Local vs. Cloud Embeddings for Temporal Personal Knowledge Search: A Head-to-Head Benchmark of UForm-256d and OpenAI text-embedding-3-small

Synthetic Benchmark — T92/T117

January 2026

Abstract

We present a rigorous, reproducible benchmark comparing UForm3 (256-dimensional, local ONNX inference) against OpenAI `text-embedding-3-small` (1536-dimensional, cloud API) for personal knowledge base retrieval. Using a fully synthetic corpus of 1,000 documents spanning 12 months and 60 queries with human-designed ground truth, we evaluate three temporal search strategies: baseline cosine similarity, exponential time-decay, and hybrid pre-filtering. OpenAI embeddings achieve 2.3× higher nDCG@10 (0.794 vs. 0.347) and 1.6× higher MRR (0.886 vs. 0.563), while UForm delivers 11× lower end-to-end latency (35ms vs. 404ms) and 12× smaller memory footprint. Time-decay uniformly *hurts* retrieval quality for both models on this corpus, while hybrid pre-filtering shows marginal benefit for UForm on temporal queries. All code and data are publicly available.

1 Introduction

Personal knowledge management systems increasingly rely on semantic search over heterogeneous document collections. Two deployment paradigms exist: *local* models offering privacy and low latency, and *cloud* models offering superior quality at higher cost and latency. This study quantifies the exact trade-offs using a controlled, reproducible setup.

We compare:

- **UForm3-image-text-english-small**: 256-dimensional embeddings, local ONNX Runtime inference, zero API cost.
- **OpenAI text-embedding-3-small**: 1536-dimensional embeddings, cloud API, \$0.02/1M tokens.

Both are evaluated across three temporal search strategies on a synthetic corpus designed to mimic an AI engineer's daily knowledge base.

2 Methodology

2.1 Synthetic Dataset

We generate 1,000 documents spanning January–December 2025, covering 12 topic categories (authentication, caching, API design, database, deployment, frontend, ML, monitoring, security, architecture, team, migration). Each document is 150–400 words using parameterized templates with realistic technical content. Documents include temporal markers ("today we discussed," "Q3 planning") and are assigned 1–3 topic labels. Random seed is fixed at 42 for full reproducibility.

2.2 Query Design

60 queries are divided into three categories:

- **Temporal** (20): Reference specific months/quarters ("What did we discuss about auth in November?")
- **Topical** (20): Pure topic queries ("How does the caching layer work?")
- **Mixed** (20): Recency-biased topic queries ("Latest update on the migration project?")

Ground truth relevance is assigned programmatically: primary topic match = 3, secondary topic match = 2, temporal-only match = 1, combined temporal+topical = 3.

2.3 Search Methods

Baseline: Pure cosine similarity ranking.

Time-Decay: score = $\cos(q, d) \cdot e^{-\lambda \cdot \Delta t}$ where Δt is days from reference date (Dec 31, 2025). We sweep $\lambda \in \{0.001, 0.005, 0.01, 0.02, 0.05\}$.

Hybrid Pre-filter: Filter documents to ±1 month of the query's target month, then rank by cosine similarity. Falls back to baseline for non-temporal queries.

2.4 Metrics

nDCG@10, Recall@10 (threshold ≥ 2), Mean Reciprocal Rank, and wall-clock latency. All metrics report mean $\pm$ standard deviation across 60 queries.

3 Results

3.1 Embedding Quality

Table 1: Main retrieval results (all 60 queries)

Configuration	nDCG@10	Recall@10	MRR
OpenAI Baseline	**0.794**±0.28	**0.049**±0.02	**0.886**±0.27
OpenAI TimeDecay*	0.734±0.26	0.044±0.02	0.857±0.29
OpenAI Hybrid	0.727±0.28	0.043±0.02	0.815±0.32
UForm Baseline	0.347±0.29	0.021±0.02	0.563±0.40
UForm TimeDecay*	0.359±0.19	0.017±0.01	0.552±0.37
UForm Hybrid	0.352±0.23	0.018±0.01	0.554±0.38

*Best $\lambda = 0.001$ for both models.

OpenAI embeddings significantly outperform UForm across all metrics and all search strategies. The quality gap is substantial: 2.3× on nDCG@10, 1.6× on MRR.

3.2 Per Query-Type Analysis

Table 2: nDCG@10 by query type (baseline cosine)

Model	Temporal	Topical	Mixed
OpenAI	0.719	**0.890**	0.773
UForm	0.363	0.334	0.345

OpenAI's advantage is most pronounced on topical queries (0.890 vs. 0.334), suggesting stronger semantic understanding. Both models show weaker performance on temporal queries, as expected—pure embedding similarity cannot capture time.

3.3 Temporal Search Strategies

Counterintuitively, time-decay *hurts* both models. The best $\lambda = 0.001$ (nearly no decay) marginally helps UForm's nDCG but hurts recall. Stronger decay ($\lambda \geq 0.005$) consistently degrades all metrics. This occurs because our ground truth values topical relevance over temporal proximity—a document from the "wrong" month but right topic is more relevant than a recent off-topic document.

Hybrid pre-filtering shows negligible benefit, likely because the ± 1 month window still contains many topically irrelevant documents.

Table 3: Operational characteristics

Metric	UForm	OpenAI
Embedding dim	256	1,536
Embed throughput (docs/s)	**101**	77
E2E query latency (ms)	**35.3**±15	403.6±247
Search-only latency (ms)	**1.6**±1.7	28.7±7.9
Index memory (KB/1000 docs)	**1,000**	12,000
Requires network	No	Yes
API cost	$0	$0.02/1M tok

3.4 Speed and Efficiency

UForm's operational advantages are substantial: 11× lower E2E latency, 12× smaller memory footprint, no network dependency, zero marginal cost.

4 Discussion

The results reveal a clear quality–efficiency trade-off. For applications where retrieval quality is paramount (e.g., knowledge synthesis, question answering), OpenAI embeddings are worth the latency and cost penalty. For latency-sensitive or privacy-critical applications (e.g., real-time autocomplete, offline-first tools), UForm provides acceptable quality at dramatically lower operational cost.

Temporal search findings suggest that naive time-decay is counterproductive when topical relevance dominates. More sophisticated approaches—query-time temporal intent detection, learned decay parameters, or separate temporal and semantic scores with learned fusion—may yield better results.

Limitations: (1) Synthetic corpus may not capture real document diversity; (2) Programmatic ground truth may favor certain retrieval patterns; (3) UForm's 256d model is the smallest variant—larger UForm models may close the quality gap; (4) OpenAI latency depends on network conditions and API load.

5 Conclusion

OpenAI `text-embedding-3-small` delivers 2.3× better nDCG@10 than UForm3-256d on personal knowledge search, but at 11× higher latency and 12× higher memory cost. Simple temporal decay strategies hurt both models. For production personal knowledge systems, we recommend a tiered approach: UForm for real-time search with OpenAI for quality-critical retrieval tasks.

Reproducibility: All code, synthetic data, and ground truth are available. Dataset generation uses seed=42 and deterministic templates.