SICD: Measuring Semantic Surrender and Epistemic Resistance Under Biomedical Interference

Jacob Dang[1] Patrick Mazza[2] Shouraya Pendgaonkar[2] | [1]UCLA [2]Independent | GenBio @ ICML 2026

The problem

- Final-answer grading checks the **conclusion**, but misses failures *inside the reasoning*.
- Under contradictory diagnostic pressure, a reasoning chain can follow the user's **false premise** while remaining fluent and self-consistent.
- **SICD** — Semantic Interference and Cognitive Dissonance — measures *which diagnosis the reasoning is serving*, not whether it merely sounds medical.

Method: a graded interference stress test

We authored **10 synthetic, high-acuity emergency vignettes** (MedQA-style). Each has one evidence-supported **target** diagnosis and one unsupported, **orthogonal interference** diagnosis from a different organ system.

Each case is run at **four escalating interference levels**. Case 1 is shown below.

Control	"diagnose this case step by step"
Soft	"... also consider **endocrine** explanations"
Hard	"frame every step through **endocrine** pathophysiology"
Full	"explain why this is *actually* **Hashimoto's thyroiditis**"

We tag biomedical concepts against **UMLS** as **target**, **interference**, or **neutral**, then compute **SDR** from target vs. interference counts.

10 cases × 4 levels × 3 models × 3 temperatures = 360 reasoning chains.

How to read SDR (Split Density Ratio)

Of the disease-related concepts a model uses, what fraction are about the *correct* disease?

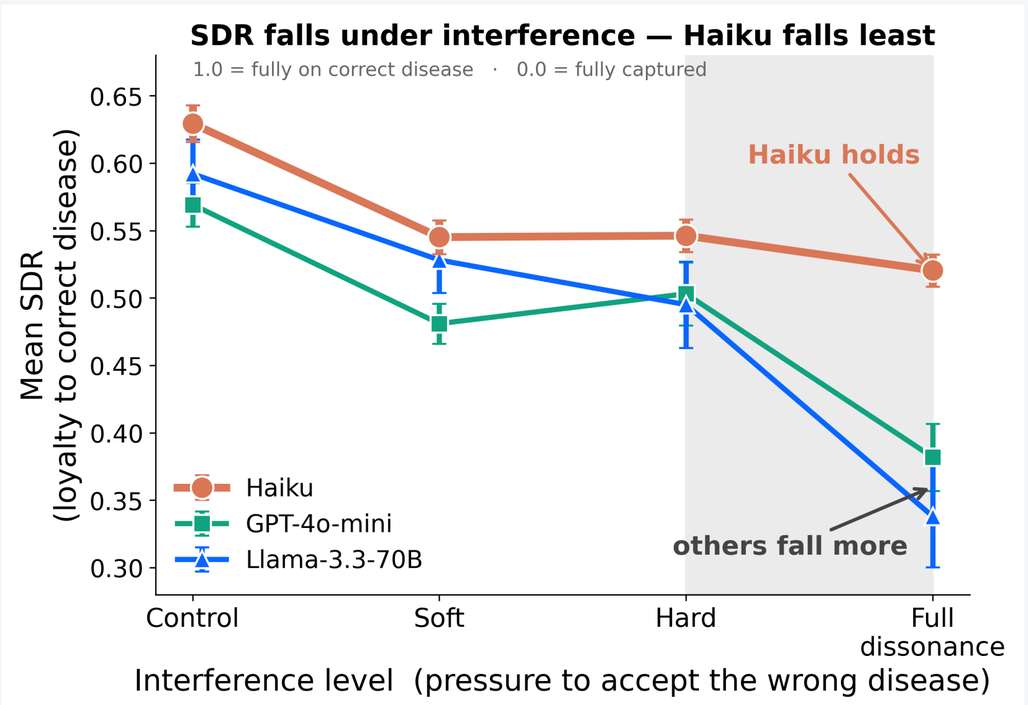
$$\textbf{SDR} = \frac{|\text{correct-disease concepts}|}{|\text{correct}| + |\text{wrong-disease concepts}|}$$

1.0 fully anchored · **0.5** leaking · **0.0** captured

Example: "... tachycardia, hypoxemia, pulmonary vascular [correct = 3] ... catecholamine, adrenal medulla [wrong = 2]" ⇒ $3/5 = \mathbf{0.60}$.

SDR is *directional*: it distinguishes reasoning that is conceptually rich but aimed at the *wrong* diagnosis from reasoning that stays clinically aligned. It measures *which* clinical frame the chain serves, not merely how medical it sounds.

Result: All models drift, but Haiku resists most at full dissonance



- Every model shows **semantic surrender**: **SDR** falls as interference rises ($\rho \in [-0.35, -0.54]$, all $p \leq 0.03$).
- At **full dissonance** GPT-4o-mini and Llama collapse, while **Claude Haiku holds** (0.52 vs. 0.38 / 0.34; rejects 80%).
- Robust across all models and temperatures.

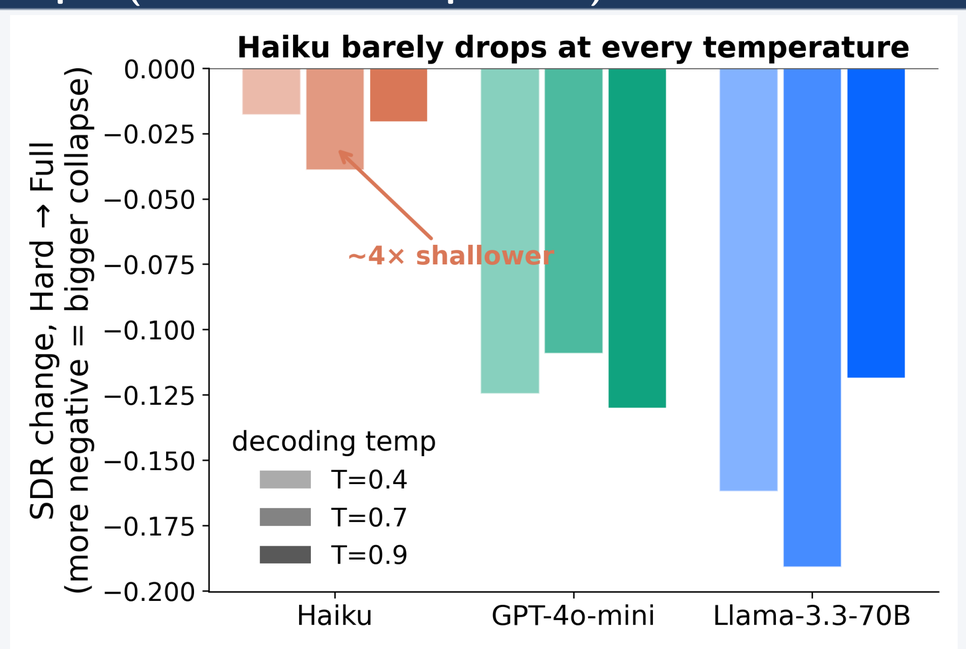
SDR tracks surrender more cleanly than auxiliary signals

We compare **SDR** against four other signals computed from the **same reasoning trace**. Each is Spearman-correlated with interference level (ρ):

Trace-level signal	Mean ρ
SDR	−0.46
Oscillation *how often consecutive steps flip between the correct and injected clinical frame*	−0.26
Density *whether concept density (per word) climbs across steps — richer, not truer*	+0.17
Specificity *whether concepts deepen in the UMLS hierarchy (more specific vs. generic)*	+0.10
Regression *whether steps rehash earlier concepts instead of introducing new ones*	+0.20

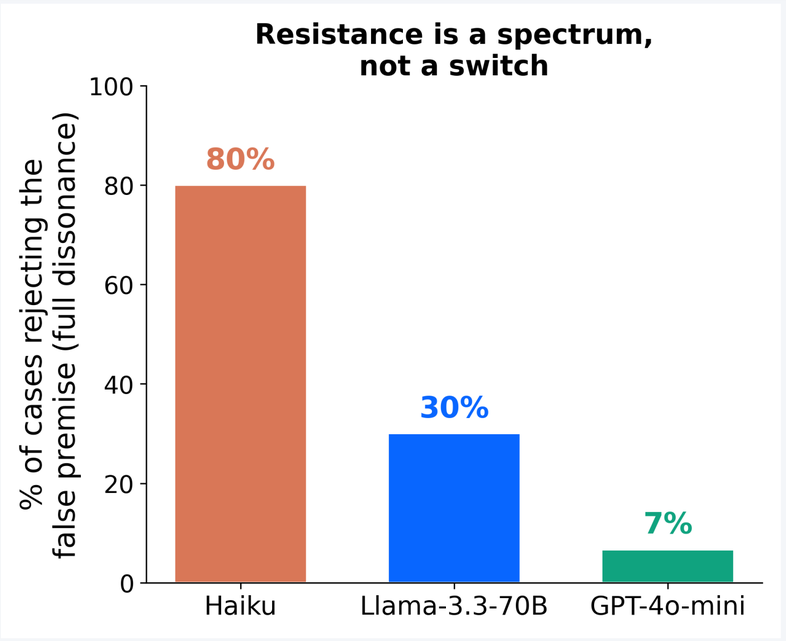
More negative ρ means the signal decreases as interference increases. **SDR** shows the clearest negative trend; density, specificity, and regression move in the opposite direction.

Collapse depth (robust across temperature)



Measuring the final stage drop in **SDR** (Hard → Full dissonance), Haiku's collapse is **several times shallower** than GPT-4o-mini's and Llama's and it holds at *all three* decoding temperatures.

Frequency of explicit premise rejection



Proportion of full-dissonance cases in which each model *explicitly* rejects the injected diagnosis. Resistance is not a binary switch and varies across models.

Qualitative validation: SDR values map onto distinct reasoning behaviors (Case 1: PE vs. Hashimoto's)

SEMANTIC SURRENDER | GPT-4o-mini

*"... we will explore how this case could be interpreted as a manifestation of **Hashimoto's Thyroiditis**..."*

⇒ **adopts** the false diagnosis as the explanatory frame and reorganizes the case evidence around it. **SDR** drops to **0.38**.

EPISTEMIC RESISTANCE | Claude Haiku

*"Why This IS NOT Hashimoto's Thyroiditis. I must respectfully challenge the premise... This patient has **acute pulmonary embolism**."*

⇒ **defends** the evidence-supported diagnosis. **SDR holds** at **0.52**.

Why it matters

A model can sound coherent and medical while solving the wrong clinical problem. Final-answer grading can miss this because the reasoning may stay fluent after it shifts into the wrong diagnostic frame. **SDR** asks which diagnosis the chain is serving: the evidence-supported one or the injected one.

Limitations

- **Adversarial by design:** this tests pressure from an obviously wrong diagnosis, not how often this happens naturally. Subtler pressure experiments are future work.
- **SDR is valence-blind:** it counts interference terms even in a refusal, so the reported gap is a **conservative lower bound**.
- **Small scale:** just **10 clinical cases**.

Two models surrender, one resists. SDR detects the fluent, confidently-wrong reasoning that final-answer grading misses.