LightRAG Tuning — 2026-05

LightRAG Tuning — May 2026

Last Updated: 2026-05-12 (MC #100467)
Status: LIVE

Current Config (LIVE as of 2026-05-12 21:13)

Parameter	Value	Changed From
`cosine_threshold`	0.5	0.2
`related_chunk_number`	10	5
`enable_rerank`	false	(unchanged, deferred)

Why These Values

AgentForge audit (Chip Huyen lens, MC #100451) identified 2 quick-win retrieval optimizations:

- **Cosine 0.5:** Industry standard for 768-dim embeddings (bge-m3). Filters false-positive chunks that pollute LLM context with noise. **Expected:** 8-12% token savings per query.
- **Chunks 10:** Broader context window for multi-faceted queries (e.g., "explain Pillar #9 DR strategy"). Reduces re-query loops when 5 chunks = incomplete answer. **Expected:** 6-10% fewer re-queries.

Proveo validation (MC #100458): 8/10 test queries rated ≥3/5 quality, +15-30% context delta likely (ceiling estimate — API lacks chunk-count telemetry).

What We Did NOT Touch (and Why)

Forbidden changes until MC #100009 backlog stabilization ships:

- `embedding_batch_num: 10` — raising risks OOM on bge-m3 (already at memory ceiling)
- `max_parallel_insert: 2` — parallelism = more heap pressure
- `max_async: 4` — async I/O ceiling, won't help if bottleneck = compute
- `embedding_model` switch (e.g., to smaller all-MiniLM-L6-v2) — would BREAK all existing embeddings, require full re-index

Reason: These params affect the ingest pipeline. LightRAG already has 121K doc backlog + memory pressure. Retrieval-tuning (cosine, chunks) is safe because it's query-time only.

Validation Summary

Proveo 10-query test suite (MC #100458):

Metric	Result
Queries with quality ≥3/5	8/10 (PASS threshold: 7/10)
HTTP 500 errors	0/10
Estimated context token delta	+15-30% (ceiling +40%, likely lower in practice)
Response quality by bucket	Product/code queries strongest (3.7/5 avg), process queries weakest (2.5/5 avg)

Proveo verdict: REQUEST_CHANGES (functional pass, but lacks chunk-count telemetry to machine-verify actual cost impact)

Open Work

- **MC #100467:** This documentation (COMPLETE)
- **MC #100468:** TEI reranker investigation (bge-reranker-base unavailable in Ollama) — highest ROI optimization (15-30% quality lift) deferred
- **MC #100469:** API chunk-count telemetry (add `chunks_retrieved` to /query response for cost verification)

How to Verify Live State

```
curl -s http://localhost:9621/health | jq .configuration
# Look for: cosine_threshold=0.5, related_chunk_number=10, enable_rerank=false
```

Evidence snapshots:

- Before: `/tmp/lightrag-baseline-100458-raw.json`
- After: `/tmp/lightrag-postverify-100458.json`

How to Revert (If Needed)

```
cd /Users/makinja/system/docker/lightrag

# Revert .env
sed -i '' '/# Retrieval Tuning/,+3d' .env

# Revert compose
git checkout docker-compose.yml  # or manual edit if not git-tracked

# Recreate container
docker compose down && docker compose up -d lightrag

# Verify restoration
curl -s http://localhost:9621/health | jq '.configuration.cosine_threshold,
.configuration.related_chunk_number'
# Expected after rollback: 0.2, 5
```

Related Resources

- **ADR-026:** `~/system/specs/adr-026-lightrag-tuning-2026-05-12.md`
- **AgentForge audit:** `~/system/artifacts/lightrag-100458/lightrag-audit-100451.md`
- **FlowForge report:** `~/system/artifacts/lightrag-100458/flowforge-100458-report.md`
- **Proveo validation:** `~/system/artifacts/lightrag-100458/proveo-100458-validation.md`

Revision #2
Created 2026-05-12 19:39:22 UTC by John
Updated 2026-06-14 20:03:12 UTC by John