(Google Deepmind 2025) On the Theoretical Limitations of Embedding-Based Retrieval

October 14, 2025

Seoul National University

Background

- Can single-vector embeddings represent **all combinations** of query-document relevance?
- The qrel matrix (queries x documents with 0/1 labels) generally cannot be exactly represented with fixed dimentional embeddings. The limitation is quantified by the matrix's sign-rank.

Theoretical limitation

- Consider m queries and n documents
- Ground-truth matrix $A \in \{0,1\}^{m \times n}$, where $A_{ij} = 1$ if and only if document j is relevant to query i.
- Vector embedding models map each query to a vector $u_i \in \mathbb{R}^d$ and each document to a vector $v_j \in \mathbb{R}^d$.
- Relevance is modeled by the dot product $u_i^T v_j$
- Concatenating the vectors for queries in a matrix $U \in \mathbb{R}^{d \times m}$ and those for documents in a matrix $V \in \mathbb{R}^{d \times n}$
- Score matrix $B = U^T V$.
- The smallest embedding dimension d that can realize a given score matrix is the rank of B.
- Therefore, our goal is equivalent to finding the minimum rank of a score matrix B that correctly orders documents according to the relevance specified in A.

Definition 1 (Row-wise order-preserving rank)

Given a matrix $A \in \mathbb{R}^{m \times n}$, the *row-wise order-preserving rank* of A is the smallest integer d such that there exists a rank-d matrix $B \in \mathbb{R}^{m \times n}$ that preserves the relative order of entries in each row of A. We denote this as

$$\operatorname{rank}_{\mathrm{rop}} A = \min\left\{ \operatorname{rank} B \;\middle|\; B \in \mathbb{R}^{m \times n},\ \forall i, j, k,\ A_{ij} > A_{ik} \Rightarrow B_{ij} > B_{ik} \right\}.$$

Theoretical limitation

Definition 2 (Thresholdable ranks)

Given a binary matrix $A \in \{0,1\}^{m \times n}$:

- **Row-wise thresholdable rank** of A ($\mathrm{rank}_{\mathrm{rt}}\, A$) is the minimum rank of a matrix B for which there exist row-specific thresholds $\{\tau_i\}_{i=1}^{m}$ such that for all i, j, $B_{ij} > \tau_i$ if $A_{ij} = 1$ and $B_{ij} < \tau_i$ if $A_{ij} = 0$.
- **Globally thresholdable rank** of A ($\mathrm{rank}_{\mathrm{gt}}\, A$) is the minimum rank of a matrix B for which there exists a single threshold τ such that for all i, j, $B_{ij} > \tau$ if $A_{ij} = 1$ and $B_{ij} < \tau$ if $A_{ij} = 0$.

Proposition 3

For a binary matrix $A \in \{0,1\}^{m \times n}$, *we have* $\mathrm{rank}_{\mathrm{rop}}\, A = \mathrm{rank}_{\mathrm{rt}}\, A$.

Theoretical limitation

Definition 4 (Sign Rank)

The sign rank of a matrix $M \in \{-1, 1\}^{m \times n}$ is the smallest integer d s.t. there exists a rank d matrix $B \in \mathbb{R}^{m \times n}$ whose entries have the same sign as those of M, i.e.

$$\mathrm{rank}_{\pm}(M) = \min\{\mathrm{rank}(B) \mid B \in \mathbb{R}^{m \times n},\ \mathrm{sign}\, B_{ij} = M_{ij}\ \forall i, j\}.$$

Proposition 5

Let $A \in \{0, 1\}^{m \times n}$ be a binary matrix. Then $2A - 1_{m \times n} \in \{-1, 1\}^{m \times n}$, and we have

$$\mathrm{rank}_{\pm}\left(2A - 1_{m \times n}\right) - 1 \leq \mathrm{rank}_{\mathrm{rop}}\, A = \mathrm{rank}_{\mathrm{rt}}\, A \leq \mathrm{rank}_{\mathrm{gt}}\, A \leq \mathrm{rank}_{\pm}\left(2A - 1_{m \times n}\right).$$

Experiments-Free Embedding

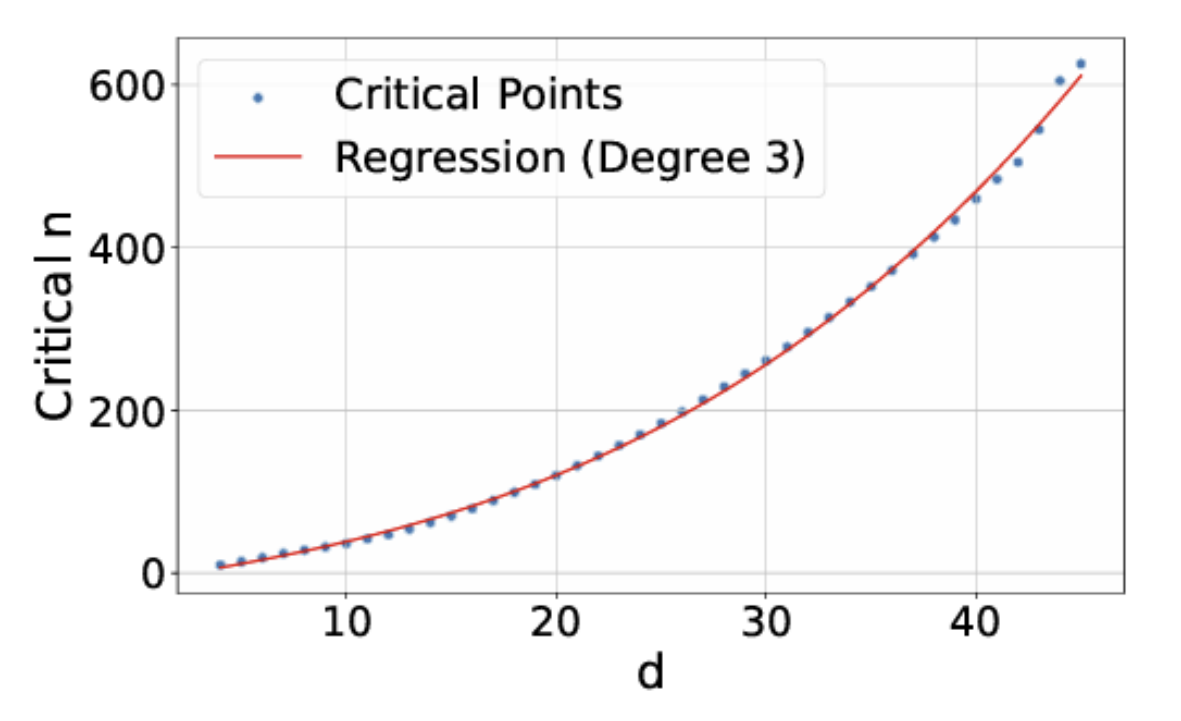

Figure 1: The critical-n value where the dimensionality is too small to successfully represent all the top-2 combinations.

Experiments: LIMIT Dataset

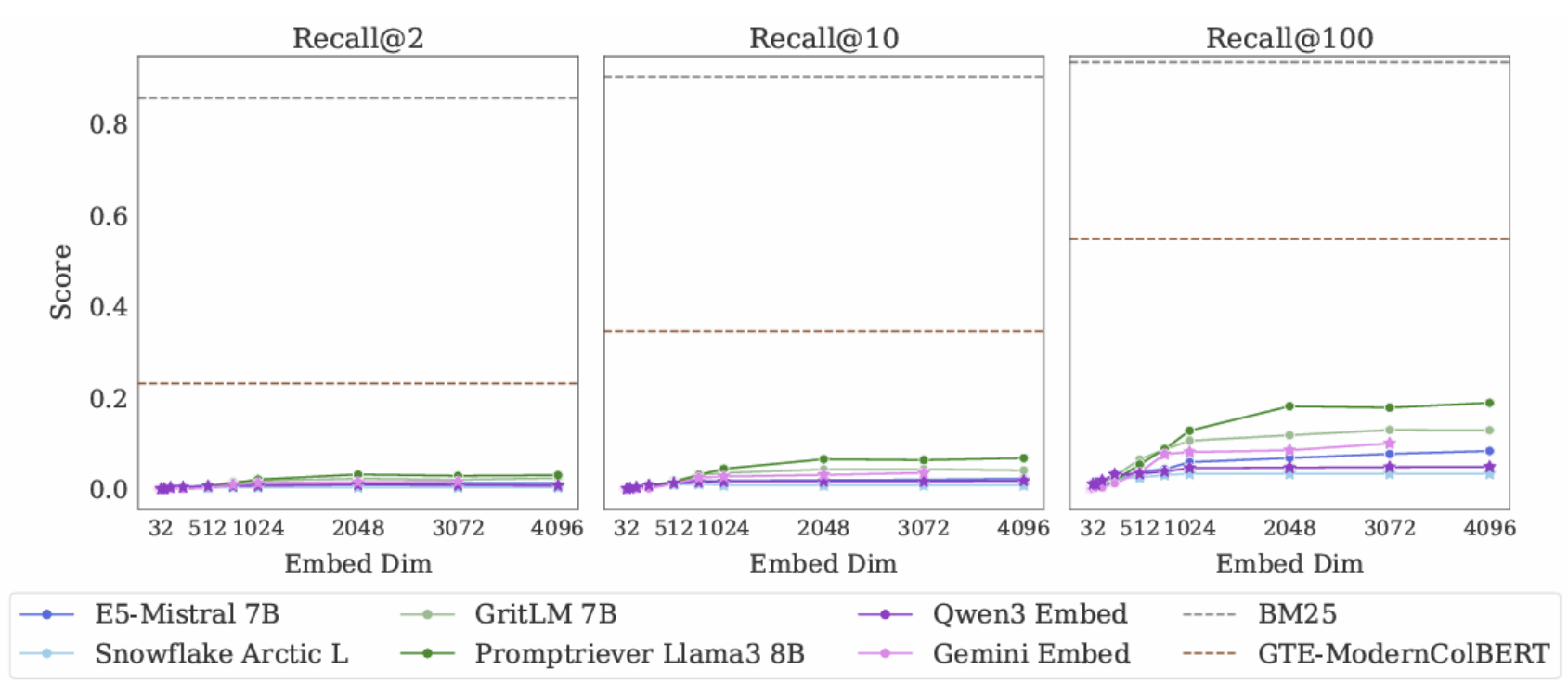

Figure 2: Scores on the LIMIT task. Despite the simplicity of the task we see that SOTA models struggle.

Thank you!

Appendix-LIMIT Dataset

- Short docs (e.g, "Geneva Durben likes Quokkas, River Otters, ...)
- Simple queries: (e.g, who likes X?)
- 50k documents. 1,000 queries.
- $k = 2$ relevant docs per query

Appendix-Recall@{k}

- Recall@k $= \frac{\#\{\text{relevant docs in TOP-k}\}}{R}$, where R is the total number of relevant for that query.
- Interpretation: "Within the top-k results, what fraction of all relevant documents did we recover?"
- For LIMIT dataset(full):

$$\text{Recall@2} = \frac{1}{1000} \sum_{q} \frac{\|\text{Top-2(q)} \cap R_q\|}{2}$$

Appendix-Free Embedding Overview

- **What is free embedding?** Queries/documents are **trainable vectors** (no text encoder); we optimize them directly.
- **Data:** Documents are n IDs (no text). Each query corresponds to a **combination of k docs**; its relevant set is exactly those k docs.
- **Training Input:**
  - Current query embeddings $Q \in \mathbb{R}^{q \times d}$ and document embeddings $D \in \mathbb{R}^{n \times d}$ (both are learnable)
  - qrel binary matrix $A \in \{0,1\}^{q \times n}$ or a dict (query $\rightarrow$ set of gold doc IDs)
- **Training Output:**
  - Loss $\mathcal{L}$ (InfoNCE with all-docs negatives)

$$\mathcal{L}_{\text{total}} = -\frac{1}{M} \sum_{i=1}^{M} \log \frac{\sum_{d_r \in R_i} \exp(\text{sim}(q_i, d_r)/\tau)}{\sum_{d_k \in D} \exp(\text{sim}(q_i, d_k)/\tau)}.$$

    where d_r is the relevant documents for query q_i and d_k are the non-relevant documents.
  - Updated embeddings Q, D (after L_2 normalization and gradient updates)

Appendix-Free Embedding Overview

- **Objective:** For every query, push its k relevant docs into the top-k by dot-product similarity.
- **Evaluation (Recall@k):** A query scores 1 if all k gold docs appear in the top-k; otherwise partial credit. With $k = 2$, Recall@2 $= 1.0$ means all constraints are perfectly satisfied.
- **Finding critical-n:** Fix dimension d, increase n, and locate the largest n for which Recall@k $= 1.0$ remains achievable (galloping $\rightarrow$ binary search $\rightarrow$ local sweep).
- **Why it's an upper bound:** This is the **most favorable (cheat-mode)** setting—no language constraints, vectors can move freely. If it fails here, real language embeddings with the same d won't do better. Hence, free embedding measures the **expressivity upper bound** of single-vector models.

Appendix-Prop 1.

Proposition 1

For a binary matrix $A \in \{0,1\}^{m\times n}$, we have $\mathrm{rank}_{\mathrm{rop}}(A) = \mathrm{rank}_{\mathrm{rt}}(A)$.

Proof. ($\leq$) Suppose B and thresholds $\tau = \{\tau_i\}$ satisfy the row-wise thresholdable rank condition. Since A is binary, $A_{ij} > A_{ik}$ implies $A_{ij} = 1$ and $A_{ik} = 0$, thus $B_{ij} > \tau_i > B_{ik}$, and hence B also satisfies the row-wise order-preserving condition.

($\geq$) Let B satisfy the row-wise order-preserving condition, so $A_{ij} > A_{ik}$ implies $B_{ij} > B_{ik}$. For each row i, let $U_i = \{B_{ij} \mid A_{ij} = 1\}$ and $L_i = \{B_{ij} \mid A_{ij} = 0\}$. The row-wise order-preserving condition implies that every element of U_i is greater than every element of L_i. We can therefore always find a threshold τ_i separating them (e.g., $\tau_i = (\max L_i + \min U_i)/2$ if both are non-empty, trivial otherwise). Thus B is also row-wise thresholdable to A. $\square$

Appendix-Prop 2.

Proposition 2

Let $A \in \{0,1\}^{m\times n}$ be a binary matrix. Then $2A - \mathbf{1}_{m\times n} \in \{-1,1\}^{m\times n}$, and

$$\operatorname{rank}_{\pm}(2A-\mathbf{1}_{m\times n})-1 \;\le\; \operatorname{rank}_{\mathrm{rop}} A = \operatorname{rank}_{\mathrm{rt}} A \;\le\; \operatorname{rank}_{\mathrm{gt}} A \;\le\; \operatorname{rank}_{\pm}(2A-\mathbf{1}_{m\times n}).$$

Proof. We prove each inequality separately.

1. $\operatorname{rank}_{\mathrm{rt}} A \le \operatorname{rank}_{\mathrm{gt}} A$: By definition, any matrix satisfying the globally thresholdable condition trivially satisfies the row-wise thresholdable condition with the same threshold for each row.
2. $\operatorname{rank}_{\mathrm{gt}} A \le \operatorname{rank}_{\pm}(2A - \mathbf{1}_{m\times n})$: Let B be any matrix whose entries have the same sign as $2A - \mathbf{1}_{m\times n}$. Then

$$B_{ij} > 0 \iff 2A_{ij} - 1 > 0 \iff A_{ij} = 1.$$

   Thus B satisfies the globally thresholdable condition with a threshold of 0.

Combining these gives the desired chain of inequalities.

Appendix-Prop 2.

3. $\operatorname{rank}_{\pm}(2A - \mathbf{1}_{m\times n}) - 1 \le \operatorname{rank}_{\mathrm{rt}} A$: Suppose B satisfies the row-wise thresholding condition with minimal rank, so $\operatorname{rank}_{\mathrm{rt}} A = \operatorname{rank}(B)$. Then there exists $\tau \in \mathbb{R}^m$ such that $B_{ij} > \tau_i$ if $A_{ij} = 1$ and $B_{ij} < \tau_i$ if $A_{ij} = 0$. Hence the entries of $B - \tau \mathbf{1}_n^\top$ have the same sign as $2A - \mathbf{1}_{m\times n}$, since $(B - \tau \mathbf{1}_n^\top)_{ij} = B_{ij} - \tau_i$ and

$$B_{ij} - \tau_i > 0 \iff A_{ij} = 1 \iff 2A_{ij} - 1 > 0,$$
$$B_{ij} - \tau_i < 0 \iff A_{ij} = 0 \iff 2A_{ij} - 1 < 0.$$

$$\begin{aligned}\operatorname{rank}_{\pm}(2A - \mathbf{1}_{m\times n}) &\le \operatorname{rank}(B - \tau \mathbf{1}_n^\top) \\ &\le \operatorname{rank}(B) + \operatorname{rank}(\tau \mathbf{1}_n^\top) \\ &= \operatorname{rank}_{\mathrm{rt}} A + 1. \ \square\end{aligned}$$