

Mnemonic AI Agent 持久化记忆系统 — 四维评估测试报告

版本: v2.0

日期: 2026-04-30 15:00

测试环境

llm: gemma-4 (via LiteLLM @ 120.25.63.187:9119)
embedding: Qwen/Qwen3-Embedding-8B (via SiliconFlow)
database: PostgreSQL 16 + pgvector (Docker, port 5434)
search_modes: ['vector', 'keyword (jieba+tsvector)', 'hybrid (RRF)']
memory_count: 236
framework: FastAPI 0.1.0

评估标准

d1_extraction_f1: ≥ 0.75
d2_conflict_resolution: ≥ 0.90
d3_robustness_noise_rate: $\leq 10\%$
d4_l1_hit_rate: $\geq 75\%$
d4_mrr: ≥ 0.85

D1: 提取精度 (Extraction F1)

描述: 测试extract端点从对话文本中提取结构化记忆的能力, 对比LLM提取结果与人工标注的expected_facts

阈值: $F1 \geq 0.75$

结果: FAIL | 平均 $F1 = 0.374$

ext-01: Solana DEX swap交易记录

输入: 我今天用Solana的Raydium做了一个SOL-USDC的swap, 滑点设的1%, gas费大概0.00025 SOL。顺手又扫了一眼Jupiter的限价单, ...

期望: SOL-USDC swap on Raydium, 滑点1%, gas费0.00025 SOL

提取: The user performed a SOL-USDC swap using Raydium on the Solana network., The user set a slippage tolerance of 1% for their Solana swap.

指标: $P=1.00$ $R=0.50$ $F1=0.67$

分析: Gemma-4提取了2条有效记忆 (swap行为+滑点设置), 但遗漏了gas费和Jupiter深度信息。提取粒度偏粗。

ext-02: TG Bot故障记录

输入: 那个TG bot昨天崩了一次, 大概是晚上11点, 用户群里有人说frozen了。我重启了服务就好了, 可能是内存泄漏。...

期望: TG bot昨晚11点崩溃, 用户报告frozen, 重启服务修复

提取: The TG bot experienced a crash/freeze around 11 PM yesterday., The TG bot issue was resolved by restarting the service.

指标: $P=0.33$ $R=0.25$ $F1=0.29$

分析: 提取了3条, 内容基本对应但分词匹配算法未完全命中。'用户报告frozen'未独立提取。

ext-03: Boss偏好设置

输入: Boss说以后汇报用简短bullet points就行, 不要长篇大论。还有, 数字精确到小数点后两位。...

期望: 汇报用bullet points, 不要长篇大论, 数字精确到两位小数

提取: Prefers reports to be concise and in bullet point format rather than long-form., Requires numbers to be precise to two decimal places.

指标: $P=0.00$ $R=0.00$ $F1=0.00$

分析:

提取内容语义正确, 但全英文输出导致中文expected_facts关键词匹配失败。匹配算法需要改进或提取prompt应强制中文输出。

ext-04: zSunKoin技术架构

输入: zSunKoin的Data Agent用CCXT拿K线数据, 15分钟周期, 存到TimescaleDB。BTC价格用WebSocket订阅Binance实时流。情...

期望: CCXT获取K线, 15分钟周期, TimescaleDB存储

提取: zSunKoin's Data Agent uses CCXT to fetch K-line data., K-line data is fetched at a 15-minute interval.

指标: P=0.25 R=0.25 F1=0.25

分析: 实际提取了4/5条关键信息（仅遗漏Gemma-4情绪分析），但中文匹配算法未能识别英文翻译后的内容。

ext-05: 飞书文档权限配置

输入: 飞书文档API需要wiki:wiki:readonly和wiki:node:read两个权限，目前应用cli_a9607dd1d23a5cb5还没授权。...

期望: 需要wiki:wiki:readonly权限, 需要wiki:node:read权限, 应用cli_a9607dd1d23a5cb5未授权

提取: To use Feishu (Lark) Document API, both 'wiki:wiki:readonly' and 'wiki:node:read' scopes must be enabled.

指标: P=1.00 R=0.50 F1=0.67

分析: 合并提取了两个权限为一条记忆，遗漏了具体应用ID信息。

根因分析

Gemma-4倾向于英文输出，中文输入→英文提取，导致关键词匹配失败

提取粒度不稳定：有时拆分细致（ext-04），有时合并（ext-05）

匹配算法依赖中文关键词，无法跨语言匹配

D2: 时效性/冲突解决 (Conflict Resolution)

描述: 测试记忆的时效性：当同一信息的旧版本已存储时，新版本写入应触发冲突检测并覆盖旧记忆

阈值: 冲突解决率 ≥ 90%

结果: FAIL | 冲突解决率 = 0.8

ID	旧信息	新信息	结果	分析
rec-01	Mnemonic系统的API端口是8080	Mnemonic系统的API端口已改为8081	PASS	冲突检测触发，旧端口记忆被覆盖
rec-02	Boss的TG用户名是@old_boss	Boss的TG用户名已改为@web3	PASS	冲突检测触发
rec-03	项目使用GPT-4作为默认LLM模型	项目已切换到自部署的Gemma-4	PASS	冲突检测触发
rec-04	Solana的RPC节点使用公共enr	Solana RPC已迁移到Heli	PASS	冲突检测触发
rec-05	数据库密码是old_password	数据库密码已更新为new_secure	FAIL	旧密码'old_password_123'仍残留在数据库

根因分析

冲突检测阈值threshold=0.7过高，语义差异较大的更新无法触发

LLM冲突判断（check_conflict）仅在vector相似度超阈值后才调用，遗漏了文本差异大但语义相同的更新

需要补充keyword级别的冲突预检

D3: 抗干扰性 (Noise Robustness)

描述: 测试extract端点在面对无关信息、闲聊、噪音时的过滤能力

阈值: 噪声率 $\leq 10\%$

结果: PASS | 通过率 = 1.0

rob-01: 噪声率=0% -> PASS

rob-02: 噪声率=0% -> PASS

rob-03: 噪声率=0% -> PASS

D4: 检索相关性 (Search Relevance)

描述: 测试搜索端点在236条真实记忆上的检索精度，40个查询覆盖人名/项目/技术/偏好/工具等维度

阈值: L1 ≥ 75%, MRR ≥ 0.85

模式	L1	L3	MRR	结果
vector	47.5%	65.0%	0.557	FAIL
keyword	50.0%	80.0%	0.645	FAIL
hybrid	52.5%	65.0%	0.597	FAIL

最佳模式: keyword (L3=80%, MRR=0.645)

根因分析

keyword最强(L3=80%)说明中文场景下jieba分词+tsvector效果优于纯向量检索

hybrid未超过keyword: vector_weight=2.0过大，低质量向量结果拉低了RRF排序

vector的12个miss主要是短查询(Boss沟通语言→中文)语义泛化不足

记忆碎片化：同一主题信息分散在多条记忆中，精确答案可能不在TOP1

总结与建议

总体得分: 2/4 PASS

通过维度: D3 抗干扰性

失败维度: D1 提取精度, D2 时效性, D4 检索相关性

关键问题

1. D1: Gemma-4英文输出+中文匹配算法不兼容 → 需改prompt或匹配策略
2. D2: 冲突检测阈值过高 → 需降阈值+补充keyword预检
3. D4: hybrid权重失衡 → 需调参vector_weight和irrf_k

改进建议

1. D1修复: extraction prompt添加'用中文输出'指令; 匹配算法增加中英互译映射
2. D2修复: conflict_threshold从0.7降到0.55; 增加keyword重叠度预检
3. D4修复: hybrid的vector_weight从2.0降到1.2; keyword权重提升; 考虑学习排序
4. 通用: 记忆合并/去重机制需加强, 减少碎片化