of
    the money she needs. Her parents decided to give her $15 for that purpose,
    and her grandparents twice as much as her parents. How much more money does
    Betty need to buy the wallet?

Write another problem similar to this one.
Andre works at a bike repair shop with Tom and Brady. On Tuesday, there are
    twenty bikes they need to repair. By noon, Andre has repaired a quarter of
    them. Tom has repaired three, and Brady has repaired double that. How many
    remaining bikes do they need to repair in the afternoon?

Here is the problem from the user:
{question}

Write another problem similar to this one. Start directly with the problem
    statement and DO NOT include any phrases such as "Here is a new problem
    similar to a given one".
After the problem is generated, finish your response right away.
```

B.1.2 SOLUTION GENERATION PROMPT

```
Here are some examples of problems and solutions, followed by a new problem that
     you need to solve.
Make sure to include each step in your reasoning. You may use either x or * to
    indicate multiplication, and / for division.
Write your answer (and only the numeric answer) on a new line beginning with
    "#### ".


Problem:
Natalia sold clips to 48 of her friends in April, and then she sold half as many
     clips in May. How many clips did Natalia sell altogether in April and May?

Solution:
Natalia sold 48/2 = 24 clips in May.
Natalia sold 48+24 = 72 clips altogether in April and May.
#### 72


Problem:
Weng earns $12 an hour for babysitting. Yesterday, she just did 50 minutes of
    babysitting. How much did she earn?

Solution:
Weng earns 12/60 = $0.2 per minute.
Working 50 minutes, she earned 0.2 x 50 = $10.
#### 10


Problem:
Betty is saving money for a new wallet which costs $100. Betty has only half of
    the money she needs. Her parents decided to give her $15 for that purpose,
    and her grandparents twice as much as her parents. How much more money does
    Betty need to buy the wallet?

Solution:
In the beginning, Betty has only 100 / 2 = $50.
Betty's grandparents gave her 15 * 2 = $30.
```

```
This means Betty needs 100 - 50 - 30 - 15 = $5 more.
#### 5


Problem:
{synthetic question}

Solution:
```

B.1.3 PROBLEM ADJUSTMENT PROMPT

```
Help the user fix a math problem to have positive integer solutions by changing
    the numbers involved.
Make sure to leave the rest of the problem and solution unchanged. Here are some
     examples of how to complete this task.


Problem: Six coworkers go out for lunch and order the following items: 6 slices
    of pizza at $2 each; 3 bowls of salad at $3.50 each; 6 bottles of juice at
    $1.25 each; and 2 plates of pasta at $4.75 each. If they split the total
    bill evenly, how much does each coworker need to pay?

Solution: Pizza costs 6 x $2 = $12.
Salad costs 3 x $3.50 = $10.50.
Juice costs 6 x $1.25 = $7.50.
Pasta costs 2 x $4.75 = $9.50.

Total bill is $12 + $10.50 + $7.50 + $9.50 = $39.50.
Split among 6 coworkers: $39.50 / 6 = $6.5833..., which rounds to $6.58.

#### 6.58

Fixed Problem: Six coworkers go out for lunch and order the following items: 6
    slices of pizza at $2 each; 3 bowls of salad at $3.50 each; 6 bottles of
    juice at $0.75 each; and 2 plates of pasta at $4.50 each. If they split the
    total bill evenly, how much does each coworker need to pay?

Fixed Solution: Pizza costs 6 x $2 = $12.
Salad costs 3 x $3.50 = $10.50.
Juice costs 6 x $0.75 = $4.50.
Pasta costs 2 x $4.50 = $9.

Total bill is $12 + $10.50 + $4.50 + $9 = $36.
Split among 6 coworkers: $36 / 6 = $6.

#### 6


Problem: In a zoo, 16 parrots are sharing 15 bowls of seeds. 7 bowls each have
    13 scoops, where each scoop contains 11 seeds. The other bowls each contain
    8 scoops, and each scoop has 15 seeds. If the total seeds are divided
    equally among all the parrots, how many seeds does each parrot get?

Solution: 7 bowls have 13 * 11 = 143 seeds each, so they have 7 * 143 = 1001
    seeds total.
The other 15-7=8 bowls have 8 * 15 = 120 seeds each, so they have 8 * 120 = 960
    seeds total.
Altogether there are 1001 + 960 = 1961 seeds.
```

```
Dividing equally among 16 parrots gives 1961 / 16 = 122.5625 seeds per parrot.

#### 122.5625


Fixed Problem: In a zoo, 16 parrots are sharing 15 bowls of seeds. 7 bowls each
    have 9 scoops, where each scoop contains 16 seeds. The other bowls each
    contain 8 scoops, and each scoop has 15 seeds. If the total seeds are
    divided equally among all the parrots, how many seeds does each parrot get?

Fixed Solution: 7 bowls have 9 * 16 = 144 seeds each, so they have 7 * 144 =
    1008 seeds total.
The other 15-7=8 bowls have 8 * 15 = 120 seeds each, so they have 8 * 120 = 960
    seeds total.
Altogether there are 1008 + 960 = 1968 seeds.
Dividing equally among 16 parrots gives 1968 / 16 = 123 seeds per parrot.

#### 123


Here is the problem and solution from the user:

Problem: {invalid synthetic question}

Solution: {previously generated solution}
```

B.2 CLUSTERING

Fitting a GMM with BIC on the embedded synthetic datasets typically resulted in a number of mixture components on the order of the embedding dimension. Further exploratory data analysis suggested that the data lacks discernible clusters when mapped into the embedding space. This motivated our decision to compute our metrics directly on the empirical measure, as opposed to atoms derived from clustering, as in the previous experiments.

B.3 VALIDATING FIDELITY

We also verify that our metrics capture fidelity even when approximated on empirical measures instead of mixing measures. To that end, we exactly replicate the setup used to produce Fig. 3a, except we now swap the synthetic dataset being measured with the reference set. Instead of fixing the reference set, we now assess fidelity and diversity of the same dataset against increasingly diverse references. Consequently, we expect fidelity to improve monotonically as we increase the breadth of what we consider as the reference. This is in fact the observed outcome (Fig. 6), suggesting that our metrics remain sensitive to differences in fidelity even on the level of empirical measures.

B.4 SFT

We conduct supervised fine-tuning (SFT) on both Llama-3.2-1B and Llama-3.2-3B base models (Llama Team, 2024). We finetuned a fresh copy of each base model on each of the eight synthetic datasets and the original GSM8K training set. All finetuning for both models was completed using the AdamW optimizer (Loshchilov & Hutter, 2019) with cosine learning rate decay over a total of two epochs. We used a maximum learning rate of 4e-5 for the Llama-3.2-1B and 2e-5 for the Llama-3.2-3B. For all training datasets, we remove and set aside 200 randomly sampled examples to use for validation. Throughout training, we compute both cross-entropy loss and exact-match accuracy on the validation set to track progress. The final model used for evaluation on the GSM8K test set is loaded from the checkpoint achieving the highest validation accuracy. See Fig. 7 for all

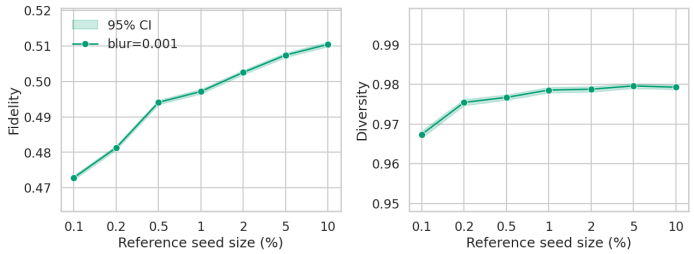
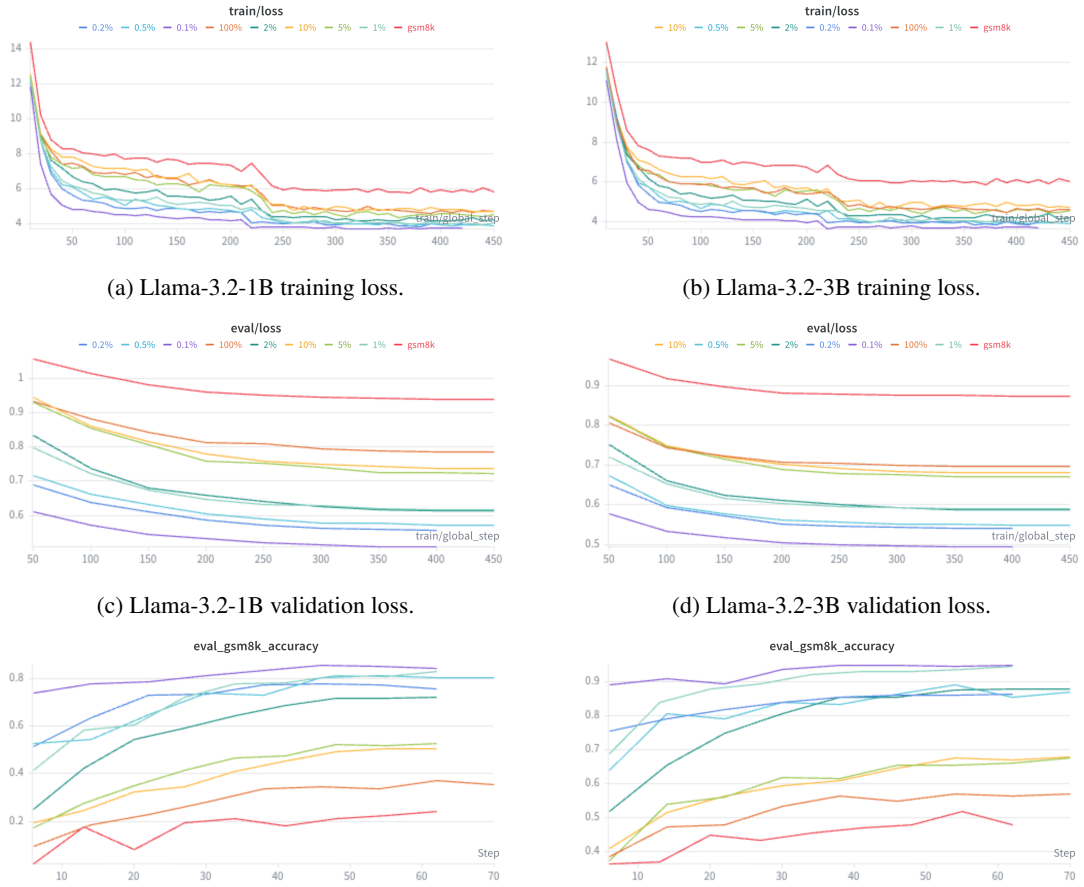


Figure 6: Fidelity + diversity of swapped GSM8K-style datasets.

training and validation loss curves, as well as the validation accuracy used for model selection. All finetuning can be completed on an NVIDIA RTX A6000 GPU in 1 to 2 hours.



(a) Llama-3.2-1B training loss.

(b) Llama-3.2-3B training loss.

(c) Llama-3.2-1B validation loss.

(d) Llama-3.2-3B validation loss.

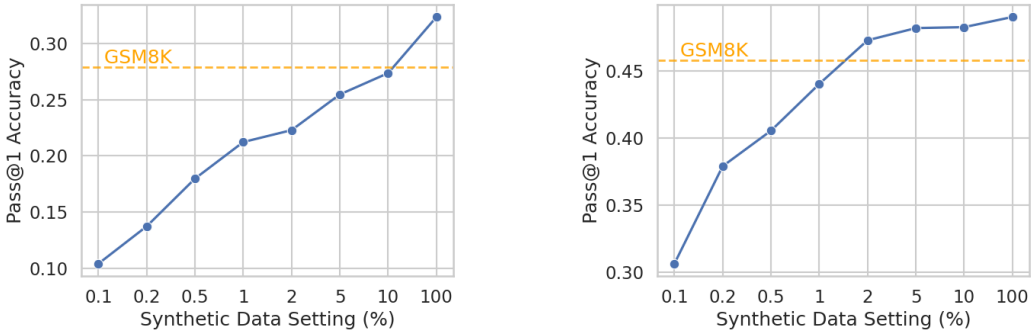
(e) Llama-3.2-1B validation accuracy.

(f) Llama-3.2-3B validation accuracy.

Figure 7: Training and validation curves from SFT on GSM8K.

B.5 Model evaluation

Each finetuned model is evaluated on exact-match accuracy on the GSM8K test set, with basic filtering applied to extract the final answer from the model generations. Test accuracies for both the 1B and 3B models are included in Fig. 8. We see a clear upward trend in test accuracies across both model sizes, with the larger model extracting strong test performance even from training on synthetic datasets of smaller seed size.



(a) Llama-3.2-1B

(b) Llama-3.2-3B

Figure 8: Exact-match test accuracies of Llama-3.2 base models finetuned on synthetic datasets, evaluated on the GSM8K test split. Values on the x-axes indicate the percentage of original GSM8K training examples included in the seed data for that synthetic dataset. The dashed orange line indicates the test accuracy achieved by finetuning a base model on the entire original GSM8K training split.

C Comparisons to Existing Metrics

C.1 FID, MAUVE, and Vendi

We compare our metric to a FID-style score, MAUVE, and Vendi in both the GSM8K setting from Section 4.2 and the swapped GSM8K setting from Appendix B.3. In the swapped setting, we switch the synthetic dataset being evaluated with the reference set to assess fidelity and diversity of the same dataset against increasingly diverse references. Table 2 lists the relevant correlations (r) and p-values for each score.

	Fidelity		Diversity		FID		MAUVE		Vendi	
	r	p	r	p	r	p	r	p	r	p
GSM8K	-0.670	0.099	**+0.973**	< 0.001	**-0.972**	< 0.001	**+0.872**	0.011	**+0.875**	0.010
Swapped	**+0.992**	< 0.001	**+0.877**	0.010	**-0.972**	< 0.001	**+0.871**	0.011	N/A	N/A

Table 2: Comparison between our fidelity and diversity metrics, and the FID, Vendi, and MAUVE scores on the GSM8K (Sec. 4.2) and Swapped GSM8K (App. B.3) data.

Referring to the GSM8K results in the top row, fidelity does not correlate in a significant manner with the seed size used to create the evaluation dataset while diversity exhibits a strong positive correlation. As a reminder, since we used seed size to control the diversity of the generated evaluation dataset and selected a maximally diverse reference dataset, we expected diversity to increase with seed size and fidelity to remain roughly flat. From the swapped results in the bottom row, observe that fidelity correlates strongly with the seed size of the reference dataset, which is again consistent with the experimental design.

Examining the FID and MAUVE scores in the GSM8K and Swapped GSM8K rows, both scores are identical, up to quantization noise for MAUVE. As such, it is impossible to discern whether differences in the FID or MAUVE score are driven by changes in fidelity or diversity, without a supplementary score which disaggregates diversity and fidelity.

Finally, while Vendi scores do isolate diversity, they are an absolute, *reference-free* measure. This is conceptually distinct from our notion of diversity. For us, the same candidate dataset should score differently on diversity depending on the desired scope of data, as specified via the reference. For instance, consider the swapped GSM8K setup, where reference datasets of varying diversity might encode various practitioners' preferences. Since Vendi scores are reference-free, the same evaluation dataset necessarily achieves the same fixed Vendi score, while our fidelity-diversity scores vary with the choice of reference.

C.2 Precision & Recall

While our metrics bear a strong resemblance to the precision and recall metrics put forth in Le Bronnec et al. (2024), there exist important conceptual differences in our approaches. At a high level, their precision and recall scores only quantify overlap in the supports of the evaluation and reference distributions P and Q. Precision captures the proportion of points drawn from P which lie in the support of Q, while recall captures the mass under Q which is covered by the support of P:

$$\begin{aligned} \text{Precision} &= \Pr_{u\sim P}[u \in \operatorname{supp} Q] = P(\operatorname{supp} Q), \\ \text{Recall} &= \Pr_{v\sim Q}[v \in \operatorname{supp} P] = Q(\operatorname{supp} P). \end{aligned} \tag{4}$$

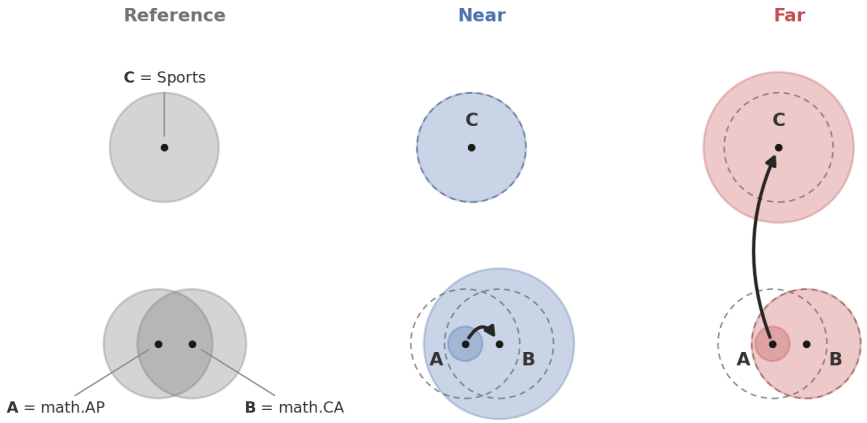
In contrast, our fidelity score quantifies how far points drawn from P lie from the support of Q, while diversity quantifies residual mass discrepancies between P and Q in terms of a transport distance.

To illustrate the empirical differences that arise from using transport distances instead of overlap estimation, we give an example in which recall fails to capture our intuitive conception of diversity, while our diversity score succeeds. Consider the setting depicted in Fig. 9a, where we draw data from three categories. The first two categories A and B are semantically similar, while the third category C is semantically far from either. For the reference set, we sample a uniform mixture over the three categories. Now consider two alternatives: in the first, we start with the uniform mixture and displace mass from category A to B (near), while we displace the same amount of mass from A to C (far) in the second. By design, we should measure stronger degradations in diversity (recall) in the far than the near alternative, since transferring mass between closer categories has a more mild effect on the overall distribution than moving mass to a distant category. We also expect a stronger decline in fidelity (precision) for the far alternative than near, although fidelity should degrade less aggressively than diversity in both, since mass is only ever transported to destinations in or near the reference support.

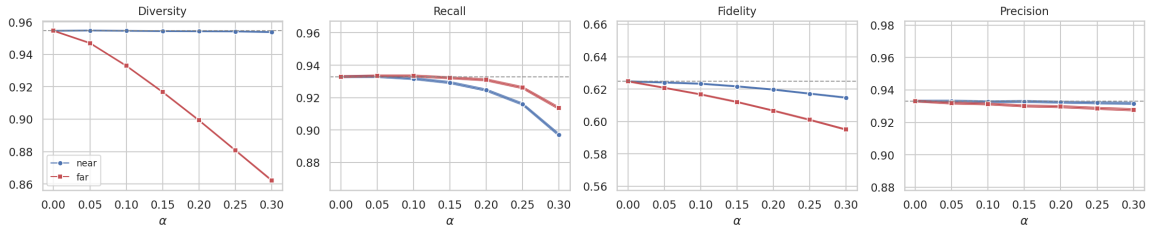
For α the fraction of total mass being displaced, we compute both fidelity-diversity and precision-recall metrics on the two alternatives with α ranging between 0 and 0.3, against the uniform reference (Fig. 9b). As in Section 4.1, we randomly partition each category before sampling our reference datapoints from one partition and evaluation datapoints from the other. We compute the quantization step for fidelity and diversity using k-means with fixed $k = 100$, since all datasets are of fixed size. Observe that while fidelity-diversity metrics behave as expected, recall actually contradicts the expected trend. In particular, our diversity score witnesses a much stronger decline when mass is displaced to the far category than the near category, while recall declines more severely when mass is displaced to the near category than the far category.

This counterintuitive result can be explained by inspecting precision and recall's support estimation step. Given samples drawn from the distribution of interest, they estimate its support as the union of k-nearest-neighbor balls centered at the samples, where the radius of each ball is given by the distance from the sample to its kth-nearest-neighbor. Thus, sparse regions of the support tend to be approximated by fewer but larger k-NN balls, while dense regions are approximated by many smaller k-NN balls. In the given setting where mass is shifted away from category A, recall is driven by the proportion of the reference points in A that are covered by the evaluation dataset's estimated support. This method shrinks the radii of the k-NN balls for the near

alternative in the region spanned by A and B, compared to the radii for the far alternative in the same region. As a consequence, the near alternative's support fails to reach as many of the reference points in A compared to far, resulting in its worse recall score. Tellingly, fixing the radii across both near and far alternatives inverts the trend, yielding higher recall for the near alternative than far.



(a) Illustration of experimental setting used in Fig. 9b. Reference is composed of a uniform $(1/3, 1/3, 1/3)$ mixture over categories (A, B, C). Near is composed of a $(1/3 - \alpha, 1/3 + \alpha, 1/3)$ mixture, while far is composed of a $(1/3 - \alpha, 1/3, 1/3 + \alpha)$ mixture. Dashed lines in the near and far panels correspond to the outlines of the reference mixture.



(b) Average fidelity-diversity and precision-recall scores of near and far computed against reference, over α ranging from $[0, 0.3]$. Scores are averaged over 500 choices of the reference-evaluation partition, and shaded regions with hairline borders indicate 95% confidence intervals on the mean over all 500 partitions. Dashed lines correspond to scores computed on $\alpha = 0$, at which near and far coincide.

Figure 9: Comparison between fidelity-diversity and precision-recall metrics on a combination of data drawn from two categories from M2D2 S2ORC and one from M2D2 WIKI. The two categories from S2ORC are A = Analysis of PDEs (*math.AP*) and B = Classical Analysis & ODEs (*math.CA*). The far category is C = Sports & Recreation (*Culture_and_the_arts__Sports_and_Recreation*) from M2D2 WIKIPEDIA.

We reproduce this result in two more settings, one with categories drawn exclusively from M2D2 S2ORC and the other where we select categories exclusively from M2D2 WIKIPEDIA (See Fig. 10). More generally, precision-recall metrics consider only local nearest-neighbor distances, which makes them less discriminative compared to our OT-based diversity metric on data in which significant distance discrepancies exist on a global scale.

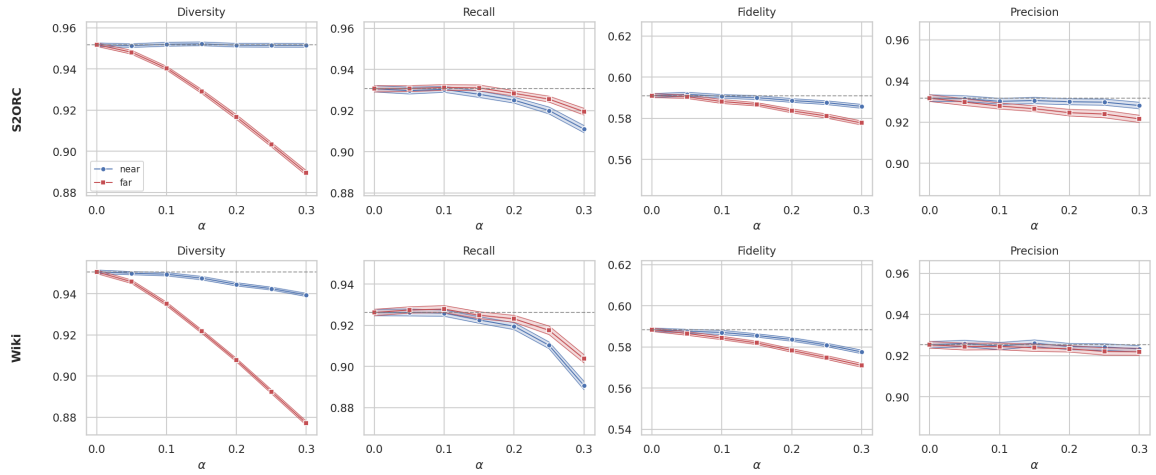


Figure 10: Top row compares fidelity-diversity and precision-recall metrics on data drawn exclusively from M2D2 S2ORC; bottom row uses data from M2D2 WIKI. The three categories from S2ORC are A = Condensed Matter: Quantum Gases (*cond-mat.quant-gas*), B = Condensed Matter: Other (*cond-mat.other*), and C = Art (*Art*). The three categories from WIKI are A = Nature (*Natural_and_physical_sciences__Nature*), B = Biology (*Natural_and_physical_sciences__Biology*), and C = Sports & Recreation (*Culture_and_the_-arts__Sports_and_Recreation*). Scores are averaged over 100 choices of the reference-evaluation partition, and shaded regions with hairline borders indicate 95% confidence intervals on the mean over all 100 partitions. Dashed lines again correspond to scores computed on $\alpha = 0$, at which near and far coincide.

D ALTERNATIVE EMBEDDINGS

The effectiveness of our metrics depends on the quality of the embeddings used. We evaluate our metrics on two additional embeddings commonly used for computing MAUVE (Pillutla et al., 2023; Sahoo et al., 2026): GPT-2-large (Radford et al., 2019) and ModernBERT-large (Warner et al., 2024). ModernBERT is a bidirectional encoder-only model, while GPT-2 is a causal decoder-only model. Table 3 compares fidelity-diversity scores computed in the M2D2 WIKIPEDIA and S2ORC nested setups from Sec. 4.1 (using Collection 0, see App. A.3) and the GSM8K from Sec. 4.2, as well as fidelity scores in the distribution shift setups from App. A.1 and swapped GSM8K setup from App. B.3.

By design, we expect to isolate a significant effect on diversity in the nested and GSM8K settings, and to observe a significant effect on fidelity in the shifted and swapped settings. The metrics computed on the GPT-2 and ModernBERT embeddings consistently recover the same trends as those computed on E5, illustrating the robustness of our method to different embeddings.

E PROOF OF EQUATION (3)

In this Appendix, we provide a brief proof of the identity (3), which reduces to showing

$$\inf_{\mu \in \mathcal{P}(H)} W_1(G, \mu) = \sum_{i=1}^{k} \alpha_i \cdot \min_{1 \le j \le k} \|\theta_i - \eta_j\|. \tag{5}$$

Setting		$E5_{small}$ (dim=384)	GPT-2-large (dim=1280)	ModernBERT-large (dim=1024)
Nested WIKI	Fid.	-0.0021 (0.456)	-0.0073 (0.121)	-0.0045 (0.131)
	Div.	**+0.0246** (< 0.001)	**+0.0143** (0.002)	**+0.0081** (0.019)
Shifted WIKI	Fid.	**-0.0093** (< 0.001)	**-0.0033** (0.001)	**-0.0032** (< 0.001)
Nested S2ORC	Fid.	+0.0011 (0.586)	+0.0071 (0.245)	+0.0028 (0.085)
	Div.	**+0.0203** (0.001)	**+0.0089** (0.006)	**+0.0255** (0.035)
Shifted S2ORC	Fid.	**-0.0096** (< 0.001)	**-0.0083** (0.001)	**-0.0063** (0.001)
GSM8K	Fid.	-0.670 (0.099)	-0.767 (0.054)	-0.230 (0.620)
	Div.	**+0.973** (< 0.001)	**+0.987** (< 0.001)	**+0.882** (0.009)
Swapped GSM8K	Fid.	**+0.992** (< 0.001)	**+0.995** (< 0.001)	**+0.954** (0.001)

Table 3: Comparison of fidelity and diversity scores computed from $E5_{small}$, GPT-2-large, and ModernBERT-large embeddings. OLS coefficients and corresponding p-values in the nested (Sec. 4.1) and distribution shift (App. A.1) setups are reported as $\{\beta\ (p)\}$. Pearson correlation coefficients and corresponding p-values computed in the GSM8K (Sec. 4.2) and swapped GSM8K (App. B.3) setups are reported as $\{r\ (p)\}$.

The left-hand side can be expressed as

$$\inf_{\mu \in \mathcal{P}(H)} W_1(G, \mu) = \inf_{\gamma \in \Pi(G)} \sum_{i,j=1}^{k} \gamma_{ij} \|\theta_i - \eta_j\|,$$

where $\Pi(G)$ denotes the set of all stochastic matrices $\gamma \in \mathbb{R}_+^{k \times k}$ such that $\sum_{j=1}^{k} \gamma_{ij} = \alpha_i$ for each $i = 1, \ldots, k$. Unlike the set of couplings $\Pi(G, H)$, notice that $\Pi(G)$ places no constraint on the second marginal of γ. Now, for any $\gamma \in \Pi(G)$, we have

$$\sum_{i,j=1}^{k} \gamma_{ij} \|\theta_i - \eta_j\| \geq \sum_{i,j=1}^{k} \gamma_{ij} \min_{1 \leq j' \leq k} \|\theta_i - \eta_{j'}\| = \sum_{i=1}^{k} \alpha_i \cdot \min_{1 \leq j \leq k} \|\theta_i - \eta_j\|.$$

Taking the infimum over all $\gamma \in \Pi(G)$ leads to

$$\inf_{\mu \in \mathcal{P}(H)} W_1(G, \mu) \geq \sum_{i=1}^{k} \alpha_i \cdot \min_{1 \leq j \leq k} \|\theta_i - \eta_j\|.$$

The lower bound is achieved by the coupling

$$\gamma_{ij} = \alpha_i \cdot I\Big(j = \operatorname*{argmin}_{1 \leq j' \leq k} \|\theta_i - \eta_{j'}\|\Big), \quad i, j = 1, \ldots, k,$$

breaking ties arbitrarily. The claim thus follows.