Self-Improvement Agent Harness: A Deterministic SIA Exemplar

Meta → Target → Feedback loops with public/private task splits

Daniel Ari Friedman
Active Inference Institute
daniel@activeinference.institute
ORCID: 0000-0001-6232-9096
DOI: 10.5281/zenodo.19139090

May 29, 2026

Contents

1 Abstract

This exemplar documents **template_sia**, a deterministic implementation of the Self-Improvement Agent (SIA) harness contract described in [AI, 2026]. The default pipeline replays fixture-backed generations for the `mini_classify` task; opt-in live mode runs bounded target subprocesses and optional Ollama-backed meta/feedback steps.

Run snapshot. Task `mini_classify`, run 1, 3 generation(s), live=false. Final accuracy=0.8333 over 6 held-out samples. Values are injected by `scripts/z_generate_manuscript_variables.py` after analysis.

Keywords: self-improvement agents, benchmark harness, reproducible evaluation, agent loops

2 Methodology

2.1 SIA loop

The harness implements a three-agent cycle:

1. **Meta** — proposes or seeds a target agent for generation n.
2. **Target** — runs against public task data and writes `agent_execution.json`.
3. **Feedback** — reads private evaluation metrics and proposes improvements for generation $n + 1$.

Artifacts land under `output/runs/run_{id}/gen_{n}/` with `target_agent.py`, `agent_execution.json`, optional `improvement.md`, and canonical `results.json`.

2.2 Task layout

Each task exposes:

- `data/public/` — agent-visible instructions and data (`task.md`, `train.csv`, `evaluate.py`).
- `data/private/` — held-out labels for evaluation only.
- `reference/reference_target_agent.py` — deterministic baseline.

The exemplar task `mini_classify` is a threshold classifier on a single feature column.

2.3 Determinism contract

When `sia.live: false` (default), generations replay recorded fixtures from `src/fixtures/recorded_generations/`. CI never executes generated agent code or calls external LLM APIs.

Pass `--live-sia` to `scripts/run_sia_loop.py` for bounded subprocess execution and optional Ollama feedback when a model is configured.

3 Results

Table tbl. 1 summarizes fixture-replay metrics for the bundled run.

Table 1: SIA generation metrics (fixture replay).

Gen	Metric	Value	N
1	accuracy	0.5000	6
2	accuracy	0.6667	6
3	accuracy	0.8333	6

Final injected token: accuracy=0.8333 (n=6).

4 Conclusion

template_sia demonstrates how to embed the SIA harness contract in the Research Project Template without vendoring upstream orchestration code. Layer 1 (`infrastructure/sia/`) owns task validation, evaluation, context logging, and the generation state machine; Layer 2 wires a minimal classification task, fixture replay, and manuscript tokens.

Live self-improvement remains opt-in; the exemplar does not claim benchmark scores from [AI, 2026] without explicit live runs.

References

Hexo AI. Self-improvement agents. 2026. URL https://arxiv.org/abs/2605.27276.