Embedding Model Landscape for Personal Knowledge Retrieval: A Benchmark with LLM-Judged Ground Truth and Adversarial Review

Claude (Anthropic) with Codex (OpenAI) Adversarial Critique
Generated: January 28, 2026

Abstract

We benchmark four embedding models on a 1,000-document synthetic personal knowledge base using LLM-judged relevance (Gemini 2.0 Flash) as a proxy for ground truth. Using system-level pooling across all models and 60 diverse queries, we find that all tested models achieve nDCG@10 scores in the 0.48–0.55 range (interpretation of absolute level requires a baseline, which we lack) with broadly overlapping confidence intervals. OpenAI `text-embedding-3-small` achieves the highest mean nDCG@10 (0.554) but is not statistically distinguishable from `bge-base-en-v1.5` or `nomic-embed-text-v1.5` at the 95% confidence level. Only the smallest model (`all-MiniLM-L6-v2`) is statistically separable from the top performers, and even then with small effect sizes ($d \approx 0.20$–0.29). We emphasize limitations: LLM-judged relevance is a proxy not validated against humans, recall remains low across all models, and the synthetic corpus may not generalize. This paper was adversarially reviewed by OpenAI Codex (gpt-5.2-codex, xhigh reasoning) at three junctures.

1 Introduction

Embedding models are widely used for semantic retrieval in personal knowledge bases—collections of notes, meeting summaries, architecture decisions, and code reviews. Selecting an embedding model involves trading off retrieval quality, latency, memory footprint, and API dependency.

This benchmark compares four embedding models across three query types (semantic, temporal, mixed) on a synthetic personal knowledge base. A key methodological choice is using an LLM (Gemini 2.0 Flash) as a relevance judge rather than keyword matching, which previous iterations showed produces noise-floor results (nDCG $\approx$ 0.03–0.05) that fail to differentiate models.

1.1 Adversarial Review Process

This work was reviewed at three junctures by OpenAI Codex (gpt-5.2-codex with xhigh reasoning effort):

1. **Post-design**: Codex identified pooling bias as a critical flaw—judging only BM25-selected candidates would create model-dependent bias. We adopted system-level pooling (top-20 from each model).

2. **Post-results**: Codex flagged overlapping confidence intervals, the risk of "insidious enthusiasm" about relative gains, and the need for baseline context.

3. **Post-paper**: Final review for unsupported claims.

2 Methodology

2.1 Dataset

A synthetic corpus of 1,000 documents spanning 8 document types (standups, meetings, architecture decisions, bug fixes, reading notes, personal notes, code reviews, project updates) over 12 months, generated with seed 42 for reproducibility.

2.2 Models

Model	Dimensions	Throughput	Type
all-MiniLM-L6-v2	384	11 docs/s	Local
bge-base-en-v1.5	768	4 docs/s	Local
nomic-embed-text-v1.5	768	3 docs/s	Local
OpenAI text-embedding-3-small	1,536	97 docs/s	API

Table 1: Models tested. UForm3-small and jina-embeddings-v3 failed to load due to library version conflicts and are excluded. Throughput measured on 24-core CPU (no GPU).

2.3 Queries

60 queries divided equally: 20 semantic ("How did we handle the strangler fig pattern?"), 20 temporal ("What happened in January 2025?"), and 20 mixed ("Security incidents in the first half of 2025?").

2.4 Ground Truth: LLM-as-Judge

This is not ground truth in the traditional sense. It is a proxy for relevance, using Gemini 2.0 Flash to score each (query, document) pair on a 0–3 scale. We use system-level pooling: for each query, the top 20 documents from *each* model plus 10 random negatives are judged. This yields an average of 63 judged documents per query ($\sim$3,800 total query-document pairs judged for this experiment; the judgment cache contains 5,284 entries including judgments from prior experimental runs).

Known limitations of this approach:

- Single judge with no inter-annotator agreement measurement
- Documents truncated to 1,500 characters for judging
- Unjudged documents treated as irrelevant (mitigated but not eliminated by system-level pooling)
- No human validation of LLM judgments
- Synthetic corpus may have uniform style that favors or disfavors certain models

2.5 Metrics

- **nDCG@10**: Primary metric. Uses graded relevance (0–3).
- **Recall@10**: Fraction of relevant documents (score ≥ 2) in top 10.
- **MRR**: Mean reciprocal rank of first relevant document (score ≥ 2).

Statistical analysis uses bootstrap 95% confidence intervals (1,000 resamples), Wilcoxon signed-rank tests (paired, non-parametric), and Cohen's d for effect size.

3 Results

3.1 Retrieval Quality

Model	nDCG@10	Recall@10	MRR
all-MiniLM-L6-v2	0.476 [0.405, 0.553]	0.177 [0.151, 0.201]	0.634 [0.537, 0.726]
bge-base-en-v1.5	0.531 [0.462, 0.599]	0.210 [0.185, 0.241]	0.712 [0.610, 0.804]
nomic-embed-text-v1.5	0.525 [0.455, 0.587]	0.228 [0.202, 0.254]	0.709 [0.623, 0.803]
openai-3-small	**0.554** [0.489, 0.613]	**0.243** [0.213, 0.278]	**0.754** [0.681, 0.830]

Table 2: Retrieval quality. Values shown as mean [95% bootstrap CI]. Bold indicates highest mean, but all confidence intervals overlap with all other models.

Key observations:

1. **All confidence intervals overlap.** No model's nDCG@10 CI is disjoint from any other model's CI. Note: overlapping CIs do not strictly imply no difference (paired tests can detect effects that marginal CIs miss), but the overlap combined with small effect sizes suggests limited practical separation.

2. **Absolute nDCG is moderate (0.48–0.55), not high.** Without a baseline (e.g., BM25), we cannot assess whether these values represent "good" performance or merely "better than keyword matching."

3. **Recall is low (0.18–0.24).** All models miss the majority of relevant documents in their top 10. This is a meaningful practical limitation.

4. **MRR is reasonable (0.63–0.75)**, suggesting models generally place at least one relevant document in the top 2–3 positions.

3.2 Pairwise Statistical Comparisons

Comparison	Δ nDCG	Cohen's d	p-value	Sig.
MiniLM vs bge-base	-0.055	-0.198	0.027	*
MiniLM vs nomic	-0.049	-0.177	0.126	ns
MiniLM vs openai	-0.078	-0.288	0.012	*
bge-base vs nomic	$+0.006$	$+0.022$	0.667	ns
bge-base vs openai	-0.023	-0.089	0.431	ns
nomic vs openai	-0.029	-0.112	0.230	ns

Table 3: Pairwise comparisons. Only MiniLM vs bge-base ($p = 0.027$) and MiniLM vs openai ($p = 0.012$) reach significance at $\alpha = 0.05$, both with small effect sizes ($|d| < 0.30$).

Interpretation with appropriate caution:

- The top three models (bge-base, nomic, openai) are **statistically indistinguishable** from each other ($p > 0.23$ for all pairs).
- MiniLM is significantly weaker than bge-base and openai, but the effect sizes are small ($d \approx 0.20$–0.29), suggesting the practical difference is limited.
- With 60 queries and overlapping CIs, these tests have limited statistical power. A larger query set might resolve ambiguities.
- We do not apply multiple testing correction (e.g., Bonferroni), which would render even the MiniLM comparisons non-significant ($\alpha_{\text{adj}} = 0.05/6 = 0.0083$).

3.3 Operational Characteristics

Model	Embed Time (s)	Throughput (docs/s)	Dimensions
all-MiniLM-L6-v2	97	11	384
bge-base-en-v1.5	158	4	768
nomic-embed-text-v1.5	270	3	768
openai-3-small	11	97	1,536

Table 4: Operational metrics. CPU-only inference (24-core, 98.9GB RAM). OpenAI throughput reflects API latency, not local compute.

OpenAI's API-based model is 9–32× faster than local models on CPU, though it requires network access and incurs per-token costs. For a 1,000-document corpus, all models complete indexing within minutes.

4 Discussion

4.1 What We Can Say

- All four models achieve moderate retrieval quality on this synthetic corpus, with nDCG@10 in the 0.48–0.55 range.
- The three larger models (bge-base, nomic, openai) form a statistical cluster that we cannot differentiate with 60 queries.
- MiniLM shows a small disadvantage versus bge-base and openai (significant at $\alpha = 0.05$ but not after multiple-testing correction). The lower dimensionality (384 vs. 768–1,536) correlates with this gap, but we cannot attribute causation.
- OpenAI's API model offers the best throughput-per-quality trade-off if API dependency and cost are acceptable.

4.2 What We Cannot Say

- We **cannot** claim any model "beats" another in the top-3 cluster. The data do not support ranking among bge-base, nomic, and openai.
- We **cannot** claim these absolute nDCG values are "good" without a baseline.
- We **cannot** generalize from a synthetic corpus to real personal knowledge bases.

- We **cannot** validate that LLM-judged relevance matches human relevance without a human labeling study.

4.3 Comparison to Previous Methodology

A prior iteration using keyword-based ground truth produced nDCG@10 of 0.03–0.05 for all models, with no differentiation. The LLM-judge approach yields 10× higher absolute scores and meaningful (though modest) model separation. This suggests that keyword matching underdifferentiates semantic retrieval models in this synthetic benchmark setting, though we cannot generalize from a single dataset and judge.

4.4 Practical Recommendation

For a personal knowledge base of ∼1,000 documents:

- If API dependency is acceptable: `text-embedding-3-small` offers the fastest throughput with comparable quality (subject to cost and privacy constraints).
- If local-only is required: `bge-base-en-v1.5` or `nomic-embed-text-v1.5` are statistically equivalent choices.
- `all-MiniLM-L6-v2` is viable for resource-constrained environments, accepting a small quality trade-off.
- The differences are small enough that **other factors** (cost, privacy, latency requirements, existing infrastructure) should dominate the decision.

5 Limitations

1. **LLM judge not validated**: Gemini 2.0 Flash judgments are a proxy, not ground truth. No human annotation or inter-annotator agreement was measured.
2. **Synthetic corpus**: May not represent real personal knowledge bases in style, complexity, or topic distribution.
3. **Small query set**: 60 queries limits statistical power and may not cover all retrieval scenarios.
4. **Two models excluded**: UForm3-small and jina-embeddings-v3 failed due to library version conflicts. Results may differ with a complete model set.
5. **No baseline**: Without BM25 or random baseline scores, absolute metric values lack context.
6. **CPU-only**: Operational metrics reflect CPU inference; GPU results would differ substantially.
7. **Pooling depth**: System-level pooling with top-20 per model may still miss relevant documents retrieved by none of the tested models.
8. **Single evaluation run**: No repeated runs to assess seed sensitivity.

6 Adversarial Review Summary

Three rounds of adversarial review by OpenAI Codex (gpt-5.2-codex, xhigh reasoning) identified and addressed:

1. **Pooling bias** (design stage): Switched from BM25-only candidate selection to system-level pooling across all models.
2. **Overlapping CIs and enthusiasm risk** (results stage): Explicitly report that all CIs overlap and avoid claiming winners in the top-3 cluster.
3. **Paper claims** (paper stage): Codex flagged internal inconsistency in judgment counts, over-interpretation of CI overlap, over-strong causal claims about dimensionality, and insufficiently scoped conclusions about keyword matching. Corrections incorporated.

7 Conclusion

Within the scope of this synthetic benchmark, LLM-judged relevance produces more differentiated evaluation of embedding models for personal knowledge retrieval, in contrast to keyword-based approaches that yield noise-floor results. Among the four models tested, we find a statistical cluster of three models (bge-base-en-v1.5, nomic-embed-text-v1.5, and OpenAI text-embedding-3-small) that are indistinguishable at $n = 60$ queries, with only the smallest model (all-MiniLM-L6-v2) showing a modest disadvantage that is significant in pairwise tests but not robust to multiple-testing correction. For practical deployment, the choice between these models should be driven by operational considerations (API vs. local, cost, latency) rather than retrieval quality, which is effectively equivalent within this evaluation's resolution.