Token Embeddings Violate the Manifold Hypothesis

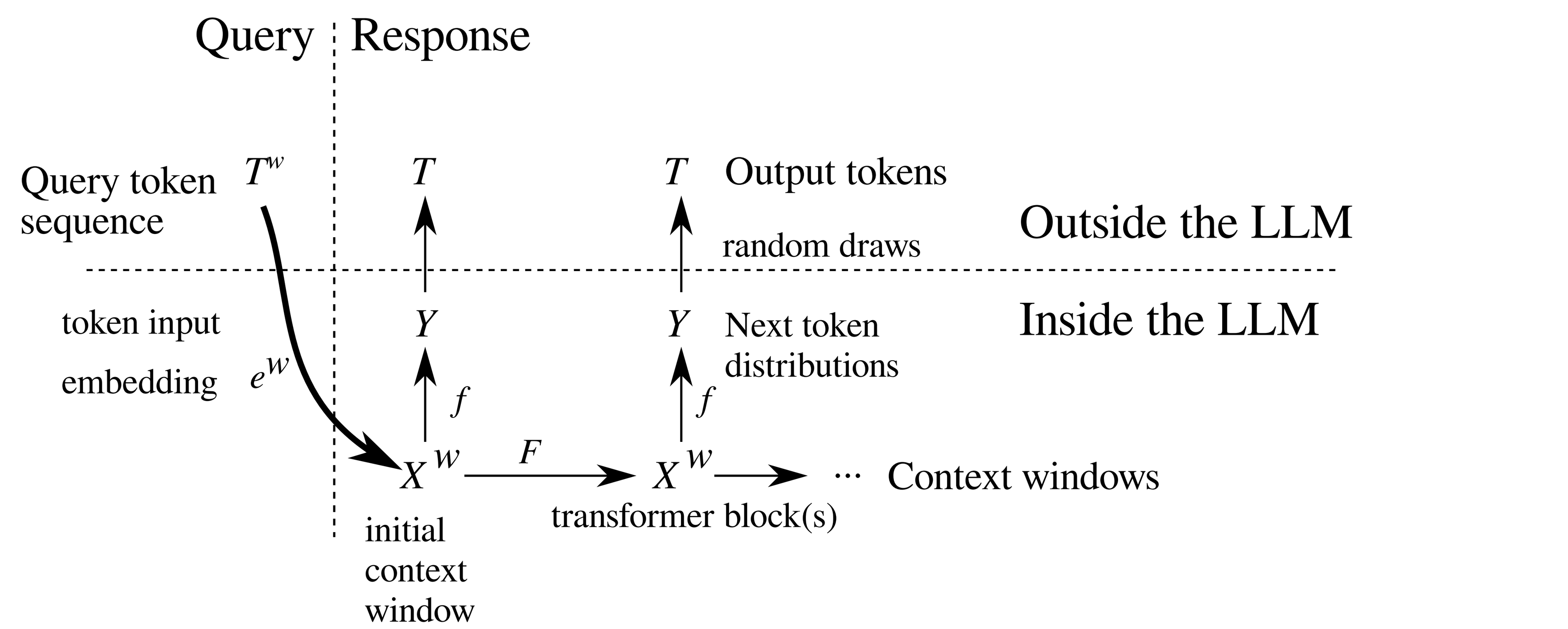
Michael Robinson · Sourya Dey · Tony Chiang

The manifold and fiber bundle hypotheses

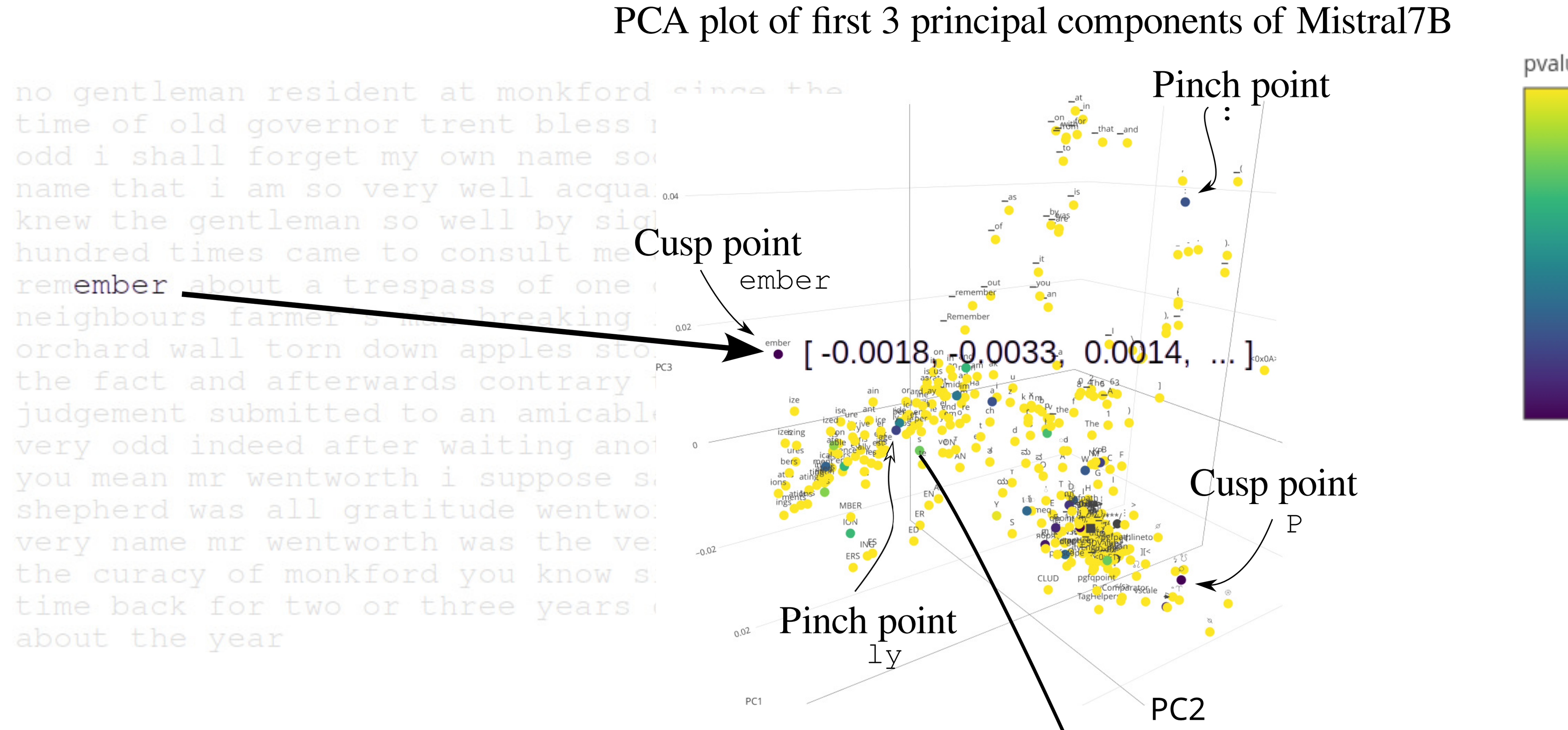
Hypotheses we test. We assume the token subspace has reach $\tau > 0$, and estimate the dimension in the ball centered at token ψ with radius $r < \tau$.

	Manifold test	Fiber bundle test
H_0	There is a unique dimension at ψ	The dimension at ψ in a ball of radius r does not increase as r increases.
H_1	There is not a unique dimension at ψ	The dimension at ψ increases at some r

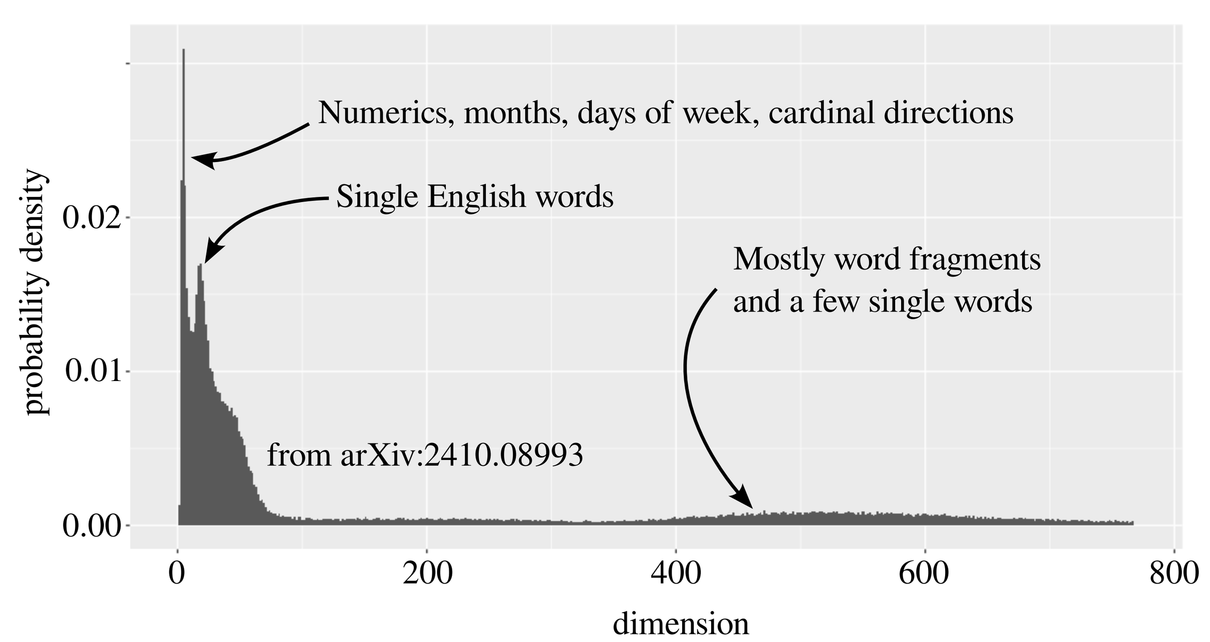
Token embeddings for LLMs



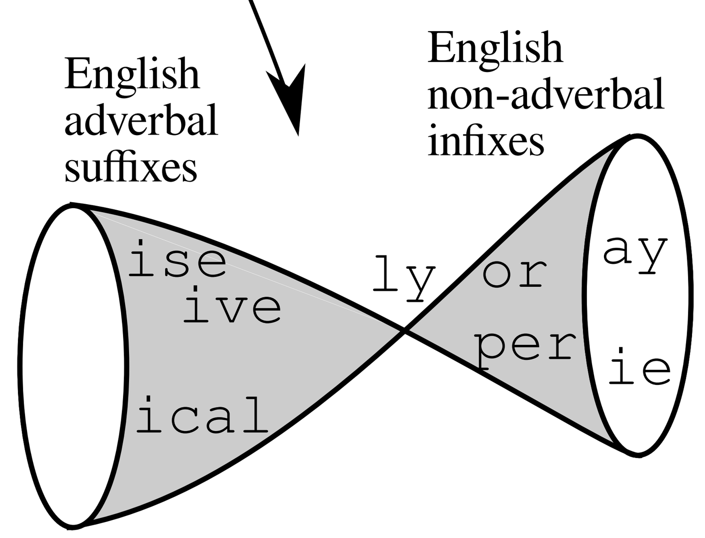
PCA plot of first 3 principal components of Mistral7B



Evidence against the manifold hypothesis



Multimodal distribution of dimension estimates



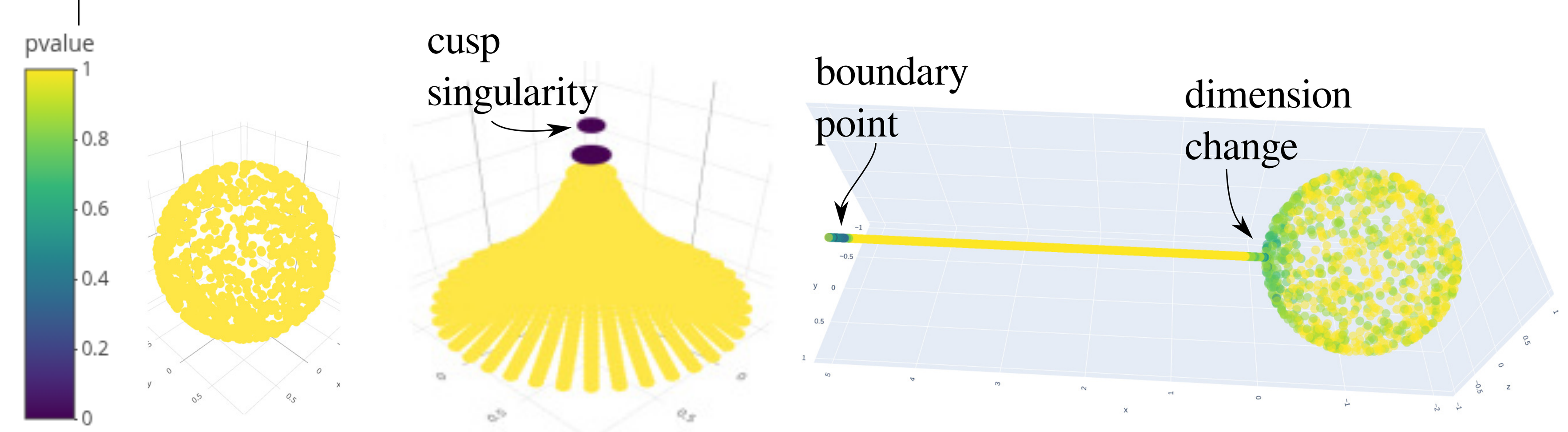
Polysemantic tokens (such as ly)
Linguistic justification for singularities as polysemantic tokens (Jakubowski 2020)

Implementing our tests

Algorithm 1 Manifold and fiber bundle tests

Require: $x_1, \dots, x_n \in \mathbb{R}^\ell$: coordinates for each point
Require: v_{min} and v_{max}: minimum and maximum number of tokens in neighborhood
Require: W: sliding window size
Require: α: significance level
Ensure: p_1: set of p values for manifold hypothesis
Ensure: p_2: set of p values for fiber bundle hypothesis
Ensure: Set of dimension estimates

procedure MANIFOLDANDFIBERBUNDLETEST($x_\bullet, v_{min}, v_{max}, W$)
  Compute $n \times n$ pairwise distance matrix D between all tokens
  for Each column of D **do** ▷ Columns correspond to token indices
    Sort the column ▷ Now row indices of distance matrix are volumes, entries are radii
    Retain rows v_{min} through v_{max}
    Compute log-log slopes (= dimension estimates) along the column
    Run two sample T-test along adjacent sliding windows of size W with level α:
    Manifold test: Append to p_1: the p-value for the hypothesis that the slope is constant
    Fiber bundle test: Append to p_2: the p-value for the hypothesis that the slope decreases with row index
    Store both p values and slope with corresponding token (column index)
  end for
  Apply Holm-Bonferroni multiple test correction to both sets of p-values
end procedure

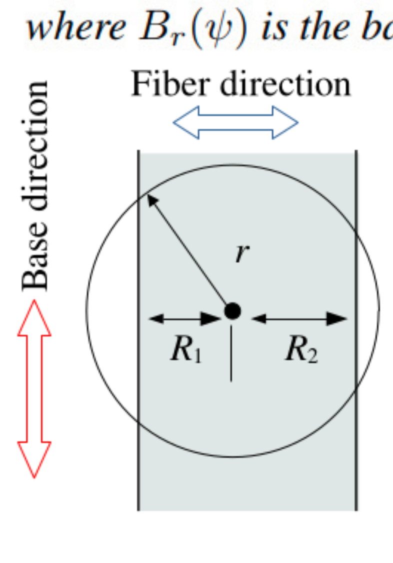


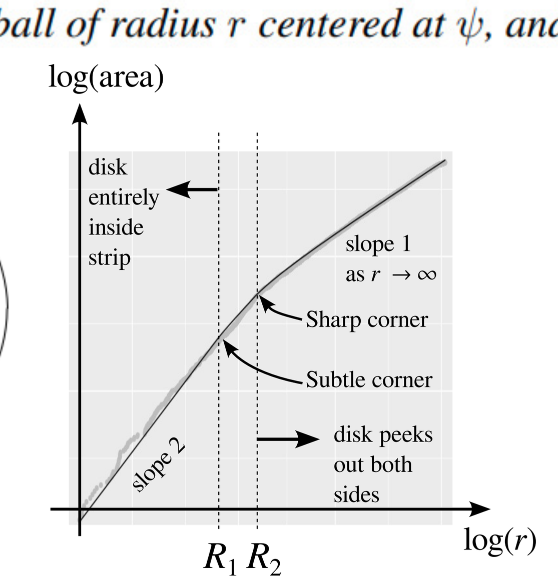
Theoretical justification

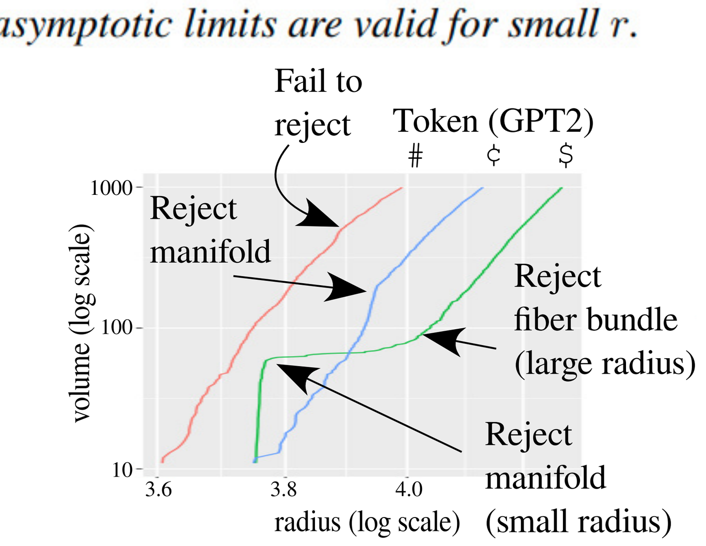
Theorem 1. *Suppose that T is a compact, finite-dimensional Riemannian manifold with boundary, with a volume form v satisfying $v(T) < \infty$, and let $p : T \to S$ be a fiber bundle. If $e : T \to \mathbb{R}^\ell$ is a smooth embedding with reach τ, then there is a function $\rho : e(T) \to [0, \tau]$ such that if $\psi \in e(T)$, the induced volume (e_*v) in $\mathbb{R}^\ell$ satisfies*

$$(e_*v)\left(B_r(\psi)\right) = \begin{cases} O\left(r^{\dim T}\right) & \text{if } 0 \le r \le \rho(\psi), \\ (e_*v)\left(B_{\rho(\psi)}(\psi)\right) + O\left((r - \rho(\psi))^{\dim S}\right) & \text{if } \rho(\psi) \le r, \end{cases}$$

where $B_r(\psi)$ is the ball of radius r centered at ψ, and the asymptotic limits are valid for small r.







Theorem 2. *Let Z be a d-dimensional bounding manifold for the token subspace, such that $T \subseteq Z$. Consider an LLM with a context window of size w, in which the latent space of tokens is $\mathbb{R}^\ell$, and we collect m tokens as output from this LLM.*

Suppose the following, (enough tokens are collected from the response) $m > \frac{2wd}{\ell}$, but (the context window is longer than the number of tokens we collected) $w \geq m$. Under these conditions, a generic set of transformers yields a topological embedding of $T^w = T \times \cdots T$ into the output of the LLM.

Experimental results

Running the two hypothesis tests on four open weight LLMs

Model	Manifold rejects	Fiber bundle: Smaller Radius dim.	Fiber bundle: Smaller Radius rejects	Fiber bundle: Larger radius dim.	Fiber bundle: Larger radius rejects
GPT2 $n = 50257$	66 $p \approx 3 \times 10^{-8}$	Q1: 20 Q2: 389 Q3: 531	12 $p \approx 9 \times 10^{-6}$	Q1: 8 Q2: 14 Q3: 32	7 $p \approx 3 \times 10^{-8}$
Llemma7B $n = 32016$	33 $p \approx 5 \times 10^{-9}$	Q1: 4096 Q2: 4096 Q3: 4096	0 N/A	Q1: 8 Q2: 11 Q3: 14	1 $p \approx 3 \times 10^{-4}$
Mistral7B $n = 32016$	40 $p \approx 3 \times 10^{-7}$	Q1: 9 Q2: 48 Q3: 220	1 $p \approx 8 \times 10^{-4}$	Q1: 5 Q2: 6 Q3: 9	2 $p \approx 8 \times 10^{-5}$
Pythia6.9B $n = 50254$	54 $p \approx 2 \times 10^{-7}$	Q1: 2 Q2: 108 Q3: 235	0 N/A	Q1: 2 Q2: 5 Q3: 145	0 N/A

Test rejected at how many tokens?
What was the smallest p-value?
Quartiles for the distribution of dimension estimates (for those tokens not rejecting manifold hypothesis)

Note: complete lists of all tokens rejecting the hypotheses are in the paper

Do the hypotheses of Theorem 2 apply to each of the four LLMs we tested?

Model	Latent dim ℓ	Bounding dim. d: Small	Bounding dim. d: Large	Context window w	Min. output tokens m such that: Small	Min. output tokens m such that: Large	Singularities persist? $w \ge m$: Small	Singularities persist? $w \ge m$: Large
GPT2	768	389	14	1024	1038	38	Maybe	Yes
Llemma7B	4096	4096	11	4096	8193	23	Maybe	Yes
Mistral7B	4096	48	6	4096	97	13	Yes	Yes
Pythia6.9B	4096	108	5	4096	217	11	Yes	Yes

"Yes" means singularities will persist into LLM output

Implications

1. Token embeddings are not samples from a low curvature manifold nor from a fiber bundle
2. Irregular tokens may result in instabilities in LLM output
   Small changes in a prompt may result in large changes in the response
3. Longer context windows and fine tuning do not resolve these instabilities (Thm 2)
4. This may explain LLM behavioral features:
   a. Glitch tokens
   b. Behavior differences between models, could be useful for model attribution

References

Vargas *et al*., "Understanding Generative AI Content with Embedding Models," *Science Advances*, Nov 2025.

Michael Robinson, Sourya Dey, and Shauna Sweet. The structure of the token space for large language models, 2024

Alfred Gray. The volume of a small geodesic ball of a Riemannian manifold. *Michigan Mathematical Journal*, 20(4):329 – 344, 1974

Alexander Jakubowski, Milica Gasic, and Marcus Zibrowius. Topology of word embeddings: Singularities reflect polysemy. *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, 2020.

Code Paper

This material is based upon work partially supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001124C0319. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of DARPA. Distribution Statement "A" (Approved for Public Release, Distribution Unlimited) applies to the portion of the work funded by DARPA.