NodeMind

Real-World Benchmark Report — 500,000 Chunks

NodeMind replaces float32 vector indexes with compact binary fingerprints and Multi-Index Hashing (MIH) search — pure integer arithmetic, no GPU, no external vector database. This report presents independently verifiable results on a real mixed corpus of 500,000 text chunks.

32×	48×	96×	1.000
vs float32 RAG	vs HNSW index	with BGE-base 256-bit	Recall@10

1. Corpus & Methodology

The benchmark corpus is intentionally mixed — it covers general knowledge, scientific literature, and long-form prose to stress-test generalization. All sources are public domain or open-access.

Source	Domain	Approx. Size	Description
Wikipedia (Simple English)	General knowledge	~100 MB	Encyclopedia articles
arXiv papers	Science / ML	~40 MB	CS & ML abstracts + intros
Project Gutenberg	Literature / Prose	~28 MB	Public domain books
Total	Mixed	~168 MB raw	642,939 paragraphs

Chunking: 400 words per chunk, 50-word overlap. Final chunk count: 500,000. Embedding: BAAI/bge-m3 (1024-dim) on NVIDIA A40 (46 GB VRAM). Total embedding time: 42.5 minutes. Recall evaluation: 1,000 queries sampled from the corpus; ground truth = exact cosine top-k on the float32 embeddings.

NOTE: Self-retrieval protocol — queries are derived from corpus chunks, which is optimistic for binary methods. See Section 5 (Caveats).

2. Retrieval Accuracy

BGE-M3 · 1024-bit binary fingerprints

Metric	NodeMind MIH	Ground Truth
Recall@1	0.999	1.000
Recall@3	0.999	1.000

Recall@5	1.000	1.000
Recall@10	1.000	1.000
Recall@20	1.000	1.000
MRR@10	0.9992	1.000

BGE-base · 768-bit and 256-bit (PCA) fingerprints

Metric	768-bit	256-bit (PCA)
Recall@1	0.999	1.000
Recall@3	1.000	1.000
Recall@5	1.000	1.000
Recall@10	1.000	1.000
Recall@20	1.000	1.000
MRR@10	0.9995	1.000

Same 500K corpus and evaluation protocol as BGE-M3. BGE-base embedding time: 21.0 minutes on the same A40.

3. Index Size Comparison

All sizes are for the index structure only — the document text corpus is stored separately and equally in all systems.

Index	Size (500K chunks)	Bytes/chunk	vs float32
NodeMind BGE-M3 (1024-bit)	64 MB	128 B	32×
Float32 RAG — BGE-M3 (baseline)	2,048 MB	4,096 B	1× (reference)
HNSW index (float32 × 1.5× overhead)	3,072 MB	6,144 B	0.67× → 48× vs NM
NodeMind BGE-base 256-bit (PCA)	16 MB	32 B	96×

HNSW 1.5× overhead is a conservative standard estimate — actual FAISS HNSW size varies with M (graph degree) and efConstruction settings. Download the HNSW Size Reference file for the exact formula.

4. How It Works

1. Embed

Text is chunked and embedded with a sentence model (BGE-M3 or BGE-base), producing a float32 vector per chunk.

2. Binarise

Each float32 embedding is converted to a compact binary fingerprint using pre-computed index metadata. The result is 1024 bits (128 bytes) per chunk for BGE-M3, or 256 bits (32 bytes) with BGE-base + PCA. The conversion is integer-only — no GPU required at query time. The exact binarisation method is patent-protected (AU 2026901656) and is a trade secret; it is not disclosed in this document.

3. Index (MIH)

Binary fingerprints are stored in a Multi-Index Hash structure. At query time, candidates are found in matching hash buckets and re-ranked by full Hamming distance. Pure integer arithmetic — runs on any CPU. MIH structure follows Norouzi et al. (CVPR 2012). The novel contribution — CTV binarisation and portable single-file format — is covered under AU 2026901657.

4. Query

A query string is embedded with the same model, binarised using the index metadata, and searched via Hamming distance. The index pkl file is self-contained — no external server required.

5. Honest Caveats

Self-retrieval benchmark: Queries are perturbed versions of corpus chunks. This is optimistic for binary methods — real end-to-end QA accuracy on BEIR or MS MARCO has not yet been measured. Results may differ on out-of-distribution queries.

HNSW comparison is size-only: Real FAISS HNSW achieves recall@10 of 0.95–0.99 on most corpora using graph traversal. NodeMind achieves recall@10 = 1.000 on this self-retrieval test — a direct neutral head-to-head benchmark has not yet been conducted.

96× requires a lighter model: The 96× figure uses BGE-base (768-dim) + PCA compression to 256 bits. If you need BGE-M3 (stronger, cross-lingual), you get 32×/48×.

Text-only corpus: Tables, code blocks, and multi-modal documents were not tested in this benchmark.

Float32 RAG download is 2 GB: Budget the bandwidth if you want to download the float32 baseline to verify compression ratios yourself.

6. Download & Verify

All indexes are hosted at nodemind.space/benchmark. Download NodeMind + float32 RAG side by side to verify compression ratios yourself.

File	Size	What it is
nm_bgem3_index.pkl	64 MB	NodeMind BGE-M3 binary index (32×)
rag_bgem3_index.pkl	2,048 MB	Float32 RAG baseline — verify 32× yourself
hnsw_size_reference.txt	<1 KB	HNSW size formula — explains the 48× number
nm_bgebase256_index.pkl	16 MB	NodeMind BGE-base 256-bit (96×)
corpus.pkl	~144 MB	500K text chunks shared by all indexes

Quick verification snippet

```
import pickle
with open('nm_bgem3_index.pkl','rb') as f: nm = pickle.load(f)
with open('rag_bgem3_index.pkl','rb') as f: rag = pickle.load(f)
nm_mb = nm['fps'].nbytes / 1e6 # → 64
rag_mb = rag['embeddings'].nbytes / 1e6 # → 2048
print(f'Compression: {rag['embeddings'].nbytes // nm["fps"].nbytes}x')
```

7. Patents & IP

AU 2026901656 — WHT Integer Codec. Integer-only binarisation without learned projection. Filed IP Australia, May 2026.

AU 2026901657 — NodeMind Centroid MIH. CTV-based binary fingerprinting and Multi-Index Hash search. Filed IP Australia, May 2026.

The binarisation algorithm and index structure are trade secrets not disclosed in this document. The index pkl files are self-contained and functional without knowledge of the underlying method.

NodeMind · nodemind.space · Built in Coleambally, NSW, Australia · Owner: Sai Kiran Bathula