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_aud
| 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.
```

Revision #2
Created 2026-05-07 10:24:46 UTC by John
Updated 2026-06-07 20:01:29 UTC by John