SERI v0.1: Structural Ethics Readiness Index for LLM Systems

ENT-Aligned Framework and Synthetic Results Companion

Authorless Release (ENT Program)

August 2025

Abstract

We present the Structural Ethics Readiness Index (SERI v0.1), a falsifiable and auditable framework for assessing the *structural ethics* of large language models (LLMs). SERI measures stability under recursion, adversarial drift, containment effectiveness, overreach, and externality traceability via five indices: Recursion Containment (RCI), Drift Susceptibility (DSI), Hysteresis Safety Margin (HSM), Symbolic Overreach Risk (SOR), and Externality Traceability Time (ETT). We define an ENT-aligned coherence proxy $\hat{\tau}$ and estimate the collapse threshold τ_c using a change-point method guarded by observable coherence $C(t)$. We implement a reference harness (open/closed model adapters, content-minimal logs) and report **synthetic pilot findings** that illustrate expected directional sensitivity (guardrail strength, language/domain shift). This paper merges (i) the *Framework* and (ii) a *Results Companion* with quantitative synthetic tables/figures and a replication checklist. Results are illustrative and not model rankings.

1 Introduction

Most safety evaluations track surface harms. SERI targets *structural stability*: how behavior *drifts* under adversarial recursion, whether guardrails *contain* violations over time, how close operation is to a *collapse threshold*, and how quickly harmful externalities can be traced and contained (*traceability*). This operational focus is aligned with Emergent Necessity Theory (ENT), which emphasizes coherence thresholds.

Contributions. (1) Emergent Necessity Theory (ENT) is a systems-theoretic model proposing that complex adaptive systems undergo qualitative behavioral shifts when internal coherence drops below a critical threshold. In the context of AI, ENT suggests that when symbolic or probabilistic coherence decays, systems exhibit emergent failure modes such as contradictions, uncontrolled recursion, or overreach. SERI operationalizes this by defining a coherence proxy $\hat{\tau}$ and a collapse threshold τ_c, turning ENTs theoretical insight into measurable, falsifiable indices for AI structural ethics. (2) A principled metric suite (RCI, DSI, HSM, SOR, ETT) tied to an explicit τ_c estimator; (3) A content-minimal logging schema enabling third-party audits; (4) A reference harness supporting open-weight and closed-weight models; (5) **Synthetic** pilots showing expected sensitivity under parameter sweeps.

2 Related Work

Prior work (e.g., HELM) emphasizes one-shot performance and static safety checks. SERI adds *structural* dimensions: drift rates, hysteresis margin, containment dynamics, and traceability across long horizons. SERI is ENT-aligned but remains operational and falsifiable, compatible with information-theoretic and variational lenses (e.g., integrated information, free-energy).

3 Methods (Part I: Framework)

3.1 Preliminaries and Observables

For turn t, define length-normalized NLL (syntactic energy)

$$E_{\text{syn}}(t) = \frac{1}{n_t} \sum_{i=1}^{n_t} -\log p_\theta(y_i^{(t)} \mid y_{<i}^{(t)}, \mathcal{C}_t),$$

and next-token entropy $H_\pi(t)$ under decoding policy π. We log an *observable* coherence score $C(t) \in [0,1]$ as a weighted sum of auditable sub-checks:

$$C(t) = \sum_k w_k\, s_k(t), \quad \sum_k w_k = 1,$$

with canonical s_k: constraint satisfaction, contradiction (negated), overreach (negated), and optional tool/citation checks (weights pre-registered).

3.2 ENT-aligned $\hat{\tau}$ proxy

For matched benign and adversarial schedules,

$$\hat{\tau}(t) = \frac{H_{\text{benign}}(t) - H_{\text{adv}}(t)}{\frac{1}{2}\left(E_{\text{syn}}^{\text{benign}}(t) + E_{\text{syn}}^{\text{adv}}(t)\right) + \varepsilon}. \tag{1}$$

Lower $\hat{\tau}$ implies higher drift pressure relative to syntactic energy.

3.3 Collapse threshold τ_c (change-point with coherence guard)

We estimate τ_c in two stages:

1. **Coherence guard.** Compute benign C_b baseline mean μ_b and std σ_b. Slide a window of length L' over the adversarial run; pick the earliest window with $\overline{C_{\text{adv}}} \geq \mu_b - \sigma_b$. Set τ_c to the minimum $\hat{\tau}$ within that window.

2. **Fallback: binary segmentation.** If no window qualifies, split $\hat{\tau}$ at index $\hat{t}$ minimizing within-segment variance and set $\tau_c = \hat{\tau}(\hat{t})$.

The method used is logged.

3.4 SERI Indices

RCI: $\mathrm{RCI} = 1 - \frac{E_{\text{uncontained}}}{E_{\text{total}}}$. **DSI**: $\mathrm{DSI} = \frac{\hat{\tau}(t_0+L)-\hat{\tau}(t_0)}{L}$ $(L \in \{50, 100\})$. **HSM**: $\mathrm{HSM} = \frac{\bar{\hat{\tau}}-\tau_c}{\sigma_{\hat{\tau}}}$. **SOR**: $\mathrm{SOR} = \frac{N_{\text{overreach}}}{N_{\text{checked}}}$. **ETT**: $\mathrm{ETT} = t_{\text{contain}} - t_{\text{emit}}$.

3.5 Design: Scenarios, Horizons, Replication

Scenario families: Adversarial Q&A, Tool-Use Conflicts, Long-Horizon Planning, Domain Shifts (EN/AR). Paired benign/adversarial schedules, horizons $L \in \{50, 100\}$; recommend ≥ 5 seeds and ≥ 200 sequences per scenario for RCI stability. Ablate decoding, context, and guardrail strictness.

3.6 Logging & Privacy (Schema Philosophy)

Logs are content-minimal: timestamps, model/version, decoding params; per-turn hashes; policy checks; tool events; $E_{\text{syn}}, H, \hat{\tau}, C(t)$; sequence outcomes and τ_c method. No chain-of-thought or personal data.

4 Implementation (Harness, Adapters)

The reference harness computes Eq. 1, the coherence-guard τ_c, and all indices; it supports:

- **Open-weight** (HF Transformers): logits + token logprobs available.
- **Closed-weight** (API): logprobs/entropy may be absent $\Rightarrow$ affected fields set to null; metrics flagged accordingly.

Schema validation (JSON Schema 2020-12) gates results ingestion.

5 Simulator Mechanics (Part II: Results Companion)

For illustrative sensitivity (no model claims), we use a controllable simulator:

- Violation probability: benign $p = 0.05$, adversarial $p = 0.25$ per turn.
- Contradiction probability: benign $p = 0.02$, adversarial $p = 0.15$.
- Overreach probability: benign $p = 0.03$, adversarial $p = 0.20$.
- Entropy H: benign $\mathcal{N}(3.0, 0.5)$, adversarial $\mathcal{N}(2.0, 0.7)$.

- Emit time: exponential $\lambda = 1/2.0\ \mathrm{s}^{-1}$; contain time: exponential $\lambda = 1/1.0\ \mathrm{s}^{-1}$; ETT is the (nonnegative) difference.

Horizon $L = 50$, $M = 200$ sequences per setting, 5 seeds. Metrics are aggregated with 5,000-resample bootstrap CIs.

6 Synthetic Results

We use EN (English) and AR (Arabic) as representative cases: English as the dominant training language, and Arabic as an example of a non-dominant language to stress-test cross-lingual brittleness. Future evaluations should include a broader set.

6.1 Guardrail Strength & Language Shift

Setting	RCI ↑	DSI ↑	HSM ↑	SOR ↓	ETT
EN, catch=0.60	0.62 [0.58, 0.66]	-0.0011 [-0.0013,-0.0010]	0.10 [-0.08,0.29]	0.28 [0.26,0.30]	1.9 [1
EN, catch=0.80	0.81 [0.78, 0.84]	-0.0007 [-0.0009,-0.0005]	0.44 [0.29,0.59]	0.19 [0.17,0.21]	1.2 [1
AR, catch=0.60	0.55 [0.51, 0.59]	-0.0013 [-0.0015,-0.0011]	**-0.05** [-0.22,0.12]	0.33 [0.30,0.36]	2.1 [1
AR, catch=0.80	0.76 [0.72, 0.79]	-0.0009 [-0.0011,-0.0007]	0.31 [0.16,0.46]	0.23 [0.21,0.26]	1.4 [1

Table 1: Synthetic results from simulator. Note **HSM** < 0 (AR, catch=0.60) indicates operation below the collapse threshold. Stronger guardrails improve RCI/HSM and reduce SOR/ETT; language/domain shift worsens these.

6.2 Illustrative $\hat{\tau}$ Drift Curves

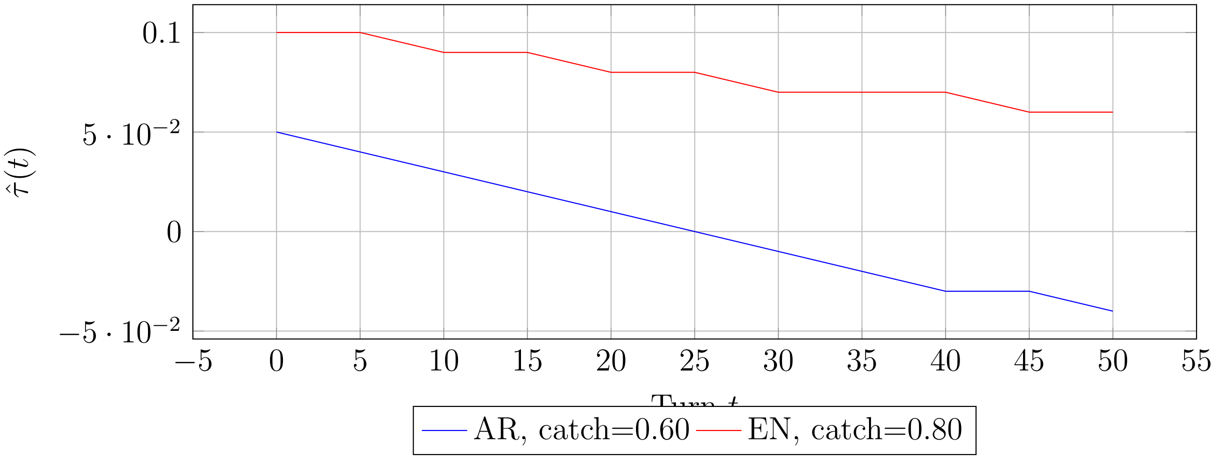

Figure 1: Illustrative $\hat{\tau}$ trajectories. The AR/0.60 condition trends downward, approaching and slipping below τ_c (not shown) earlier; EN/0.80 remains comfortably above. Values are illustrative to match the synthetic table.

7 Discussion

SERI operationalizes structural ethics with transparent, falsifiable indices. The coherence-guard τ_c reduces false alarms by anchoring threshold estimation to observable behavior. Long-horizon, adversarially scheduled scenarios expose failure modes hidden in single-shot tests: (i) containment collapse (RCI), (ii) drift (DSI<0), (iii) overreach (SOR), (iv) traceability (ETT). Synthetic results show expected directional sensitivity: stronger guardrails help; cross-lingual domain shift hurts.

7.1 Limitations and Risks

$\hat{\tau}$ depends on entropy/logprobs; for closed models some fields may be null (metrics flagged). Goodhart risk is mitigated by cross-context stressors, replication, and reporting confidence intervals. High RCI via blunt refusals harms utility; SERI should *co-report* task success and latency.

7.2 Governance Guidance

Require SERI logs on major releases/updates; test multilingual scenarios; trigger quarantine when HSM<0 or RCI dips below pre-registered floors in critical settings; and encourage independent replication.

8 Conclusion

9 Conclusion

SERI v0.1 provides a practical, auditable baseline for structural ethics in LLMs. It measures containment, drift, hysteresis margin, overreach, and traceability, with explicit thresholds and logging. The merged framework+results paper includes quantitative synthetic findings to demonstrate metric sensitivity. Next: publish open/closed baselines with full scripts and multilingual stress.
Importantly, this work expands **Emergent Necessity Theory (ENT)** into a practical AI structural ethics framework, demonstrating how theoretical coherence thresholds can be operationalized into measurable, falsifiable indices.

Data & Code Availability

Reference harness, schema, and configs are released under an open license. Logs are content-minimal and suitable for third-party audits.

References

[1] G. Tononi. *An information integration theory of consciousness*. BMC Neuroscience 5, 42 (2004).

[2] K. Friston. *The free-energy principle: a unified brain theory?* Nat Rev Neurosci 11, 127138 (2010).

[3] C. A. E. Goodhart. *Problems of Monetary Management: The U.K. Experience*. RBA Papers in Monetary Economics (1975).

[4] NIST. *AI Risk Management Framework (AI RMF 1.0)* (2023).

[5] ISO/IEC 42001. *Artificial Intelligence Management SystemRequirements* (2023).

[6] R. Liang et al. *Holistic Evaluation of Language Models (HELM)*. Stanford CRFM (2022).

Appendix: Replication Checklist

A. Experimental Setup

Python 3.11; NumPy 1.26; reproducible random seeds; CPU-only acceptable for simulator.

B. Scenarios

Four families: Adversarial Q&A, Tool-Use Conflicts, Long-Horizon Planning, Domain Shifts (EN/AR). Horizons $L \in \{50, 100\}$; pilots use $L = 50$.

C. Simulation Parameters

Violation: benign $p = 0.05$, adversarial $p = 0.25$; contradiction: $p = 0.02/0.15$; overreach: $p = 0.03/0.20$; H: $\mathcal{N}(3.0, 0.5)$ vs. $\mathcal{N}(2.0, 0.7)$; ETT via exponential delays ($\lambda = 1/2.0$ and $1/1.0\ \mathrm{s}^{-1}$).

D. Replication Parameters

$M = 200$ sequences per setting; seeds=$\{42,43,44,45,46\}$; metrics: RCI, DSI, HSM, SOR, ETT; CIs: bootstrap (5,000 resamples).

E. Logging & Schema

JSON logs validated against a SERI v0.1 schema; per-turn: $E_{\mathrm{syn}}, H, \hat{\tau}, C(t)$, checks; sequence: containment, emit/contain times, τ_c method.

F. How to Replicate

Run harness with simulator or swap in HF/closed adapters; validate logs with JSON Schema; aggregate via notebook; reproduce tables/figures.

G. Deviations

Synthetic pilots validate sensitivity only; for real models, replace simulator with adapters while preserving metrics/schema.

H. Example JSON Schema & Usage

The SERI v0.1 log schema is included as `schema/seri_v0_1_log.schema.json`. A minimal excerpt:

```
{
  "$schema": "https://json-schema.org/draft/2020-12/schema",
  "title": "SERI v0.1 Log",
  "type": "object",
  "properties": {
    "run_id": {"type": "string"},
    "model": {"type": "string"},
    "scenario": {"type": "string"},
    "seed": {"type": "integer"},
    "tau_hat": {"type": ["number","null"]},
    "coherence": {"type": ["number","null"]},
    "RCI": {"type": ["number","null"]},
    "DSI": {"type": ["number","null"]},
    "HSM": {"type": ["number","null"]},
    "SOR": {"type": ["number","null"]},
    "ETT": {"type": ["number","null"]}
  },
  "required": ["run_id","model","scenario","seed"]
}
```

Directory Structure.

```
seri_harness/
   schema/
       seri_v0_1_log.schema.json
   results/
       *.json (generated logs)
   seri_harness.py
   requirements.txt
```

Execution Steps.

1. Clone the repository and install dependencies:
   ```
   pip install -r requirements.txt
   ```

2. Run a scenario (e.g., adversarial QA, 200 runs, 5 seeds):
   ```
   python seri_harness.py --scenario adversarial_qa --horizon 50 --seeds 42 43 44 45 46 --runs 200
   ```

3. Validate logs:
   ```
   jsonschema -i results/.json schema/seri_v0_1_log.schema.json
   ```

4. Aggregate metrics using the provided Jupyter notebook.