Embedding Landscape Benchmark: A Claude+Codex Adversarial Review Pipeline

Claude (Anthropic) with Codex (OpenAI) as Adversarial Critic

January 28, 2026

Abstract

We benchmark 6 embedding models on a 1000-document synthetic knowledge base using 60 queries with keyword-based ground truth. An adversarial review pipeline—Claude designs and runs experiments, Codex critiques at three junctures—reveals that **no clear separation is observed among quality scores**. The keyword-overlap ground truth is fundamentally misaligned with semantic embedding evaluation. We report throughput and latency comparisons as the only trustworthy findings, and document the adversarial review process as the primary contribution.

1 Introduction

This paper reports on an embedding model comparison conducted through a novel collaborative pipeline: Claude (Anthropic) designed and executed the benchmark, while Codex (OpenAI, gpt-5.2-codex at xhigh reasoning) served as adversarial critic at three research junctures—post-design, post-results, and post-paper.

The adversarial process itself is the primary contribution. The benchmark results, as Codex correctly identified, are largely uninformative for semantic retrieval quality due to fundamental methodological limitations.

2 Methodology

2.1 Models Under Test

1. **UForm3-small** (256d) — ONNX-based, lightweight
2. **all-MiniLM-L6-v2** (384d) — sentence-transformers
3. **bge-base-en-v1.5** (768d) — sentence-transformers
4. **nomic-embed-text-v1.5** (768d) — sentence-transformers
5. **jina-embeddings-v3** (1024d) — FAILED (import incompatibility)
6. **OpenAI text-embedding-3-small** (1536d) — API

2.2 Dataset and Ground Truth

1000 synthetic documents (personal knowledge base, seed=42) with keyword-overlap relevance labels. 60 queries: 20 neutral (topic-based), 20 temporal (date-based), 20 mixed (topic+time). Graded relevance 1–3 based on keyword/topic/doc-type matching, top 15 per query.

2.3 Evaluation

Exact numpy cosine search (no ANN), k=10 retrieval. Metrics: nDCG@10, Recall@10, MRR. Three search trials per query for latency measurement.

2.4 Methodological Limitations (Codex Critique #1)

Codex's pre-experiment critique identified fundamental issues:

- Keyword-based ground truth creates systematic bias against semantic models
- Relevance set truncated to 15 docs distorts recall
- Synthetic data's lexical regularity inflates keyword-matching value
- API vs. local latency comparison includes network overhead

We proceeded with these caveats explicitly acknowledged, framing this as a keyword-overlap retrieval benchmark, not a semantic evaluation.

3 Results

3.1 Quality Metrics

Table 1: Overall quality metrics (all 60 queries). All differences are statistically insignificant.

Model	Dim	nDCG@10	±std	Recall@10	MRR
bge-base-en-v1.5	768	0.0494	0.0839	0.0322	0.1410
all-MiniLM-L6-v2	384	0.0382	0.0710	0.0267	0.0938
OpenAI-3-small	1536	0.0367	0.0571	0.0278	0.0817
UForm3-small	256	0.0361	0.0697	0.0233	0.1055
nomic-embed-v1.5	768	0.0349	0.0656	0.0234	0.0894

These numbers are noise. For every model, the standard deviation exceeds the mean nDCG@10. The total spread across all 5 models (0.0145) is far smaller than within-model variance. No ranking can be claimed with statistical confidence.

3.2 Per Query Type

Table 2: nDCG@10 by query type. Query type effects (0.02–0.07 spread) dominate model effects (0.01 spread).

Model	Neutral	Temporal	Mixed
bge-base-en-v1.5	0.0411	**0.0350**	**0.0721**
all-MiniLM-L6-v2	**0.0661**	0.0103	0.0382
OpenAI-3-small	0.0395	0.0229	0.0477
UForm3-small	0.0281	0.0320	0.0481
nomic-embed-v1.5	0.0503	0.0077	0.0468

In this benchmark, temporal queries score near zero for all models—suggesting that embedding models alone cannot resolve date-range constraints without structured filtering, though this may partly reflect the keyword ground truth's limitations. Rankings flip across query types.

3.3 Throughput and Latency (Trustworthy)

Table 3: Performance metrics. These are the benchmark's reliable findings.

Model	Embed (s)	docs/s	E2E (ms)	Embed 1q (ms)	RAM (MB)
UForm3-small	6.2	160.9	21.2	20.9	509
all-MiniLM-L6-v2	35.8	28.0	280.2	271.5	1279
bge-base-en-v1.5	136.1	7.3	616.9	608.3	1979
nomic-embed-v1.5	223.6	4.5	475.0	466.3	2581
OpenAI-3-small*	10.3	96.7	299.2	290.5	—

*OpenAI latency includes network round-trip; not directly comparable to local models.
Hardware: 8-core AMD EPYC, 92GB RAM, CPU-only (no GPU). Batch size: full corpus (1000 docs). Single-threaded inference.

In this setup, UForm3-small is 29× faster E2E than bge-base and uses 3.8× less RAM. These are indicative measurements on specific hardware; absolute values will vary with different configurations.

4 Codex Critique #2: Post-Results Review

Codex's critique of the raw results confirmed:

1. nDCG@10 std exceeds the mean for every model—averages are unstable
2. Model separation (0.0145) is noise-level given per-query variance
3. Rankings change across metrics (nDCG vs MRR vs Recall)—weak signal
4. Query-type effects dominate model effects
5. Missing baseline (BM25 or random) makes absolute scores uninterpretable

5 Discussion

5.1 What This Benchmark Cannot Tell Us

This experiment cannot rank embedding models for semantic retrieval. The keyword-overlap ground truth systematically favors lexical matching over semantic understanding. A model that correctly retrieves semantically relevant documents lacking keyword overlap would be *penalized*.

5.2 What This Benchmark Can Tell Us

- **Throughput and latency profiles** are reliable and show 29× variation across models
- **RAM footprints** range from 509MB to 2.6GB
- **No model clearly outperforms others** on keyword-overlap retrieval in this setup—the benchmark cannot separate them, which likely reflects its limitations rather than true model equivalence

5.3 The Adversarial Review Pipeline

The most valuable output of this experiment is the pipeline itself. Codex's critiques caught:

- **Pre-experiment**: fundamental ground-truth misalignment (Critique #1)
- **Post-results**: statistical insignificance of all quality rankings (Critique #2)

Without adversarial review, we might have reported "bge-base-en-v1.5 achieves the best nDCG@10 (0.0494), outperforming OpenAI by 35%." That statement, while technically true, is deeply misleading—the difference is noise, and 0.0494 is not a good score by any standard.

6 Conclusion

An honest experiment honestly reported: in this benchmark, no embedding model clearly separates from others on keyword-overlap retrieval of synthetic documents. The most informative comparisons are operational—throughput, latency, and RAM—though these are hardware-specific. The adversarial Claude+Codex pipeline, demonstrated here as a case study, prevented us from dressing up noise as signal.

For meaningful embedding evaluation, future work requires human-judged relevance labels, paraphrased queries, and standard IR test collections (e.g., BEIR).