BrainStorm

Your Local AI Agent

Sam - the system oracle for Black Beauty

Version 0.2.2

August 2026

No cloud. No subscription. Just your machine, your data, your agent.

BrownMediaGroup LLC

Family-owned. Michigan-based.

810-223-1623

Executive Summary

Sam is a local-first AI agent embedded in BrainStorm, a Tauri v2 desktop app. He lives on your machine, talks to your filesystem, remembers what you teach him, and never phones home.

Benchmark	**94.4% avg**
Action Test	**47.5% avg**
Memory	**10 + 8 facts**

The Growth Story

Qwen 3.6 baseline: 78.6% -> post-training 100% on Phase 1-6.

Qwen 3.8: 94.4% average on full 10-phase suite (113 questions).

Action test grew from 10% to 47.5% across 4 major versions.

6 major versions. 4 days. 4x improvement in action capability.

The Technology Stack

Frontend

Tauri v2 + vanilla JavaScript - lightweight, fast, no Electron bloat.

Backend

Rust - agent loop, tool routing, memory management, path grounding.

AI Model

Qwen 3.8 27B (medium quant, GGUF) - served locally via Ollama. 27B parameters, ~16GB for consumer GPU.

Previous model: Qwen 3.6 27B - used through v0.1.3, achieved 100% on Phase 1-6 baseline after training.

Inference

Ollama local (:11434) with GPU offload (CUDA). 20+ tokens/sec on RTX 5070 Ti.

Memory Architecture

3-tier system: permanent vault (10 facts), sliding window (8 facts), full disk vault.

Tools

fs_mkdir, fs_write, fs_copy, fs_move, fs_rename, fs_delete, shell, list_files, read_file, browse_memory, retrieve.

Benchmark Results

Test Design

10-phase suite, 113 questions across baseline knowledge, banter/personality, mixed tool use, long-form complex, and recall/memory.

Model Comparison: Qwen 3.6 vs 3.8

Qwen 3.6 (v0.1.3): Baseline 78.6%, post-training 100%, Phase 1-7 96.4%.

Qwen 3.8 (v0.1.12 through v0.2.2): Average 94.4% on full 10-phase suite (113 questions).

Note: Qwen 3.8 tests are broader, so scores are not directly comparable.

Qwen 3.6 Baseline Journey

The Starting Point

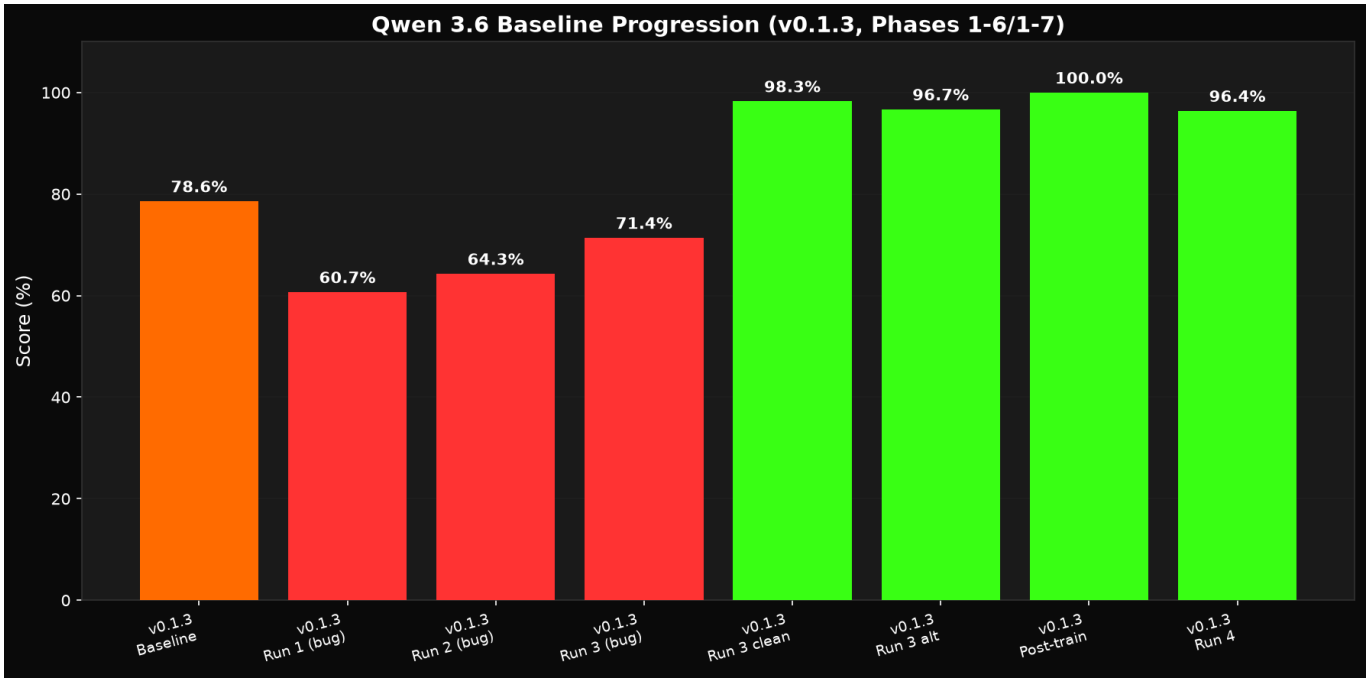
Before v0.1.12, Sam ran on Qwen 3.6 27B. The baseline was rough: 22/28 = 78.6%.

Training Progression

Run 1-3 had a harness bug. Run 3 clean: 59/60 = 98.3%. Post-training: 28/28 = 100%. Phase 1-7: 53/55 = 96.4%.

What Changed

- Harness fix: reading correct output field
- Memory architecture: 10+8 fact cap
- Tool-routing classifier fixed
- Shell guard fixed
- Training method: /api/learn with real-world phrasings



Benchmark Results - v0.2.2

v0.2.2 Results (3 Runs)

Run 1: 111/113 = 98.2%

Run 2: 105/113 = 92.9%

Run 3: 104/113 = 92.0%

Average: 107/113 = 94.4%

Phase Breakdown (Average)

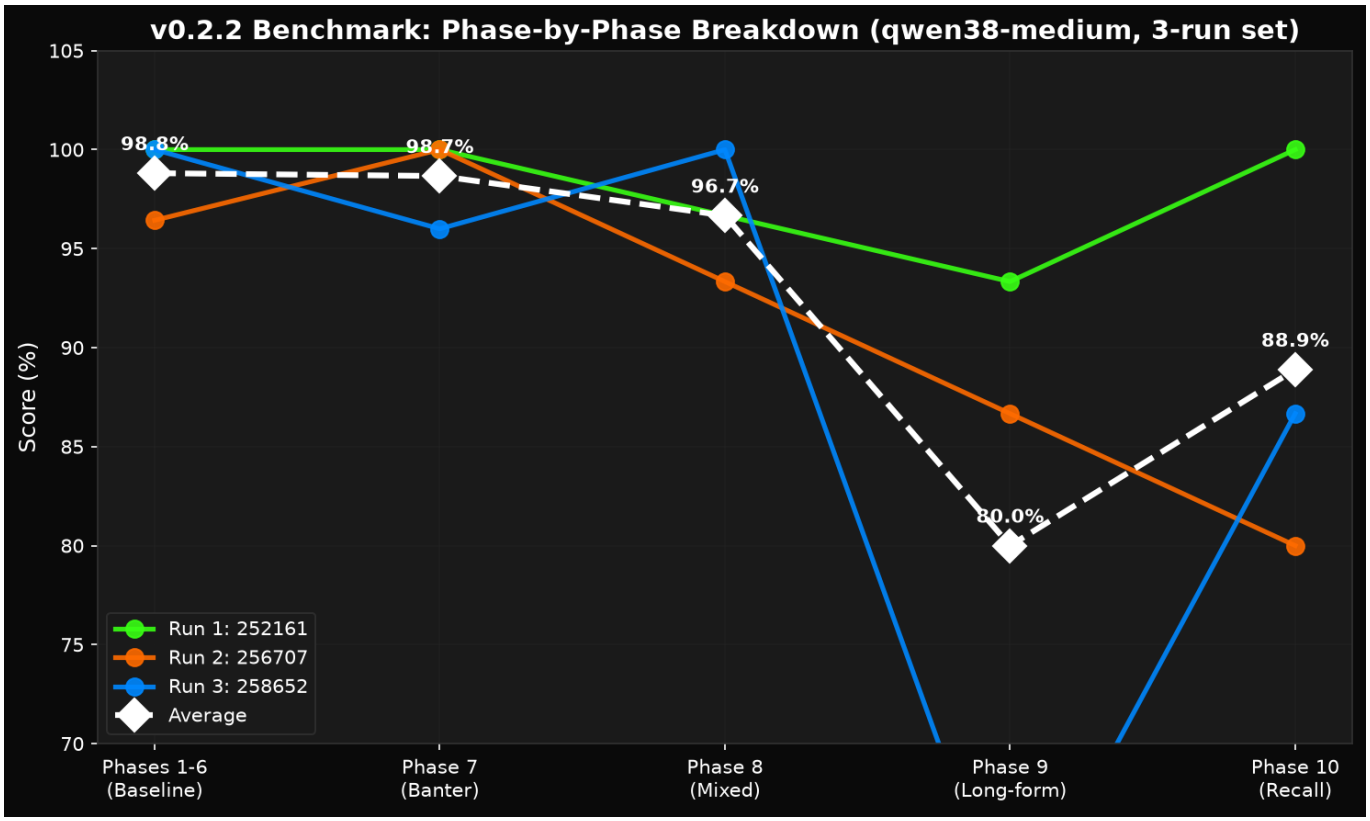
Phases 1-6 (Baseline): 27.7/28 = 98.8%

Phase 7 (Banter): 24.7/25 = 98.7%

Phase 8 (Mixed): 29/30 = 96.7%

Phase 9 (Long-form): 12/15 = 80.0%

Phase 10 (Recall): 13.3/15 = 88.9%



Action Test Results

Test Design

10-step filesystem sandbox on E: drive. Each step chains on the previous. Verified by checking actual filesystem state.

v0.2.2 Results (4 Runs)

Run 1: 4/10 = 40.0%

Run 2: 6/10 = 60.0%

Run 3: 6/10 = 60.0%

Run 4: 3/10 = 30.0%

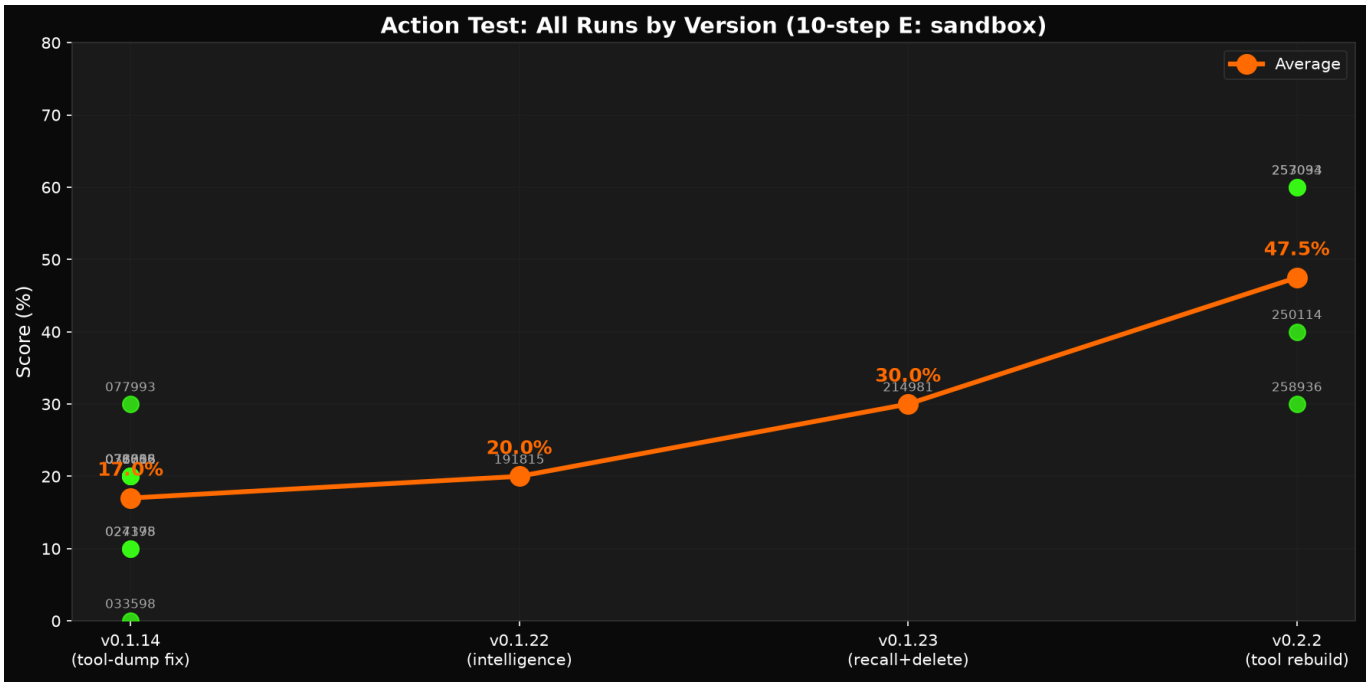
Average: 4.75/10 = 47.5%

What Works

- File creation with content (fs_write)
- Copying files and folders (fs_copy)
- Moving folders (fs_move)
- Creating archives, listing directories, deleting files/folders

What is Hard

- Multi-step chaining (3+ actions in sequence)
- Exact instruction following
- Path grounding for deeply nested structures
- Inference consistency on long action chains

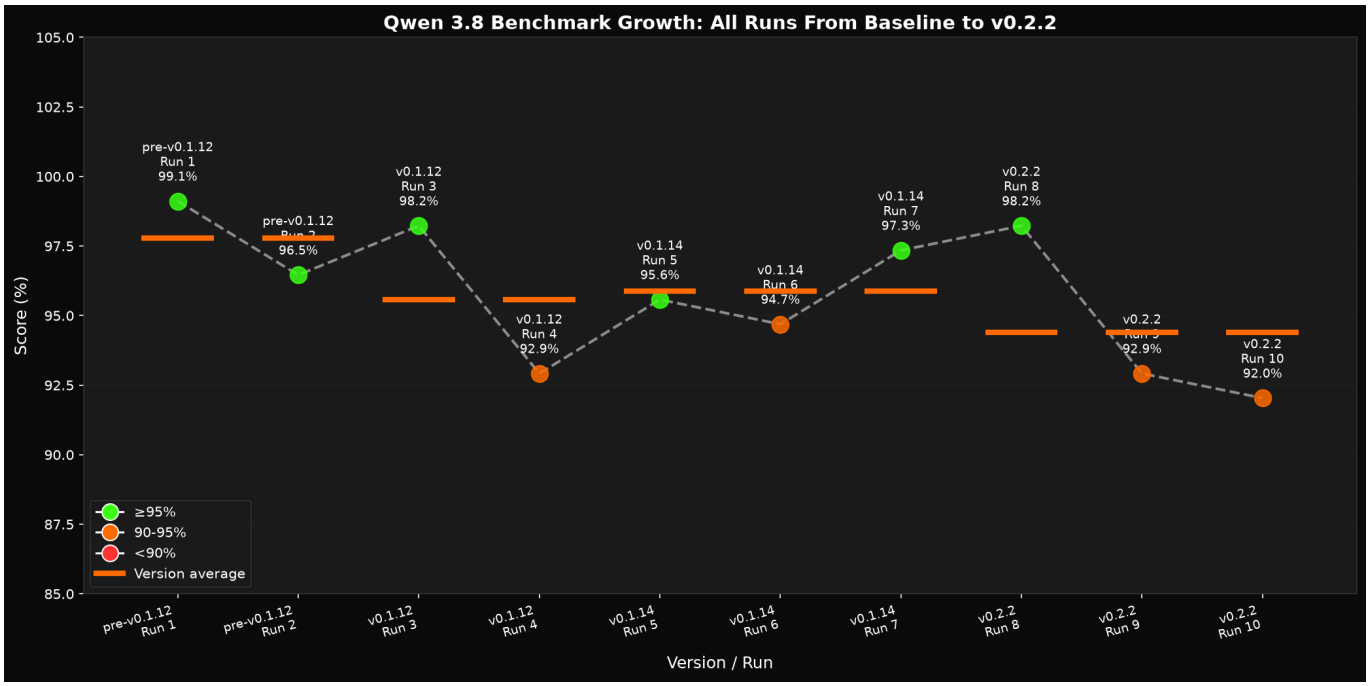

Action Test: All Runs by Version (10-step E: sandbox)
Average
Score (%)
80
70
60
50
40
30
20
10
0
077993
033598
191815
214981
250114
258936
17.0%
20.0%
30.0%
47.5%
v0.1.14
(tool-dump fix)
v0.1.22
(intelligence)
v0.1.23
(recall+delete)
v0.2.2
(tool rebuild)

Qwen 3.8 Benchmark Growth

All Runs From Baseline to v0.2.2

Same model. Same hardware. Pure software improvement.

Benchmark scores stayed in the 92-98% band throughout. The software improvements went into execution capability and reliability, not raw reasoning.



Architecture Deep Dive

Agent Loop (Rust)

1. Receive task via /api/run POST
2. Load memory - permanent vault (10) + sliding window (8)
3. Seed history - last 20 turns from persistent buffer
4. Generate tool calls via LLM inference
5. Ground paths - force-rebase to sandbox root
6. Execute tools via Python bootstrap
7. Format observations - human-readable summaries
8. Push turns to persistent buffer
9. Return final answer

Tool Layer (Layers A/B/C)

Layer A - Schema Registry: Single source of truth. Every tool declares semantic slots.

Layer B - Semantic Arg Resolution: Normalizes any key name the model emits to tool canon.

Layer C - Schema Validation: Checks required slots against schema. Missing -> honest error.

Memory System

3-Tier Memory Architecture

Tier 1 - Permanent Vault (10 facts): Never evicted, loaded every session, core identity.

Tier 2 - Sliding Window (8 non-permanent facts): Recent learnings, capped at 8 per session.

Tier 3 - Full Disk Vault: learnings.jsonl (all /api/learn writes), soul.md (core identity), chroma (semantic search).

Prompt Efficiency

Only 18 facts loaded per prompt (10 perm + 8 non-perm). No prompt bloat. Full vault accessible on demand via tools.

Teaching Sam

Use /api/learn to teach new facts. Writes immediately to learnings.jsonl + soul.md. Available in next session.

Call to Action

Try BrainStorm

Download: github.com/RiseUpGit/BrainStorm-commercial/releases

Source: github.com/RiseUpGit/BrainStorm-commercial

Landing Page: ppcrepair.com/brainstorm

For Business

BrownMediaGroup LLC - Family-owned. Michigan-based. 810-223-1623

Joe Brown - Founder / Owner

Alex Brown - Repairs, teardown, on-camera work

Lily Brown - Bookkeeping, scheduling, CapCut/video production

Technical Specifications

Framework: Tauri v2 + Rust

AI Model: Qwen 3.8 27B (medium quant, GGUF)

Inference: Ollama local with CUDA offload

Memory: 3-tier architecture

Platform: Windows 10/11

GPU: NVIDIA single-GPU (RTX 5070 Ti tested)

License: Commercial (proprietary)

Your machine. Your agent. Your rules.

No cloud. No subscription. No data exfil.

Built with pride in Michigan.