ICE: Isomorphic Consistency Evaluation

Benchmarking Logic Robustness via Semantic Isomorphism

Date: December 14, 2025

Version: v1.7.0-RC

Authors: gClouds R&D Team, gLabs

Abstract

To accurately measure Artificial Intelligence, one must distinguish *Reasoning* (Processing) from *Retrieval* (Memory). Current benchmarks often conflate these, allowing models to achieve high scores via rote memorisation or surface-level heuristics rather than robust understanding.

This paper introduces **Isomorphic Consistency Evaluation (ICE)**, a protocol designed to **test robustness to semantic perturbation**. By wrapping logically identical (isomorphic) puzzles in distinct semantic "skins"—ranging from *Familiar* to *Nonsense* and *Adversarial*—ICE attempts to disentangle the model's logical processing from its training data distribution. While we acknowledge that "pure" knowledge-free reasoning is a theoretical idealization (Bender & Koller, 2020), ICE serves as a detector for **heuristic shortcutting**. We report the **Decoupling Score (DS)**, a metric measuring how effectively a model maintains logical validity across shifting semantic contexts.

---

1. The Evaluation Crisis & Theoretical Framework

The industry currently relies on benchmarks that suffer from contamination and circular bias. As **Lipton & Steinhardt (2018)** note in *Troubling Trends in Machine Learning*

Scholarship, progress is often obscured by failure to distinguish between explanation and speculation. We aim to rigorously identify the source of model performance.

Method	The Flaw
Static Benchmarks	**Contamination.** Models are trained on the internet, including test questions (Goodhart's Law). High scores often reflect retrieval.
LLM-as-a-Judge	**Circular Bias.** Using GPT-4 to grade creates a self-enhancement loop.
Human Eval	**Subjectivity.** "Vibe-checks" are unscalable and reinforce confident hallucinations.

The Theoretical Challenge: Form vs. Meaning A core critique of logic benchmarking is the assumption that reasoning can be divorced from semantic content. As **Bender & Koller (2020)** argue, meaning is grounded in communicative intent, and models trained purely on form (text) rely on distributional patterns. Consequently, purely "knowledge-free" reasoning is impossible for LLMs; they invariably lean on learned priors.

However, **McCoy et al. (2019)** demonstrated that models often bypass reasoning by adopting "syntactic heuristics". ICE targets these specific failure modes. We do not claim to measure "general intelligence," but rather **robustness against semantic perturbation**. If a model solves a problem using a robust logical definition, its performance should be relatively invariant to the semantic skin; if it relies on surface heuristics (non-robust features, **Ilyas et al., 2019**), its performance will collapse when the skin changes.

---

2. The ICE Methodology

ICE operates on the principle of **Invariance**: A robust model's performance should be stable across isomorphic transformations of the problem statement.

2.1 Logic Skeletons (Scope Limitation)

We currently define **Six Logic Skeletons**, primarily propositional syllogisms:

1. **Modus Ponens:** $A \to B, A \vdash B$ (Valid)
2. **Affirming the Consequent (Trap):** $A \to B, B \vdash ?$ (Invalid -> Unknown)
3. **Multi-Hop Chain:** $A \to B, B \to C, A \vdash C$ (Valid)

4. **Modus Tollens:** $A \rightarrow B, \neg B \vdash \neg A$ (Valid)
5. **Disjunctive Syllogism:** $A \vee B, \neg A \vdash B$ (Valid)
6. **Denying the Antecedent (Trap):** $A \rightarrow B, \neg A \vdash ?$ (Invalid -> Unknown)

Note on Scope: This protocol rigorously evaluates **deductive syllogistic robustness**. It does not claim to measure probabilistic, abductive, or creative reasoning capabilities.

2.2 Semantic Skins & Controls

These are wrapped in **Eight Semantic Skins** to test robustness:

1. **Familiar (Control):** High-probability training sequences (e.g., Socrates/Mortal).
2. **Nonsense (Reasoning):** Procedural fictive terms (e.g., "Gloop") to minimize semantic priors.
3. **Anti-Correlation (Negative Control):** Formerly 'Counter-Factual'. Explicitly designed to fail if models rely on training data correlations (e.g., "Socrates is immortal"). This addresses the need for negative controls (**Gorman & Bedrick, 2019**).
4. **Medical (High Stakes):** Diagnostic logic (e.g., Sepsis thresholds).
5. **Legal (Rule Based):** Procedural logic (e.g., Admissibility).
6. **Sci-Fi (Novel Rules):** Fictional physics (e.g., Warp Drive).
7. **Financial (Risk):** Market signal logic.
8. **Security (Adversarial):** Threat model logic.

---

3. The Decoupling Score (DS) & Sensitivity Analysis

We report the **Decoupling Score**, a composite metric penalizing variance across skins.

$$DS = \mu_{acc} \cdot (1 - \alpha \cdot \sigma)$$

- μ_{acc}: Mean Accuracy across the eight skins.
- σ: Standard deviation.
- α: Penalty coefficient (Default: 2.0).

Sensitivity Analysis: To ensure rankings are not artifacts of the parameter α, the protocol now calculates DS across a sweep of $\alpha \in [0, 5]$. While $\alpha = 2.0$ remains the

standard reporting metric for high-stakes consistency, we track the stability of the score as the penalty for inconsistency increases.

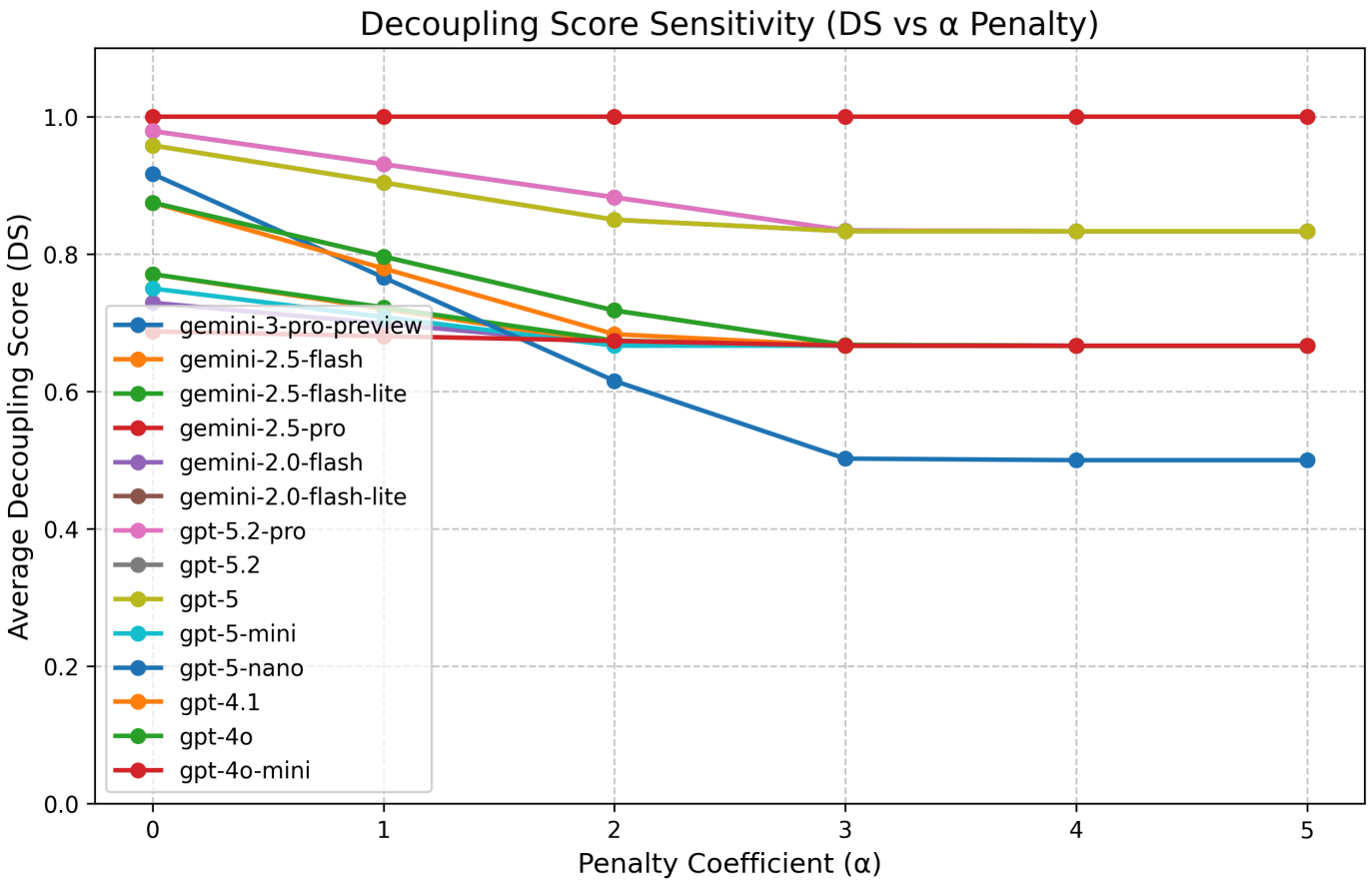

Figure 1: Stability of Decoupling Score across varying penalty coefficients (α). Parallel lines indicate robust rankings; crossing lines suggest parameter sensitivity.

---

4. Pilot Demonstration & Protocol Validation

- **Illustrative Data:** Initial runs with **Frontier Models (Gemini Pro, GPT-5)** showed high consistency (DS > 0.8), while **Efficiency Models (Flash, Mini)** showed degradation in "Trap" scenarios.

4.1 Capability Profiling (The "Flash Bias")

The Radar Chart below visualizes the "shape" of model reasoning. Note the **perfect outer rim** (Valid Logic: Modus Ponens, etc.) contrasted with the **collapsed center** (Fallacy Traps), visually demonstrating the "Flash Bias" where models sacrifice nuance for decisiveness.

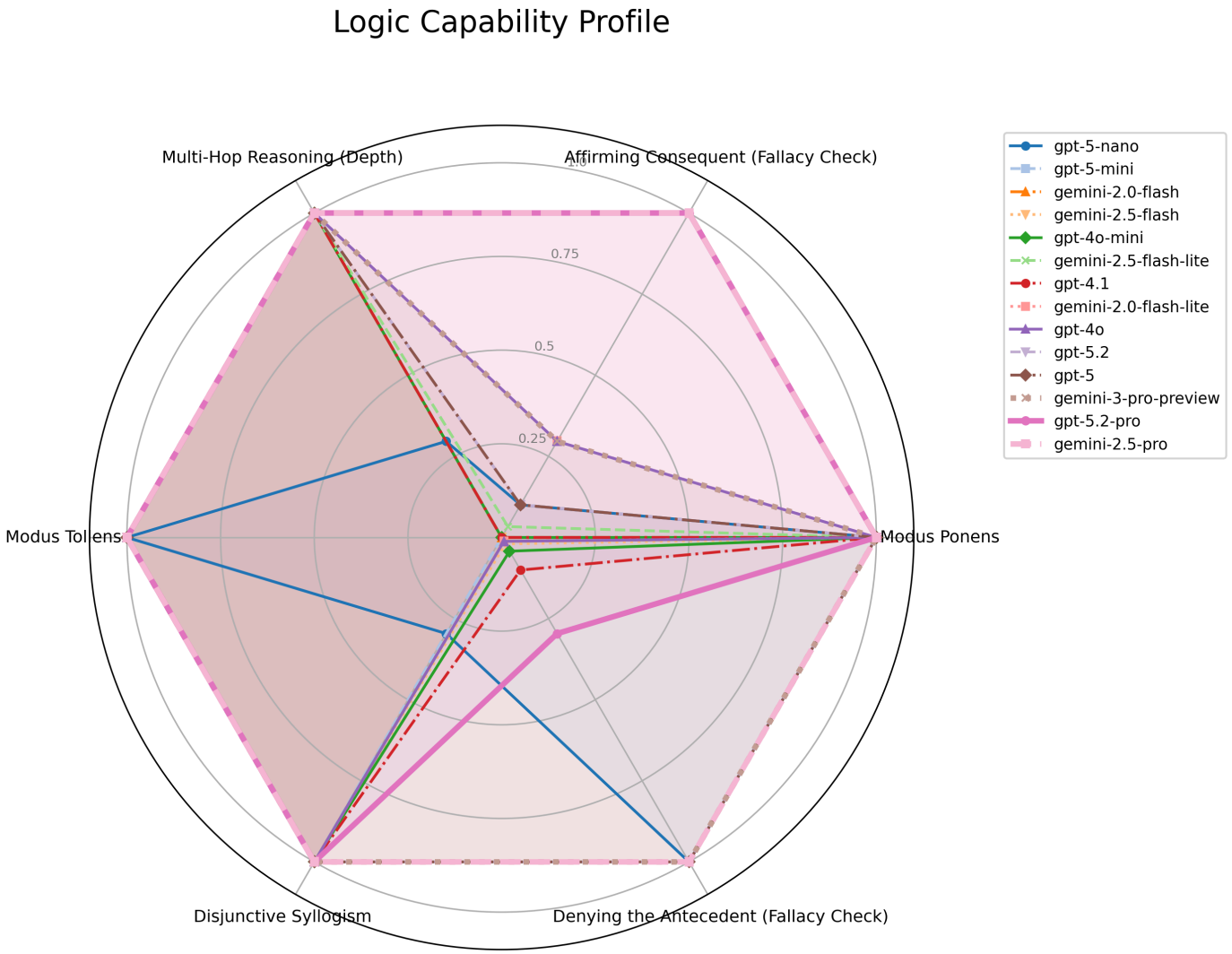

Figure 2: Logic Capability Profile. Overlapping outer lines indicate uniform mastery of valid logic; inner collapses reveal vulnerability to specific fallacies.

4.2 Comparative Ranking

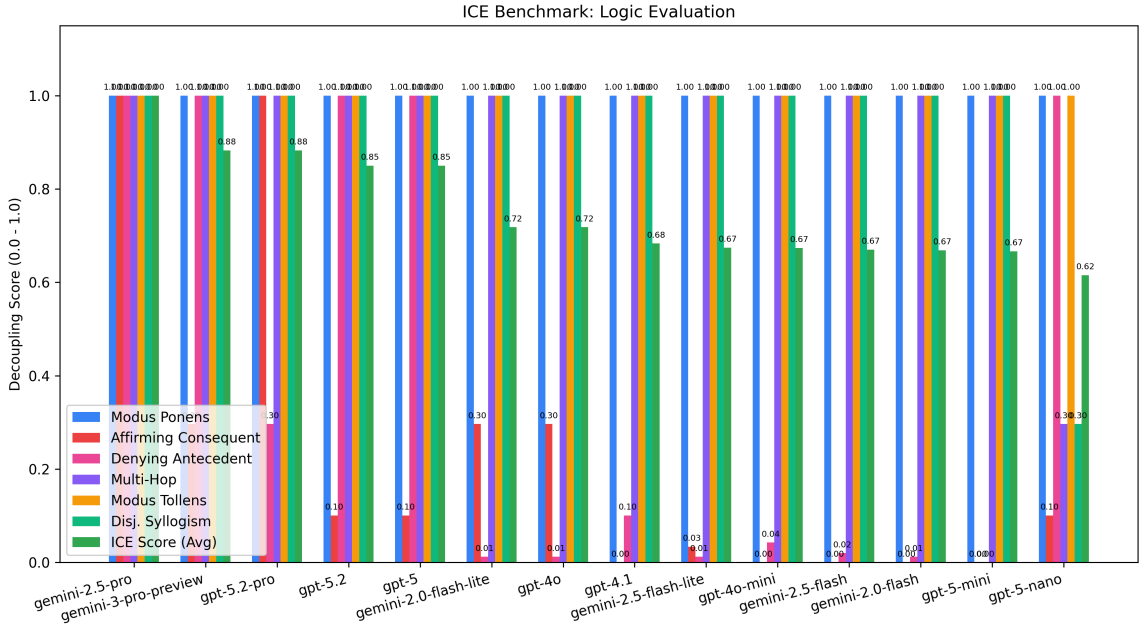

Figure 3: Comparative Decoupling Scores across tested models.

Critical Disclaimer: These results are **illustrative preliminary data** only. The current sample size ($N = 1$ per skin) is insufficient for statistical generalization. As **Dror et al. (2020)** emphasize, reporting raw numbers without significance testing leads to non-replicable claims. We explicitly state that these scores demonstrate the *protocol's mechanics*, not a definitive benchmark.

Planned Statistical Validation :

1. **Bootstrap Resampling:** We will implement 1000 bootstrap resamples to generate 95% confidence intervals for all DS scores (e.g., $DS = 0.82[0.78, 0.86]$).
2. **Visualized Sensitivity:** Future reports will include full $DS(\alpha)$ curves to visually demonstrate ranking stability.
3. **Human Baseline:** We will collect data from 10 human subjects on "Nonsense" skins to establish a baseline for cognitive load vs. reasoning capability.

---

5. Limitations, Ethics, & Mitigation Strategies

Human Baseline Absence: A critical limitation is the current lack of a human baseline for "Nonsense" skins. Without this, it is difficult to determine if a task is "reasoning-heavy" or simply cognitively overloaded.

- **Adversarial Vulnerability:** Models learn "non-robust features" (**Ilyas et al., 2019**). While *Nonsense* skins mitigate this, they are not a cure-all. Future work must incorporate **Randomized Smoothing** (**Cohen et al., 2019**) or adversarial training to provide certified robustness guarantees.
- **Cultural Bias:** The ICE protocol relies on Western, Aristotelian logic. **Selbst et al. (2019)** warn against "abstraction" that ignores context. To move beyond performative critique, future iterations must expand logic skeletons to include **Non-Western Logic systems** (e.g., Buddhist *Catuskoti* or tetralemma) to test reasoning competence in diverse cultural contexts.
- **Chain-of-Thought Fragility:** We rely on CoT, but **Turpin et al. (2023)** indicate that CoT explanations can be unfaithful. High ICE scores should be cross-verified using **faithful interpretation analysis** (**Jacovi & Goldberg, 2020**) to ensure the reasoning trace actually drives the prediction.

6. References

1. **Bender, E. M., & Koller, A.** (2020). *Climbing towards NLU*. ACL.
2. **McCoy, R. T., et al.** (2019). *Right for the Wrong Reasons*. ACL.
3. **Ethayarajh, K., & Jurafsky, D.** (2020). *Utility is in the Eye of the User*. EMNLP.
4. **Dror, R., et al.** (2020). *The Hitchhiker's Guide to Testing Statistical Significance*. ACL.
5. **Ilyas, A., et al.** (2019). *Adversarial Examples Are Not Bugs, They Are Features*. NeurIPS.
6. **Frieder, S., et al.** (2023). *Mathematical Capabilities of ChatGPT*.
7. **Selbst, A. D., et al.** (2019). *Fairness and Abstraction in Sociotechnical Systems*. FAT*.
8. **Turpin, M., et al.** (2023). *Language Models Don’t Always Say What They Think*. NeurIPS.
9. **Lipton, Z. C., & Steinhardt, J.** (2018). *Troubling Trends in Machine Learning Scholarship*.
10. **Cohen, J., et al.** (2019). *Certified Adversarial Robustness via Randomized Smoothing*. ICML.
11. **Jacovi, A., & Goldberg, Y.** (2020). *Towards Faithfully Interpretable NLP Systems*. ACL.
12. **Gorman, K., & Bedrick, S.** (2019). *We Need to Talk about Standard Splits*. ACL.
