Beyond Scale: the Diversity Coefficient as a Data Quality Metric Demonstrates LLMs are Pre-trained on Formally Diverse Data

Alycia Lee [*1] **Brando Miranda** [*1] **Sanmi Koyeyo** [1]

Abstract

Current trends to pre-train capable Large Language Models (LLMs) mostly focus on scaling of model and dataset size. However, the *quality* of pre-training data is an important factor for training powerful LLMs, yet it is a nebulous concept that has not been fully characterized. Therefore, we use the recently proposed Task2Vec diversity coefficient to ground and understand formal aspects of data quality, to go beyond scale alone. Specifically, we measure the diversity coefficient of publicly available pre-training datasets to demonstrate that their formal diversity is high when compared to theoretical lower and upper bounds. In addition, to build confidence in the diversity coefficient, we conduct interpretability experiments and find that the coefficient aligns with intuitive properties of diversity, e.g., it increases as the number of latent concepts increases. We conclude the diversity coefficient is reliable, show it's high for publicly available LLM datasets, and conjecture it can be used to build useful diverse datasets for LLMs.

1. Introduction

Current trends in pre-training Large Language Models (LLMs) tend to concentrate on model and dataset size scaling (Chowdhery et al., 2022; Nostalgebraist, 2022; OpenAI, 2023; Google, 2023). Therefore, vast amounts of effort have been invested in understanding neural scaling laws – the power-law relationship between the loss of deep artificial networks and the *size* of the pre-training dataset and model for a fixed compute budget (Hestness et al., 2017; Rosenfeld et al., 2019; Henighan et al., 2020; Kaplan et al., 2020; Gordon et al., 2021; Hernandez et al., 2021; Jones, 2021; Zhai et al., 2022; Hoffmann et al., 2022; Clark et al., 2022; Neumann & Gros, 2022). In addition, recent work focuses on training a fixed model but using *more* tokens (Touvron et al., 2023). However, the effectiveness of these systems also fundamentally relies on the quality (Longpre et al., 2023) and coverage of the pre-training data (Hashimoto, 2021; David et al., 2010) and not only the *size*. Unfortunately, data quality and coverage (David et al., 2010) are often overlooked or discussed in vague and imprecise ways (Longpre et al., 2023). Hence, we propose to ground the discussion of data quality through the diversity coefficient (Miranda et al., 2022a), a data coverage metric that moves beyond scale alone. We extend the diversity coefficient to formally quantify data diversity of publicly available datasets and discover that LLMs are pre-trained on formally diverse data. We demonstrate the diversity coefficient is *high* for these pre-training datasets by comparing their formal diversity to the non-vacuous conceptually well-motivated lower and upper bounds of the diversity coefficient. In addition, to instill confidence in the usage of the diversity coefficient, we assess the interpretability of the coefficient as it relates to intuitive and expected properties of such a diversity metric. Concretely, we demonstrate:

1. The diversity coefficient increases as one concatenates more pre-training datasets of different sources.
2. We show the task embedding distances used in the diversity coefficient groups in a meaningful way, reflecting the conceptual and semantic information humans expect.
3. Using the Generative IN-Context Learning (GINC) (Xie et al., 2021) dataset, we show that as the number of latent concepts[1] increases the diversity coefficient increases.
4. We show that a larger, more diverse vocabulary leads to a higher diversity coefficient in the Generative IN-Context Learning (GINC) (Xie et al., 2021) dataset.

Our key **contributions** are:

1. A paradigm shift beyond dataset scale to a data-centric

[*]Equal contribution [1]Department of Computer Science, Stanford University, Palo Alto, USA. Correspondence to: Alycia Lee <alylee15@stanford.edu>, Brando Miranda <brando9@stanford.edu>.

Preliminary work. Under review. Do not distribute.

[1]Latent concepts represent document-level features such as semantics, structure, and style (Xie et al., 2021).

machine learning perspective through a formal data quality metric – the diversity coefficient.

2. We advance discussions on data quality by measuring an aspect of quality – data diversity – using the diversity coefficient.

3. We further validate the diversity coefficient by demonstrating its interpretability and correlation with intuitive diversity properties aligned with human intuitions, e.g., the coefficient increases as more datasets are concatenated, the number of latent concepts increases, and a richer vocabulary is used.

4. We formally demonstrate the high diversity of public datasets for LLM pre-training is *high* using well-motivated lower and upper bounds.

5. Lastly, for ease of usage of our method, we also study properties of different parameters for computing the formal diversity and therefore provide practitioners with simpler ways to evaluate the diversity coefficient.

Therefore, we conclude the diversity coefficient is reliable, and conjecture the diversity coefficient can be used to build quality diverse datasets for capable LLMs. In doing so, we hope this work inspires more systematic and effective techniques for dataset design beyond simply increasing the number of data points, sequences, or tokens.

2. Methods

2.1. Task2Vec Embeddings for Sequence Data

We use the Task2Vec diversity coefficient (Miranda et al., 2022a) to compute the formal diversity of a dataset, The first step is to compute Task2Vec (vectorial) embeddings of a batch of sequences. The original Task2Vec method (Achille et al., 2019) embeds data (e.g. few-shot learning task) using the diagonal entries of the Fisher Information Matrix (FIM) that result from (partially) fine-tuning the final layer of a fixed neural network (also called a *probe network*) to solve the current task (or batch). We implement this framework by fine-tuning GPT-2 (Radford et al., 2019) to predict the next token for each sequence in the current batch B, then compute the FIM as follows:

$$\hat{F}_B = \mathbb{E}_{x,t,\hat{x}_t} \nabla_w \log \hat{p}_w(\hat{x}_t \mid x_{t-1:1}) \nabla_w \log \hat{p}_w(\hat{x}_t \mid x_{t-1:1})^\top$$

The Task2Vec embedding $\vec{f}_B$ is the diagonal ($Diag$) of the FIM:

$$\vec{f}_B = Diag(F_B)$$

where x is a sequence of length T_x sampled from a batch B i.e. $x \in B$, $\hat{x}$ is a sequence of tokens sampled from the fine-tune probe network f_w (with weights w) conditioned on the real sequence x i.e. $\hat{x} \sim \hat{p}_w(\hat{x}_t \mid x_{t-1:1})$, t indicates taking the average across the sequence length when computing the (log) loss.

To better understand the Task2Vec embedding, observe that the (diagonal) of the FIM can be interpreted as a measure of the information that a given parameter contains about the generative distribution $p_w(\hat{x}_t \mid x_{t-1:1})$. Therefore, it serves as a unique fingerprint, or feature vector, for a batch, which defines a task distribution. Empirical findings in (Achille et al., 2019) show that Task2Vec embeddings cluster in a way that reflects semantics between different visual concepts and that Task2Vec cosine distances are positively correlated with taxonomical distances.

2.2. Diversity Coefficient Computation for Natural Language Datasets

Using our extension of Task2Vec for sequence data, we explain how to compute the Task2Vec diversity coefficient (Miranda et al., 2022a) for natural language datasets using GPT-2 as a probe network. We compute the Task2Vec diversity coefficient as the expected cosine distance d between pairs of Task2Vec embeddings of batches:

$$\hat{d}\text{iv}(D) = \mathbb{E}_{B_1, B_2} d(\vec{f}_{B_1}, \vec{f}_{B_2})$$

where D is the natural language dataset from which we sample batches B_1, B_2, and $\vec{f}_{B_i}$ is the Task2Vec embedding of a batch B_i using the diagonal of the FIM matrix $\hat{F}_{B_i}$.

To compute Task2Vec embeddings, we use GPT-2 (Radford et al., 2019) pre-trained on the English language as the probe network f_w. Following Task2Vec, we fine-tune only the final layer (a language modeling head) on each batch. Figure 5 demonstrates our pipeline.

By measuring the distance between FIMs, the diversity coefficient captures the average intrinsic variability of batches in the underlying data distribution as a proxy for data coverage or information contained in the dataset. Another interpretation is that dataset diversity reflects how different batches are from each other. Therefore, a low diversity coefficient implies that batches are not very different.

2.3. Recipe for Establishing if a Diversity Coefficient is High via the Conceptual Lower and Upper Bounds

To establish if a diversity coefficient $\hat{div}(D)$ of a dataset D is high (or low), we use two conceptually well-motivated reference values. We call them the lower and upper bounds of the diversity coefficient. There, we explain the conceptually motivated lower and upper bounds of the diversity coefficient. Consider a dataset constructed by sampling with most of the probability mass concentrated on some arbitrary token. This is a good candidate for a dataset with minimum diversity. On the other extreme, a dataset constructed by sampling any token uniformly at random given a fixed vocabulary (in our case, the GPT-2 tokenizer vocabulary) is a good candidate to create a dataset with maximum diversity.

Therefore, we measure a conceptual lower bound on a dataset with a vocabulary size of 2: <eos> token and a randomly selected non-special token from the GPT-2 tokenizer vocabulary. The <eos> token was assigned a probability weight of $1/\{\text{GPT-2 vocab size}\}$. The non-special token was assigned the remaining weight. Similarly, a high or maximum diversity dataset would consist of random sequences of all possible tokens, with no underlying order to semantics, formatting, etc. The upper bound of the diversity coefficient was therefore measured on a synthetic dataset with an equal probability of occurrence assigned to all tokens in the GPT-2 tokenizer vocabulary.

2.4. LLM Pre-training Datasets

Since LLMs are often trained on internal, non-public datasets[2], we used publicly available language datasets from the same sources as LLM pre-training data:

C4, a 305GB cleaned version of Common Crawl's web crawl corpus in English (Raffel et al., 2019). Sequences in C4 were extracted from the web via de-duplication methods and heuristics to remove boiler-plate and gibberish.

WikiText-103, a 500MB collection of over 100 million tokens extracted from the set of verified Good and Featured articles on Wikipedia (Merity et al., 2016).

The Pile, a 825 GiB open-source English-text corpus for language modeling that combines 22 smaller, high-quality datasets from diverse sources (Gao et al., 2020). These sources include Pile-CC (Common Crawl), PubMed Abstracts, Books3, OpenWebText2, ArXiv, and GitHub.

For instance, GPT-3 was trained on a filtered Common Crawl dataset and Wikipedia (Brown et al., 2020), which are represented by C4 and WikiText-103. It was also trained on WebText2 and Books, which are sub-datasets of The Pile.

We also evaluate the diversity coefficient of the following five sub-datasets of The Pile:

Pile-CC, a 227 GiB preprocessed version of Common Crawl's web crawl corpus (Gao et al., 2020). While both Pile-CC and C4 are sourced from Common Crawl, Pile-CC was preprocessed from Web Archive files, which are raw HTTP responses and page HTML, whereas C4 was preprocessed from WET files, which consist of plaintext. Nonetheless, we expect that both datasets are non-mutually-exclusive.

HackerNews, a 4 GiB scraped and parsed dataset of comment trees from Hacker News, a social news website that aggregates article links (Gao et al., 2020). Articles are generally focused on topics in computer science and entrepreneurship.

[2]For instance, Gopher was trained on Google's internal dataset MassiveText.

Table 1. **Diversity coefficients of LLM pre-training datasets with 95% confidence intervals are 3-5 times higher than the conceptual lower bound and more than half that of the upper bound.**

Dataset	Diversity Coeff.
Lower Bound	$\mathbf{0.0525} \pm 3.41\text{E-}4$
NIH ExPorter	$0.15 \pm 3.218\text{E-}5$
USPTO	$0.1582 \pm 4.09\text{E-}5$
PubMed Abstracts	$0.168 \pm 2.63\text{E-}5$
HackerNews	$0.201 \pm 4.52\text{E-}5$
WikiText-103	$0.2140 \pm 7.93\text{E-}5$
C4	$0.2374 \pm 2.785\text{E-}5$
The Pile	$0.2463 \pm 3.034\text{E-}5$
Pile-CC	$0.2497 \pm 3.41\text{E-}5$
C4 and WikiText-103	$0.2711 \pm 3.22\text{E-}4$
Concatenation of five datasets	$\mathbf{0.2939} \pm 2.03\text{E-}4$
Upper Bound	$\mathbf{0.4037} \pm 1.932\text{E-}5$

NIH ExPorter, a 1.9 GiB dataset of NIH Grant abstracts for awarded applications from 1985-present hosted on the ExPORTER initiative (Gao et al., 2020).

PubMed Abstracts, a 19 GiB dataset of abstracts from 30 million publications in PubMed (Gao et al., 2020).

USPTO Backgrounds, a 23 GiB dataset of background sections from patents granted by the United States Patent and Trademark Office (USPTO) (Gao et al., 2020).

3. Experiments & Results

In this section, we describe the experiments and results supporting the contributions outlined in the introduction.

3.1. Diversity Coefficients of Pre-training Data shows LLMs are Pre-trained on Formally Highly Diverse Data

Experiments: We evaluate the diversity coefficient (described in section 2) of eight publicly available LLM pre-training datasets (described in section 2.4). We also compute the diversity coefficient of two concatenated datasets: 1) C4 and WikiText-103, and 2) five sub-datasets of The Pile: Pile-CC, HackerNews, NIH ExPorter, PubMed, and USPTO (section D.4). In addition, we compute our conceptually well-motivated lower and upper bounds on the diversity coefficient (section 2.3).

Results: Table 1 reports the measured diversity coefficients of eight publicly available LLM pre-training datasets, in addition to the conceptually well-motivated lower and upper bounds. Table 1 also reports the measured diversity coefficients of the concatenation of different publicly available datasets. The key observations from our results are:

- The diversity coefficients of pre-training datasets tend to be **3-5 times greater than the theoretical lower bound and, on average, half the upper bound.** Prominently, WikiText-103, C4, The Pile, and Pile-CC exhibit high diversity coefficients (0.21, 0.25).
- The measured diversity of Pile-CC is higher than that of C4, indicating a potentially more stringent preprocessing method applied to the Common Crawl corpus for Pile-CC, which contributes to enhanced data diversity.
- Three sub-datasets of The Pile, namely NIH ExPorter, PubMed Abstracts, and USPTO, show relatively low diversity (0.15-0.17), approximately half of the upper bound (0.4). The nature of these datasets, curated from specialized fields, may account for this observation. For instance, patent backgrounds in USPTO may share similar formatting and semantics as do abstracts in NIH ExPorter or PubMed Abstracts.
- However, we observe that Pile-CC and HackerNews have higher diversity, which may be attributed to their coverage of a broad range of topics. Among these, Pile-CC exhibits higher diversity, in line with its heterogeneous content composition.

3.2. Concatenation of Datasets of Different Sources Produces Higher Measured Diversity

Experiments: To show that the concatenation of different datasets produces high diversity datasets, we measure the diversity coefficient of C4 plus WikiText-103, as well as the diversity coefficient of the five sub-datasets of The Pile in Table 1. To understand the source of this increased diversity, we plot the Task2Vec (cosine) distances between batches from individual datasets and distances of batches from the different datasets. We report these distances in Figure 1.

Results: Our key observations are:

- The diversity coefficient for the C4 and WikiText-103 concatenated dataset is 0.2711, about +0.03-0.05 higher than that of each individual dataset.
- The diversity coefficient for the concatenation of the five sub-datasets of the Pile is 0.2939 (Table 1), which is about +0.04-0.1 (Figure 1) that of each individual dataset.
- The concatenation of the five sub-datasets of The Pile achieves the highest diversity coefficient in Table 1.

This increase in diversity occurs because concatenating datasets produces higher pairwise Task2Vec distances between batches from different datasets (see Figure 1). This results in a higher diversity coefficient, since the coefficient is an average of all pairwise Task2Vec distances. Note that, this aligns with human intuition that combining data from heterogeneous sources increases the overall diversity of the data.

3.3. Distribution of Pairwise Batch Distances Reflects Conceptual and Semantic Dataset Information

To increase our confidence in the diversity coefficient as a diversity metric, we study distributions of the Task2Vec (cosine) distances used to compute the coefficient. In particular, we examine the alignment of the grouping of these distances with (human) conceptual and semantic understanding.

Experiments: Therefore, we analyze Task2Vec (cosine) distances between batches from five sub-datasets of The Pile. In particular, we compare distances between batches of individual sub-datasets and distances across different sub-datasets. We show the resulting histograms and violin plots in Figure 1. We also segment these distances between batches across C4 and WikiText-103 in Figure 1.

Results: Our key observations are:

- Figure 1 (top, left) shows 3 modes. We confirm that the modes correspond to pairings of datasets in Figure 1 (top, right). For instance, the right-most mode, corresponding to distances with values higher than the diversity coefficient, consists of pairwise distances between C4 and WikiText-103 batches. This confirms intuitive properties we'd expect, i.e. we'd expect 3 modes given 2 datasets ($C_2^2 + 2 = 3$).
- Similarly to the preceding point, Figure 1 (bottom, left) shows 15 modes, which is exactly the number expected in enumerating all possible pairs of batches from 5 datasets.[3] Due to overlaps in distance values we only see 11 modes in the Figure 1 (bottom, right).
- We also observe that the combined datasets have an increased diversity coefficient compared to the individual data sets. We outlined this in the previous section, but we underscore it here to emphasize this semantic property.
- We expect pairings of unrelated datasets to have higher diversity compared to pairings of related datasets. We observe this in Figure 1 (right). For the concatenated dataset of C4 and WikiText-103, the distribution of pairwise distances where one batch is from C4 and one is from WikiText-103 (right-most violin) is higher than that of individual datasets. For the concatenated sub-datasets of The Pile, the violin plots for combinations of conceptually unrelated datasets group above the dotted line (e.g. Hacker News and PubMed), while the violin plots of technical subjects written in a similar style[4] are below the dotted line (e.g. PubMed and USPTO). Note however that all combined diversities always increased after a concatenation.
- We expect Pile-CC and HackerNews to cover the most

[3] Given a 5 by 5 distance matrix, we'd expect the lower triangular portion plus the diagonal to be the number of pairings, so $C_2^5 + 5 = 15$.

[4] e.g. NIH ExPorter and PubMed Abstracts both contain medical abstracts, and have the lowest distances (third violin from the right) among combinations of different datasets.

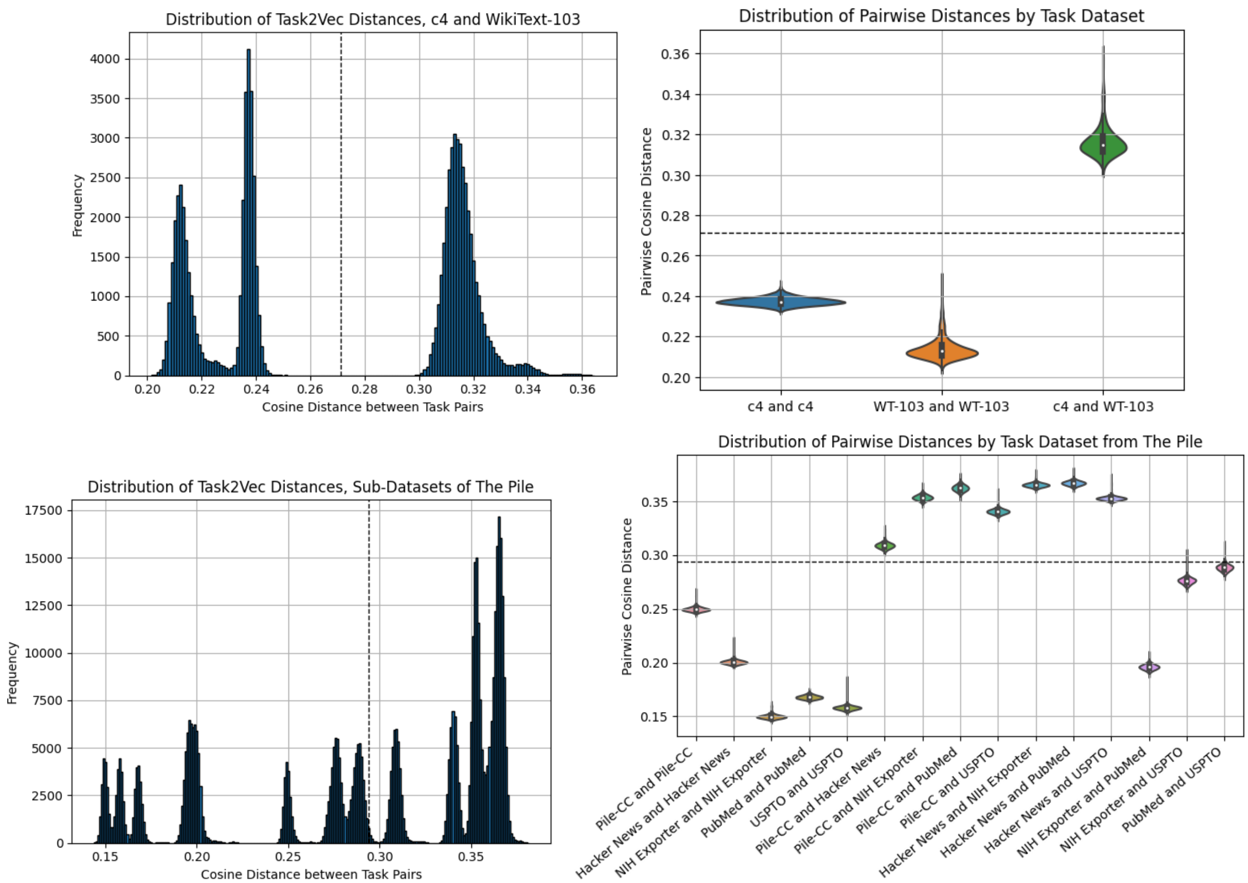

Figure 1. **Distribution of pairwise batch distances reflect conceptual and semantic dataset properties, therefore increasing trust in the diversity coefficient.** Pairwise task distances from concatenated C4 and WikiText-103 dataset (top) and concatenated five sub-datasets of The Pile (bottom) take on a multi-modal form according to dataset comparisons. Pairwise distances are segmented by source datasets for each pair of batches (right), where each sub-distribution corresponds to a mode from the histograms (left). Dotted lines denote the diversity coefficient of the concatenated C4 and WikiText-103 dataset (top) and concatenation of five sub-datasets of The Pile (bottom). These results show that combining batches from two different datasets computes a higher diversity, as expected. Therefore, these results align with human intuition, increasing the confidence in the diversity coefficient as a diversity metric.

diverse topics since they are broad web-scale datasets, unlike the remaining which are technical in nature. Therefore, we anticipate 1) these two to have the highest individual diversities, as shown in the first two violin plots in Figure 1, and 2) to have the highest increase when combined with other datasets, as shown in the 6th to the 12th violin plots when counting from the left, in Figure 1.

- Distances between batches from Pile-CC and HackerNews (sixth violin from the left) are the lowest among pairwise distances of concatenated datasets above the diversity coefficient. This aligns with human conceptual intuition because the Pile-CC and HackerNews are the most similar in those sub-datasets, since they are both web-scale datasets.

These findings build trust in the diversity coefficient as a dataset diversity metric, since the coefficient and underlying Task2Vec distances of batches behave in interpretable ways that align with human intuition.

3.4. Diversity Coefficient Captures LLM Pre-training Data Distributional Properties

To instill further confidence in the diversity coefficient, we perform a correlation analysis with data distributional properties on a synthetic language dataset. We use the GINC dataset (Xie et al., 2021) method, which generates sequences by modeling how real documents are generated given a fixed number of latent document concepts. It achieves this through a mixture of Hidden Markov Models (HMM) where each HHM has a latent concept that models document statistics, e.g. wiki bio. Further details on GINC can be found in section F.

Experiments: Given that each GINC dataset is a mixture of HMMs with a fixed number of latent concepts (1-10,000), we plot how the diversity coefficient varies as the number of latent concepts increases for each dataset. We plot this in Figure 2 (top) and fit a curve for GINC datasets with fixed vocabulary sizes of 50 and 150. Then we fix the number of latent concepts at 5 and 5000 and similarly plot how increasing the vocabulary size for the GINC dataset (50-10,000 unique tokens) increases the diversity coefficient. We plot this in Figure 2 (bottom) and fit a curve for GINC datasets with 5 latent concepts and 5000 latent concepts.

Results: Our observations are as follows:

- **Diversity coefficient increases with greater number of latent concepts.** Figure 2 (top) shows adding more latent concepts increases the diversity coefficient with diminishing returns. We hypothesize that additional latent concepts introduce new and varied document-level statistics, resulting in an increase in the diversity coefficient. The R^2 is high with values 0.952 and 0.898.
- The diversity coefficient saturates as more latent concepts are added. We hypothesize this may be due to marginal increases in variation from increased overlap, e.g. wiki bios and autobiographical web pages may have syntactical and semantic similarities.
- **Diversity coefficient increases with larger vocabularies.** Figure 2 (bottom) shows the measured diversity coefficient increases at a seemingly exponential pace for larger vocab sizes. The R^2 is high with values 0.993 and 0.984.
- We hypothesize the growth might be exponential because scaling the number of tokens produces a more diverse dataset by vastly increasing the number of ways to represent any sequence. More formally, given a sequence x of length T_x and vocab size $|V|$, the number of ways to represent that sequence is approximately $|V|^{T_x}$. Therefore, as $|V|$ increases, the growth rate of the exponential increases.

These results show the diversity coefficient successfully captures different distributional sources of variation of the data.

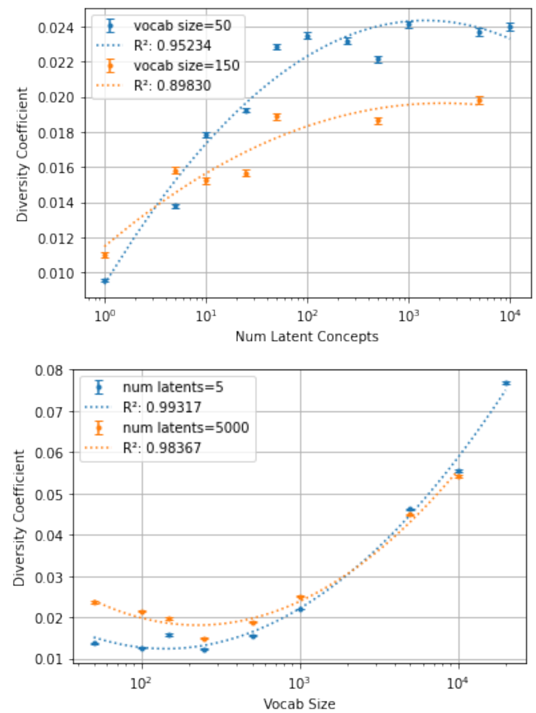

Figure 2. **Diversity coefficient of GINC datasets with varying number of latent concepts and vocab sizes shows the diversity coefficient behaves as expected.** The diversity coefficient increases and saturates with an increasing number of latent concepts (top) and exponentially increases with increasing vocab size (bottom). This implies that increases in the measured diversity coefficient correspond to changes in LM pre-training data distributional properties that intuitively enable more diverse data.

4. Using the Diversity Coefficient in Practice: Setting Batch Size and Network Parameters

Experiments: We test the sensitivity of the computed diversity coefficient value to changes in batch size and probe network parameters in order to gauge how these parameters should be set in practice for natural language datasets.

We vary the batch size and observe the impact on the diversity coefficient. For the same number of batches (200) and probe network (pretrained, fine-tuned GPT-2), we computed the diversity coefficient of C4 for batch sizes of 128, 256, 512, and 1024, and plot the results in Figure 3 (left).

We test the following probe network configurations to measure the diversity coefficient of C4 and of WikiText-103: 1. Pretrained GPT-2 with fine-tuning, 2. Pretrained GPT-2 without fine-tuning, 3. Randomly initialized GPT-2 with fine-tuning, 4. Randomly initialized GPT-2 without fine-tuning. Since using a random and/or non fine-tuned network is more resource efficient and easily accessible in practice, our motivation is to assess the necessity of using pre-trained and fine-tuned probe network, which is the original configuration used for Task2Vec in (Achille et al., 2019). We aim to determine if a good approximation of diversity can be computed without fine-tuning. We plot the diversity of coefficients measured using each of the four probe network configurations in Figure 3 (right).

Results: We observe that

- **Diversity coefficient increases with task batch size, but with diminishing returns.** Figure 3 (left) shows positive correlation between the diversity coefficient and batch size. T his may be because larger batch sizes enable more unique tokens per batch, which may result in higher distances between batches.
- However, we observe diminishing returns to the increase in diversity coefficient with increasing batch size. We hypothesize that as the batch size continues to increase,

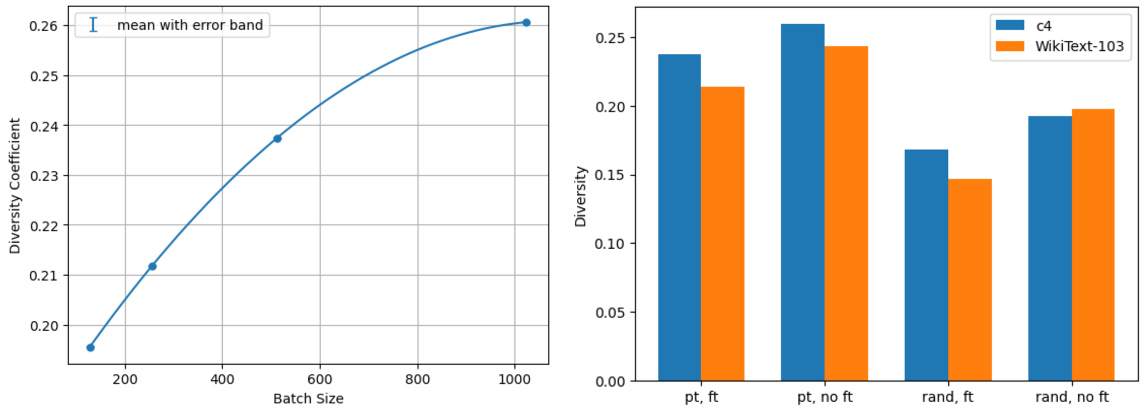

Figure 3. **Diversity coefficients of C4 computed using different task batch sizes show positive and diminishing returns with increasing batch size (left). Diversity coefficients of C4 and WikiText-103 computed using different GPT-2 probe network configurations show that random networks underestimate diversity vs. pretrained networks, and non-finetuned networks overestimate diversity vs. finetuned networks (right).** 95% confidence intervals for diversity coefficients are plotted, but are so small that they do not show. ”pt” refers to pretrained network and ”rand” refers to randomly initialized network. ”ft” refers to a network that was finetuned per task and ”no ft” refers to no finetuning performed.

there is greater coverage in tokens, topics, document formats, etc. between batches, so the increase in the diversity coefficient saturates.

- **Using a random probe network underestimates diversity.** Since the Task2Vec method (Achille et al., 2019) uses a pretrained and fine-tuned network, we consider the diversity computed using this configuration as a source of truth. Figure 3 (left) shows that using random probe networks underestimates diversity compared to pretrained networks, which is in accordance with results from (Miranda et al., 2022b) on vision datasets. We hypothesize that for random networks, the probe network parameters are not as calibrated to performing autoregressive language modeling, so batch representations from model parameters are similar, and the diversity is underestimated compared to pretrained networks.
- **Using a non fine-tuned network overestimates diversity.** Fine-tuning ensures the final Task2Vec embedding is more faithful to the dataset in question, as it adjusts the batch/task representation to a more similar distribution. This is due to batches – while different content-wise – being conditioned on the same dataset. On the other hand, a non-fine-tuned network may have more variable representations across batches, as it is not well-adapted to the dataset. This may explain the overestimation of the diversity coefficient that we observe.
- Trends in diversity coefficient overestimation vs. underestimation for different probe network configurations are consistent across C4 and WikiText-103.

Based on these findings, we recommend using a batch size of 512 sequences for faster computations and fewer out of memory issues. Our proposed diversity coefficient can be computed more efficiently using random and non fine-tuned networks, as eliminating pre-training and fine-tuning saves computational costs. While the absolute diversity coefficient values differ compared to values computed using a pre-trained and fine-tuned network, this is not a serious issue as long as the same network configuration is used consistently (see section G).

5. Related Work

Existing diversity metrics have concentrated on data produced by General Adversarial Networks (GANs) and involve variations of a precision- and recall-based framework originally proposed in (Sajjadi et al., 2018) to measure quality and diversity, respectively (Kynkäänniemi et al., 2019; Simon et al., 2019; Naeem et al., 2020). Similar to the Task2Vec diversity coefficient, these methods utilize embedding functions, These methods argue that data quality is not synonymous with data diversity in the context of GANs (Fowl et al., 2020) and hence take a two-metric approach. In the context of LLMs, we argue that data diversity is a subset of data quality, which is demonstrably important to enable capabilities not explicitly trained for such as in-context learning. Therefore, a diversity metric is sufficient to capture an important aspect of data quality. In addition, a diverse enough dataset increases the coverage and likelihood that a task in the test dataset is covered. Furthermore, large LLMs are robust to noise and therefore even if the diversity is made high, the models might still generalize. Therefore, we conjecture that high diversity is preferred and provide evidence that current datasets for open LLMs do have that property.

A recently proposed diversity metric that does not rely on an embedding function is the Vendi Score (Friedman & Dieng, 2022). The Vendi Score is given by the exponential of the Shannon entropy of the eigenvalues of a similarity matrix or kernel. However, the benefits of this more sophisticated aggregation method are not clear, and its computation $O(n^3)$ is more expensive than the diversity coefficient $O(n^2)$, as it requires eigenvalue decomposition. Moreover, the Vendi Score assumes the availability of a suitable similarity function (or kernel) for the data, and thus does not provide guidance on data representation – which is arguably the most challenging and important ingredient in machine learning. Furthermore, they suggest that utilizing data representational methods such as embedding networks that require pretrained models may be limiting. We argue instead that data representation is a fundamental property of data processing that has led to the overwhelming success in machine learning due to deep learning, e.g. in computer vision (Krizhevsky et al., 2012; He et al., 2015), natural language processing (Devlin et al., 2018; Brown et al., 2020; Chowdhery et al., 2022; OpenAI, 2023; Google, 2023), game playing (Silver et al., 2016; Mnih et al., 2013; Ye et al., 2021), theorem proving (Rabe et al.; Polu & Sutskever, 2020; Han et al.), code (Chen et al.) and more. Given the success of deep learning data representations and our work, we demonstrate deep learning is a strong way to create dataset/task embeddings. In contrast to the Vendi Score, our approach learns effective embeddings of tasks, batches, and datasets in an end-to-end manner, whereas the Vendi Score is focused on measuring diversity between specific data points. Since many canonical datasets already exist and are publicly available (e.g. Common Crawl, Wikipedia), data used to train new models may be curated from such datasets, necessitating a metric that captures overall dataset diversity. These scenarios are thus in favor of using the Task2Vec diversity coefficient. Therefore, our method is more general, flexible, and scalable than the Vendi Score. We leave a detailed comparison with the Vendi Score as future work.

6. Discussion

Our work extends, examines, and thus validates the application of the Task2Vec diversity coefficient to a new modality – natural language data – and demonstrates that open LLMs are pre-trained on formally diverse data. Our approach has a number of advantages. Through an extensive set of experiments that verifies intuitive properties of a diversity metric, we instill confidence in the diversity coefficient method, and therefore effectively concretize/ground the concept of data diversity. Our conceptually well-motivated lower and upper bounds on the diversity coefficient aid in the understanding of the magnitude of the diversity coefficient. However, the bounds we propose only apply to sequence data with a symbolic vocabulary. Using a multi-modal embedding method that embeds our proposed lower & upper bounds across modalities would solve this limitation by providing aligned comparable embedding distances. Another benefit is that our method does not rely on activations from an arbitrarily selected layer in a network. Lastly, note that activations may be unreliable for embedding dataset/tasks because large distances between datasets/tasks may be due to well-separated decision boundaries instead of intrinsic semantic properties of the dataset/task. In contrast, the diversity coefficient is well-justified, extensively tested in our work and previous work, e.g. the diversity coefficient correlates with ground truth diversities, cluster according to semantics, taxonomy etc. (see section B and (Achille et al., 2019; Miranda et al., 2022a)). In short, FIM-based representations are motivated by information theory (e.g. FIMs are metrics in distributions) and have been extensively tested by independent sources (Miranda et al., 2022a; Achille et al., 2019; Vu et al., 2020).

One potential limitation of our method is the need for a data representation. Although the requirement for a data representation might seem restrictive, we argue that it is an inherent aspect of data processing. Choosing symbols or raw pixels (or anything else) **is** a choice of data representation. We suggest deep learning representations due to their overwhelming success in machine learning, e.g. in computer vision (Krizhevsky et al., 2012; He et al., 2015), natural language processing (Devlin et al., 2018; Brown et al., 2020; Chowdhery et al., 2022; OpenAI, 2023; Google, 2023), game playing (Silver et al., 2016; Mnih et al., 2013; Ye et al., 2021), theorem proving (Rabe et al.; Polu & Sutskever, 2020; Han et al.), code (Chen et al.) and more. In addition, widely available open-source pre-trained models (e.g. CLIP (Radford et al., 2021), LLaMA (Touvron et al., 2023), etc.) has made choosing a good embedding method easier. In addition, we explore random networks and models with no fine-tuning, to make our method more accessible 4. We hypothesize that as long a consistent model/method is used to create the task embeddings, the exact model/method might not play a crucial role – because we only need comparable distances that depend on the data/task.

Data has taken a central role in the success of modern machine learning methods – like GPT4 (OpenAI, 2023), CLIP (Radford et al., 2021), and PaLM 2 (Google, 2023). This seems especially relevant for architectures with few inductive biases, like the popular Transformer (Vaswani et al., 2017). Therefore, it has become paramount to understand the pre-training data we use beyond scale alone. We conclude the diversity coefficient is a reliable trustworthy metric, and conjecture the diversity coefficient can be used to build quality diverse datasets for capable LLMs. We hope our contributions inspire more effective and quantitative data collection and curation processes in machine learning that go beyond scale alone, yet improve performance.

References

Achille, A., Lam, M., Tewari, R., Ravichandran, A., Maji, S., Fowlkes, C. C., Soatto, S., and Perona, P. Task2vec: Task embedding for meta-learning. *CoRR*, abs/1902.03545, 2019. URL http://arxiv.org/abs/1902.03545.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. *CoRR*, abs/2005.14165, 2020. URL https://arxiv.org/abs/2005.14165.

Chen, M., Tworek, J., Jun, H., Yuan, Q., Ponde de Oliveira Pinto, H., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., Ray, A., Puri, R., Krueger, G., Petrov, M., Khlaaf, H., Sastry, G., Mishkin, P., Chan, B., Gray, S., Ryder, N., Pavlov, M., Power, A., Kaiser, L., Bavarian, M., Winter, C., Tillet, P., Petroski Such, F., Cummings, D., Plappert, M., Chantzis, F., Barnes, E., Herbert-Voss, A., Hebgen Guss, W., Nichol, A., Paino, A., Tezak, N., Tang, J., Babuschkin, I., Balaji, S., Jain, S., Saunders, W., Hesse, C., Carr, A. N., Leike, J., Achiam, J., Misra, V., Morikawa, E., Radford, A., Knight, M., Brundage, M., Murati, M., Mayer, K., Welinder, P., McGrew, B., Amodei, D., McCandlish, S., Sutskever, I., and Zaremba, W. Evaluating Large Language Models Trained on Code. URL https://www.github.com/openai/human-eval.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

Clark, A., De Las Casas, D., Guy, A., Mensch, A., Paganini, M., Hoffmann, J., Damoc, B., Hechtman, B., Cai, T., Borgeaud, S., et al. Unified scaling laws for routed language models. In *International Conference on Machine Learning*, pp. 4057–4086. PMLR, 2022.

David, S. B., Lu, T., Luu, T., and Pal, D. Impossibility theorems for domain adaptation. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 129–136, 2010. URL https://proceedings.mlr.press/v9/david10a.html.

Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 1:4171–4186, oct 2018. URL https://arxiv.org/abs/1810.04805v2.

Fowl, L., Goldblum, M., Gupta, A., Sharaf, A., and Goldstein, T. Random network distillation as a diversity metric for both image and text generation, 2020.

Friedman, D. and Dieng, A. B. The vendi score: A diversity evaluation metric for machine learning, 2022.

Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., Presser, S., and Leahy, C. The pile: An 800gb dataset of diverse text for language modeling, 2020.

Google. Palm 2 technical report. Technical report, 2023. URL https://ai.google/static/documents/palm2techreport.pdf.

Gordon, M. A., Duh, K., and Kaplan, J. Data and parameter scaling laws for neural machine translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5915–5922, 2021.

Han, J. M., Rute, J., Wu, Y., Ayers, E. W., and Polu, S. Proof artifact co-training for theorem proving with language models.

Hashimoto, T. Model performance scaling with multiple data sources. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4107–4116. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/hashimoto21a.html.

He, K., Zhang, X., Ren, S., and Sun, J. Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016-December:770–778, dec 2015. ISSN 10636919. doi: 10.1109/CVPR.2016.90. URL https://arxiv.org/abs/1512.03385v1.

Henighan, T., Kaplan, J., Katz, M., Chen, M., Hesse, C., Jackson, J., Jun, H., Brown, T. B., Dhariwal, P., Gray, S., et al. Scaling laws for autoregressive generative modeling. *arXiv preprint arXiv:2010.14701*, 2020.

Hernandez, D., Kaplan, J., Henighan, T., and McCandlish, S. Scaling laws for transfer. *arXiv preprint arXiv:2102.01293*, 2021.

Hestness, J., Narang, S., Ardalani, N., Diamos, G., Jun, H., Kianinejad, H., Patwary, M., Ali, M., Yang, Y., and Zhou, Y. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. d. L., Hendricks, L. A., Welbl, J., Clark, A., et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Jones, A. L. Scaling scaling laws with board games. *arXiv preprint arXiv:2104.03113*, 2021.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. 2012. URL http://code.google.com/p/cuda-convnet/.

Kynkäänniemi, T., Karras, T., Laine, S., Lehtinen, J., and Aila, T. Improved precision and recall metric for assessing generative models, 2019.

Longpre, S., Yauney, G., Reif, E., Lee, K., Roberts, A., Zoph, B., Zhou, D., Wei, J., Robinson, K., Mimno, D., and Ippolito, D. A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity. *arXiv preprint arXiv:2305.13169*, 2023. URL https://doi.org/10.48550/arXiv.2305.13169.

Merity, S., Xiong, C., Bradbury, J., and Socher, R. Pointer sentinel mixture models, 2016.

Miranda, B., Yu, P., Wang, Y.-X., and Koyejo, S. The Curse of Low Task Diversity: On the Failure of Transfer Learning to Outperform MAML and Their Empirical Equivalence. *arXiv*, 2022a. doi: 10.48550/arXiv.2208.01545. URL https://arxiv.org/abs/2208.01545.

Miranda, B., Yu, P., Wang, Y.-X., and Koyejo, S. The curse of low task diversity: On the failure of transfer learning to outperform maml and their empirical equivalence, 2022b. URL https://arxiv.org/abs/2208.01545.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. Playing Atari with Deep Reinforcement Learning. 2013.

Naeem, M. F., Oh, S. J., Uh, Y., Choi, Y., and Yoo, J. Reliable fidelity and diversity metrics for generative models. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7176–7185. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/naeem20a.html.

Neumann, O. and Gros, C. Scaling laws for a multi-agent reinforcement learning model. *arXiv preprint arXiv:2210.00849*, 2022.

Nostalgebraist. Chinchilla's wild implications. *AI Alignment Forum*, 2022.

OpenAI. Gpt-4 technical report. 2023.

Polu, S. and Sutskever, I. Generative Language Modeling for Automated Theorem Proving. sep 2020. URL http://arxiv.org/abs/2009.03393.

Rabe, M. N., Research, G., Lee, D., Bansal, K., and Szegedy, C. Mathematical Reasoning via Self-supervised Skip-tree Training. Technical report.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners. 2019.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning Transferable Visual Models From Natural Language Supervision. feb 2021. URL https://arxiv.org/abs/2103.00020v1.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv e-prints*, 2019.

Rosenfeld, J. S., Rosenfeld, A., Belinkov, Y., and Shavit, N. A constructive prediction of the generalization error across scales. *arXiv preprint arXiv:1909.12673*, 2019.

Sajjadi, M. S. M., Bachem, O., Lucic, M., Bousquet, O., and Gelly, S. Assessing generative models via precision and recall, 2018.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., and Hassabis, D. Mastering the game of Go with deep neural networks and tree search. *Nature 2016 529:7587*, 529(7587):484–489, jan 2016. ISSN 1476-4687. doi: 10.1038/nature16961. URL https://www.nature.com/articles/nature16961.

Simon, L., Webster, R., and Rabin, J. Revisiting precision and recall definition for generative model evaluation, 2019.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. Llama: Open and efficient foundation language models, 2023.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Vu, T., Wang, T., Munkhdalai, T., Sordoni, A., Trischler, A., Mattarella-Micke, A., Maji, S., and Iyyer, M. Exploring and predicting transferability across NLP tasks. *CoRR*, abs/2005.00770, 2020. URL https://arxiv.org/abs/2005.00770.

Xie, S. M., Raghunathan, A., Liang, P., and Ma, T. An explanation of in-context learning as implicit bayesian inference. *CoRR*, abs/2111.02080, 2021. URL https://arxiv.org/abs/2111.02080.

Ye, W., Liu, S., Kurutach, T., Abbeel, P., Gao, Y., University, T., Berkeley, U. C., Qi, S., and Institute, Z. Mastering Atari Games with Limited Data. oct 2021. URL https://arxiv.org/abs/2111.00210v1.

Zhai, X., Kolesnikov, A., Houlsby, N., and Beyer, L. Scaling vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12104–12113, 2022.

A. Future Work

Our future research will explore the potential of the Task2Vec distance function for pre-training dataset curation. Given that the objective of pre-training is to maximize downstream task performance, we define high-quality training data as data that facilitates the best achievable performance on such tasks. We anticipate that higher diversity in the dataset will increase the likelihood of achieving this objective. The rationale is that a higher data diversity implies a broader coverage of tasks or batches, thereby increasing the probability of training the model on tasks or data representations that are relevant to evaluation tasks. Our focus will be to leverage Task2Vec to assess the similarity between individual data points, batches, or datasets to a target task. This assessment will enable us to curate the training data by selectively removing tasks that resemble random, noisy, or irrelevant sequences, which may adversely affect downstream performance.

B. Task2Vec Diversity Coefficient Correlates with Ground Truth Diversity

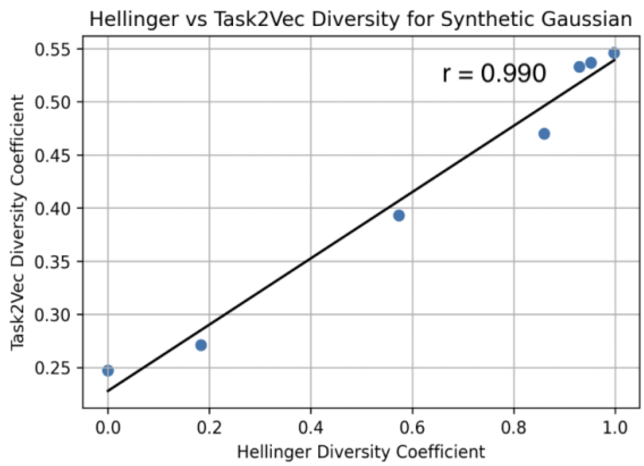

Figure 4. **Task2Vec diversity coefficient correlates with ground truth diversity for synthetic Gaussian benchmark.** Source: (Miranda et al., 2022b)

As shown in (Miranda et al., 2022b), when the ground truth diversity is available for a synthetic Gaussian benchmark, the Task2Vec diversity coefficient correlates with the ground truth diversity. These results provide confidence in the Task2Vec diversity coefficient as diversity metric.

C. Pipeline for Diversity Coefficient Computation of Natural Language Datasets

Figure 5 shows our pipeline for computing the diversity coefficient of large scale, natural language datasets. See section 2.2 for more details on our method.

D. Experimental Details

D.1. Dataset Preprocessing

In accordance with (Achille et al., 2019), we used the training split of datasets to finetune the probe network when computing Task2Vec embeddings per dataset. Sequences were tokenized using a pre-trained HuggingFace GPT-2 tokenizer based on byte-level Byte-Pair-Encoding, and padded or truncated to a max length of 128. Because the WikiText-103 dataset contained empty text examples, we removed these examples before sampling batches to compute embeddings.

D.2. Model Architecture and Finetuning

We used a pre-trained GPT-2 model with a language modeling (LM) head on top. The pre-trained GPT-2 model itself has 12 layers, 12 heads, 768-d hidden size, and 117M total parameters. The LM head is a linear layer with weights corresponding to the input embedding layers. The model was pre-trained on the English language and the pre-trained GPT-2 tokenizer

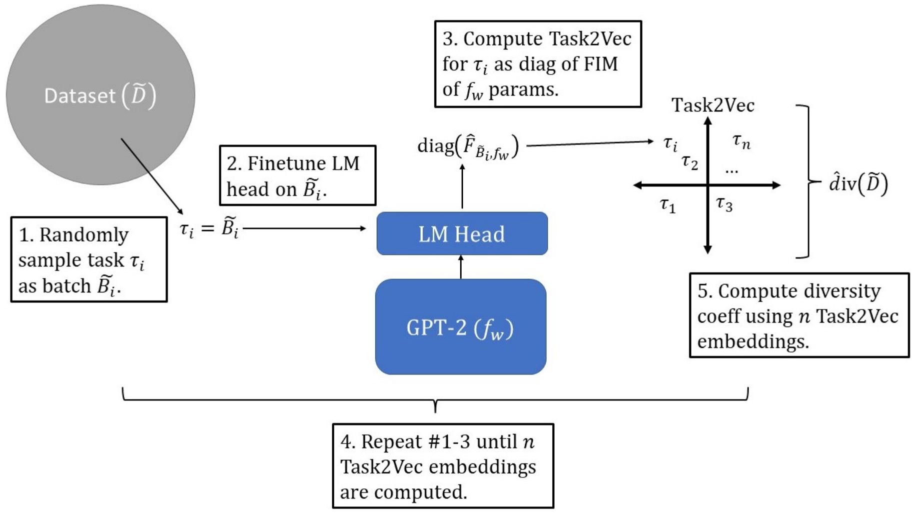

Figure 5. **A depiction of a pipeline to compute the Task2Vec diversity coefficient for a natural language dataset.**

has a vocab size of $\approx$ 50k tokens. For all finetuning experiments, we fine-tuned only the LM head for 10 epochs. We used no learning rate scheduler and no gradient accumulation. We used the AdamW optimizer, since AdamW has been shown empirically to give better training loss and improved generalization.

D.3. Number of Batches and Batch Size Selection

Diversity coefficients in Table 1 were computed using randomly selected batches of size 512 sequences and a pre-trained, finetuned GPT-2 probe network. Diversity coefficients of C4, WikiText-103, The Pile, Pile-CC, HackerNews, NIH ExPorter, PubMed Abstracts, and USPTO were each computed using 200 sampled batches. Given resource constraints, we found 200 batches[5] to be a sufficiently large number of batches to estimate the diversity coefficient with tight 95% confidence intervals on the order of 1e-5. We chose 512 as the batch size, since it is a relatively large and feasible batch size to fine-tune the probe network on 200 batches using Azure NV12s_v3 instances equipped with Tesla M60 GPUs in a reasonable amount of time (30+ hours).

D.4. Diversity Coefficient Computation of Concatenated Datasets

The diversity coefficient of a concatenated dataset of C4 and WikiText-103 was measured over a combined set of batches. Each batch consisted of sequences sampled from one of these datasets, e.g. a batch could have sequences randomly sampled from C4 or WikiText-103 but not both. The coefficient was computed over 400 batches of batch size 512 (200 batches from each dataset). Note that for the concatenated dataset, we utilized the same 200 batches per dataset that were used to compute the coefficients of C4 and of WikiText-103 individually.

The diversity coefficient of concatenated five sub-datasets of The Pile was computed over 1000 batches (200 batches from each dataset) of batch size 512. Similarly to the concatenated dataset of C4 and WikiText-103, we utilized the same 200 batches per dataset that were used to compute the coefficients of each individual sub-dataset.

D.5. Diversity Coefficient of The Pile vs. Concatenation of Five Sub-Datasets

We make a clarification on the approach taken to evaluate the diversity coefficient for The Pile vs. for concatenation of its five sub-datasets.

The diversity coefficient of The Pile was computed over 200 batches sampled across all 22 sub-datasets of The Pile. This means that any given batch could contain sequences across all 22 sub-datasets, i.e. a batch could have sequences from Pile-CC, HackerNews, and NIH ExPorter.

[5] This results in $(200^2 - 200)/2 = 19,900$ pairwise distances used to compute the diversity coefficient.

The diversity coefficient of the concatenated dataset was computed over 1000 batches comprised of 200 batches separately sampled from each of the five sub-datasets. Each batch contained sequences from only one sub-dataset, i.e. a batch could only have sequences from Pile-CC or HackerNews or NIH ExPorter.

We hypothesize this distinction in the diversity coefficient computation explains why the concatenated dataset has higher diversity, even though it consists of only five of the 22 sub-datasets of The Pile. For the diversity coefficient of The Pile, because batches were sampled such that any batch contains sequences from across the 22 sub-datasets, the batch representations learned by the probe network may have been more similar, resulting in lower diversity relative to the concatenated dataset.

E. Pairwise Distance Distributions of C4, WikiText-103, and The Pile

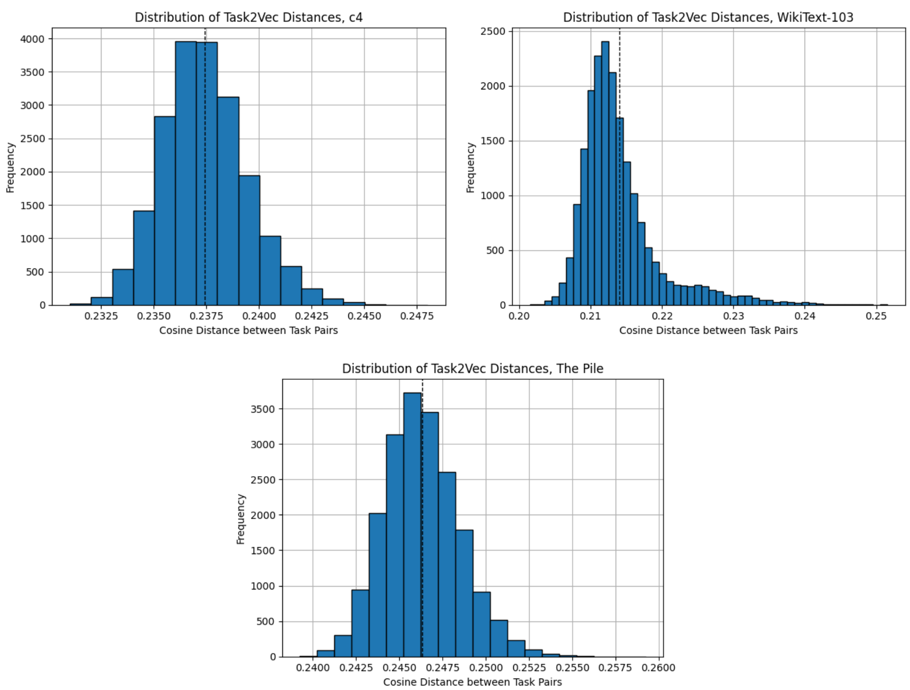

Figure 6. **Distributions of pairwise batch distances from C4 (top left), WikiText-103 (top right), and The Pile (bottom) are approximately Gaussian, which justifies the use of a sample of batches to measure the diversity coefficient.** Dotted lines indicate the average distance, i.e. the diversity coefficient, for each dataset.

Experiments: To provide confidence in the magnitude of the coefficient values of C4, WikiText-103, and The Pile, we plot the distribution of distances per dataset in Figure 6. We aim to show that a subsample of batches can provide a good estimation of population statistics, such as the diversity coefficient, which measures the expected Task2Vec (cosine) distance between batches.

Results: For each dataset, the pairwise distances take on unimodal and approximately Gaussian distributions with few outliers. These results suggest the Task2Vec distances are approximately normally distributed. This suggests we can make strong inferences about the population. Specifically, we are able to compute a good estimate of the diversity coefficient using 200 batches using the mean. This is in fact the same argument from (Miranda et al., 2022a) – but we verified it applied in our setting. Figure 6 also shows few outlier batches – the presence of which could influence the computed diversity coefficient. This provides further confidence in the coefficient values computed and justifies our use of a sample of batches to estimate diversity.

F. Generative IN-Context Learning (GINC) Dataset

F.1. Background

The GINC dataset is generated using the latent concept framework proposed in (Xie et al., 2021), where language models condition on a prompt to infer latent document concepts learned during pre-training. The pretraining distribution is defined using a uniform mixture of Hidden Markov Models (HMMs) parameterized over a family Θ of latent concepts.

F.2. Definitions of GINC Dataset Parameters

Number of latent concepts: A latent concept θ parameterizes the transitions of a HMM in the mixture. A latent concept (e.g. a wiki bio) contains document statistics, such as semantics, syntax, and the formatting of and distribution of tokens.

Vocabulary size: Each HMM in a given mixture outputs a fixed number of tokens, defined as the vocabulary size. The vocabulary is generated by enumerating combinations of letters from a to z, aa to az, etc. The delimiter token is designated by a backslash. Sequences are tokenized by whitespace.

F.3. Supplemental Figures for Diversity Coefficient vs. GINC Parameters

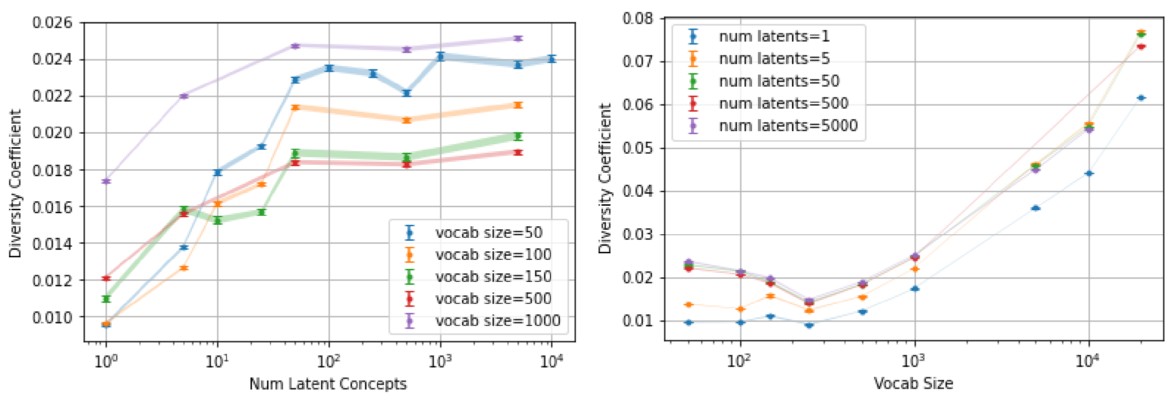

Figure 7. **Trends noted in Section 3.4 are consistent for diversity coefficient vs. number of latent concepts (left) and coefficient vs. vocab size (right) when the other parameter changes.** The diversity coefficient with 95% confidence intervals saturates with increasing number of latent concepts (left) even as vocab size is varied between 50-1000. Larger vocab sizes generally produce higher diversity coefficients (right) even as the number of latent concepts is varied between 1-5000.

Figure 7 confirms that the trends between the diversity coefficient and number of latent concepts (left) hold even as vocab size is varied. Similarly, trends between the diversity coefficient and the vocabulary size (right) hold as the number of latent concepts is varied. These trends were noted in Section 3.4.

G. Discussion (cont.)

Our paper introduces a metric that leverages tunable parameters, such as the number of batches, batch size, probe network configuration (pre-trained vs. random, fine-tuned vs. not) and depth. While these elements influence the diversity coefficient's absolute value and necessitate the recalibration of lower and upper bounds (see sections D.3 and 4), a consistent choice of hyperparameters can mitigate these effects.

Intriguingly, our proposed diversity may not always correlate with model performance, as high diversity could simply be due to uniform noise. Nevertheless, we contend that a higher diversity, in the context of a sufficiently large model, likely indicates superior performance and data quality. Furthermore, our diversity metric is intentionally designed to be widely applicable, albeit concealing causal factors, rendering it an effective tool for ablation studies.

Despite our diversity metric's broader applicability, it may obscure certain causal factors. This limitation is intentional to enhance its practical usage – since causality is often difficult to infer and is out of scope. This can be overcome with data property ablation studies, as we showed in our GINC dataset experiments.

Currently, our proposed bounds are specific to sequence data with a symbolic vocabulary, limiting their applicability across different modalities. To overcome this limitation, we suggest using a multimodal embedding method for embedding diversity coefficients and lower/upper bounds across tasks.

To really clarify why FIM is better than activations, we provide this intuitive explanation. FIM gives a weight/feature of which parameter of the generative distribution matters, e.g. the first coordinate of Task2Vec corresponds to how artsy the text sequence is. This is a feature of a task or dataset itself. Therefore, FIM exactly approximate the (task) data generative distribution we are trying to embed. Therefore, we conjecture it results in superior representations for datasets compared to activations since it directly approximates the data (or task) generative distribution. Our study, and references, provide positive evidence in favor of this argument.

The strength of embeddings is their ability to approximate **semantics** in a way that symbols may struggle with, such as distinguishing the equivalence of two sentences with different symbols but identical meanings. In NLP there is no easy way to determine this equivalence. In formal mathematics, symbolic semantics and thus equivalence can sometimes be done exactly. Though it doesn't come without its costs, e.g. requires expert knowledge, computationally demanding or (approximately) exhaustive representations like e-graphs. Therefore, embedding methods for data diversity, quality, etc. have a unique advantage being more generally applicable.