Learning from Fewer Prompts: Dimensionality Reduction and Optimal Coverage in LLM Prompt Embedding Space

2025-04-02

Introduction

Large Language Models (LLMs) have demonstrated exceptional capabilities in performing diverse tasks ranging from natural language understanding to program synthesis and logical reasoning. Yet, despite their staggering parameter counts and high-dimensional latent spaces, their ability to generalize from examples provided during inference—so-called *in-context learning*—remains a probabilistic endeavor. The problem we tackle here is foundational and broadly relevant: **given a function** $f(x)$, how do we select a minimal and optimal set of example pairs (x_i, y_i) such that a language model can best infer $f(x_n) \approx y_n$ for a new input x_n?

This is not merely a question of sample size or model capacity. Rather, it is a question of *information geometry*—how examples are represented in the model's internal embedding space and how efficiently these examples span that space to allow extrapolation or interpolation. Our research frames this inquiry as a problem in dimension reduction and optimal coverage: **How do we select a subset of prompts whose internal embeddings are maximally informative and sufficiently orthogonal to each other, so as to span the model's functional understanding of a domain?**

Our research develops a principled approach to prompt selection grounded in statistical learning theory and empirical embeddings. We propose an iterative, embedding-driven framework to select a subset of prompt examples whose final-token embeddings (denoted $\mathbf{e}_i$) are representative of the broader function space. We examine reduction techniques such as PCA and SVD to analyze and visualize these embeddings and explore how we might use gradients in reduced space to iteratively discover new, maximally informative prompts.

Motivating Example: Natural Language to SQL

To make this abstract framing concrete, consider the task of translating natural language queries into SQL queries for a specific database. A user may ask:

> "List all customers who made a purchase in the past month."

This request is passed to an LLM which generates the corresponding SQL:

```sql
SELECT * FROM Customers WHERE purchase_date >= DATE_SUB(CURDATE(), INTERVAL 1 MONTH);
```

Now suppose we have access to a large dataset of 10,000 such input-output pairs. Each input is a natural language question x_i, and each output $y_i = f(x_i)$ is a valid SQL translation. We aim to use a small number of these as in-context examples in a prompt that allows the LLM to generalize and correctly map a new natural language request x_n to $y_n \approx f(x_n)$.

The central insight is that the LLM processes the entire prompt into a high-dimensional internal representation. After processing the sequence of tokens in the prompt, the model produces a final-token embedding vector $\mathbf{e}_i \in \mathbb{R}^d$, where d is typically 2048 in models like LLaMA-3 1B. These embeddings capture the model's distilled understanding of the prompt and its influence on the next predicted token.

If we can extract these $\mathbf{e}_i$ vectors, we can project them into a lower-dimensional space using techniques like PCA or SVD. In this space, we can evaluate how "spread out" or "redundant" the prompt examples are, and select those that maximize diversity and coverage of the function space.

Mathematical Framework

Let us formally define our problem. Suppose we have:

- A target function $f : \mathcal{X} \to \mathcal{Y}$ which maps natural language inputs $x \in \mathcal{X}$ to structured outputs $y \in \mathcal{Y}$ (e.g., SQL queries).
- A dataset $\{(x_i, y_i)\}_{i=1}^N$, where N is large (e.g., 10,000).
- An LLM that, given a prompt $P = \{(x_{j_1}, y_{j_1}), \ldots, (x_{j_k}, y_{j_k})\}$, attempts to infer y_n for a new input x_n.

Let $\mathbf{e}_i \in \mathbb{R}^d$ be the final-token embedding for the full prompt that includes the pair (x_i, y_i). Our goal is to find a set $S = \{j_1, \ldots, j_k\} \subset \{1, \ldots, N\}$, such that:

1. The span of $\{\mathbf{e}_j\}_{j \in S}$ covers the main axes of variation in the full dataset $\{\mathbf{e}_i\}_{i=1}^N$,
2. The examples in S are minimally redundant, ideally close to orthogonal in the reduced space,
3. The model's performance on predicting $y_n = f(x_n)$ is maximized when conditioned on P_S.

This is analogous to active learning and optimal experimental design: we seek the smallest set of examples that best characterize the underlying structure of the task.

Johnson–Lindenstrauss lemma

The lemma states that a set of points in a high-dimensional space can be embedded into a space of much lower dimension in such a way that distances between the points are nearly preserved. In the classical proof of the lemma, the embedding is a random orthogonal projection.

Given $0 < \varepsilon < 1$, a set X of N points in $\mathbb{R}^n$, and an integer $k > 8(\ln N)/\varepsilon^2$, there is a linear map $f : \mathbb{R}^n \to \mathbb{R}^k$ such that

$$(1 - \varepsilon)\|u - v\|^2 \leq \|f(u) - f(v)\|^2 \leq (1 + \varepsilon)\|u - v\|^2$$

for all $u, v \in X$.

The formula can be rearranged:

$$(1 + \varepsilon)^{-1}\|f(u) - f(v)\|^2 \leq \|u - v\|^2 \leq (1 - \varepsilon)^{-1}\|f(u) - f(v)\|^2$$

Alternatively, for any $\varepsilon \in (0, 1)$ and any integer $k \geq 15(\ln N)/\varepsilon^2$ there exists a linear function $f : \mathbb{R}^n \to \mathbb{R}^k$ such that the restriction $f|_X$ is $(1 + \varepsilon)$-bi-Lipschitz.

Also, the lemma is tight up to a constant factor, i.e. there exists a set of points of size N that needs dimension

$$\Omega\left(\frac{\log(N)}{\varepsilon^2}\right)$$

in order to preserve the distances between all pairs of points within a factor of $(1 \pm \varepsilon)$.

Dimensionality Reduction and Coverage

We apply PCA or SVD to the matrix $E \in \mathbb{R}^{N \times d}$, where each row is $\mathbf{e}_i^\top$. The principal components give us a lower-dimensional representation that preserves maximal variance. In this reduced space, we define:

- **Coverage**: the proportion of variance explained by a subset S,
- **Diversity**: measured by average pairwise cosine distance among $\mathbf{e}_j \in S$,
- **Redundancy**: minimized when eigenvalue decay is steep, indicating that only a few components dominate.

Reservoir Sampling and Iterative Selection

We develop a reservoir-style iterative selection algorithm:

1. Sample an initial subset S_0 of m prompt pairs.
2. Compute PCA on $\{\mathbf{e}_j\}_{j \in S_0}$ and retain the top k components.
3. Iterate through the remaining $\mathbf{e}_i$ vectors:
    - If including $\mathbf{e}_i$ improves variance explained or increases orthogonality, include it and remove a redundant member of S.
4. Terminate when a fixed number of iterations pass without improvement.

This procedure approximates optimal subset selection under constraints.

Beyond PCA: Gradient-Based Search

Alternatively, we may interpret the reduced space as a landscape. Given a current subset S, we compute the gradient of a coverage or loss metric with respect to the prompt embeddings and identify directions in the reduced space where we lack coverage. We then identify $\mathbf{e}_i$ vectors in the dataset that lie in these unexplored directions. This suggests a greedy approach for discovering new prompts that expand the LLM's learned function representation most effectively.

Pipeline

1. **Data Embedding via LLM**: the pipeline starts by embedding natural language inputs paired with their structured outputs (x_i, y_i) using a large language model (LLM). These embeddings $\mathbf{e}_i \in \mathbb{R}^d$ represent the final-token embeddings, capturing the semantic content of the prompt-response pair.

2. **Dimensionality Reduction with Johnson-Lindenstrauss (JL) Lemma**: A key step is dimensionality reduction using a random projection or similar linear transformation (e.g., PCA or JL-based). This maps $\mathbf{e}_i$ to a lower-dimensional space $\mathbb{R}^k$ while preserving pairwise distances within a $(1 \pm \varepsilon)$ factor.JL lemma guarantees that if $k = O(\log N/\varepsilon^2)$, then distance preservation holds with high probability.

3. **Selection via Diversity and Coverage Criteria**: In the reduced space, select a subset $S = j_1, \ldots, j_k$ of examples that span the main variations of the dataset and are minimally redundant (Method: K-means, PCA, Greedy submodular selection).

4. **Prompt Construction**: The selected examples S are used to construct a prompt $P_S = (x_{j_1}, y_{j_1}), \ldots, (x_{j_k}, y_{j_k})$. This prompt is fed into the LLM to generate a prediction $\hat{y}_n$ for a new query x_n.

5. **Evaluation**: The performance is evaluated (likely using accuracy or execution match for structured outputs), validating whether the selected subset S is effective.

Conclusion and Broader Significance

This research combines classical optimal design theory and modern in-context learning in LLMs. Rather than treating prompts as heuristic and ad hoc, we propose a framework to make them analytically tractable, geometrically interpretable, and empirically efficient. The broader implications are substantial. If a small set of prompts can be chosen to maximize coverage of a task's embedding space, we gain a principled method for constructing teaching sets, improving model performance on specific domains, and reducing reliance on brute-force prompt enumeration. For practitioners, this offers not only efficiency but clarity: a model that can learn from fewer, better examples is a model we can better understand and trust.