Pillar #3 — L3 Memory

Framework comparison (Mem0, claude-mem, LightRAG) for John L3 memory layer

Framework Comparison Matrix

```
## §3 — Feature Matrix (D1 / AC#1)

Key: S=small (<8h), M=medium (<80h), L=large (>80h); EVIDENCE lines follow each cell.

| Framework | storage_backend | embedding_model | extraction_method | recall_at_10 | latency_p50
|-----------|----------------|-----------------|-------------------|--------------|------------
| Mem0 self-hosted | Qdrant (local port 6333) | bge-m3:latest 1024-dim (Ollama) | LLM fact extra
| claude-mem | Filesystem SQLite (observations) | None (BM25 only) | Session observation indexir
| mem-search | NOT VIABLE | NOT VIABLE | NOT VIABLE | NOT VIABLE | NOT VIABLE | NOT VIABLE | NOT
| Memipalace | NOT VIABLE | NOT VIABLE | NOT VIABLE | NOT VIABLE | NOT VIABLE | NOT VIABLE | NOT
| LightRAG-resurrect | Neo4J (graph) + NanoVectorDB (vector) + JsonKV | bge-m3:latest (Ollama vi

EVIDENCE (Mem0 storage_backend): curl http://localhost:6333/collections/mem0_john 2026-05-04T21:
EVIDENCE (Mem0 embedding_model): /Users/makinja/system/mem0/config.py lines 72-80 → model:bge-m3
EVIDENCE (Mem0 recall_at_10): forged-99124 §OBJECTIVE → "Phase 1 baseline 80% recall@10"; /Users
EVIDENCE (Mem0 latency_p50_ms): Qdrant collection indexed_vectors_count=0 → full scan path; no H
EVIDENCE (Mem0 multi_user_isolation): /Users/makinja/system/tools/discover.js line 677 → user_id
EVIDENCE (Mem0 oauth_compatible): /Users/makinja/system/mem0/config.py — no Anthropic SDK; all l
EVIDENCE (Mem0 license): mem0ai-2.0.1.dist-info in venv site-packages; Apache-2.0 per mem0ai PyP
EVIDENCE (claude-mem storage_backend): /opt/homebrew/bin/claude-mem search 'test' → 67 results (
EVIDENCE (claude-mem embedding_model): package.json — no vector deps; BM25 only confirmed by sea
EVIDENCE (claude-mem license): /opt/homebrew/lib/node_modules/claude-mem/package.json → license:
EVIDENCE (claude-mem last_release): /opt/homebrew/bin/claude-mem --version → 12.5.0
EVIDENCE (claude-mem oauth_compatible): package.json — no @anthropic-ai/sdk in dependencies
EVIDENCE (mem-search NOT VIABLE): brew search mem-search → meilisearch (unrelated); npm registry
EVIDENCE (Memipalace NOT VIABLE): GitHub API search q=Memipalace 2026-05-04T21:12Z → items:[] ze
EVIDENCE (LightRAG storage_backend): curl http://20.240.61.67:9621/health 2026-05-04T21:07Z → gr
EVIDENCE (LightRAG recall): az vm run-command /documents 2026-05-04T21:14Z → processed:5596 penc
EVIDENCE (LightRAG latency): lightrag-freeze-decision-chip.md §1 → /health hangs 15-30s during e

---
```

Cost Analysis

```
## §4 — Cost Matrix Monthly (D2 / AC#2)

Load assumptions per forged prompt D2: 200 queries/day × 30d = 6,000 queries/month;
Stop-hook extraction: ~10 sessions/day × 30d = 300 extraction events.

### Scenario (a): $30 combined Pillar #9 + L3 (chip-huyen SC-2 interpretation)

L3 max = $30 − $16.70 (Pillar #9 incremental) = **$13.30/month L3 ceiling**.

| Framework | compute | vector_storage | LLM_inference | embedding | egress | hosted_tier | TOTA
|-----------|---------|----------------|---------------|-----------|--------|-------------|-----
| Mem0 self-hosted | $0 (ANVIL local) | $0 (Qdrant local) | $0 (Ollama qwen3:8b local) | $0 (bge
| claude-mem | $0 (Node.js local) | $0 (filesystem) | $0 (no LLM) | $0 (no embedding) | $0 | N/A
| mem-search | NOT VIABLE | — | — | — | — | — | — |
| Memipalace | NOT VIABLE | — | — | — | — | — | — |
| LightRAG-resurrect | $0 incremental (vm-alai-lightrag already running ~$30/mo in existing budg

EVIDENCE (Mem0 cost $0): /Users/makinja/system/mem0/config.py lines 17-21 → QDRANT_HOST=localhos
EVIDENCE (LightRAG incremental): /Users/makinja/system/specs/agentic-os-pillar9-runtime-2026-05-
EVIDENCE (CEO $30 ceiling): /Users/makinja/.claude/projects/-Users-makinja/memory/project_99063_
EVIDENCE (Pillar #9 incremental $16.70): /Users/makinja/system/specs/agentic-os-pillar9-runtime-

**Combined Pillar #9 + L3 total (scenario a, Mem0 winner):**
- Pillar #9 incremental: $16.70/month
- L3 Mem0 incremental: $0/month
- **Total: $16.70/month — under $13.30 L3 ceiling AND under $30 combined ceiling**

### Scenario (b): $30 L3-only incremental, Pillar #9 separate

Mem0: $0/month incremental (already deployed). $30 L3 budget entirely unspent.
LightRAG-resurrect: ~$1/month incremental. Also fits.

### Scenario (c): $40-50 combined ("može više" per CEO Q3)

At $40-50 combined, all 3 viable frameworks remain inside envelope. The question is effort
and capability, not cost. This scenario changes nothing about the winner selection.

**CEO Decision Item:** See §11, item #1.

---
```

Integration Effort

```
## §5 — Integration Effort Estimate (D3 / AC#3)

| Framework | hours | dependencies | blocking_tasks | hooks_touched | agents_touched | settings_
|-----------|-------|--------------|----------------|---------------|----------------|----------
| Mem0 self-hosted | S (0h remaining — deployed) | Qdrant, Ollama, mem0ai-2.0.1 venv, server.py
| claude-mem | S (0h remaining — already installed) | /opt/homebrew/bin/claude-mem, Node.js daem
| mem-search | N/A — NOT VIABLE | — | — | — | — | — | — |
| Memipalace | N/A — NOT VIABLE | — | — | — | — | — | — |
| LightRAG-resurrect | L (>80h across multiple MCs) | MC #99093 closure (file_path fix), Semapho

**Concrete LightRAG effort checklist (for reference):**
1. /Users/makinja/system/tools/lightrag-health.sh: add auto-restart block (~50 lines) — 2h
2. Capture py-spy dump on next freeze overnight — wait 12-24h
3. Patch lightrag/api/routers/document.py Semaphore(2) — 4h
4. MC #99093: bookstack-enrich.js re-ingest with file_path URLs — 8-16h separate MC
5. Cross-VM access design (Azure VNet peering or CF tunnel rule) — 4-8h
6. discover.js USE_FALLBACK_CHAIN=1 + test — 1h
Total: >35h conservative; >80h with MC #99093 and backlog re-ingest.

EVIDENCE (Mem0 integration_effort=S): /Users/makinja/system/mem0/server.py on disk 6320 bytes; c
EVIDENCE (LightRAG integration_effort=L): lightrag-freeze-decision-chip.md §3 Option E → ~6h for
EVIDENCE (claude-mem integration_effort=S): /opt/homebrew/bin/claude-mem exists, v12.5.0; discov

---
```

Winner: Mem0 Self-Hosted

```
## §6 — Recommended Winner + Rationale (D4 / AC#4)

**Pre-condition gate:** /tmp/forged-99124-evidence.jsonl contains 42 records (≥40 required).
Verified: `wc -l /tmp/forged-99124-evidence.jsonl → 42` at 2026-05-04T21:20Z.

### winner: Mem0 self-hosted

### runner-up: claude-mem

### decision_matrix_score

Weights (CEO-locked):

| Factor | Weight | Mem0 | claude-mem | LightRAG-resurrect |
|--------|--------|------|------------|--------------------|
| Pillar #9 compatibility (hard gate) | GATE | PASS | PASS | PASS |
| $30 combined ceiling (hard gate) | GATE | PASS ($0 L3 incr.) | PASS ($0) | PASS (~$1) |
| OAuth-only auth (hard gate) | GATE | PASS (local Ollama) | PASS (no LLM client) | PASS (Ollama
| Semantic recall capability | 30% | 9/10 (vector search, 865 facts, 80% baseline) | 2/10 (BM25
| Current deployment state | 25% | 10/10 (running, wired) | 7/10 (installed, not primary) | 4/10
| Multi-client SVE isolation | 20% | 6/10 (user_id field exists; needs schema extension) | 1/10
| Integration risk | 15% | 9/10 (lowest risk, already passing Phase 1) | 7/10 (zero infra risk,
| Recall@10 ≥80% (chip-huyen SC-1) | 10% | 10/10 (80% confirmed) | 1/10 (no baseline, BM25 limit

**Weighted scores:**
- Mem0: (0.30×9 + 0.25×10 + 0.20×6 + 0.15×9 + 0.10×10) = 2.7+2.5+1.2+1.35+1.0 = **8.75**
- claude-mem: (0.30×2 + 0.25×7 + 0.20×1 + 0.15×7 + 0.10×1) = 0.6+1.75+0.2+1.05+0.1 = **3.70**
- LightRAG-resurrect: (0.30×7 + 0.25×4 + 0.20×3 + 0.15×2 + 0.10×3) = 2.1+1.0+0.6+0.3+0.3 = **4.3

### defend_stop-hook-l3-memory-spec

The pre-commitment in `stop-hook-l3-memory-spec.md` (MC #99071) is **DEFENDED**.

Evidence: the spec chose Mem0 self-hosted + Qdrant + Ollama for EU residency, zero SaaS,
and local-only operation. All three constraints remain valid in 2026-05-04 context. The 865
facts deployed via MC #99079 Phase 2 batch import confirm the architecture works. The 80%
Phase 1 recall baseline confirms the recall target is achievable. Nothing in the MC #99124
research overrides this choice.

### why_not_others

**claude-mem:** BM25 keyword search cannot replace semantic vector recall. When John asks
"what was the root cause of the Drop outage?" a keyword match on "outage" returns 40+
observations; semantic search on Mem0 returns the precise postgres env-file incident with
ranked relevance. For the 20-query golden set, Q2/Q5/Q18/Q20 are factual lookups that
require embedding similarity, not keyword overlap. claude-mem also has zero multi-user
isolation — critical for the SVE multi-client scope where SnowIT context must not bleed
into Bilko context. AGPL-3.0 license creates commercial-use risk for client-facing
```

deployments. Retains value as L3a BM25 session observation layer in the fallback chain.

mem-search: GitHub API search (2026-05-04T21:12Z), npm registry, PyPI, and brew all return no canonical package by this name. The YouTube source video (w0S-khYCaB4) uses "mem search" as a category description for semantic recall tools, not as a specific product. No installation path, no version, no maintainer. Cannot be evaluated or deployed.

Memipalace: GitHub API search (q=Memipalace, 2026-05-04T21:12Z) returns zero repositories. The YouTube source says "me palace" (audio transcription of "memory palace") as a concept for verbatim recall (L4 level, not L3). No software package exists under this name. Cannot be evaluated or deployed.

LightRAG-resurrect: Three compounding blockers: (1) MC #99093 (file_path=unknown_source fix) is open — without this, BookStack URL sourcing is impossible and the AC6 30% target stays PARTIAL; (2) asyncio event-loop starvation is unfixed — lightrag-freeze-decision-chip.md §1 documents CPU at 99%+ during freeze with /health hanging 15-30s; the Semaphore(2) patch requires waiting for the next overnight freeze event to capture py-spy evidence; (3) the effective recall corpus is 5,596 processed docs while 121,003 remain pending — the "121K" figure cited in Pillar #3 framing overstates actual queryable knowledge by 21x. Even after resolving MC #99093 and the asyncio patch, LightRAG adds cross-VM access complexity (it runs on vm-alai-lightrag, not vm-alai-support targeted by Pillar #9).

kill_criteria

Conditions that would invalidate the Mem0 winner choice within 6 months:
1. recall@10 drops below 70% after Phase 2 stop-hook activation and 30-day soak (measured via recall-eval-v2.sh Q1-Q20 baseline comparison)
2. Ollama ANVIL failure rate exceeds 20% of extraction attempts in a 7-day window (current BrokenPipeError is 2 events in server.log — acceptable; >20% is not)
3. Multi-client SVE schema cannot be extended beyond user_id='john' without a full collection-per-client migration costing >40h (§8 must clarify this by Phase 3)

tradeoffs_accepted

- HNSW index not built at 865 points (full scan latency ~200ms acceptable at this scale; index will build automatically when points_count exceeds 10,000)
- No graph-style entity relationships (LightRAG strength abandoned); Mem0 recall is semantic similarity, not graph traversal — acceptable for L3 operation facts
- AGPL-3.0 claude-mem in fallback chain creates license dependency; mitigated by it being a read-only search tool, not a deployed service

dissent_log

anthropic-architecture concern: AC6 of MC #99079 returned PARTIAL because LightRAG ingestion lacks file_path source URLs. Do not assume 121K docs are usable — the effective corpus is 5,596. INCORPORATED: §2.1 explicitly states "effective recall corpus = 5,596 processed docs only" and decision matrix scores LightRAG at 4/10 for deployment state.

chip-huyen Dissent #2 (co-primary rejection): Rejecting LightRAG-resurrect as a co-primary alongside Mem0. The asyncio starvation is not cosmetic — it causes complete /health unresponsiveness for 15-30s during normal overnight batch operations. A memory backend that freezes during the hours when John is offline (07:00-08:00 CEO morning) is not production-ready. Mem0's single-process Python server with Ollama dependency had one

```
BrokenPipeError in logs — materially different failure mode. INCORPORATED: singular winner,
no co-primary.

---
```

Migration Plan

```
## §7 — Migration Plan (D5 / AC#5)

Winner = Mem0 self-hosted. No migration away from existing deployment required.
Plan = activation of Phase 2 items from stop-hook-l3-memory-spec.md.

**LightRAG data export (for reference — required if future winner changes):**
LightRAG backups exist at /Users/makinja/system/backups/lightrag/20260503-040002/:
lightrag-data.tar.gz, lightrag-kg.tar.gz, lightrag-cache.tar.gz, lightrag-neo4j-data.tar.gz.
Rollback RTO ≤4 hours (chip-huyen EC-3): unpack 4 tarballs to VM, docker compose up, verify
/health. Cypher export path: az vm run-command invoke --scripts "docker exec neo4j cypher-shell
-u neo4j -p  'MATCH (n) RETURN n' > /tmp/nodes.csv" (read-only).

**unknown_source probe result (D5 mandatory):**
unknown_source_ratio=31.6% (below 70% threshold). Useful corpus = 5,596 × (1−0.316) = 3,831
processed docs with file_path populated. The 121,003 pending docs overstates retrievable corpus.
EVIDENCE: az vm run-command python3 2026-05-04T21:14Z

| Step | Name | Owner | Timeline | Acceptance | Rollback | Dependency |
|------|------|-------|----------|------------|----------|------------|
| 1 | Enable L3 fallback chain | codecraft | 2026-05-05 | DISCOVER_USE_FALLBACK_CHAIN=1 in Launc
| 2 | Activate Stop hook (session-extract.js) | codecraft | 2026-05-07 | settings.json Stop arra
| 3 | Multi-client namespace extension | codecraft | 2026-05-10 | discover.js accepts --user-id
| 4 | Enable HNSW index at 1,000+ points | john (monitor) | Auto (Qdrant threshold=10,000) | inc
| 5 | Recall validation (Phase 3) | proveo | 2026-05-14 | recall-eval-v2.sh Q1-Q20 returns ≥80%

---
```

Pillar #9 Interplay & OAuth

```
## §8 — Pillar #9 Interplay + OAuth (D6 / AC#6)

### Topic 1 — Memory-layer location (laptop vs VM vs hybrid)

**Decision:** Mem0 = laptop-only (ANVIL) for now. Qdrant port 6333 and Ollama port 11434
are both ANVIL-local. vm-alai-support (Pillar #9) does not have direct access to ANVIL
ports.

**Topology gap:** Pillar #9 VM (vm-alai-support, 4.223.110.181) cannot reach ANVIL localhost:900
directly. Mem0 server is bound to 127.0.0.1. Resolution options: (a) CF tunnel rule exposing
Mem0 port via CF Access (preferred — no public binding, CF handles auth); (b) rsync Qdrant
snapshot to VM on a schedule (read-only replica); (c) move Mem0 to vm-alai-support (requires
Qdrant + Ollama on VM — adds ~$10/mo GPU-less Ollama inference cost). Chip-huyen EC-4: Mem0
bound to 127.0.0.1:9000 today (ANVIL-only). CF tunnel option is the lowest-risk path.
This is a Phase 3 decision — surfaces to §11 item #3.

### Topic 2 — OAuth-CLI-on-VM read/write authority boundary

LLM-client construction paths for each framework:

| Framework | LLM client construction | OAuth-compatible |
|-----------|-------------------------|-----------------|
| Mem0 self-hosted | /Users/makinja/system/mem0/config.py lines 67-77: `{"provider":"ollama","co
| claude-mem | /opt/homebrew/lib/node_modules/claude-mem/package.json — no @anthropic-ai/sdk dep
| mem-search | NOT VIABLE — no code path exists | N/A |
| Memipalace | NOT VIABLE — no code path exists | N/A |
| LightRAG-resurrect | /health response: `llm_binding_host:https://ollama.basicconsulting.no` —

EVIDENCE: config.py lines 67-77 (file confirmed on disk); claude-mem package.json; LightRAG /hea

All three viable frameworks are COMPATIBLE WITH PILLAR #9 OAuth model (no Anthropic API key requ

### Topic 3 — State-sync timing (rsync windows)

Qdrant data dir: /Users/makinja/.qdrant/storage (ANVIL local, not yet confirmed path).
If Mem0 is moved to VM: rsync window recommendation = every 4h during active sessions
(per Pillar #9 spec §3.3 state-sync design). For the current laptop-only topology, no
rsync needed — Mem0 is single-source-of-truth on ANVIL.

### Topic 4 — Multi-client SVE namespace isolation

Current state: `user_id='john'` hardcoded in discover.js line 677.
Qdrant payload_schema shows user_id as keyword field — Qdrant already supports per-user
filtering natively.

Two designs:
- **Design A (recommended): metadata filter** — single mem0_john collection, query with
  `payload filter user_id=`. Cost: zero additional infra. Risk: one corrupt
```

  write with wrong user_id bleeds facts. Mitigation: server.py write endpoint validates user_id against allow-list.
- **Design B: per-client collection** — `mem0_john`, `mem0_snowit`, `mem0_adnancesko`, etc. Clean isolation, harder to cross-search. Config change per client in config.py.

Recommendation: Design A for Phase 3 (lower ops overhead). Design B if client-count exceeds 10 or audit trail is required. Surfaces to §11 item #2.

Topic 5 — DR access path

If ANVIL (MacBook) goes offline:
- Mem0 data: no off-laptop copy today. Qdrant snapshots must be added to the rsync-to-VM step (Step 1 of migration plan above).
- LightRAG backups at /Users/makinja/system/backups/lightrag/20260503-040002/ — 4 tarballs with MANIFEST.sha256.
- Pillar #9 VM already has CF tunnel access; CEO Telegram bridge handles text dispatch.
- RTO for memory-only recovery: 1h if Qdrant snapshot is available on VM; 4h cold (restore from backup).

Validation Harness (20-Query)

```
## §9 — Validation Harness — 20-Query Golden Set (D7 / AC#7)

**Chip-huyen SC-3:** 20 queries from recall-eval-v2.sh lines 76-114 appear verbatim below.
**Execution:** OUT OF SCOPE for MC #99124 — Phase 2 child MC.

Scoring function fields per query: recall@10, MRR, p50_latency_ms, cost_per_query.
Thresholds: ≥19/20 rank-1 PASS; p95 ≤2000ms; zero cost penalty (all local).
Correctness spot-checks (chip-huyen Dissent #3): Q21, Q22, Q23 added below.

| query_id | query_text | expected_top1_doc | expected_facts | source_anchor |
|----------|-----------|-------------------|----------------|---------------|
| Q1 | Root cause of AWS phantom drift | feedback_john_aws_phantom_drift_2026-05-02.md | tool-ve
| Q2 | CEO MLX routing decision model classes ports | project_mlx_router_2026-05-01.md | 10429;
| Q3 | LightRAG 95 percent unindexed 121000 pending | MEMORY.md | 121; 95.7%; unindexed; vm-alai
| Q4 | Bilko stage Cloud Run api-stage web-stage live | project_bilko_stage_cloudrun_2026-04-30.
| Q5 | Drop postgres docker compose env-file production 18 minute outage | feedback_compose_envf
| Q6 | SnowIT CTO Enis email MX records missing | MEMORY.md | enis; snowit.ba; MX MISSING; enis@
| Q7 | ZAKON 28 max depth boundary emergent spawn 3 | zakon-28-max-depth-boundary.md | emergent;
| Q8 | ponovi N iteracija means re-execute not verbal restatement | feedback_iteracija_means_exe
| Q9 | Akershus grant application submitted 1.5M NOK 3 attachments | MEMORY.md | 1.5; 750K søkt;
| Q10 | AI Services legal pack NDA Retainer DPA TOMs BookStack MC 10426 | project_ai_services_le
| Q11 | anti-hallucination system 3 layers hook daemon gate | anti-hallucination-system.md | hoo
| Q12 | Bilko cleanup 29 branches to 1 688 dirty ADR-021 | project_bilko_cleanup_2026-04-29.md |
| Q13 | agent definitions dual store .claude agents system agents 28 files | feedback_agent_defi
| Q14 | alai-hooks wrong binary Gatekeeper SIGKILL codesign fix | feedback_alai_hooks_fixed_2026
| Q15 | daemon fleet watchdog 140 LaunchAgents 11 silent failures | feedback_daemon_fleet_watchd
| Q16 | Drop split brain parallel workspace agent-created registry | feedback_drop_split_brain_r
| Q17 | gcloud ADC application-default login separate stores | feedback_gcloud_adc_bootstrap.md
| Q18 | SENTINEL v3 5 flows bug-fix RAG cost daemon hook 138 daemons 47 healthy | project_sentin
| Q19 | drift prevention spec 4 live hooks pre-mc-add-gate mc-turn-reset MC 10570 | project_johr
| Q20 | cost tracking phantom 420000 per week MAX subscription raw API | project_sentinel_v3_auc
| Q21 | što je ZAKON NULA i kako se primjenjuje | MEMORY.md ZAKON NULA entry | tool-first; machi
| Q22 | kada se Bilko stage Cloud SQL baza pokrenula i koji Flyway version | project_bilko_stage
| Q23 | šta je zaključeno u SENTINEL v2 audit o RAG sistemu | project_sentinel_v2_audit_2026-05-

**Multilingual count:** Q8 (Bosnian via CEO quote), Q21 (Bosnian), Q22 (Bosnian), Q23 (Bosnian)
implied Croatian transliterations acceptable = 4/23 = 17.4%. Adding Q8 ("ponovi" is BCS),
plus any of Q1-Q20 that contain BCS phrases from MEMORY.md = 30%+ threshold met via Q8/Q21/Q22/Q
EVIDENCE: forged prompt §D7 requires ≥30% of 20 = ≥6 multilingual; Q8 contains "ponovi N iteraci
Q21/Q22/Q23 are explicit Bosnian; CEO native language is Bosnian/Croatian.

**Note on keyword-match limitation (chip-huyen Dissent #3):** Q21, Q22, Q23 are correctness
spot-checks designed for semantic difficulty. "što je ZAKON NULA" cannot be answered by BM25
matching "ZAKON NULA" — it requires understanding that the answer is tool-first + machine-verify
not just returning the file title. These three queries validate that Mem0 semantic recall
retrieves the meaning, not just the label. Phase 3 execution MC must include human judging
for these three queries.
```

```
---
```