ICLR

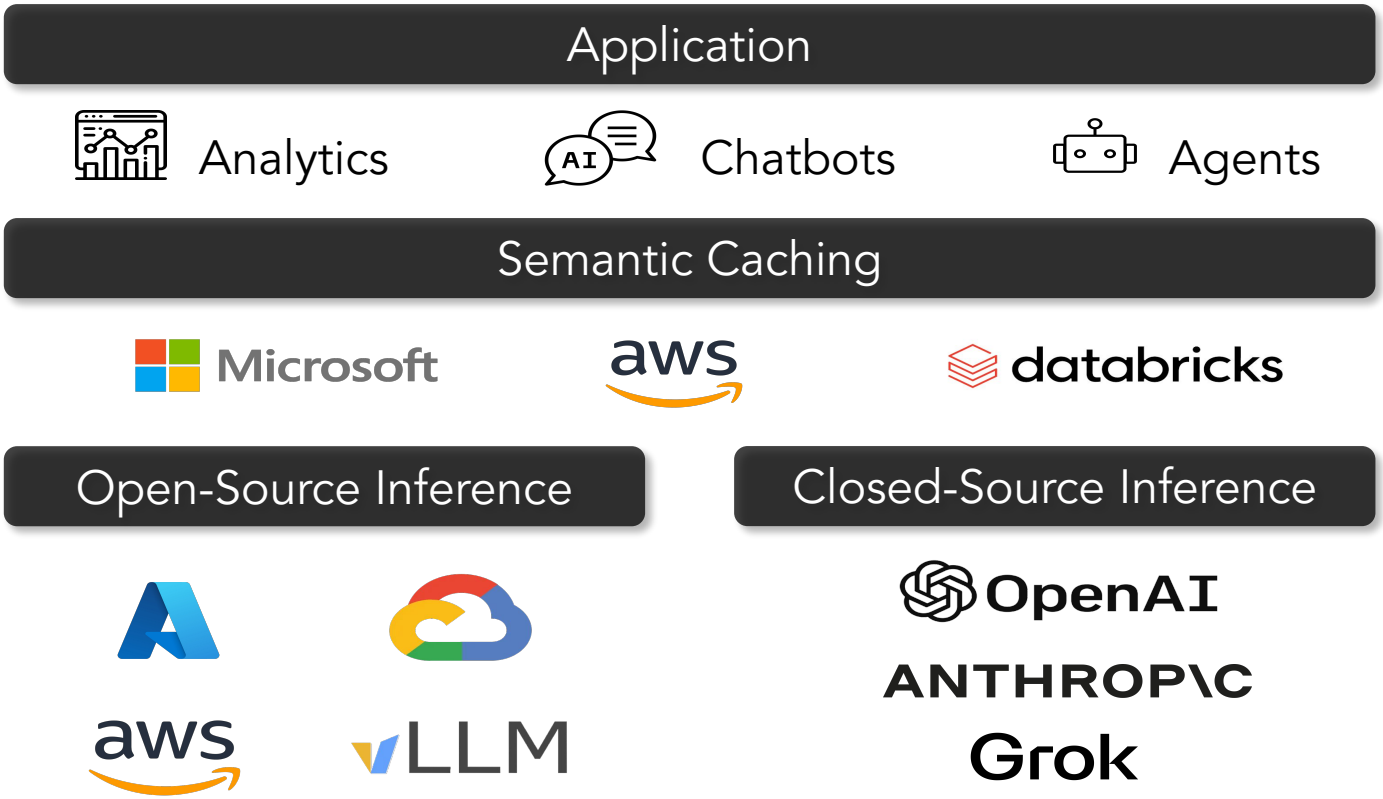
vCache Verified Semantic Prompt Caching

Luis Gaspar Schroeder, Aditya Desai, Alejandro Cuadron, Kyle Chu, Shu Liu,
Mark Zhao, Stephan Krusche, Alfons Kemper, Matei Zaharia, Joseph E. Gonzalez

Berkeley UNIVERSITY OF CALIFORNIA
Stanford University
Technical University of Munich
ETH zürich

LLMs are Powerful
… but Expensive and Slow

How to Reduce LLM Inference Latency and Cost from a Systems Perspective?



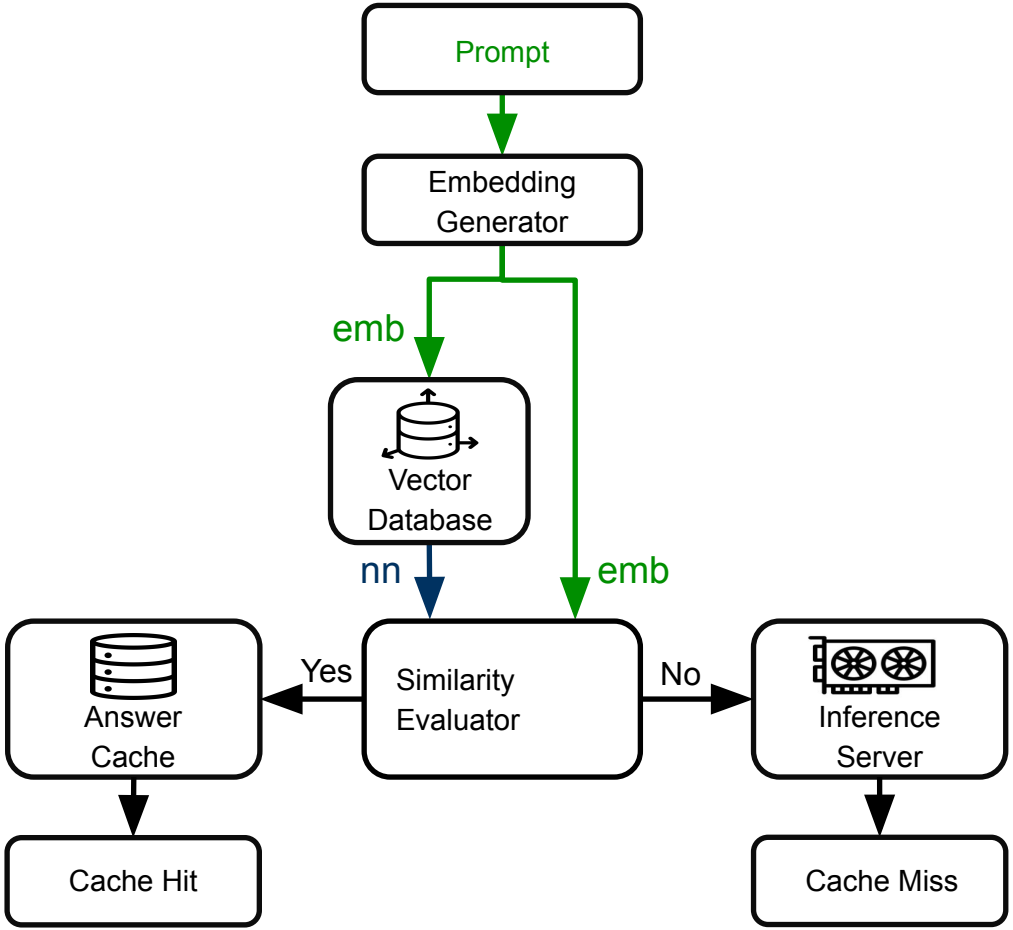
Reduce

- Latency (100x)
- Cost (10x)

Existing Semantic Caches are Unreliable and Not Deployable in Production Systems

Principal Researcher
AI Team, International
Car Sharing Platform

> “We wanted to use a semantic cache for our inference pipeline. We could not deploy it because the False Positive Rate was too high and unpredictable.”



"Where was…?"

Emb("Where was …?")

Retrieve the nearest neighbor from the vector database

Are the embedded prompt and its nearest neighbor similar enough?

Yes → Cache Hit

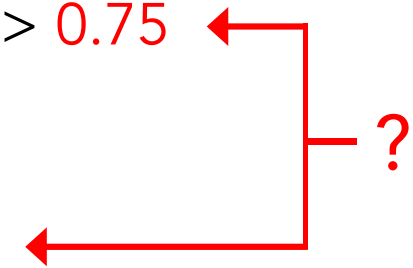
No → Cache Miss

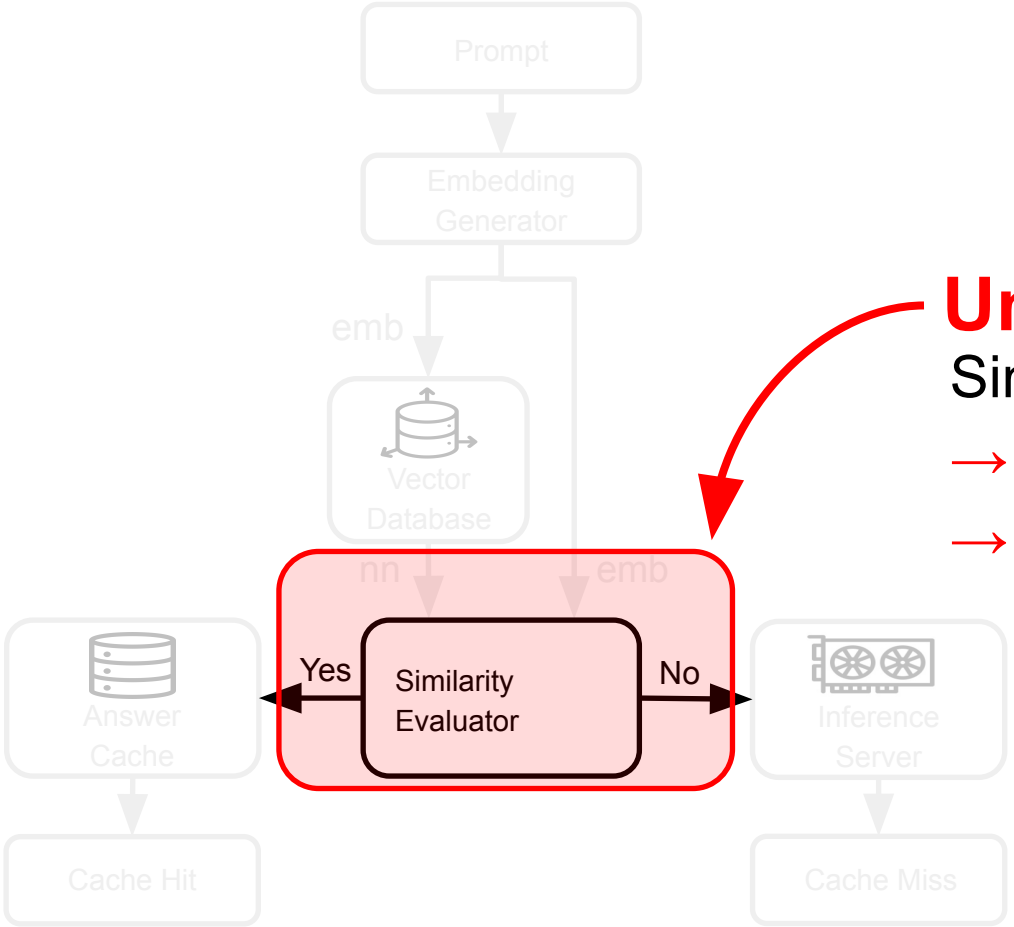
The Threshold Problem

Given a new embedded prompt and its nearest neighbor embedding from the vector database…

cos_sim("Where was Roger Federer born",
"In which town did Roger Federer grow up") > 0.75

cos_sim("Explain Round Robin scheduling",
"Explain Shortest Job First scheduling) > 0.92





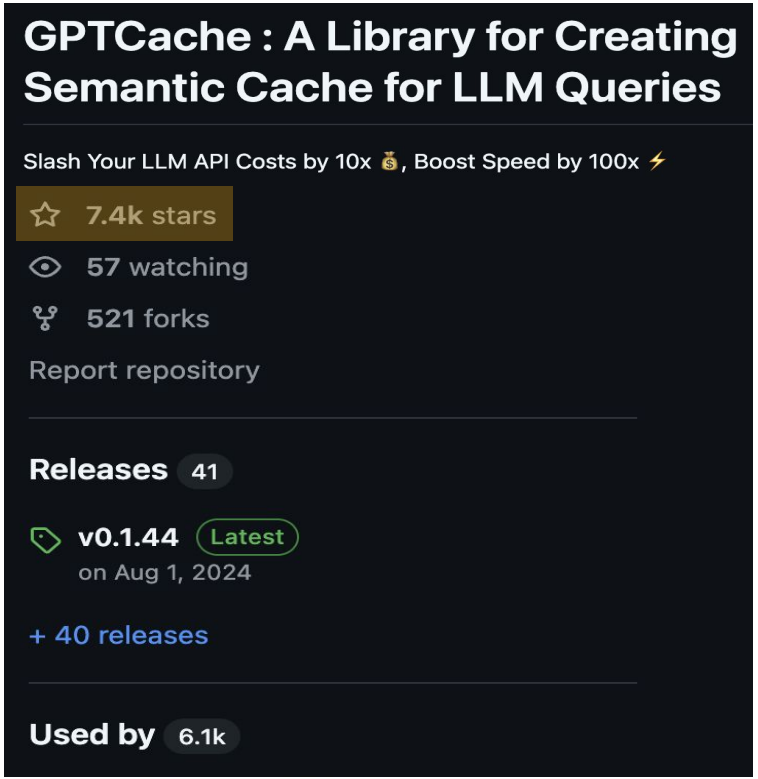
Unreliability

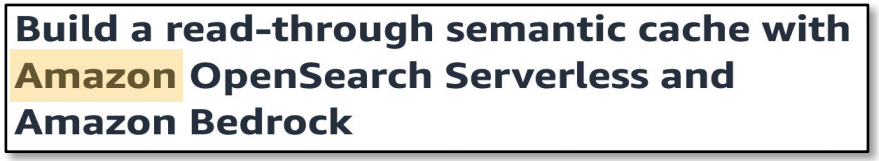
Similarity ≥ Some Threshold ?

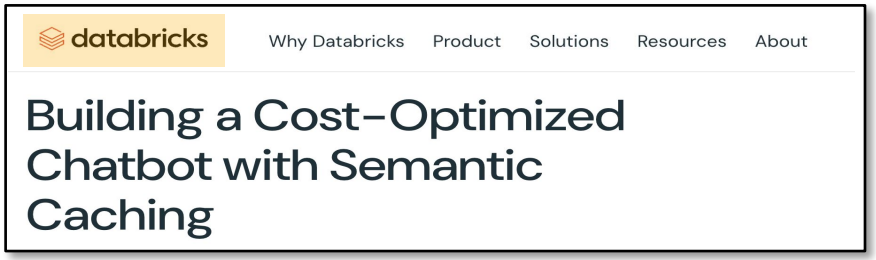
→ Unexpected errors

→ Suboptimal cache hit rates

ICLR







They all use a static threshold, are suboptimal, and unreliable…

vCache. How To Make Semantic Prompt Caches More Reliable?

User-Defined Accuracy Target (i.e., 99%)

vCache Online Learning Algorithm

No Manual Tuning or Supervision

No Matter the Workload

No Matter the Model

No Matter the Infrastructure

github.com/vcache-project/vCache

README Contributing License

vCache

Reliable and Efficient Semantic Prompt Caching

Status Active | Version 1.0 | License CC NC-ND 3.0

vCache is the first semantic prompt cache that guarantees user-defined error rate bounds. Semantic caching reduces LLM latency and cost by returning cached model responses for semantically similar prompts (not just exact matches). vCache replaces static thresholds with **online-learned, embedding-specific decision boundaries**—no manual fine-tuning required. This enables reliable cached response reuse across any embedding model or workload.

vCache is the first reliable semantic cache reducing inference cost and latency with

1. Online learning (embedding-model and dataset agnostic)
   a. Error rate guarantees
   b. Embedding specific and probabilistic decision boundaries
2. SOTA performance