Memory Retrieval Eval Harness

Memory Retrieval Eval Harness — MC #105224

Purpose

Regression-test `discover.js memory "<query>"` retrieval quality with a labeled set of real incident queries mapped to known memo files.

Artifacts

- Harness: `~/system/tools/memory-eval-harness.js`
- Test set: `~/system/specs/memory-retrieval-eval/test-set-v1.json`
- Query-log audit: `~/system/specs/memory-retrieval-eval/query-log-miss-audit-2026-07-10.md`
- Historical evidence: `~/system/evidence/memory-eval-105224-verify-20260710/`
- P0-A evidence: `~/system/evidence/memory-eval-p0a-20260721/`

How to run

```
node ~/system/tools/memory-eval-harness.js run \
  --test-set ~/system/specs/memory-retrieval-eval/test-set-v1.json \
  --output ~/system/evidence/memory-eval-$(date +%Y%m%d-%H%M%S)
```

Dry-run schema validation:

```
node ~/system/tools/memory-eval-harness.js run \
  --test-set ~/system/specs/memory-retrieval-eval/test-set-v1.json \
  --dry-run
```

Durable P0-A regression test (default/top-N prefixes, argument validation, fail-closed privacy, and metric recomputation):

```
node ~/system/tools/tests/memory-retrieval-p0a.test.js
```

Metrics

- Recall@3/5/10/20: correct memo appears at rank <= k in the 20 candidates requested by the eval harness.
- MRR@20: reciprocal rank of the correct memo; 0 when not returned in the top 20.

Normal `discover.js memory "<topic>"` calls remain capped at top 3 for backward compatibility. The eval harness explicitly uses `--top 20`, which changes only truncation depth, not scoring or ordering.

Baseline run — 2026-07-10 (historical, before P0-A)

Fresh verification output: `~/system/evidence/memory-eval-105224-verify-20260710/`

- Labeled pairs: 40
- Recall@3: 0.600 (24/40)
- MRR: 0.546
- Errors: 0/40
- p50 latency: 174ms
- p95 latency: 243ms

P0-A verification — 2026-07-21

Evidence: `~/system/evidence/memory-eval-p0a-20260721/`

- Labeled pairs: 40
- Recall@3: 0.750 (30/40)
- Recall@5: 0.825 (33/40)
- Recall@10: 0.875 (35/40)
- Recall@20: 0.925 (37/40)
- MRR@20: 0.695
- Errors: 0/40
- p50 latency: 125ms
- p95 latency: 138ms

The apparent Recall@3 change versus the historical baseline reflects corrected handling of documented fail-closed exit code 1 and evaluation over one top-20 candidate list. The `--top` path changes only truncation, not scoring or ordering.

Privacy-safe retrieval telemetry

Every `discover.js memory` invocation appends aggregate metadata to `~/system/logs/discover-queries.jsonl` (mode 0600). It records timestamp, query length/token count, requested top-N, result/confidence counts, latency, fail-closed state, and coarse error kind. It never records query text or hashes, memo filenames/content, secrets, or PII.

Revision #3
Created 2026-07-10 20:19:02 UTC by John
Updated 2026-07-21 17:58:56 UTC by John