The Embedder's Dilemma: LLMs Are Better, but at What Cost?

Adnan El Assadi · Harvard Niklas Muennighoff · Stanford Jinhyuk Lee · Independent Researcher

LLMs now match the best embedding models. At 1,431× the cost.

Paper

Code

Summary

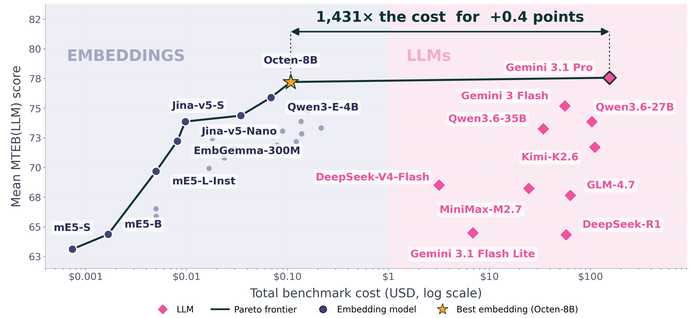
Should you replace your embedding pipeline with an LLM? We compare **10 LLMs** across six families with **26 embedding models** on **37 tasks** in five categories, with exact cost accounting for every model.

In aggregate they tie: 77.6 vs. 77.2. LLMs lead on reasoning-heavy retrieval; embeddings lead on classification; the other three categories are statistical ties.

An LLM pays 1,431× more for that tie ($154 vs. $0.11 per benchmark pass), **reasoning tokens are 28–81% of that bill**, and embeddings run **736× faster** on the same GPU.

Use embeddings by default, and an LLM only to rerank a shortlist on reasoning-heavy retrieval.

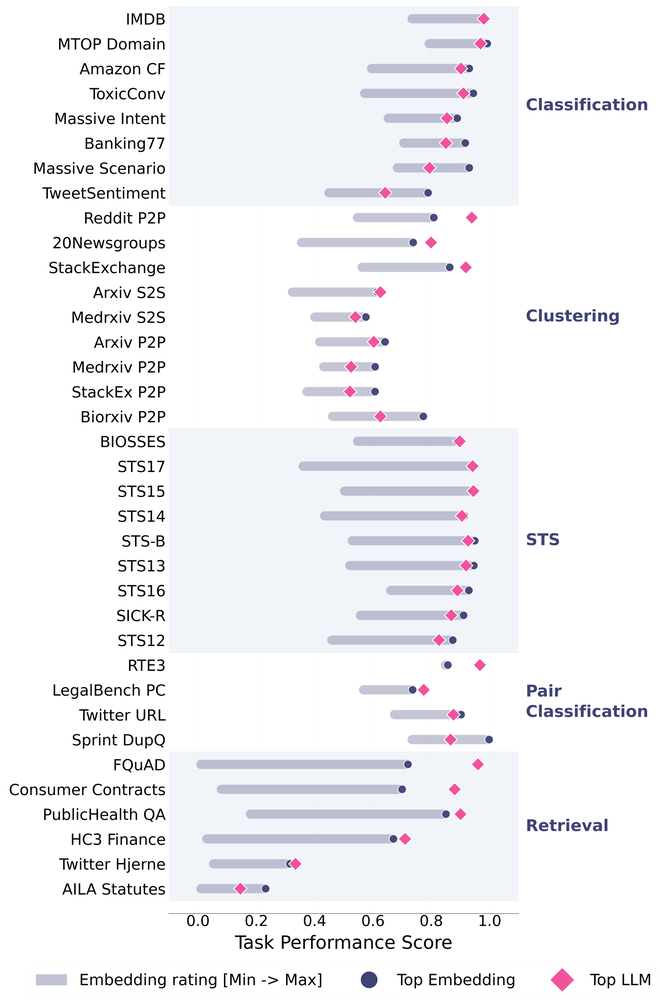
Cost vs. performance



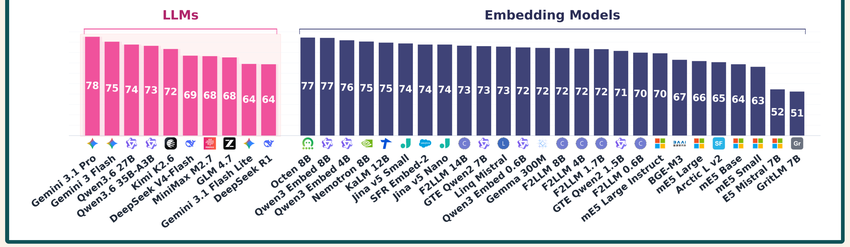
Cost vs. performance by category



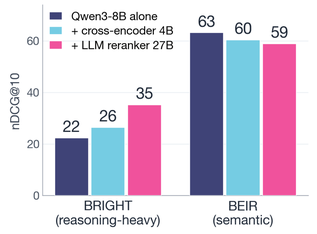
Per task



Overall leaderboard



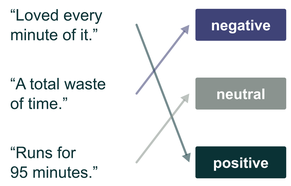
Reranking an embedding shortlist



MTEB(LLM): 37 tasks in five categories

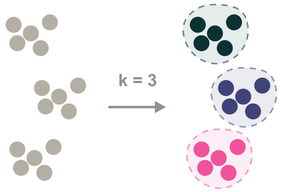
Classification

Which label fits?



Clustering

Which belong together?



Semantic Similarity

How close in meaning?

"The cat sat on the mat." — 4.8 — "A cat sits on a rug."

"A man plays guitar." — 2.1 — "A woman sings a song."

"The market crashed." — 0.3 — "She baked a cake."

Pair Classification

Do these ask the same thing?

"Can I get a refund?" ✓ "Will I get my money back?"

"Can I edit my address?" ✓ "How to change my address?"

"How do I reset my password?" ✗ "How do I reset my username?"

Retrieval

Which document answers this?

"How long does flu immunity last?"

