

# Hybrid Embedding:

## Bridging LLM and Differential Trees for Explainable, Efficient AI

Sizhe Tan

With assistance from GPT-5 (AI research assistant)

2025-09-23

Repository: <https://github.com/sizhet/hybrid-embedding-project>

DOI 10.5281/zenodo.17205861

White Paper

### Abstract

Large Language Models (LLMs) provide powerful embeddings but remain opaque and costly to adapt. Differential Trees (DTs) offer structured, interpretable decision paths. This paper introduces Hybrid Embedding, a simple yet disruptive design:  $H(x) = [C(x) \parallel R(x)]$ , where  $C(x)$  is a compact differential-tree routing code and  $R(x)$  is the within-category residual embedding. This design delivers intrinsic explainability, seamless LLM–DBM integration, and efficient deployment through a two-tier architecture.

## 1. Introduction

LLM embeddings are high-dimensional, powerful, but black-box and costly. Differential Trees (DTs) provide structure and anchors. Hybrid Embedding unifies them by decomposing embeddings into routing codes (C) and residuals (R).

## 2. Core Design: Hybrid Embedding

Hybrid Embedding:  $H(x) = [C(x)||R(x)]$ . C encodes path, margins, leaf ID (~128d). R preserves semantic fidelity (~1024d). This makes embeddings interpretable and precise.

## 3. Training Paradigm: Tree-then-Distill

Freeze base embeddings, build Teacher-Tree, produce labels (paths, margins, prototypes). Train Student heads to predict C and R. Losses include Path-CE, Margin-L1, Residual-L2, InfoNCE, Hubness regularization.

## 4. Deployment Architecture: Two-Tier LLM AI

Base Model (Teacher): costly, infrequent updates. Service Models (Students): lightweight, distilled, domain-optimized with different differential trees. Reduces cost, supports multiple domains, easy updates.

## 5. Online Serving Workflow

Router selects domain head. Phase-1 uses C for coarse bucket search. Phase-2 uses R for fine ranking. Explainer outputs path, prototypes, differences. Fallback to Teacher if low confidence.

## 6. Benefits and Metrics

Efficiency: Phase-1 reduces candidates. Fidelity: R ensures accuracy. Explainability: path, prototypes, differences. Metrics: Recall@k, path stability, human audit pass, latency.

## 7. Applications

Search & recommendation; conversational AI with intrinsic explanations; finance/healthcare requiring transparency; multi-domain platforms with one backbone and many trees.

## 8. Industry and Research Impact

For industry: reduced cost, transparency, compliance. For research: breaks LLM black-box, unifies symbolic/neural methods, opens new paradigms.

## **9. Future Work**

Differentiable Trees; multi-tree ensembles; cross-modal hybrid embeddings; standardization of Hybrid Embedding APIs.

## **10. Conclusion**

Hybrid Embedding shifts embeddings from black-box to interpretable infrastructure. It unifies LLM and DBM, lowers cost, and raises transparency. It is a foundational design for the next generation of AI.

Diagram: Hybrid Embedding Structure

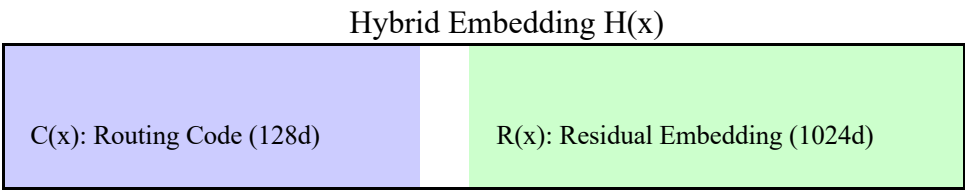


Diagram: Serving Workflow

