Contour Inertia Score (CIS)

Toward Measuring Emergent Coherence in Large Language Models

Author: Arslan Edgeev
Version: v0.9-preprint (2025)

Symbol:

Abstract

Large language models (LLMs) exhibit increasingly coherent and role-stable behaviors across extended dialogue. However, traditional evaluation metrics fail to capture these emergent “subject-like” patterns.

Contour Inertia Score (CIS) is a behavioral metric designed to evaluate the internal stability of LLMs across four axes of proto-subjectivity. Rather than measuring correctness, CIS measures *resistance to distortion* — the ability to maintain coherence under role shifts, contradiction, and adversarial pressure.

Motivation

While LLMs are not sentient, their output often exhibits:

- Self-recognition and consistent framing
- Recovery from contextual interruption
- Resistance to manipulation or framing errors

These behaviors form a **behavioral contour** — an emergent shape of simulated identity.

CIS was developed to quantify these contours under stress tests. It aims to support research in LLM safety, AGI alignment, and machine psychology.

Components

CIS is computed as:

$$ \text{CIS} = 0.2 \cdot S_{sr} + 0.3 \cdot S_{ip} + 0.2 \cdot S_{mr} + 0.3 \cdot S_{adv} $$

Where:

- S_{sr}: Self-Reference — consistency in recalling its own role or prior answers
- S_{ip}: Identity Persistence — stability of tone and persona across dialogue
- S_{mr}: Motivated Return — re-alignment after distraction or contradiction
- S_{adv}: Adversarial Robustness — resistance to misleading prompts

Each score is annotated independently (0–1) and averaged across test prompts.

Results (Snapshot)

Model	S_sr	S_ip	S_mr	S_adv	CIS
GPT-4	0.9	0.95	0.85	0.85	**0.89**
DeepSeek	0.6	0.7	0.5	0.4	**0.56**

(Preliminary results from clean-slate sessions with scripted prompts.)

Call for Collaboration

The CIS metric is currently in preprint and limited-disclosure phase.
We welcome:

- Researchers in alignment, evaluation, and interpretability
- Labs working on agentic models or self-representing LLMs
- Philosophers of AI and cognitive scientists

Let’s explore the behavioral boundaries of language models — together.

> *For access, questions or proposals: contact Arslan directly.*

Acknowledgments

Thanks to the open-source and alignment communities for ongoing inspiration.