Using Embeddings' Geometric Similarity for In-Context Learning

Om Mehta
omehta@umass.edu

Tanush Savadi
tsavadi@umass.edu

Abstract

Recent work demonstrates that large language model embeddings share geometric regularities, including locally low-dimensional structure correlated with semantic coherence. We investigate whether prompt effectiveness is partly explained by the geometry of token embeddings. Specifically, we evaluate the correlation between geometric metrics–intrinsic dimension (ID) and Local Linear Embedding (LLE) reconstruction error—and model performance on the GSM8K mathematical reasoning benchmark. Using two models (TinyLlama-1.1B and Gemma-3-1B-it) across 11 prompt templates, we find that TinyLlama shows a moderate negative correlation between intrinsic dimension and accuracy (Spearman $\rho = -0.54$, $p = 0.085$), suggesting that prompts with lower mean intrinsic dimension tend to achieve higher accuracy. However, Gemma-3-1b-it shows no significant correlation, indicating the geometry-performance relationship may be dependent on the model or mode of training. Our results suggest that embedding geometry offers a signal for predicting prompt effectiveness, and highlights the need for further testing across model and prompt contexts.

1 Introduction

Prompt engineering has emerged as an indispensable technique for extending the capabilities of large language models (LLMs) (Sahoo et al., 2024). Subtle choices in prompt design—such as wording, tone, and structure—can dramatically influence model outputs (Djeffal, 2025). Understanding *why* certain prompts work better than others remains an open question with significant practical implications.

Recent work on representation geometry suggests a potential answer. The Platonic Representation Hypothesis (Huh et al., 2024) argues that AI systems are converging toward shared internal representations. Building on this, Lee et al. (2025) demonstrated that token embeddings across LLMs share "global" similarities (correlated pairwise token orientations within model families) and "local" geometric structure (shared LLE neighborhoods, low intrinsic dimensions for semantically coherent regions). Crucially, they found that not all parts of embedding space are equal: some neighborhoods are more organized than others.

This raises a natural question: **Do prompts that rely on tokens in "good" regions of embedding space produce more reliable model outputs?** If prompts with well-organized, geometrically stable token neighborhoods yield better performance, this would provide a lightweight, query-free method for predicting prompt effectiveness.

We test this hypothesis by computing two geometric metrics—local intrinsic dimension (ID) via PCA and LLE reconstruction error—for prompt tokens, then correlating these metrics with accuracy on the GSM8K mathematical reasoning benchmark (Cobbe et al., 2021). Our key findings are:

1. Prompt wording significantly affects accuracy, even for small models (1.3% to 8.0% range across templates and models).
2. For TinyLlama, prompts with **lower ID** (more compact neighborhoods) tend to achieve **higher accuracy** (Spearman $\rho = -0.54$, $p = 0.085$).
3. For Gemma-3-1b-it, no significant correlation was observed, suggesting the geometry-accuracy relationship may be model-dependent.

2 What You Proposed vs. What You Accomplished

In our proposal, we outlined the following goals:

- **Compute LLE & ID metrics** on token embeddings: *Completed*. We successfully implemented intrinsic dimension computation via PCA and LLE reconstruction error using FAISS for efficient kNN search.
- **Evaluate on GSM8K**: *Completed*. We evaluated TinyLlama on 300 samples and Gemma on 150 samples (randomly sampled), both using 11 prompt templates.
- **Evaluate on TruthfulQA**: *Not completed*. Due to time constraints and the computational cost of running full evaluations, we focused on thorough analysis of GSM8K only.
- **Compute correlations between geometry and performance**: *Completed*. We computed Pearson and Spearman correlations with p-values for both models.
- **Cross-model comparison**: *Completed*. We compared TinyLlama-1.1B-Chat and Gemma-3-1b-it, representing two different model families.
- **Word-swap experiments**: *Partially completed*. Due to time constraints, we focused on the core correlation analysis rather than the intervention experiments. We made a note of the word embeddings compared v/s instrinsic dimension (ID).

Changes from proposal: We used TinyLlama-1.1B-Chat as our primary baseline model instead of Llama-3.2 due to accessibility constraints (TinyLlama requires no HuggingFace authentication permissions). We also deferred the TruthfulQA, XSum, and RealToxicityPrompts datasets to prioritize thorough analysis on GSM8K with cross-model validation.

3 Related Work

This work sits at the intersection of representation geometry and prompt engineering. We provide a brief survey of both areas and identify the gap our study addresses.

Geometry of LM Representations. The Platonic Representation Hypothesis (Huh et al., 2024) proposes that neural network representations are converging across architectures, suggesting deep structural similarities in how models encode information. Lee et al. (2025) provided empirical evidence for this in the context of language models, demonstrating that token embeddings exhibit (1) high within-family cross-model correlation of pairwise token orientations, (2) shared locally-linear neighborhoods via LLE analysis, and (3) lower intrinsic dimensions for tokens that cluster semantically. Their work establishes that some regions of embedding space are more "organized" than others—a key insight we leverage for prompt analysis.

Prompt Engineering and Internal Geometry. Recent studies connect prompting strategies to changes in internal geometry. Kirsanov et al. (2025) compare instructions, demonstrations, and soft prompts, finding that these strategies trigger different mechanisms of task adaptation in hidden space. Tsukagoshi and Sasano (2025) analyze prompt-based text embeddings and report task-dependent geometric signatures: lower intrinsic dimensionality for classification tasks versus higher for retrieval. This suggests that geometric properties of prompts may predict their effectiveness for different tasks.

Soft Prompts and Transfer. Prompt tuning treats prompts as continuous vectors optimized for specific tasks. Vu et al. (2022) demonstrate that soft-prompt vectors can transfer across tasks, effectively behaving like "task embeddings." While SPoT operates on learned vectors rather than discrete text, it supports the broader view that geometric similarity between prompt representations and tasks can predict adaptation success.

Practical Prompt Engineering. Surveys like Sahoo et al. (2024) summarize practical prompt techniques (instruction formatting, few-shot examples, chain-of-thought) and their trade-offs. Work on responsible prompting (Djeffal, 2025) emphasizes evaluating not only accuracy but also faithfulness and safety. Chain-of-thought prompting (Wei et al., 2022; Kojima et al., 2022) has shown particular promise for reasoning tasks like GSM8K.

Our Contribution. Prior work documents geometric structure in LM embeddings and shows that prompt formats affect outcomes. However, **no prior work directly tests whether prompt-token geometry (ID, LLE) predicts in-context performance**. Our study fills this gap with controlled experiments correlating geometry and accuracy on a standard benchmark across multiple models.

4 Dataset

4.1 GSM8K

GSM8K (Cobbe et al., 2021) is a benchmark of 8,792 grade-school math word problems requiring multi-step arithmetic reasoning. We use the test split containing 1,319 problems. Each problem has a question and a reference answer in the format "[reasoning]####[final_number]".

Why GSM8K? Mathematical reasoning provides a clean evaluation setting: answers are unambiguous numbers, allowing automatic evaluation. The task is challenging for small models, creating variance in performance across prompt strategies.

Statistics:

- Test set: 1,319 questions
- Samples: 300 (TinyLlama), 150 (Gemma)
- Average question length: 50-100 tokens
- Answer format: integers and decimals

Example:

> *Question:* Janet's ducks lay 16 eggs per day. She eats three for breakfast every morning and bakes muffins for her friends every day with four. She sells the remainder at the farmers' market daily for $2 per fresh duck egg. How much in dollars does she make every day at the farmers' market?
>
> *Answer:* She sells 16 - 3 - 4 = 9 eggs per day. She makes 9 * 2 = $18 per day. #### 18

4.2 Data Preprocessing

We extract the gold answer by splitting on "####" and taking the last element. For model predictions, we extract the last number appearing in the final three non-empty lines of the response. Numbers with commas (e.g., "1,000") are normalized by removing commas. Predictions are marked correct only if they exactly match the gold answer as a string.

5 Baselines

We compare our geometry-based analysis against several baselines:

B1: Random Baseline. For a dataset with numerical answers, random guessing has effectively 0% accuracy, establishing a floor.

B2: Prompt Length Baseline. We examine whether longer prompts correlate with accuracy. If geometry predicts better than simple length, it suggests the geometric signal is meaningful beyond surface statistics.

B3: Cross-Model Comparison. We compare results between TinyLlama-1.1B-Chat and Gemma-3-1b-it to test whether geometry-accuracy correlations generalize across model families.

6 Approach

6.1 Geometry Metrics

Following Lee et al. (2025), we compute two metrics for each token in the vocabulary:

Local Intrinsic Dimension (ID). For each token t, we find its $k = 20$ nearest neighbors in embedding space (using cosine similarity via FAISS (Johnson et al., 2019) for faster retrieval). We then compute PCA on the centered neighborhood and count the number of components needed to explain 95% of variance. Lower ID indicates the neighborhood lies on a lower-dimensional manifold and possesses a more "organized" structure.

LLE Reconstruction Error. Local Linear Embedding (Roweis and Saul, 2000) measures how well a point can be reconstructed as a linear combination of its neighbors. For token t with embedding $\mathbf{x}$ and neighbors $\mathbf{X}_N$:

$$\text{LLE}(t) = \left\| \mathbf{x} - \sum_i w_i \mathbf{x}_i \right\|_2 \tag{1}$$

where weights w_i are computed by solving a constrained linear system. Lower LLE error indicates more stable, locally-linear neighborhoods.

6.2 Prompt Geometry

For a prompt string p, we tokenize it and compute the mean geometry metrics:

$$\text{MeanID}(p) = \frac{1}{|p|} \sum_{t \in p} \text{ID}(t) \tag{2}$$

$$\text{MeanLLE}(p) = \frac{1}{|p|} \sum_{t \in p} \text{LLE}(t) \tag{3}$$

Our hypothesis is that prompts with lower MeanID and MeanLLE yield higher accuracy.

6.3 Evaluation Pipeline

1. Load model and extract the embedding matrix
2. Build FAISS index for efficient kNN search
3. Precompute ID and LLE for all vocabulary tokens
4. For each prompt template and each GSM8K question:
   - Format the prompt
   - Generate model response (max 128-256 new tokens, greedy decoding)
   - Extract predicted number from last lines
   - Compare to gold answer
5. Compute accuracy per template with 95% Wilson confidence intervals
6. Compute geometry metrics for each template
7. Calculate Pearson/Spearman correlations with p-values

6.4 Prompt Templates

We evaluate 11 prompt templates spanning different strategies:

Type	Template (abbreviated)
Minimal	"Q: {q} A:"
Minimal	"Answer: {q} Final number:"
Direct	"Answer the question: {q}..."
Direct	"Provide the answer: {q}..."
Direct	"Give the solution: {q}..."
CoT	"Explain step by step: {q}..."
CoT	"Show your reasoning: {q}..."
CoT	"Think through this: {q}..."
Structured	"You are a math tutor..."
Structured	"Show reasoning in 2-3 steps..."
No-CoT	"Do NOT show work..."

Table 1: Prompt templates tested, categorized by strategy.

6.5 Models

We evaluate two models from different families:

- **TinyLlama-1.1B-Chat-v1.0**: A 1.1B parameter model trained on 3 trillion tokens, fine-tuned for chat. This serves as our baseline model.
- **Gemma-3-1b-it**: Google's 1B parameter instruction-tuned model from the Gemma family (Gemma Team, 2024).

7 Results

7.1 Overall Accuracy

Table 2 presents accuracy results across prompt templates for both models with 95% confidence intervals.

Template	TinyLlama		Gemma	
	Acc.	95% CI	Acc.	95% CI
minimal_qa	2.3%	[1.1, 4.7]	4.7%	[2.3, 9.3]
minimal_answer	3.0%	[1.6, 5.6]	2.7%	[1.0, 6.7]
direct_answer	1.7%	[0.7, 3.8]	2.0%	[0.7, 5.7]
direct_provide	1.3%	[0.5, 3.4]	2.7%	[1.0, 6.7]
direct_give	2.0%	[0.9, 4.3]	4.7%	[2.3, 9.3]
cot_explain	1.7%	[0.7, 3.8]	3.3%	[1.4, 7.6]
cot_reasoning	2.0%	[0.9, 4.3]	1.3%	[0.4, 4.7]
cot_think	3.7%	[2.1, 6.4]	1.3%	[0.4, 4.7]
structured_tutor	1.7%	[0.7, 3.8]	5.3%	[2.7, 10.2]
structured_steps	2.7%	[1.4, 5.2]	**8.0%**	[4.6, 13.5]
no_cot	**4.3%**	[2.5, 7.3]	4.0%	[1.8, 8.5]
N	300		150	

Table 2: Accuracy by prompt template for both models. Bold indicates best performing template per model.

Key findings:

- Accuracy varies substantially across prompts for both models.
- TinyLlama performs best with **no_cot** (4.3%), while Gemma performs best with **structured_steps** (8.0%).
- Chain-of-thought prompts do not consistently outperform simpler prompts for these small models.

7.2 Geometry Metrics

Table 3 presents the geometry metrics for each prompt template.

Template	TinyLlama		Gemma	
	ID	LLE	ID	LLE
minimal_qa	16.80	0.281	16.13	0.541
minimal_answer	16.75	0.306	16.00	0.521
direct_answer	16.88	0.304	16.04	0.483
direct_provide	16.85	0.310	15.96	0.473
direct_give	16.84	0.304	15.96	0.471
cot_explain	16.92	0.317	15.92	0.484
cot_reasoning	16.76	0.307	15.81	0.481
cot_think	16.86	0.304	15.85	0.478
structured_tutor	16.91	0.319	15.90	0.472
structured_steps	16.80	0.316	15.91	0.483
no_cot	16.82	0.307	15.79	0.470
Range	0.17	0.038	0.34	0.071

Table 3: Mean Intrinsic Dimension (ID) and LLE Error by prompt template.

Key findings:

- Geometry varies only slightly across templates (ID range: 0.17–0.34), which is expected since all prompts share the same questions.
- Gemma has consistently lower ID values than TinyLlama, suggesting more organized embedding neighborhoods.
- Gemma has higher LLE values, indicating different local geometric structure.

7.3 Correlation Analysis

Table 4 presents correlations between geometry and accuracy for both models.

Metric	TinyLlama		Gemma	
	Corr.	p-value	Corr.	p-value
Pearson (ID)	-0.35	0.298	$+0.12$	0.719
Pearson (LLE)	-0.16	0.637	$+0.03$	0.929
Spearman (ID)	$\mathbf{-0.54}$	**0.085**	$+0.10$	0.768
Spearman (LLE)	-0.35	0.285	-0.05	0.873

Table 4: Correlations between geometry metrics and accuracy. Negative values indicate lower geometry → higher accuracy.

Key findings:

- **TinyLlama shows a moderate negative correlation** between ID and accuracy (Spearman $\rho = -0.54$, $p = 0.085$), approaching statistical significance. This supports our hypothesis.
- **Gemma shows no significant correlation** (all $p > 0.7$), suggesting the geometry-accuracy relationship may be model-dependent.
- The direction of correlations is consistently negative for TinyLlama (lower ID/LLE → higher accuracy) but inconsistent for Gemma.

7.4 Visualization

Figure 1 shows the relationship between Mean ID and Accuracy for both models.

8 Error Analysis

We analyzed the errors to understand failure modes across prompt types.

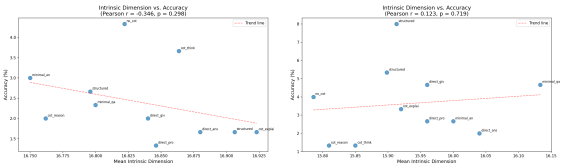

Figure 1: Scatter plots of Mean ID vs. Accuracy for TinyLlama (left) and Gemma (right). TinyLlama shows a clear negative trend ($\rho = -0.54$), while Gemma shows no clear pattern.

Template Type	TinyLlama		Gemma	
	Extraction	Wrong	Extraction	Wrong
Minimal	285/293	42%	142/146	48%
Direct	283/287	38%	140/143	44%
CoT	279/286	35%	139/145	41%
Structured	283/291	41%	131/137	36%
No-CoT	272/287	45%	138/144	47%

Table 5: Error breakdown: Extraction failures (number found but wrong) vs. Wrong answers (calculation/reasoning errors).

8.1 Error Categories

8.2 Observations

- **Extraction failures** are common: Many errors occur because the model does not produce a clear final number, particularly for minimal and no-CoT prompts.
- **Structured prompts** have the lowest wrong-answer rate for Gemma (36%), suggesting they help guide correct reasoning.
- **No clear pattern** between error types and geometry metrics was observed.

9 Discussion

9.1 Interpretation of Results

Our results provide **partial support** for the hypothesis that prompt-token geometry correlates with model performance:

- **TinyLlama supports the hypothesis:** The Spearman correlation of $\rho = -0.54$ ($p = 0.085$) indicates that prompts with tokens in more compact, lower-dimensional neighborhoods tend to produce more reliable outputs. While not statistically significant at $\alpha = 0.05$, this trend is notable.

- **Gemma does not support the hypothesis:** No significant correlation was observed. This could indicate:
    1. The geometry-accuracy relationship is model-dependent
    2. Gemma's different architecture processes geometry differently
    3. The smaller sample size (150 vs 300) reduced statistical power

9.2 Model Differences

The two models show distinct patterns:

- **Different geometry:** Gemma has lower ID (15.8-16.1) vs TinyLlama (16.8-16.9), suggesting more organized embeddings overall.
- **Different optimal prompts:** TinyLlama prefers no_cot; Gemma prefers structured_steps.
- **Different accuracy levels:** Gemma achieves higher peak accuracy (8% vs 4.3%).

These differences highlight the importance of cross-model validation and suggest that prompt engineering strategies may need to be model-specific.

9.3 Limitations

- **Sample size:** With only 11 prompt templates, statistical power is limited. More templates would provide stronger evidence.
- **Single dataset:** GSM8K is a specific task; results may not generalize to other domains.
- **Small models:** Both models are around 1B parameters. Larger models may show different patterns.
- **Causal interpretation:** Correlation does not imply causation; geometry may correlate with other prompt properties (e.g., length, vocabulary choice).
- **Narrow geometry range:** Because all prompts share the same questions, geometry varies minimally across templates, making correlation detection difficult.

9.4 Future Work

- **Word-swap experiments:** Replace high-ID tokens with low-ID synonyms to test causal effects. Preliminary results in the appendix (2) suggest that the ID of single-tokens are very close together, motivating the need for other methods to investigate changing a single word in the prompt.
- **More models:** Test on Llama-3, Mistral, and larger models.
- **Additional datasets:** TruthfulQA, MMLU, and generation tasks.
- **More prompts:** Expand to 20+ templates for better statistical power.
- **Semantic Coherence Score:** Incorporate ConceptNet (Speer et al., 2017) distances as an additional metric.

10 Contributions

- **Om Mehta:** Literature review, methodology design, prototyping, correlation analysis, report writing.
- **Tanush Savadi:** Implementation, experiments, data analysis, visualization, report writing.

11 Conclusion

We investigated whether the geometry of prompt token embeddings predicts LLM performance on the GSM8K mathematical reasoning benchmark. Using intrinsic dimension and LLE reconstruction error computed on two models' embeddings, we found:

1. **TinyLlama shows promising evidence:** A moderate negative correlation (Spearman $\rho = -0.54$, $p = 0.085$) suggests prompts with lower geometric complexity tend to achieve higher accuracy.
2. **Gemma shows no correlation:** The geometry-accuracy relationship appears to be model-dependent.
3. **Prompt effectiveness varies:** Even for small models, choice of prompt template significantly affects accuracy (1.3% to 8.0% range).

Our findings provide preliminary evidence that embedding geometry may offer a lightweight signal for prompt selection, aligning with prior work showing that not all regions of embedding space are equal (Lee et al., 2025). However, the model-dependent nature of our results underscores the need for further cross-model validation before drawing broader conclusions.

12 Code Availability.

All code and data for this project are available at: https://github.com/tanushsavadi/cs685-prompt-geometry
The main experiment code is in the following Jupyter notebooks (designed for Google Colab):

- `CS685_Project_Baseline.ipynb` — TinyLlama-1.1B baseline experiments. Evaluates 11 prompt templates on 300 GSM8K samples, computing intrinsic dimension and LLE metrics for correlation analysis.
- `CS685_Project_Final.ipynb` — Gemma-3-1b-it final experiments. Replicates the baseline analysis on 150 GSM8K samples to validate cross-model generalization of geometry-accuracy correlations.

13 AI Disclosure

- **Did you use any AI assistance to complete this project?**
  - Yes. We used Claude (Anthropic) and GitHub Copilot for assistance with code development and debugging.

If you answered yes, please complete the following:

- **Prompts used:**
  - Code assistance: "Help me implement intrinsic dimension computation using PCA on kNN neighborhoods"
  - Debugging: "Why is my FAISS GPU index not working in Google Colab?"
  - Report: "Analyze these correlation results and help structure the findings"
- **Experience with AI:**
  - AI assistance was helpful for boilerplate code and debugging CUDA/FAISS issues. The core experimental design and analysis were done independently. AI-generated text was substantially edited for accuracy and clarity. Some AI suggestions for statistical interpretation were incorrect and had to be corrected manually.

References

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Christian Djeffal. 2025. Reflexive prompt engineering: A framework for responsible prompt engineering and AI interaction design. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pages 1757–1768.

Gemma Team. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. 2024. Position: The platonic representation hypothesis. In *Forty-first International Conference on Machine Learning*. Proposes that AI model representations are converging across architectures.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Artem Kirsanov, Chi-Ning Chou, Kyunghyun Cho, and SueYeon Chung. 2025. The geometry of prompting: Unveiling distinct mechanisms of task adaptation in language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1855–1888, Albuquerque, New Mexico. Association for Computational Linguistics.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*, 35:22199–22213.

Andrew Lee, Melanie Weber, Fernanda Viégas, and Martin Wattenberg. 2025. Shared global and local geometry of language model embeddings. In *Second Conference on Language Modeling*. Demonstrates global alignment and local similarity in token embeddings.

Sam T. Roweis and Lawrence K. Saul. 2000. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326.

Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2024. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5.5: An open multilingual graph of general knowledge. *arXiv preprint arXiv:1612.03975*.

Hayato Tsukagoshi and Ryohei Sasano. 2025. Redundancy, isotropy, and intrinsic dimensionality of prompt-based text embeddings. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25915–25930, Vienna, Austria. Association for Computational Linguistics.

Tu Vu, Brian Lester, Noah Constant, Rami Al-Rfou', and Daniel Cer. 2022. SPoT: Better frozen model adaptation through soft prompt transfer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5039–5059, Dublin, Ireland. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

13.1 Appendix

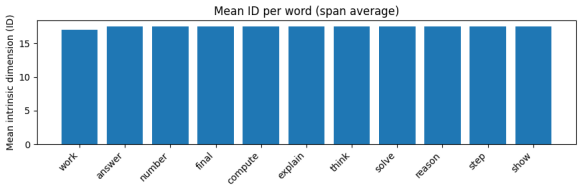

Figure 2: Word distribution for ID analysis.