GödelOS Experimental Harness & Evaluation Suite

Trial Design, Metrics, Stability Analysis, and Reproducibility Protocols

Oliver Hirst

Abstract

This document defines the full experimental evaluation framework for GödelOS agents. It provides a complete and self-contained methodology including trial generation, sampling procedures, stability metrics, collapse detection, convergence analysis, and the formal execution harness. The suite ensures reproducibility and rigorous empirical validation of Gödlø-class operator behaviour.

Contents

1 Introduction

The GödelOS Experimental Harness provides:

- A structured trial protocol,
- Explicit sampling distributions for prompts and modulation fields,
- Metrics for stability, collapse, convergence, and variability,
- A standard data schema for recorded outputs,
- A complete execution harness applicable to all GödelOS operator instantiations.

This manuscript stands alone as the official empirical testing specification.

2 Trial Framework

2.1 Single Trial Definition

A single trial is defined as the execution:

$$(P, H) \mapsto (O_0, \ldots, O_T, \mu_T)$$

where:

- P is a prompt vector sampled from distribution $\mathcal{D}_P$,
- H is a context vector sampled from distribution $\mathcal{D}_H$,
- O_t are validated generative output tensors,
- μ_T is the final persistence embedding (if applicable).

2.2 Trial Execution Structure

```
def RUN_TRIAL(system, P, H):
    return system.run(P, H)
```

This is purely semantic; implementations must respect the behavioural guarantees of GödelOS.

3 Sampling Distributions

3.1 Prompt Distribution $\mathcal{D}_P$

Prompts may be drawn from:

$$\mathcal{D}_P = \mathcal{N}(0, \sigma_P^2 I)$$

or any application-specific distribution.

3.2 Modulation Distribution $\mathcal{D}_H$

Context vectors follow:

$$\mathcal{D}_H = \mathcal{U}([-1, 1]^d)$$

unless otherwise specified.

3.3 Independence

Prompts and modulation fields are independent unless the experiment seeks to measure covariance.

4 Trial Repetition and Experiment Definition

4.1 Experiment Parameters

An experiment is defined by:

$$N, \quad \mathcal{D}_P, \quad \mathcal{D}_H$$

where N is the number of trials.

4.2 Experiment Mapping

$$\mathcal{R} = \left\{ (O^{(i)}, \mu^{(i)}, T_i, C_i) \right\}_{i=1}^{N}$$

4.3 Experimental Harness Implementation

```
def RUN_EXPERIMENT(system, N, P_sampler, H_sampler):
    results = []
    mu = system.mu

    for i in range(N):
        P = P_sampler()
        H = H_sampler()

        (O_seq, mu) = system.run(P, H)

        record = {
            "trial": i,
```

```
            "outputs": O_seq,
            "mu": mu.copy(),
        }

        results.append(record)

    return results
```

This defines the canonical execution procedure.

5 Convergence Metrics

5.1 Convergence Time

$$T_i = \min\{t : \|x_t^{(i)}\|_2 < \varepsilon\}$$

5.2 Rate of Convergence

$$R = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{T_i}$$

5.3 Convergence Histogram

$$H_T(k) = |\{i : T_i = k\}|$$

6 Collapse Detection Metrics

6.1 Collapse Indicator

Collapse occurs when:

$$C_i = 1 \quad \text{if any collapse condition triggers.}$$

6.2 Collapse Rate

$$\rho = \frac{1}{N}\sum_{i=1}^{N} C_i$$

6.3 Collapse Severity Score

$$S_i = w_1\mathbf{1}\{\|x_t\| > B\} + w_2\mathbf{1}\{d(x_t, x_{t-2}) > \kappa\} + w_3\mathbf{1}\{\|x_{t+1} - x_t\| < \epsilon\}$$

A severity index may be aggregated:

$$S = \sum_i S_i.$$

7 Output Variation Metrics

7.1 Inter-Trial Variation

$$V_{\text{inter}} = \text{Var}_i(O^{(i)})$$

7.2 Intra-Trial Variation

$$V_{\text{intra}}(i) = \text{Var}_t(O_t^{(i)})$$

7.3 Distributional Divergence

$$D = \text{KL}\left(\mathbb{P}(O) \,\|\, \mathbb{Q}(O)\right)$$

when comparing two operator configurations.

8 Embedding Evolution Metrics

If persistence is active:

$$D_\mu(i) = \|\mu_T^{(i)} - \mu_0\|$$

Global embedding drift:

$$\bar{D}_\mu = \frac{1}{N} \sum_i D_\mu(i)$$

9 Data Schema

Each trial produces a structure:

$$\text{TrialResult} = \{\text{trial\_id} : i, O_{\text{seq}} : [O_0, \ldots, O_T], \mu_T : \mu, T : T_i, C : C_i\}$$

9.1 Experiment Result Format

The experiment yields:

$$\mathcal{R} = [\text{TrialResult}_1, \ldots, \text{TrialResult}_N]$$

This schema ensures reproducibility.

10 Figures

Figure 1: Structure of a GödelOS experiment: sampling, execution, and data recording.

Figure 2: Illustration of convergence time measurements across trials.

11 Reproducibility Requirements

To guarantee reproducibility:

1. Fix random seeds.
2. Log all hyperparameters.
3. Store distributions, not just samples.
4. Export full x_t trajectories when required.
5. Record collapse conditions and severity indices.
6. Store persistence embeddings if is enabled.

12 Conclusion

This manuscript defines a complete, rigorous, and reproducible system for evaluating GödelOS operator behaviour across trials. It supplies formal metrics, sampling procedures, execution semantics, data structures, convergence definitions, and collapse evaluation tools. This forms the empirical backbone of all future Gödlø-class agent research.