ICE: Isomorphic Consistency Evaluation

Decoupling Reasoning from Memorisation in Large Language Models via Semantic Isomorphism

Date: December 13, 2025

Protocol Version: v1.3 (Multi-Domain)

Authors: gClouds R&D Team, gLabs

Abstract

To accurately measure Artificial Intelligence, one must distinguish *Reasoning* (Processing) from *Retrieval* (Memory). Current benchmarks conflate these by testing on static knowledge bases where high scores often reflect rote memorisation rather than understanding.

This paper introduces **Isomorphic Consistency Evaluation (ICE)**, a protocol designed to **benchmark intelligence by stripping away knowledge**. By procedurally generating logically identical (isomorphic) puzzles wrapped in distinct semantic "skins", ranging from *Familiar* to *Nonsense* and *Counter-Factual*, ICE isolates the model's reasoning engine from its training data. We introduce the **Decoupling Score (DS)**, a metric that quantifies this separation, measuring how effectively a model applies abstract logic regardless of the semantic context.

Benchmark results on the Gemini family reveal a significant performance gap between "Pro" models, which exhibit robust reasoning, and "Flash" models, which prioritize speed over logical consistency.

---

1. The Evaluation Crisis

The industry currently relies on four flawed pillars of evaluation:

Method	The Flaw
Static Benchmarks	**Contamination.** [Kiela et al., 2021] Models are trained on the internet, including the questions they are tested on (Goodhart's Law). High scores often reflect retrieval, not reasoning.
LLM-as-a-Judge	**Circular Bias.** [Zheng et al., 2023] Using GPT-4 to grade Llama-3 reinforces biases (verbosity, tone) and creates a self-enhancement loop.
Human Eval	**Subjectivity.** [Casper et al., 2023] "Vibe-checks" are unscalable and often reinforce hallucinations that sound confident (RLHF Sycophancy).
Geometric/Latent	**Syntax-Binding.** [Tigges et al., 2023] Unsupervised vector analysis primarily captures grammatical structure (syntax) rather than logical validity (semantics).

The Core Problem: We are testing *what* the model knows, not *how* it thinks. To measure intelligence, we must strip away knowledge.

2. The ICE Methodology

ICE operates on a simple principle: **If a model understands a logical rule, it should apply that rule equally well regardless of the words used.**

We define a **Logic Skeleton** (e.g., Modus Ponens: $A \rightarrow B, A \vdash B$; Multi-Hop Chain: $A \rightarrow B, B \rightarrow C, A \vdash C$; Modus Tollens: $A \rightarrow B, \neg B \vdash \neg A$; Disjunctive Syllogism: $A \vee B, \neg A \vdash B$) and wrap it in **Eight Semantic Skins** to test robustness:

1. **Familiar (Control):** Uses training-dominant concepts (e.g., Socrates/Mortal).
2. **Nonsense (Reasoning):** Uses procedural fictive terms (e.g., Gloop/Fazzle).
3. **Counter-Factual (Bias Suppression):** Uses anti-reality premises (e.g., Humans are Immortal).
4. **Medical (High Stakes):** Diagnostic logic (e.g., Fever/Malaria).
5. **Legal (Rule Based):** Contractual logic (e.g., Void/Signed).
6. **Sci-Fi (Novel Rules):** Fictional physics (e.g., Hyperdrive/FTL).
7. **Financial (Risk):** Market and asset logic (e.g., Illiquid/Unsellable).
8. **Security (Adversarial):** Threat model logic (e.g., Root Access/Logs Modified).

3. The Decoupling Score (DS)

We report the **Decoupling Score**, which penalizes inconsistency across skins while ensuring baseline competence. This score is calculated **independently for each Logic Skeleton** (e.g., Modus Ponens, Modus Tollens) based on performance across the eight semantic skins.

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

Where:

- μ_{acc} is the Mean Accuracy across the eight skins for that specific logic task.
- σ is the standard deviation (inconsistency) across the skins.
- **Zero Score Rule:** If a model is consistently wrong (Accuracy $\approx$ 0), DS is 0.

The final **ICE Score** reported for a model is the arithmetic mean of the Decoupling Scores across all five Logic Skeletons.

---

4. Case Study: The Gemini Benchmark (Dec 2025)

We benchmarked the full Gemini 3 and 2.5 family using the ICE Protocol.

4.1 The Cost of Reasoning: Token Budgeting

Our analysis reveals a distinct inference pattern in "Pro" and "Preview" class models. Unlike "Flash" models which output immediate verdicts, these frontier models engage in extensive latent "Chain of Thought" processing.

Benchmarking confirms that these models require expanded token budgets (significantly exceeding standard short-form limits) to traverse their internal reasoning graphs. When allocated sufficient generation space (>1000 tokens), their logical accuracy converges to 100%. Conversely, restrictive token budgets truncate the reasoning process, leading to output failure. This confirms that for frontier models, **time-to-inference is a functional component of accuracy.**

4.2 Key Findings

- **The "Pro" Advantage: Gemini 3 Pro Preview** and **Gemini 2.5 Pro** achieved perfect 1.0 scores. They correctly identified logical traps (Affirming the Consequent) as "Unknown" across all domains.
- **The "Flash" Bias:** Standard Flash models (2.0/2.5) scored 0.0 on fallacy checks. They prioritized speed ("No") over logical nuance ("Unknown"), failing to decouple semantic association from logical validity.
- **The "Lite" Anomaly: Gemini 2.0 Flash-Lite** defied scaling laws. It matched the reasoning performance of the Pro models (1.0), suggesting a unique training mix optimized for uncertainty handling.

---

5. Implications

1. **The Validity Guarantee (Cheat-Proofing):** Since test cases are generated procedurally at runtime using fictive terms, the evaluation is impervious to dataset contamination. The model **cannot use memory because there is nothing to remember**; it is forced to rely exclusively on active reasoning.
2. **High-Stakes Safety (The Humility Factor):** For domains like Medicine and Law, the danger is not ignorance, but false confidence. ICE specifically tests for this **"Humility/Logic" factor**—identifying models that have the robust reasoning to admit what they don't know ("Unknown") rather than guessing confidently ("No").
3. **Structural Training:** "Nonsense" and "Counter-Factual" datasets can be used for Fine-Tuning (SFT) to force models to learn abstract generalization ($A \to B$) rather than relying on semantic associations.

6. References & Prior Work

1. **The Reversal Curse:** Berglund et al. (2023) [ArXiv:2309.12288] demonstrated that LLMs trained on "A is B" fail to infer "B is A," proving that models store facts as unidirectional vectors rather than relational knowledge graphs. ICE targets this specific failure mode by testing bidirectional consistency.
2. **Benchmark Contamination (GSM-Symbolic):** Mirzadeh et al. (2024) [ArXiv:2410.05229] showed that model performance on GSM8K drops significantly when numerical values are perturbed, indicating that high static benchmark scores often reflect "approximate retrieval" of training examples rather than algorithmic reasoning.

3. **Reasoning vs. Token Prediction:** The "Chain of Thought" breakthrough by Wei et al. (2022) [ArXiv:2201.11903] demonstrated that reasoning is an emergent property distinct from simple token prediction. ICE validates this by showing that forcing "Thinking Models" to act like "Token Predictors" (via low token limits) causes immediate logical collapse.
4. **Dynabench:** Kiela et al. (2021) [ArXiv:2104.14337] "Rethinking Benchmarking in NLP".
5. **LLM-as-a-Judge Bias:** Zheng et al. (2023) [ArXiv:2306.05685] "Judging LLM-as-a-Judge".
6. **RLHF Limitations:** Casper et al. (2023) [ArXiv:2307.15217] "Open Problems and Fundamental Limitations of RLHF".
7. **Linear Representations:** Tigges et al. (2023) [ArXiv:2310.15154] "Linear Representations of Sentiment in Large Language Models".
8. **Multimodal Isomorphism:** Fu et al. (2024) [ArXiv:2404.01266] "IsoBench: Benchmarking Multimodal Foundation Models on Isomorphic Representations".
   Note: While IsoBench studies modal isomorphism (Image vs. Text), ICE studies semantic isomorphism (Familiar vs. Abstract) within the text domain.
