Novelty Assessment Report

Paper: Learning to Recall with Transformers Beyond Orthogonal Embeddings
PDF URL: https://openreview.net/pdf?id=CfFj68C9Cn
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-30

Abstract

Modern large language models (LLMs) excel at tasks that require storing and retrieving knowledge, such as factual recall and question answering. Transformers are central to this capability, thanks to their ability to encode information during training and retrieve it at inference. Existing theoretical analyses typically study transformers under idealized assumptions such as infinite data or orthogonal embeddings. In realistic settings, however, models are trained on finite datasets with non-orthogonal (random) embeddings. We address this gap by analyzing a single-layer transformer with random embeddings trained with (empirical) gradient descent on a simple token-retrieval task, where the model must identify an informative token within a length-L sequence and learn a one-to-one mapping from tokens to labels. Our analysis tracks the "early phase" of gradient descent and yields explicit formulas for the model's storage capacity—revealing a multiplicative dependence between sample size N, embedding dimension D, and sequence length L. We complement this with a lower bound for the statistical problem, showing that this multiplicative scaling is inherent under non-orthogonal embeddings.

Core Task Landscape

This paper addresses: **Token Retrieval and Factual Recall in Transformers with Random Embeddings**

A total of **4 papers** were analyzed and organized into a taxonomy with **4 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Theoretical Analysis of Transformer Learning Dynamics**
- **Application-Driven Retrieval Systems**

Complete Taxonomy Tree

- Token Retrieval and Factual Recall in Transformers with Random Embeddings Survey Taxonomy
- Theoretical Analysis of Transformer Learning Dynamics
  - Gradient Descent Analysis with Non-Orthogonal Embeddings ★ (1 papers)
  - [0] Learning to Recall with Transformers Beyond Orthogonal Embeddings (Anon et al., 2026) View paper
  - Memory Mechanisms and Topological Embeddings (1 papers)
  - [4] Stochastic Topological Memory Embedding in Large Language Models: An Empirical Analysis Using Open-Source Neural Architectures (Connor Travis, n.d.) View paper
- Application-Driven Retrieval Systems
  - Text-Based Information Retrieval (2 papers)
  - [1] Domain-Specific Text Embedding Models for Information Retrieval (Kasmaee, 2025) View paper
  - [3] Enhancing Term-Based Document Retrieval by Word Embedding and Transformer Models (Farzana, 2021) View paper
  - Multimodal and Vision-Based Retrieval (1 papers)
  - [2] Vision Transformer-based Context-Aware System for Lingual Ultrasound in Digital Health Ecosystem (Al-hammuri, 2024) View paper

Narrative

Core task: token retrieval and factual recall in transformers with random embeddings. The field divides into two main branches. The first, Theoretical Analysis of Transformer Learning Dynamics, investigates how transformers learn to retrieve tokens and recall facts when embeddings are not carefully initialized or orthogonal, focusing on gradient descent behavior, convergence properties, and the interplay between attention mechanisms and non-orthogonal representations. The second branch, Application-Driven Retrieval Systems, emphasizes practical embedding methods and retrieval architectures for domain-specific tasks, including specialized text embedding approaches and vision-language integration. While the theoretical branch seeks to understand learning dynamics from first principles, the application-driven branch prioritizes performance on real-world retrieval benchmarks, often leveraging pre-trained or domain-adapted embeddings.

A central tension in the field concerns whether theoretical insights into random or non-orthogonal embeddings can inform practical retrieval system design. Works such as Domain-Specific Text Embedding[1] and Enhancing Retrieval Word Embedding[3] demonstrate that carefully tuned embeddings yield strong empirical results, yet they typically assume well-structured representations rather than random initialization. In contrast, Transformers Beyond Orthogonal[0] sits squarely within the theoretical branch, analyzing gradient descent dynamics when embeddings lack orthogonality—a setting that challenges standard assumptions but may reveal how transformers adapt to less ideal initialization. This work complements studies like Stochastic Topological Memory[4], which explores memory structures in neural systems, by providing a rigorous lens on how attention layers cope with embedding geometry. The interplay between these theoretical foundations and application-driven methods remains an open question, with potential for cross-pollination as practitioners seek principled initialization strategies.

Related Works in Same Category

No sibling papers were found in the same taxonomy leaf. A taxonomy-subtopic-level comparison will be produced instead.

Taxonomy-Level Summary

Both subtopics examine how transformers handle token retrieval and factual recall when using random or non-standard embeddings, but from complementary angles. The original leaf focuses on theoretical guarantees through gradient descent analysis, capacity bounds, and sample complexity. The sibling focuses on empirical memory behavior, sequential recall fidelity, and topological/stochastic embedding structures.

Similarities: - Both address transformers with random or non-orthogonal embeddings rather than learned representations - Both are concerned with factual recall and token retrieval tasks - Both explore how embedding structure affects model performance

Differences: - Original leaf emphasizes theoretical analysis (gradient descent dynamics, capacity bounds, sample complexity); sibling emphasizes empirical memory studies - Original leaf focuses on training dynamics and convergence guarantees; sibling focuses on memory fidelity and sequential information recall - Original leaf excludes empirical memory studies without gradient analysis; sibling excludes theoretical gradient descent work - Sibling explicitly mentions topological and stochastic embedding structures; original leaf focuses on non-orthogonality and randomness in a gradient analysis context

Suggested Search Directions: - Theoretical analysis of memory capacity in transformers with structured random embeddings - Sample complexity bounds for sequential recall tasks - Connections between topological embedding properties and gradient descent convergence

Sibling Subtopics

- **Memory Mechanisms and Topological Embeddings** (leaves: 1, papers: 1)
- Scope: Studies investigating memory fidelity and sequential information recall using stochastic or topologically-structured embeddings in LLMs.
- Exclude: Excludes theoretical gradient descent analysis; that belongs to Gradient Descent Analysis with Non-Orthogonal Embeddings.

Contributions Analysis

Overall novelty summary. The paper analyzes gradient descent dynamics for single-layer transformers trained on token retrieval with random (non-orthogonal) embeddings, deriving explicit capacity formulas that reveal multiplicative scaling between sample size, embedding dimension, and sequence length. Within the taxonomy, it occupies the 'Gradient Descent Analysis with Non-Orthogonal Embeddings' leaf under 'Theoretical Analysis of Transformer Learning Dynamics'. This leaf contains only the original paper itself, indicating a sparse research direction. The broader parent branch ('Theoretical Analysis of Transformer Learning Dynamics') contains just one sibling leaf ('Memory Mechanisms and Topological Embeddings'), suggesting that theoretical work on transformer learning dynamics with non-ideal embeddings remains relatively underdeveloped.

The taxonomy reveals a clear structural divide: theoretical analysis of learning dynamics versus application-driven retrieval systems. The original paper's branch focuses on understanding how transformers learn under non-orthogonal embeddings, examining gradient descent convergence and capacity bounds. The neighboring 'Memory Mechanisms and Topological Embeddings' leaf investigates memory fidelity and sequential recall using stochastic or topologically-structured representations, complementing but not directly overlapping with gradient descent analysis. Meanwhile, the 'Application-Driven Retrieval Systems' branch emphasizes practical embedding methods for text and multimodal retrieval, operating under different assumptions (often pre-trained or domain-adapted embeddings) and prioritizing empirical performance over theoretical characterization of learning dynamics.

The literature search examined only one candidate paper across three identified contributions, finding zero refutable pairs. Contribution A (gradient descent analysis with random embeddings) examined zero candidates; Contribution B (explicit multiplicative scaling formulas) also examined zero candidates; Contribution C (statistical lower bound for multiplicative scaling) examined one candidate but found it non-refutable or unclear. This extremely limited search scope—one candidate total—means the analysis provides minimal evidence about prior work overlap. The absence of refutable pairs among this single candidate suggests no immediate overlap was detected, but the search scale is too small to draw strong conclusions about novelty relative to the broader literature.

Given the sparse taxonomy structure (only one paper in the target leaf, minimal theoretical work in the parent branch) and the extremely limited literature search (one candidate examined), the analysis suggests the work addresses an underexplored theoretical direction. However, the search scope is insufficient to assess whether related gradient descent analyses exist outside the top-1 semantic matches. A more comprehensive search would be needed to confidently characterize the contribution's novelty relative to the full landscape of transformer theory and non-orthogonal embedding studies.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Analysis of gradient descent on transformers with random embeddings for factual recall

Description: The authors analyze a single-layer transformer trained via gradient descent on finite datasets with non-orthogonal (random) embeddings for a token-retrieval task. This addresses a gap in existing theory which typically assumes infinite data or orthogonal embeddings.

This contribution was assessed against **0 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

Contribution 2: Explicit formulas revealing multiplicative scaling between sample size, embedding dimension, and sequence length

Description: The authors derive explicit formulas characterizing the model's storage capacity during the early phase of gradient descent. These formulas reveal a multiplicative relationship among sample size, embedding dimension, and sequence length, showing how these parameters jointly govern learning efficiency.

This contribution was assessed against **0 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

Contribution 3: Statistical lower bound demonstrating intrinsic multiplicative scaling under non-orthogonal embeddings

Description: The authors provide a statistical lower bound for the factual recall problem, showing that the multiplicative scaling between parameters is not merely an artifact of their training algorithm but is intrinsic to the statistical problem when using non-orthogonal embeddings.

This contribution was assessed against **1 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. On the Black-Box Complexity of Private-Key Inner-Product Functional

URL: View paper

Brief Assessment

The candidate paper (Private-Key Inner-Product Complexity[5]) focuses on black-box complexity of private-key inner-product functional encryption, which is a cryptographic primitive. This is fundamentally different from the original paper's focus on transformer learning dynamics and factual recall with non-orthogonal embeddings in neural networks.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] Learning to Recall with Transformers Beyond Orthogonal Embeddings View paper
- [1] Domain-Specific Text Embedding Models for Information Retrieval View paper
- [2] Vision Transformer-based Context-Aware System for Lingual Ultrasound in Digital Health Ecosystem View paper
- [3] Enhancing Term-Based Document Retrieval by Word Embedding and Transformer Models View paper
- [4] Stochastic Topological Memory Embedding in Large Language Models: An Empirical Analysis Using Open-Source Neural Architectures View paper
- [5] On the Black-Box Complexity of Private-Key Inner-Product Functional View paper
- [6] Is Random Attention Sufficient for Sequence Modeling? View paper