ICML
International Conference
On Machine Learning

NUS
National University
of Singapore

Tencent
AI Lab

Measuring Diversity in Synthetic Datasets

Yuchang Zhu, Huizhe Zhang, Bingzhe Wu, Jintang Li,

Zibin Zheng, Peilin Zhao, Liang Chen, Yatao Bian

zhuych27@mail2.sysu.edu.cn

https://github.com/bluewhalelab/dcscore

ICML
International Conference
On Machine Learning

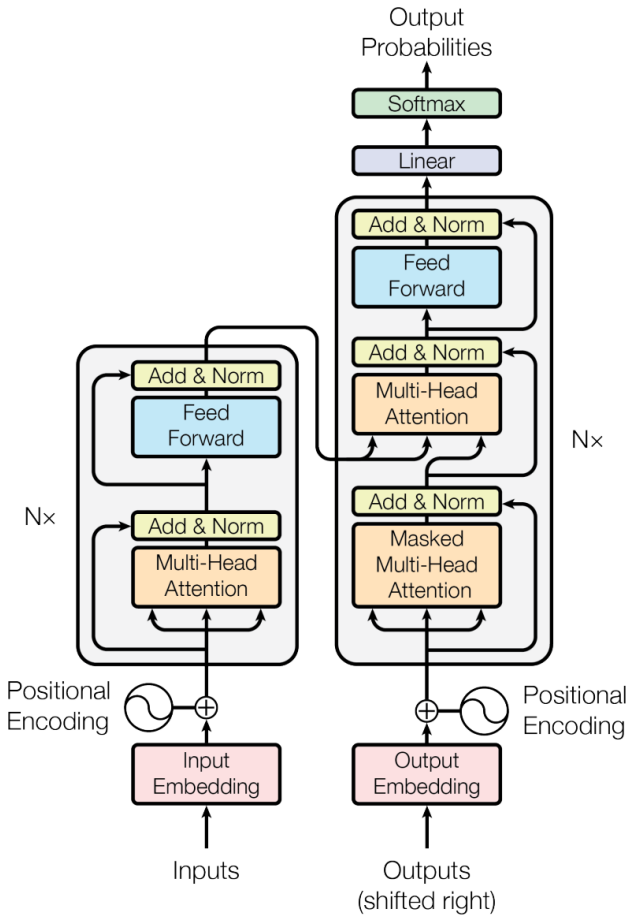
CONTENTS

- Background & Motivation
- Method: DCScore
- Experiments & Discussions
- Conclusion

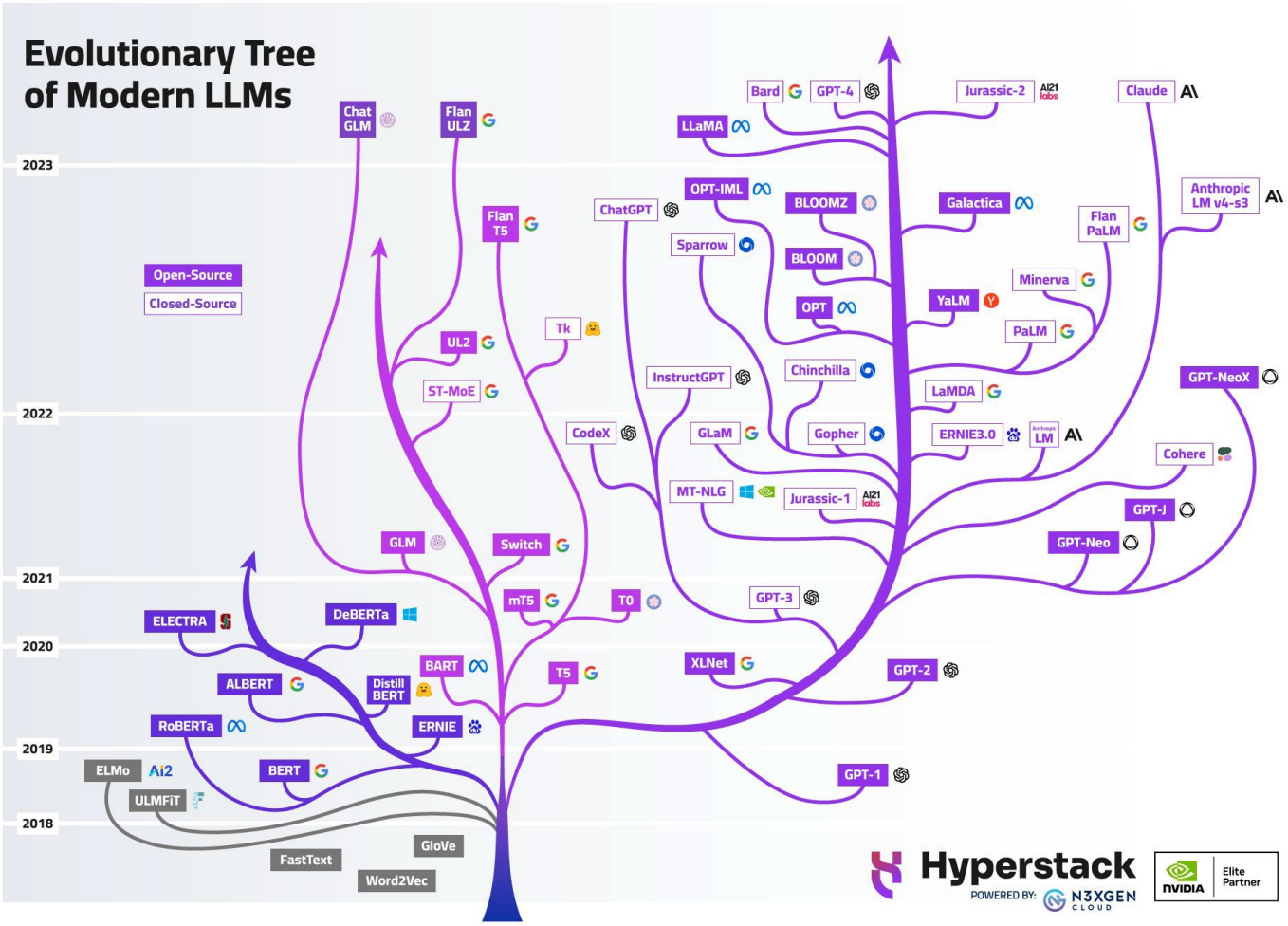
BACKGROUND & MOTIVATION

ICML International Conference On Machine Learning

Large language models (LLMs)



Transformer



LLMs Tree

BACKGROUND & MOTIVATION

Application of LLMs

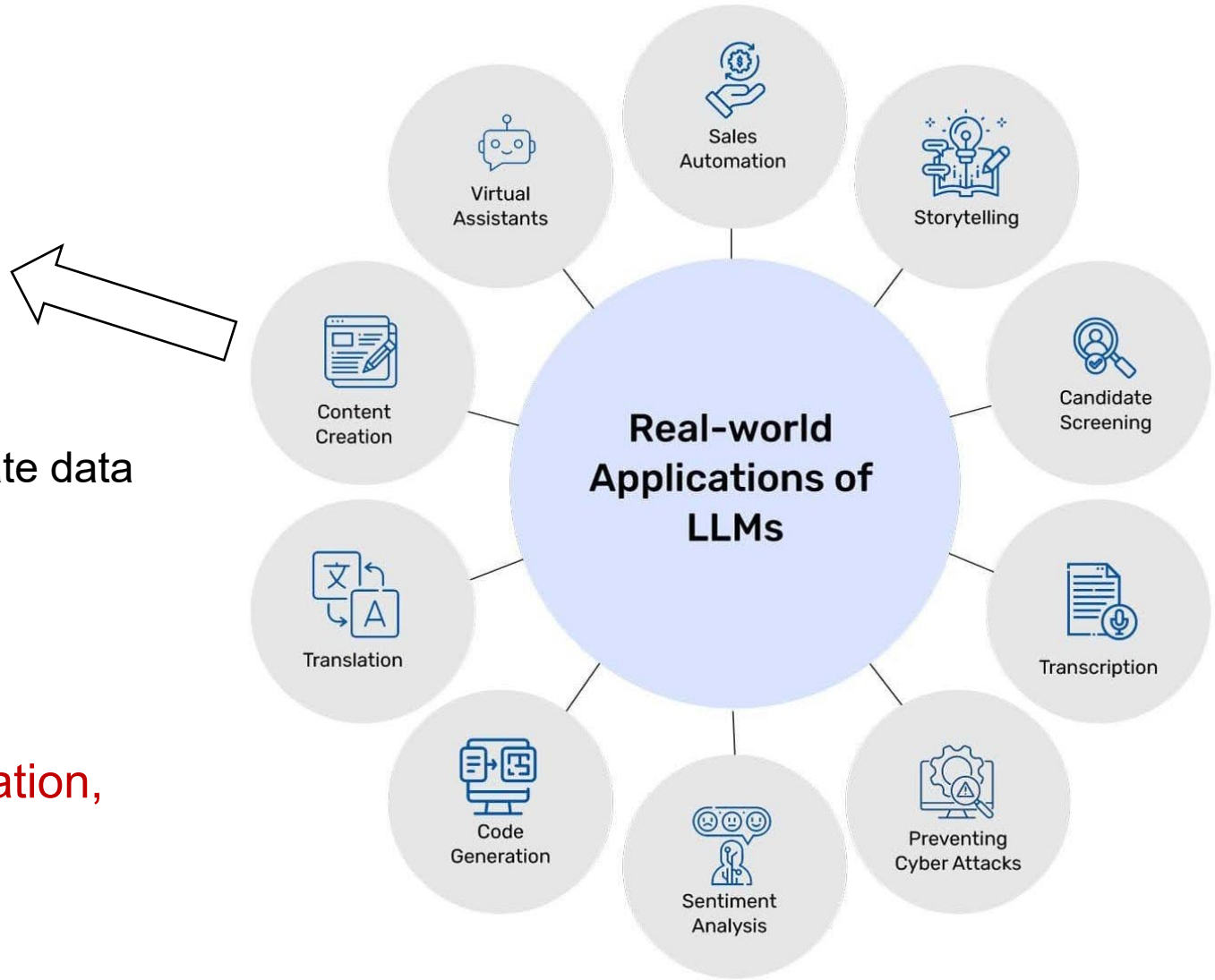
Dataset generators

$$\mathcal{D} \leftarrow \mathcal{M}(T, \mathcal{D}_{sup})$$

$\mathcal{D}$: LLM-generated dataset

T: generation task

$\mathcal{D}_{sup}$: supplementary materials to facilitate data augmentation



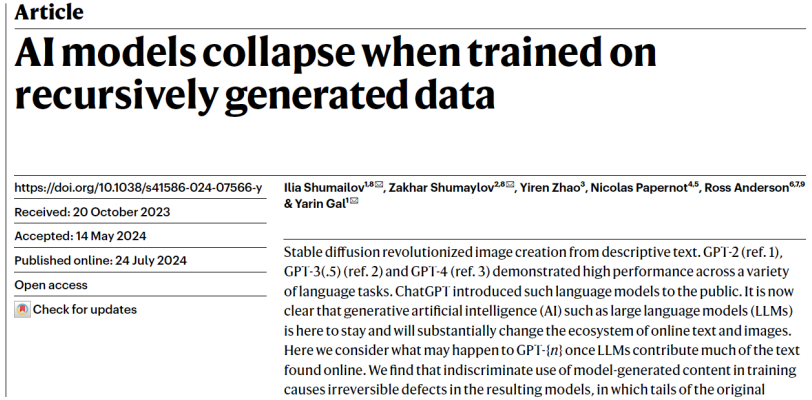
LLM as dataset generators

- Data augmentation (e.g., annotation, augment samples)
- Generate datasets from scratch

BACKGROUND & MOTIVATION

Synthetic datasets and its evaluation

With the rise of synthetic data and its impact on next-gen models, its evaluation is often neglected.



Article

AI models collapse when trained on recursively generated data

https://doi.org/10.1038/s41586-024-07566-y
Received: 20 October 2023
Accepted: 14 May 2024
Published online: 24 July 2024
Open access
Check for updates

Ilia Shumailov[1,8], Zakhar Shumaylov[2,8], Yiren Zhao[3], Nicolas Papernot[4,5], Ross Anderson[6,7,9] & Yarin Gal[1]

Stable diffusion revolutionized image creation from descriptive text. GPT-2 (ref. 1), GPT-3(.5) (ref. 2) and GPT-4 (ref. 3) demonstrated high performance across a variety of language tasks. ChatGPT introduced such language models to the public. It is now clear that generative artificial intelligence (AI) such as large language models (LLMs) is here to stay and will substantially change the ecosystem of online text and images. Here we consider what may happen to GPT-{n} once LLMs contribute much of the text found online. We find that indiscriminate use of model-generated content in training causes irreversible defects in the resulting models, in which tails of the original

Model collapse

Quality (Diversity) of synthetic datasets

- **Data diversity** [1]: the richness and variety of data.
- Recent studies [2] suggest a lack of diversity within the dataset may lead to performance degradation in some scenarios.

Diversity evaluation:

Given an LLM-generated dataset $\mathcal{D} = \{\mathcal{T}_i\}_{i=1}^{n}$, $\{\tilde{\mathcal{T}}_i\}_{i=1}^{n}$ represents a collection of diversity-sensitive components:

$$\text{DiversityScore} \leftarrow \text{Eval}(\{\tilde{\mathcal{T}}_i\}_{i=1}^{n})$$

[1] Beyond scale: the diversity coefficient as a data quality metric demonstrates llms are pre-trained on formally diverse data.
[2] Large language model as attributed training data generator: A tale of diversity and bias (Neurips 2023)

BACKGROUND & MOTIVATION

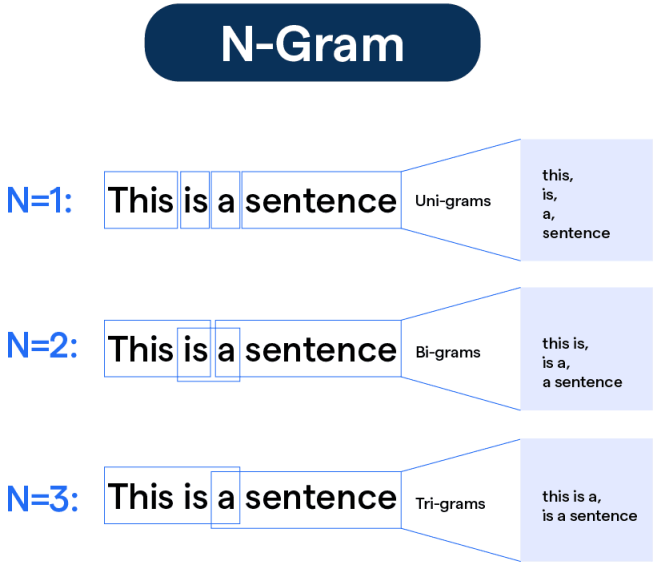
Previous Works (ML&NLP)

- **N-gram-based method**

N-gram: a sequence of *n* adjacent symbols in particular order.

$$\text{Distinct-n}(\mathcal{D}) = \frac{|\,\text{Unique}(\text{n-grams}(\text{Concat}(\mathcal{D})))|}{|\,\text{n-grams}(\text{Concat}(\mathcal{D}))|},$$

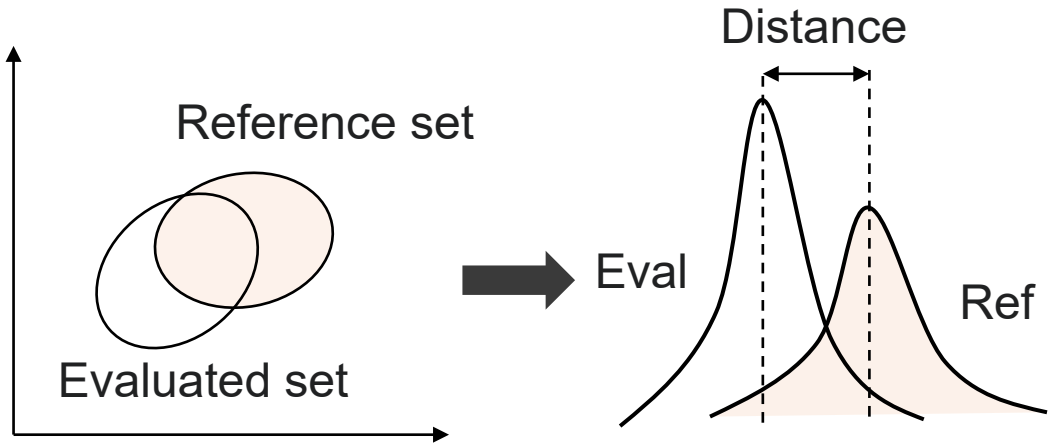
Drawback: Focus on the form difference



- **Reference-based method**

Core Idea: Introduce a reference distribution to evaluate diversity.

Drawback: Over-reliance on reference distribution



BACKGROUND & MOTIVATION

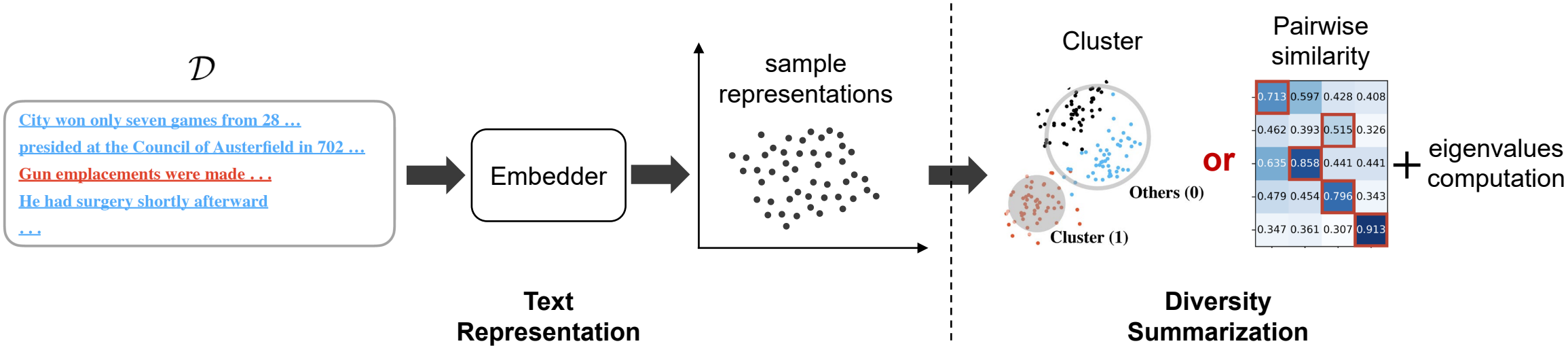
Previous Works (ML&NLP)

- **Transformation-based method**

Core Idea: leverage well-designed models to generate representations and then summarize diversity through clustering or eigenvalue computation.

Drawback: high computational costs in diversity summarization.

Transformation-based method also includes the entropy-based method, e.g., Vendi Score [1].



[1] The Vendi Score: A Diversity Evaluation Metric for Machine Learning, TMLR 2023.

BACKGROUND & MOTIVATION

Challenges

Holistic Analysis

Diversity evaluation is a holistic analysis task, considering each sample.

Axiomatic Requirements

A principled diversity evaluation method should satisfy some intuitive axioms.

Lower Computational Costs

With growing synthetic datasets, a lower-cost diversity evaluation method is needed.

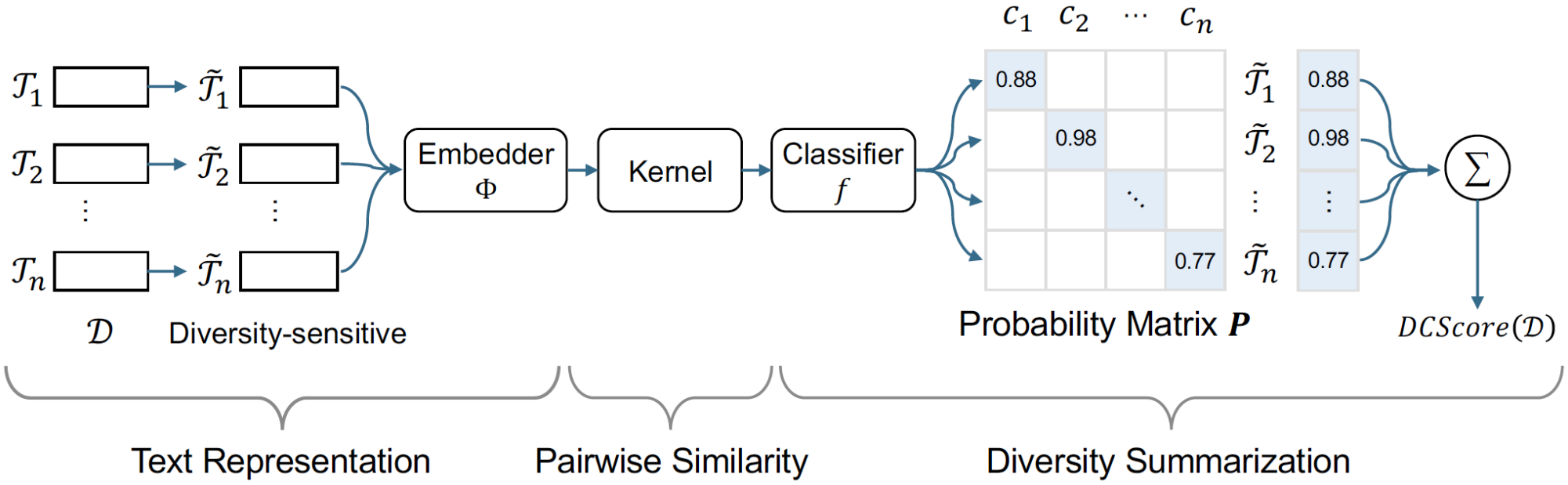
Technical Insight

Our method treats the diversity evaluation as a sample classification task, considering mutual relationships among samples.

METHOD: DCScore

Overview

- **Core idea:** regard the holistic diversity evaluation as **the classification task** at the sample level.



Our method: DCScore

METHOD: DCScore

Details of DCScore

Evaluate $\mathcal{D} = \{\mathcal{T}_i\}_{i=1}^n = \{(x_i, y_i)\}_{i=1}^n \Longrightarrow$ Evaluate $\{\tilde{\mathcal{T}}_i\}_{i=1}^n$

① Text Representation

$$\mathbf{H} = \Phi(\{\tilde{\mathcal{T}}_i\}_{i=1}^n),$$

② Pairwise Similarity

$$\mathbf{K} = \text{Kernel}(\mathbf{H}),$$

③ Diversity Summarization

Classification: $$P(c = c_j|\tilde{\mathcal{T}}_i) = \mathbf{P}[i,j] = f_{\mathbf{K}}(\mathbf{K}[i,j]) = \frac{\exp \mathbf{K}[i,j]/\tau}{\sum_j \exp \mathbf{K}[i,j]/\tau},$$

Definition 1 (`DCScore`). *Let $\mathcal{D} = \{\mathcal{T}_i\}_{i=1}^n$ denote the LLM-generated dataset with n samples, and let $\{\tilde{\mathcal{T}}_i\}_{i=1}^n$ represent a set of diversity-sensitive components within $\{\mathcal{T}_i\}_{i=1}^n$. Denote P_i as the classification probability vector of $\tilde{\mathcal{T}}_i$. By conducting the classification task for all $\tilde{\mathcal{T}}_i$ and obtaining the probability matrix $\boldsymbol{P} = [P_1, P_2, ..., P_n]$, `DCScore` for $\mathcal{D}$ is defined as the trace of $\boldsymbol{P}$:*

$$\text{DCScore}(\mathcal{D}) = \text{tr}(\boldsymbol{P}) = \sum_{i=1}^{n} \boldsymbol{P}[i,i]. \tag{5}$$

METHOD: DCScore

Properties of DCScore

- **Effective number**

  Diversity should be defined as the effective number of samples in a dataset, ranging from 1 to n.

- **Identical samples**

  Identical samples never increases diversity: $\mathrm{DCScore}(\mathcal{D}_1) = \mathrm{DCScore}(\mathcal{D}_2) = \mathrm{DCScore}(\mathcal{D}^{'})$.

- **Symmetry**

  Diversity remains constant regardless of the order of the samples: $\mathrm{DCScore}(\mathcal{D}) = \mathrm{DCScore}(\pi(\mathcal{D}))$.

- **Monotonicity**

  The diversity of a dataset decreases as the similarity between its samples increases.

$$\mathrm{DCScore}(\mathcal{D}_1^{'}) > \mathrm{DCScore}(\mathcal{D}_2^{'}).$$

METHOD: DCScore

Complexity Analysis

Table: Complexity analysis of DCScore and VendiScore

		General Kernels	Inner Product
Pairwise Similarity	VendiScore	$\mathcal{O}(n^2 \cdot \mathcal{O}_{kernel})$	$\mathcal{O}(d^2 n)$
	DCScore		$\mathcal{O}(n^2 d)$
Summarization	VendiScore	$\mathcal{O}(n^3)$	$\mathcal{O}(d^3)$
	DCScore	$\mathcal{O}(n^2)$	$\mathcal{O}(n^2)$
Total	VendiScore	$\mathcal{O}(n^2 \cdot \mathcal{O}_{kernel} + n^3)$	$\mathcal{O}(d^2 n + d^3) = \mathcal{O}(d^2 n)$
	DCScore	$\mathcal{O}(n^2 \cdot \mathcal{O}_{kernel} + n^2)$	$\mathcal{O}(n^2 d + n^2)$

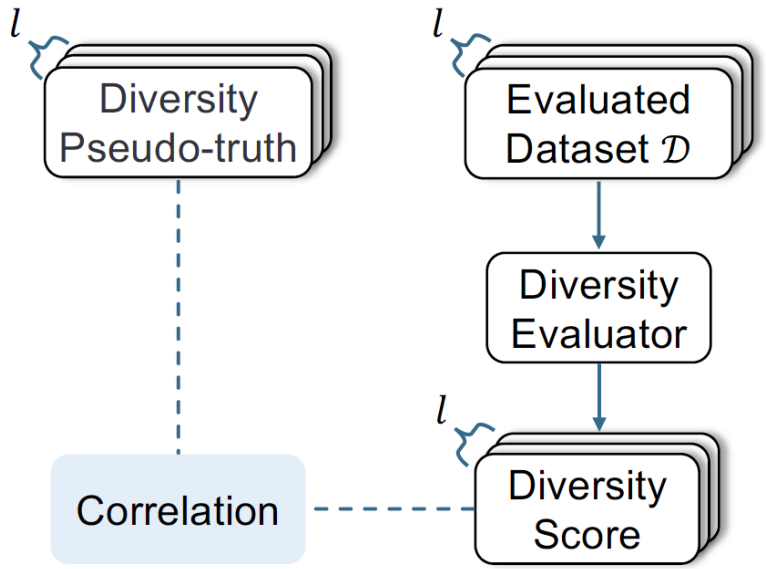
- DCScore has lower computational complexity with non-linear kernels.
- When using the inner product kernel with $n \gg d$, VendiScore's calculation simplifies but yields only marginal time savings.

EXPERIMENTS & DICUSSIONS

Experimental Settings

How to verify the effectiveness of DCScore?

The correlation between these diversity scores and the corresponding diversity pseudo-truth for each dataset.



Evaluation Metrics: Spearman's ρ

Baselines: Distinct-n, K-means Inertia, VendiScore.

Datasets

- Self-generated datasets
- Existing generated datasets

Experiments

- Correlation evaluation
- Computational costs

EXPERIMENTS & DICUSSIONS

Correlation with generation temperature τ_g

Table: Correlation (Spearman's ρ) with τ_g

Methods	*Zero-shot setting*				*Few-shot setting*			
	Text classification		Story completion		Text classification		Story completion	
	13B	70B	13B	70B	13B	70B	13B	70B
Distinct-n	0.9909	**0.9870**	0.9766	0.9701	0.9857	0.9766	0.9779	0.9935
K-means Inertia	-0.1143	0.9688	0.9454	0.8727	0.7104	0.7273	0.9662	0.9662
VendiScore	**0.9961**	0.9818	**0.9870**	**0.9922**	**0.9909**	0.9857	**0.9857**	0.9961
DCScore	**0.9961**	0.9779	0.9844	0.9792	**0.9909**	**0.9883**	**0.9857**	**0.9974**

Observation:

- DCScore outperforms all baseline methods in most cases.
- DCScore exhibits a strong correlation with the diversity pseudo-truth.

Correlation with human judgment / LLM evaluation

Table: Correlation (Spearman's ρ) with human judgment

	Story-Few	Story-Zero	Text-Few	Text-Zero
Human-DCScore	$0.9040_{\pm 0.04}$	$0.7870_{\pm 0.10}$	$0.7915_{\pm 0.16}$	$0.8798_{\pm 0.10}$
τ_g-DCScore	$0.9086_{\pm 0.07}$	$0.7829_{\pm 0.16}$	$0.8400_{\pm 0.16}$	$0.8971_{\pm 0.07}$
τ_g-Human	$0.9276_{\pm 0.02}$	$0.9194_{\pm 0.06}$	$0.9770_{\pm 0.02}$	$0.9255_{\pm 0.08}$

Table: Correlation (Spearman's ρ) with GPT-4 evaluation

	Story-Few	Story-Zero	Text-Few	Text-Zero
GPT-4-DCScore	$0.6057_{\pm 0.30}$	$0.9010_{\pm 0.04}$	$0.6131_{\pm 0.18}$	$0.9052_{\pm 0.09}$
τ_g-DCScore	$0.6757_{\pm 0.30}$	$0.8782_{\pm 0.08}$	$0.5714_{\pm 0.27}$	$0.9336_{\pm 0.06}$
τ_g-GPT-4	$0.9086_{\pm 0.07}$	$0.7829_{\pm 0.16}$	$0.8400_{\pm 0.16}$	$0.8971_{\pm 0.07}$

Observation:

- DCScore presents strong correlations with human judgement / LLM evaluation.

EXPERIMENTS & DICUSSIONS

Computational cost

Observation:

DCScore offers significantly lower time complexity than VendiScore while sacrificing little in diversity evaluation performance.

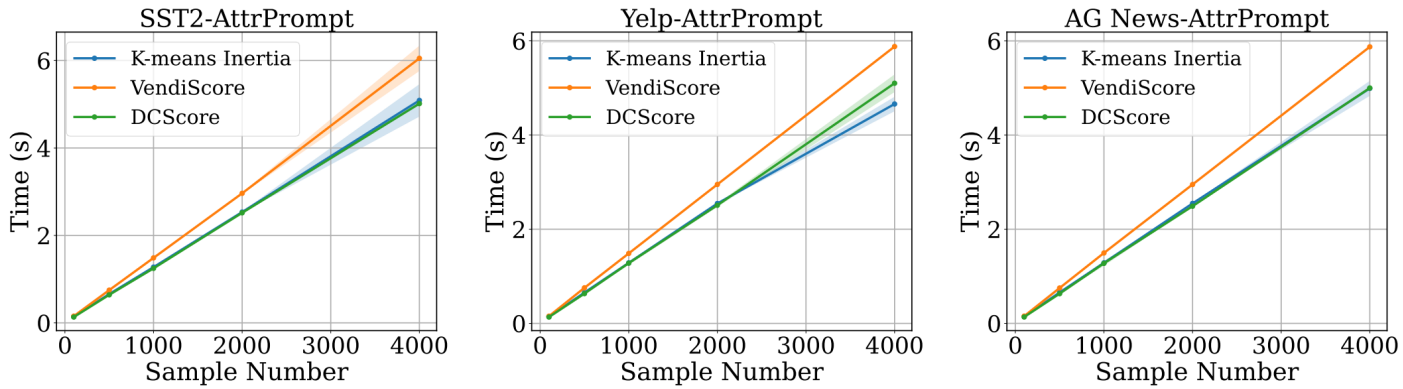


Figure: Computation times on small datasets

Observation:

For nonlinear kernels, DCScore shows substantial computation time advantages, while slightly underperforming VendiScore with linear kernels (Inner product).

Table: Computation times on a large dataset (Yelp)

Kernels		Yelp				
	Sample num	4k	8k	16k	32k	64k
Inner product	VendiScore	57.96$_{\pm 0.35}$	**114.64**$_{\pm 1.63}$	**227.76**$_{\pm 7.04}$	**451.49**$_{\pm 19.73}$	**912.60**$_{\pm 25.69}$
	DCScore	**57.95**$_{\pm 0.31}$	115.35$_{\pm 1.16}$	232.49$_{\pm 1.34}$	448.98$_{\pm 23.94}$	961.29$_{\pm 2.86}$
RBF kernel	VendiScore	59.31$_{\pm 0.06}$	118.15$_{\pm 0.91}$	242.06$_{\pm 7.60}$	527.99$_{\pm 2.89}$	1272.93$_{\pm 21.15}$
	DCScore	**58.49**$_{\pm 0.14}$	**116.29**$_{\pm 0.92}$	**232.94**$_{\pm 3.09}$	**471.18**$_{\pm 7.80}$	**953.62**$_{\pm 17.21}$
Poly kernel	VendiScore	59.48$_{\pm 0.05}$	118.94$_{\pm 0.95}$	234.08$_{\pm 11.72}$	522.82$_{\pm 3.04}$	1313.55$_{\pm 12.64}$
	DCScore	**58.73**$_{\pm 0.08}$	**117.02**$_{\pm 0.90}$	**227.72**$_{\pm 9.51}$	**462.45**$_{\pm 13.91}$	**988.53**$_{\pm 1.10}$

CONCLUSION

- We investigate the diversity evaluation of synthetic datasets, a topic systematically under-explored in existing research.
- We present DCScore, a diversity evaluation method from a classification perspective.
- We provide theoretical guarantees demonstrating that DCScore meets the axiom requirements (Leinster & Cobbold, 2012) for a principled diversity evaluation method.
- DCScore exhibits a better correlation with diversity pseudo-truths. DCScore exhibits significantly lower computational cost compared to transformation-based counterparts.

Thanks !


zhuychz27@mail2.sysu.edu.cn


https://github.com/bluewhalelab/dcscore

ICML
International Conference
On Machine Learning