# Brando Miranda, Alycia Lee

Dept. of Computer Science
Advised by Sanmi Koyejo, Stanford Computer Science (CS)
In Collaboration with Patrick Yu

Brando's website

# The Diversity Coefficient: A Data Quality Metric that shows LLMs are Pretrained on Formally Diverse Data

*Keywords: large language models (LLMs), data quality, metrics, diversity*

## Summary

- We develop a data quality metric to measure the formal diversity of the pretraining data of Large Language Models (LLMS).
- Diversity coefficient of LLM pretraining datasets are high compared to previous work on vision datasets.
- We test that the diversity coefficient correlates with the ground truth diversity (when known).
- The diversity coefficient passes important sanity checks:
  - The diversity coefficient correlates with latent concepts and vocab size in the synthetic GINC (language) datasets.
  - The diversity coefficient estimates low diversity when comparing tasks of the same data set, but high diversity in cross task comparisons is across different data sets.

## More about Brando Miranda

- Current EDGE Scholar at Stanford University.
- Research interests lie in meta-learning, foundation models for theorem proving, and human & brain inspired AI.
- M.Eng. in Electrical Engineering and Computer Science at MIT.

Stanford | Data Science

NSF III 2046795, IIS 1909577, CCF 1934986, NIH 1R01MH116226-01A, NIFA award 2020-67021-32799, the Alfred P. Sloan Foundation, and Google Inc.

## Methods

**Task2Vec-based diversity coefficient** (Miranda, Yu, et al 2022) approximately measures the intrinsic variability of tasks in a few-shot learning benchmark.

Ground Truth Diversity Coefficient: expected distance between pairs of tasks τ1, τ2:

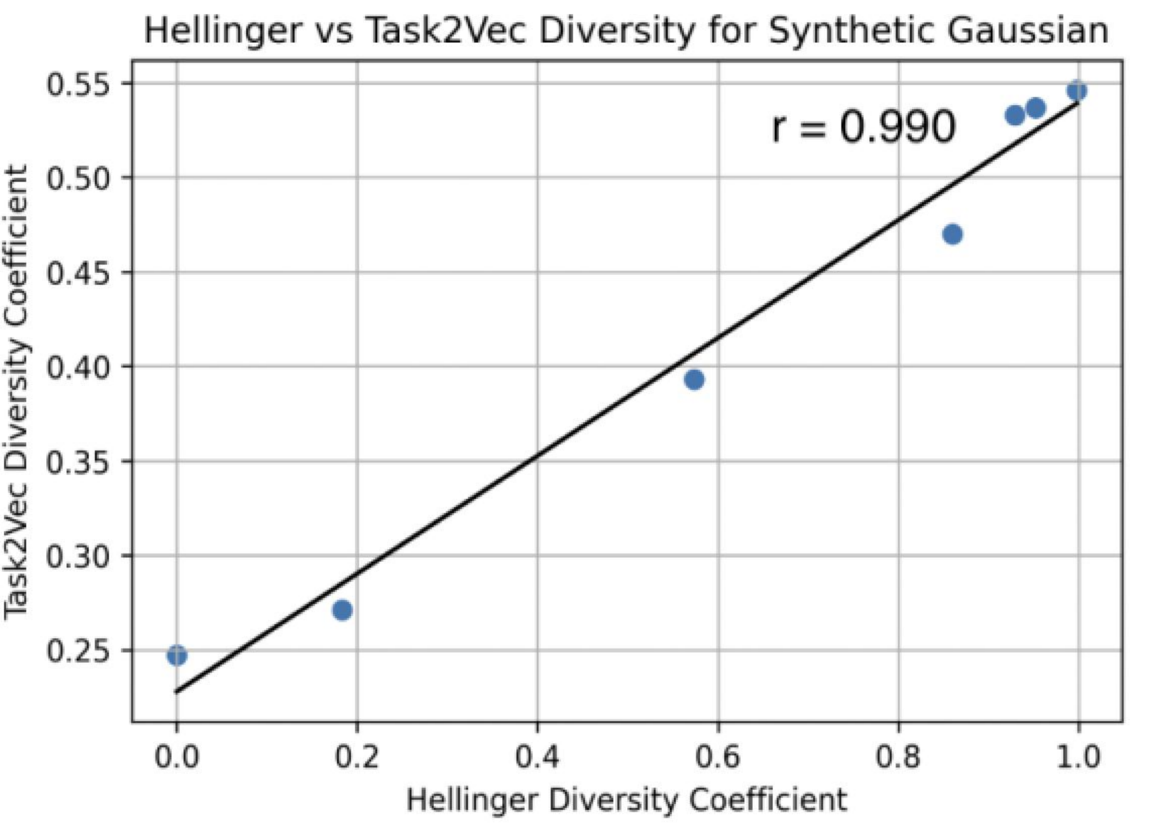
$$\hat{div}(B) = \mathbb{E}_{\tau_1,\tau_2 \sim \hat{p}(\tau|B):\tau_1 \neq \tau_2} \left[ d(p(x_1, y_1 \mid \tau_1), p(x_2, y_2 \mid \tau_2)) \right]$$

Definition: expected distance between pairs of tasks τ1, τ2 as Task2Vec embeddings:

$$\hat{div}(B) = \mathbb{E}_{\tau_1,\tau_2 \sim \hat{p}(\tau|B):\tau_1 \neq \tau_2} \mathbb{E}_{D_1 \sim \hat{p}(x_1,y_1|\tau_1), D_2 \sim \hat{p}(x_2,y_2|\tau_2)} \Big[ d(\mathrm{diag}(\hat{F}_{D_1,f_w}), \mathrm{diag}(\hat{F}_{D_2,f_w})$$

## Experiments & Results

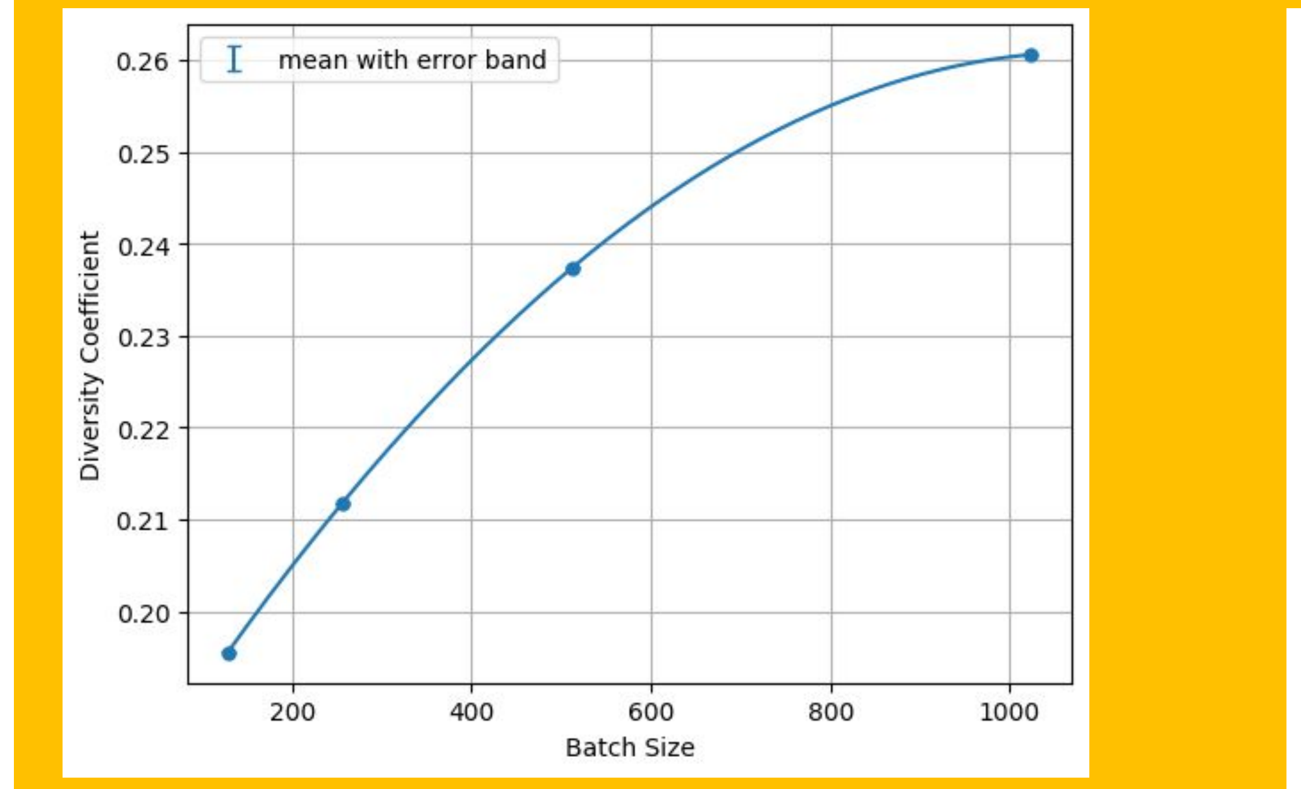
### Task2Vec diversity coefficient correlates with ground truth diversity for synthetic data



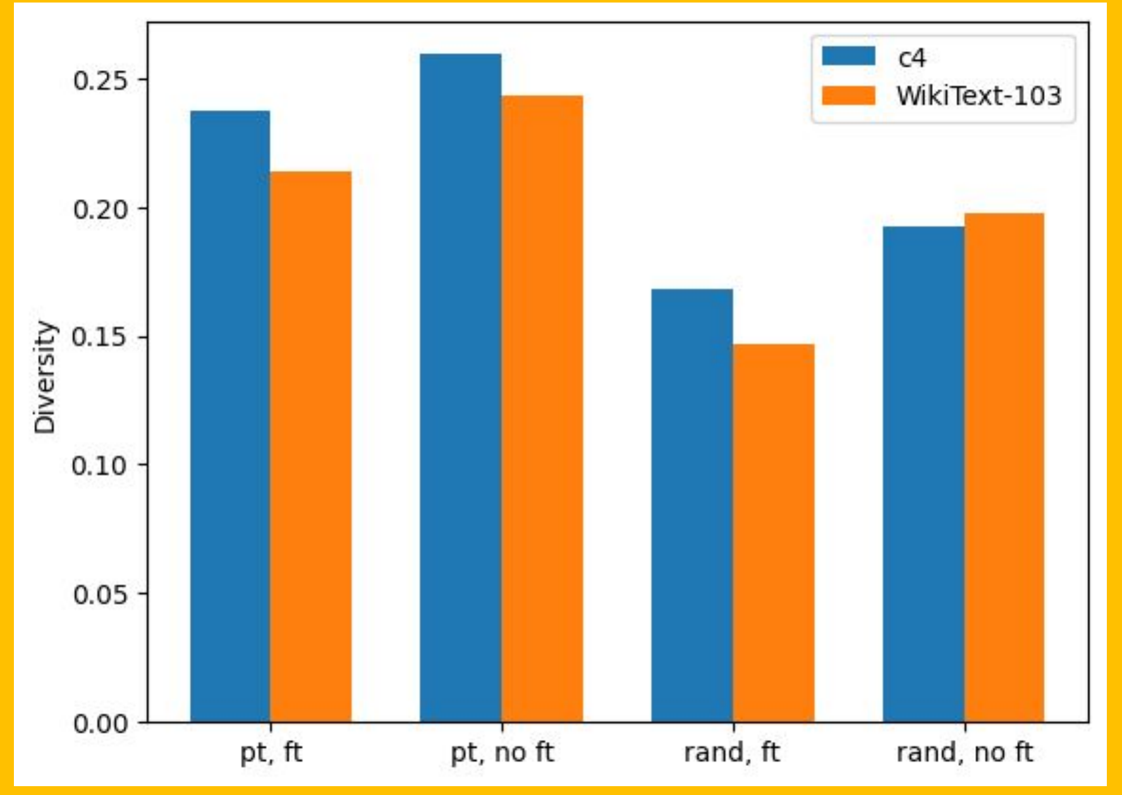
### Diversity of c4, Wikitext-103, and The Pile are twice as high vs. vision benchmarks

| Dataset | Probe Network | Diversity Coeff |
|---|---|---|
| MiniImagenet | Resnet18 | 0.117 ±2.098e-5 |
| Cifar-fs | Resnet18 | 0.100 ±2.18e-5 |
| c4 | GPT-2 | 0.2374 ±2.785e-5 |
| WikiText-103 | GPT-2 | 0.2140 ±7.93e-5 |
| The Pile | GPT-2 | 0.2463 ±3.034-5 |

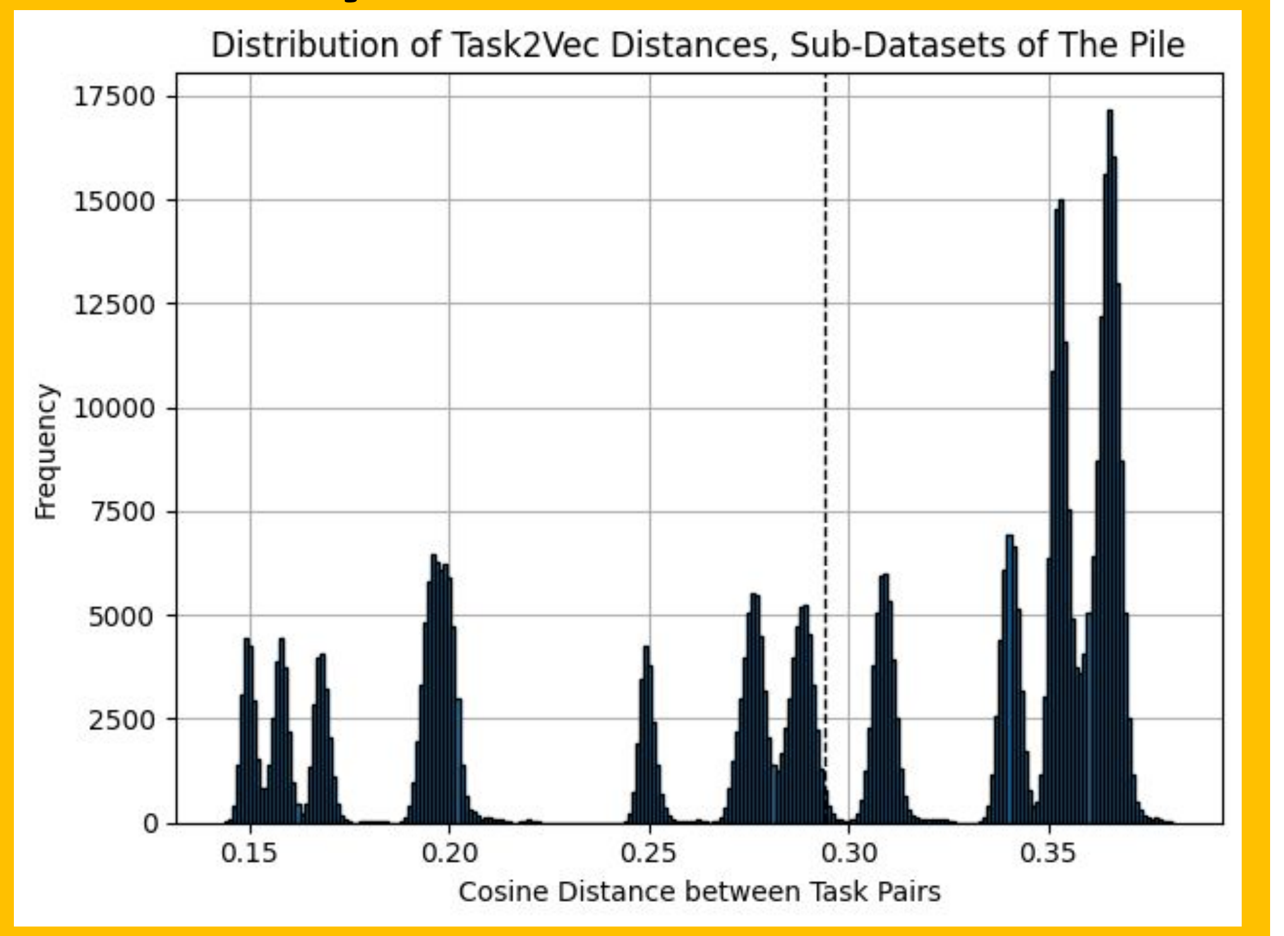
### Batch size correlates with diversity

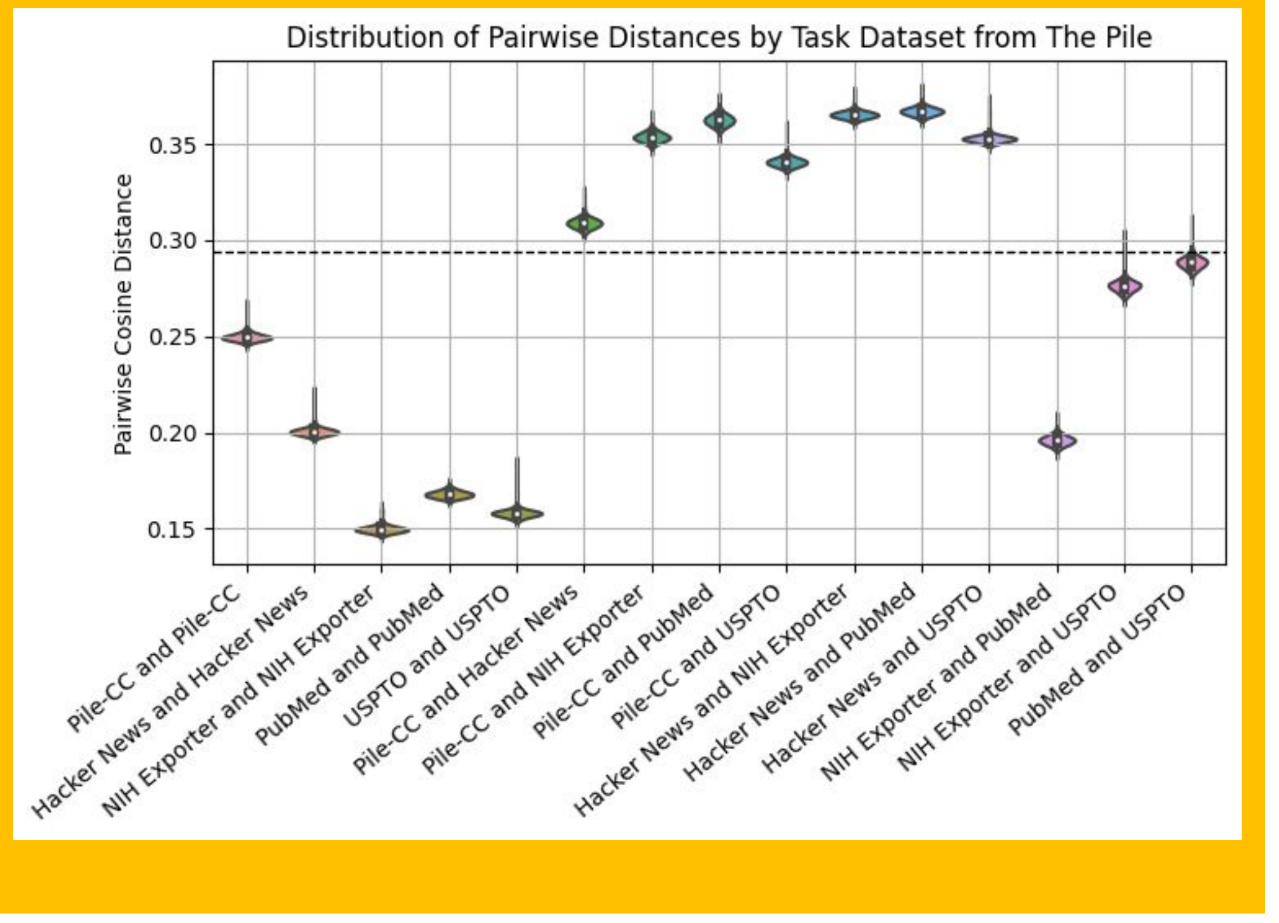


### Random probe underestimates diversity, non fine-tuned overestimates diversity

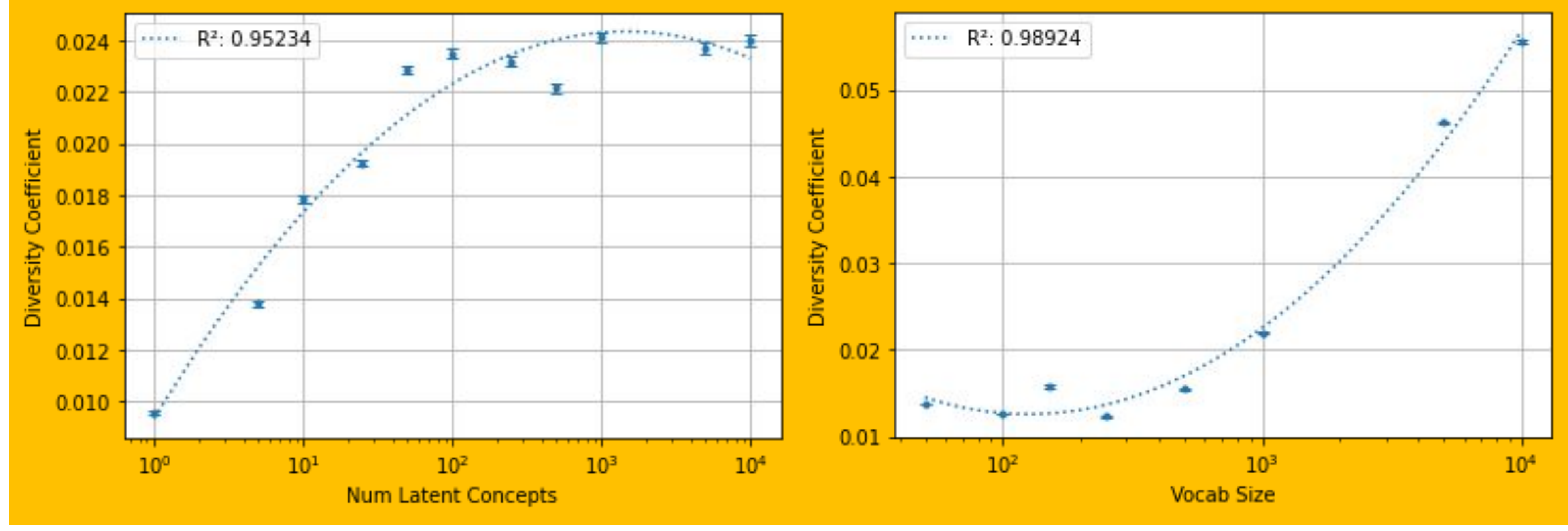


### Pairwise combinations of The Pile datasets have higher diversity vs. individual datasets





### Diversity correlates with # latent concepts (left) and vocab size (right) in GINC


R²: 0.95234
Diversity Coefficient
$10^0$
$10^1$
$10^2$
$10^3$
$10^4$
Num Latent Concepts
R²: 0.98924
Diversity Coefficient
$10^2$
$10^3$
$10^4$
Vocab Size