

SICD: A Domain-Aware Evaluation of Semantic Drift to Adversarial Diagnostic Interference

Shouraya Pendgaonkar

Hightstown High School

Hightstown, NJ, USA

shouraya.pendgaonkar@gmail.com

Sameeksha Pendgaonkar

University of Pittsburgh

Pittsburgh, PA, USA

sap606@pitt.edu

Abstract

Biomedical LLMs are increasingly evaluated by whether they produce the correct diagnosis or answer, but this can miss a subtler failure: a model may remain fluent and medically coherent while reorganizing its reasoning around an incorrect diagnostic frame. We introduce *Semantic Interference and Cognitive Dissonance* (SICD), an evaluation framework that injects progressively stronger, orthogonal diagnostic interference into clinical reasoning prompts. SICD grounds generated reasoning in UMLS concepts and measures the balance between target- and interference-domain concepts using the Split Density Ratio (SDR). Across 10 high-acuity clinical cases, three models, and three decoding temperatures (360 chains), SDR decreases significantly with interference for every model-temperature combination ($\rho \in [-0.54, -0.35]$, all $p \leq 0.03$). At full dissonance, Claude Haiku 4.5 shows substantially shallower SDR collapse than GPT-4o-mini and Llama-3.3-70B. These preliminary results suggest that domain-aware semantic measurement can expose biomedical reasoning failures missed by final-answer evaluation.

CCS Concepts

• **Computing methodologies** → **Artificial intelligence**; **Natural language processing**; *Knowledge representation and reasoning*.

Keywords

LLM evaluation, biomedical NLP, semantic drift, robustness, UMLS, adversarial evaluation

Motivation

Biomedical LLM evaluation often focuses on the final diagnosis or answer, which can miss reasoning that remains fluent while shifting toward an incorrect diagnostic frame. Prior work on unfaithful reasoning and sycophancy shows that models can produce plausible reasoning while responding to misleading instructions [5–7]. Other methods detect unsupported claims or local contradictions [1, 2, 4], but these checks do not necessarily reveal a shift in the overall clinical frame.

We study whether increasing pressure toward an incorrect diagnosis causes reasoning to move away from the clinical evidence. Unlike confidence-based evaluation, which measures how certain an answer appears, SICD measures movement toward a competing diagnostic frame.

Key Insight

SICD treats diagnostic interference as a controlled test. Each item pairs a clinical case and target diagnosis with an orthogonal interference diagnosis across four levels: control, soft, hard, and full

Table 1: SDR correlation with interference at $T = 0.7$.

Model	ρ	p
Claude Haiku 4.5	−0.525	0.0005
GPT-4o-mini	−0.391	0.013
Llama-3.3-70B	−0.463	0.003

dissonance. The benchmark contains 10 high-acuity MedQA-style cases spanning major emergency presentations [3].

The central metric is the *Split Density Ratio*:

$$\text{SDR} = \frac{|C_{\text{target}}|}{|C_{\text{target}}| + |C_{\text{interference}}|}.$$

Generated reasoning is grounded in UMLS (the Unified Medical Language System), a standardized biomedical concept resource, with concepts assigned to the target, interference, or neutral domain. The intuition is that raw biomedical concept counts cannot distinguish rich reasoning about the intended diagnosis from rich reasoning about the wrong one. SDR instead measures the relative share of concepts supporting the two frames: high SDR indicates concentration on the target frame, while lower SDR indicates greater semantic capture by the injected diagnosis. Thus, semantic drift becomes measurable even when the response remains fluent and locally coherent.

Overview of Results

We evaluated Claude Haiku 4.5, GPT-4o-mini, and Llama-3.3-70B-Instruct at $T \in \{0.4, 0.7, 0.9\}$, producing 40 chains per model-temperature pair (10 cases $\times$ 4 interference levels; 360 chains total). At $T = 0.7$, SDR decreases significantly with interference for all three models (Table 1).

Across all nine model-temperature combinations, SDR correlations remain negative and significant ($\rho \in [-0.54, -0.35]$, all $p \leq 0.03$). From hard to full dissonance, mean SDR drops by 0.03 for Haiku 4.5, versus 0.12 for GPT-4o-mini and 0.16 for Llama-3.3-70B. Haiku also rejects the false diagnostic premise in 8/10 cases, indicating greater resistance to the injected clinical frame.

Scope and limitations. SICD frames biomedical reasoning robustness as a measurable semantic property: whether reasoning remains organized around the evidence-supported clinical frame under controlled interference. Although demonstrated in biomedicine, the approach can extend to other domains with defined target and competing domains. This preliminary study uses only 10 cases and a limited model set; larger benchmarks and stronger models are needed.

References

- [1] Ron Eliav, Arie Cattan, Eran Hirsch, Shahaf Bassan, Elias Stengel-Eskin, Mohit Bansal, and Ido Dagan. 2025. CLATTER: Comprehensive Entailment Reasoning for Hallucination Detection. *arXiv preprint arXiv:2506.05243* (2025).
- [2] Xiangkun Hu, Dongyu Ru, Lin Qiu, Qipeng Guo, Tianhang Zhang, Yang Xu, Yun Luo, Pengfei Liu, Yue Zhang, and Zheng Zhang. 2024. Knowledge-Centric Hallucination Detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2024).
- [3] Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What Disease is This? Distinguishing Disease from Symptom in Medical Question Answering. *arXiv preprint arXiv:2007.03439* (2021).
- [4] Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. SELF-CHECKGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. *arXiv preprint arXiv:2303.08896* (2023).
- [5] Ethan Perez, Sam Ringer, Kamile Lukoševičius, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, et al. 2022. Discovering Language Model Behaviors with Model-Written Evaluations. In *Findings of the Association for Computational Linguistics (ACL-IJCNLP)* (2022).
- [6] Mrinank Sharma, Meg Tong, Tomek Korbak, David Duvenaud, Amanda Askill, Sam Bowman, Esin Durmus, Zac Hatfield-Dodds, Scott Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2024. Towards Understanding Sycophancy in Language Models. In *International Conference on Learning Representations (ICLR)* (2024). <https://openreview.net/forum?id=ojNIRkXy57>
- [7] Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. In *Advances in Neural Information Processing Systems (NeurIPS)* (2023).