

LightRAG G1 complete blue-green rebuild — MC #900195

LightRAG G1 complete blue/green rebuild — MC #900195

Status: IN PROGRESS / resumable final G1 ingest supervisor running; cutover blocked **Date:** 2026-08-24 **Worktree:** `/Users/makinja/system/.claude/worktrees/lightrag-g1-900195`

Decision

Do not repair or drain the existing G0 corpus in place. Build a separate G1 generation, validate it, then cut over with G0 preserved as rollback. Pi and Claude Code continue using retrieval-only `safe-query`; server-generated answers and automatic per-prompt injection remain disabled.

Verified baseline

- Live G0: LightRAG core `1.3.4`, healthy, 28,749 processed statuses, 28,751 full-doc records, 21,267 text chunks, 85,807 entities.
- Attribution: only 546/28,751 full documents (`1.899%`) and 387/21,267 chunks (`1.82%`) are attributed.
- Frozen queue: 14,006 unique hashes / 37,466,092 content bytes, all `queued`, dated 2026-04-16 through 2026-07-31.
- The frozen queue contains 1,193 superseded source versions; among 4,500 latest local evidence/spec/rule identities, 343 files are gone and 31 hashes are stale.
- All six broad ingest/integrity LaunchAgents are `NOT_LOADED`.
- Safe on-demand retrieval is live from MC #900176, but the historical corpus remains too unattributed for complete coverage.

Inventory: `/Users/makinja/system/evidence/900176/full-recovery-inventory-2026-08-24.json`.

Upgrade target

Use the official stable HKUDS release, not `latest` and not the `v1.5.7rc2` release candidate:

- Version: `v1.5.6`
- Image: `ghcr.io/hkuds/lightrag:v1.5.6`
- Multi-arch digest: `sha256:ab23a9c83a735901b18c8960b6b482b602d5b6291abb7e07c5776f7bb2da504e`
- ARM64 manifest: `sha256:65f210a9f5fb41d64c24ca4979aa329f617b1fac26105c370853bfcf2cb046bf`

The source contract was verified directly at tag `v1.5.6`: text insertion requires valid `file_source`, context-only returns before answer-cache handling, query results include structured raw data/references, and custom chunking is supported.

Build waves

G1.1 — Official isolated runtime

- Dedicated `lr-g1-900195-*` volumes/network and loopback port.
- Official image pinned by digest; no G0 runtime monkey patches.
- Existing local Ollama and BGE-M3; Neo4j isolated from G0.
- Prove health, auth mode, native `file_source`, structured context, cache separation, restart persistence, and zero G0 mutation.

G1.2 — Canonical source registry

Versioned policy with include/exclude rules and stable source IDs:

- Include current BookStack pages, active rules/specs, completed MC outcomes, and explicitly curated durable evidence.
- Exclude sessions, runtime/generated evidence, logs, backups, archives, secrets, vendor/build output, and stale/superseded versions.
- Each source gets native `file_source`, in-text provenance envelope, content hash, source timestamp, and authority class.

G1.3 — Regenerated latest-only queue

- Build `ingest-queue-g1.sqlite` from current authoritative sources.
- One latest row per canonical source identity/chunk.
- Never copy or drain G0's 14,006 queue.
- Deterministic manifest and source counts; queue remains reproducible.

G1.4 — Model and throughput benchmark

- Small representative corpus first.
- Compare available local extraction models only if necessary; keep BGE-M3 fixed.
- Select the fastest model that passes extraction and retrieval gates.
- Serial/backpressured ingestion with restart/resume evidence.

G1.5 — Full curated ingest and evaluation

- 35–50 gold queries across rules, infrastructure, products, MC outcomes, and specs.
- Negative tests for stale GCP Bilko facts, unknown sources, session noise, auth errors, and cache poisoning.
- Retrieval-only output remains the agent contract.

G1.6 — Cutover and monitoring

- Switch local/Cloudflare origin only after independent PASS.
- Verify health plus authenticated real safe-query, source attribution, latency, freshness, and no queue growth.
- Preserve G0 stopped with all volumes for immediate rollback.

Current progress — 2026-08-24

- BookStack-first takeover verified at 13:55 CEST. MC #900195 lease is held by session `01a03398-47e3-7a20-8186-71d1f3c0205b`; current files and machine evidence, not prior transcripts, are authoritative.
- Fusion Harness completed two bounded builder/validator loops. Commits `502b512c4` and `d0d6e272f` close archive leakage, observer-timeout misclassification, fixed-date freshness aging, sensitive-data gaps, proxy readiness, bounded prechunking, stale queue defaults, and automatic truncation-log preflight. Follow-up commits `ed0df54ba` and `244146959` apply measured extraction-limit iterations.
- Benchmark v2 completed all 20 documents with zero failed/timeouts, but preflight correctly blocked it on five `done_reason=length` warnings at output cap `2,400`. Benchmark v3 at output cap `4,096` again completed 20/20 with zero failed/timeouts, but remained blocked on four length warnings. Neither result was promoted as PASS.
- The v3 warnings were sanitized and correlated to three dense source documents without recording model/document bodies. Because raising the output cap did not eliminate truncation, the per-request server chunk size was reduced from `800` to `400` tokens.
- Canonical v7 queue at `/Users/makinja/system/state/lightrag-g1/ingest-queue-g1-900195-v7.sqlite` has 4,407 pending rows and verifier PASS: unique source/native/hash IDs `4,407/4,407/4,407`; archive `0`; oversized `0`; unknown source `0`; sensitive `0`; disallowed

BookStack books 0; maximum final content length 13,875; every row uses server chunk size 400.

- Benchmark v4 finished after 2,799 seconds: 19 processed, 1 failed, 0 observation timeouts. The one remote failure was a separate embedding-worker timeout; extraction preflight also found three done_reason=length warnings. Reducing server chunk size from 800 to 400 lowered truncation from four to three warnings but did not clear the gate. Evidence directory: /Users/makinja/system/evidence/900195/benchmark-v4/.
- Root-cause correction now uses official LightRAG 1.5.6 structured extraction: ENTITY_EXTRACTION_USE_JSON=true, maximum 40 total records / 20 entities, think:false, output cap 4,096, and embedding timeout 120s. A targeted three-document probe covering both previously truncating sources and the embedding-timeout source passed 3/3 in 271 seconds with zero failures, timeouts, or truncation/format warnings.
- Fresh 20-document benchmark v6 passed after 2,889 seconds: attempted 20, accepted 20, processed 20, failed 0, observation timeouts 0, circuit open false. Extraction preflight PASS: done_reason=length 0, incomplete delimiters 0, partial extraction 0; log fetch and runner exit both 0. Evidence directory: /Users/makinja/system/evidence/900195/benchmark-v6/.
- G1 remains isolated and healthy on official core 1.5.6; think:false, output cap 4,096, and sidecar upstream readiness remain active. Frozen G0 remains exactly 14,006 rows / 37,466,092 content bytes, all queued; no G0 queue, container, or volume was drained or deleted.
- R2-F1 is closed by benchmark v6. The benchmark generation is stopped-not-removed; all seven named volumes were snapshotted read-only with SHA-256 manifests (preservation-verdict: PASS) while G0 stayed healthy and unchanged.
- Metadata-only curation found 123 candidates, below the 221-row rebuild threshold. Eleven non-contiguous retained chunk-index groups were adjudicated as expected post-numbering exact-content dedup gaps: 129 missing positions fit inside 154 measured exact-content drops; source/native duplicate counters remain zero. Queue v7 is retained without age, fuzzy-title, or speculative exclusions.
- Isolated concurrency-2 benchmark passed: two workers processed 10/10 each, total 20/20, failures/timeouts/warnings 0, OOM/restarts 0, elapsed 1,397s versus serial 2,889s (~2.07x; also above the de-throttled 1.5x gate). G0 remained exactly 14,006 rows / 37,466,092 bytes.
- Fresh final generation lightrag-g1-900195-final-v7 is healthy and empty on 127.0.0.1:19641, with disjoint containers/network/seven volumes and promoted concurrency 2. Pristine v7 was cloned to a mutable run queue with logical SHA-256 equality: 4,407 pending and verifier PASS.
- Controlled ingest wave 001 passed: two workers processed 50/50 each, total 100/100, failures/timeouts/circuit/truncation warnings 0, verifier PASS, remote inventory 100, container ID unchanged, restarts/OOM 0, G0 unchanged. Elapsed 5,940s.
- Durable orchestration is committed in 0aa42a1fe plus resume-safety correction 63e03eff3: dynamic 100-row waves and final 7-row wave, partial-wave resume without duplicate upload, peer termination on failure, nested reconciliation/status checks, per-wave verifier/preflight/container/G0 gates, signed PASS summaries, and read-only seven-volume snapshots after waves 11/22/33/44. Offline orchestration tests 12/12 and related

G1 tests 41/41 pass.

- Supervisor tmux lr-g1-final-supervisor-900195 is running wave 002. At the 02:20 CEST checkpoint: 104 processed, 2 processing/accepted, 4,301 pending. Any failure, timeout, warning, OOM/restart, verifier regression, or G0 drift stops the supervisor.
- Full ingest must finish all 4,407 rows with zero failure/timeout/warning before the stricter 35–50 query evaluation, negative tests, rollback demonstration, independent verification, and any staged cutover. Expected ingest time is roughly 3–4 days at validated concurrency. Edge exposure and cutover remain **BLOCKED**.

Acceptance

1. Official pinned G1 runtime starts without G0 patches and survives restart.
2. G0 container IDs, volumes, corpus counts, and 14,006-row queue remain unchanged during G1 build.
3. New corpus has 100% native file-source/provenance coverage and zero unknown_source in top-5.
4. Latest-only queue contains no missing files, stale hashes, or superseded source versions.
5. Recall@5 ≥ 0.85 , MRR ≥ 0.65 , exact-term coverage ≥ 0.90 , stale-contradiction rate 0, p95 retrieval $\leq 3s$.
6. BookStack/MC freshness canaries complete within 15 minutes after steady-state adapters are enabled.
7. Fresh Pi and Claude Code use only attributed safe-query; generated-answer modes and --allow-unknown are not used for answers.
8. Independent implementation and post-cutover verification pass with P0=0/P1=0.
9. G0 rollback is demonstrated before G1 becomes the sole live generation.

Hard safety boundaries

- Never run the existing 14,006-row drain.
- Never attach G1 to lr106747-* volumes.
- Never delete G0 volumes during this task.
- Never expose credentials in compose, evidence, logs, or prompts.
- No production cutover before the full evaluation and independent verifier pass.

Rollback

Before cutover, stop and remove only G1 containers/network; retain its volumes for evidence. After cutover, revert Cloudflare/local origin to G0 and start the unchanged G0 compose. Volume deletion is out of scope and requires separate explicit approval.

