

BHATT CONJECTURES: ON NECESSARY-BUT-NOT-SUFFICIENT BENCHMARK TAUTOLOGY FOR HUMAN LIKE REASONING

Manish Bhatt*

AI Security Researcher

manish.bhatt13212@gmail.com

June 2025

1 Introduction

Debates about whether Large Language/Reasoning Models (LLMs/LRMs) truly *reason* or merely *pattern-match* (Marcus, 2023; Chollet, 2019) suffer from shifting goal-posts. In my opinion, two analytic—hence “tautological”—benchmarks cut through that fog in my mental model:

- \mathbf{T}_1 : *Reasoning-Capability Tautology*
- \mathbf{T}_U : *Understanding-Capability Tautology*

Because each definition states *necessary and sufficient* conditions **by construction**, the challenge is **not** to “prove the tautologies are true”—they are definitions—but to determine whether a *concrete system* meets them.

This paper first restates \mathbf{T}_1 and \mathbf{T}_U with surgical precision, then extends \mathbf{T}_U to a richer \mathbf{T}_U^* that incorporates causal modelling (Pearl, 2009), metacognition (Flavell, 1979), the fast/slow thinking dichotomy (Kahneman, 2011), and the (currently untestable) question of phenomenal awareness (Baars, 1988).

2 \mathbf{T}_1 — Reasoning-Capability Tautology

Definition 2.1 (\mathbf{T}_1). A system **reasons** with respect to a problem class C **if and only if**:

(R1) it produces a correct solution from *any* logically equivalent representation of

*This work is not related to author’s position at Amazon.

the formally specified premises, and

- (R2) its success probability remains high when the surface form of the problem is *outside* its training distribution (no isomorphic instance or step-by-step demonstration seen in training or finetuning data).

2.1 Canonical Corollaries

ID	Corollary	Example Diagnostic Test
C1	Representation Invariance	
	Restate the task in a different natural language, formal notation, diagram, or noisy encoding; accuracy must persist.	Paraphrase a Sudoku constraint set into first-order logic; solver must still succeed.
C2	Complexity Scaling	
	Hold logical structure fixed, increase instance size (e.g., more discs in Tower of Hanoi). Accuracy should not collapse.	Apple LRMs drop to $\approx 0\%$ beyond 5 discs despite flawless small- n performance (Shojaee et al., 2025).
C3	Zero-Shot Robustness	
	Inject <i>novel</i> but logically equivalent surface patterns absent from training.	Swap “ $A \rightarrow B$ ” syntax for λ -calculus encoding; solution quality must persist.

Table 1: Corollaries and diagnostic probes for \mathbf{T}_1 .

3 \mathbf{T}_U — Understanding-Capability Tautology

Definition 3.1 (\mathbf{T}_U). A system **understands** a knowledge domain D **if and only if** for every proposition φ about D :

- (U1) it maps *any* truth-preserving representation $\rho(\varphi)$ to an internal state $I(\varphi)$ that recovers the truth value of φ , and
- (U2) conditions U1 hold when $\rho(\varphi)$ stems from a distribution statistically independent of the training data, and when queried with *unseen* operations derivable

solely from $I(\varphi)$ (e.g., counterfactuals, entailments).

3.1 Canonical Corollaries

ID	Corollary	Example Diagnostic Test
C4	Modal Invariance	
	Understanding survives cross-modal transfer (text \rightarrow image schema, Braille, speech).	Encode “water freezes at 0°C” as a line in a phase diagram; model must still answer “True/False?”
C5	Counterfactual Competence	
	Given $I(\varphi)$, the system answers <i>derived</i> queries (inference, contradiction, analogy).	From “All swans are birds,” infer “If X is a swan, X can fly <i>unless</i> flightless-bird exceptions apply.”
C6	Distribution Shift	
	Replace everyday examples with rare or synthetic ones; truth evaluations remain stable.	Test chemical facts using IUPAC strings never seen during training.

Table 2: Corollaries and diagnostic probes for $\mathbf{T_U}$.

3.2 Relation to $\mathbf{T_1}$

$\mathbf{T_U}$ *strictly contains* $\mathbf{T_1}$ on factual domains: correct reasoning over φ demands first mapping its meaning robustly; hence $\mathbf{T_U} \Rightarrow \mathbf{T_1}$ (but not conversely).

4 $\mathbf{T_U^*}$ — Extended Understanding

$\mathbf{T_U}$ is intentionally behaviourist. To bridge explanatory gaps without leaping to metaphysics, we propose an *Extended Understanding-Capability Tautology*, $\mathbf{T_U^*}$:

Definition 4.1 ($\mathbf{T_U^*}$). A system **deeply understands** a domain D if and only if it satisfies $\mathbf{T_U}$ *and* the following premises:

- (E1) **Causal Structural Fidelity** — Its internal representation mirrors the *causal* graph of D , enabling do-calculus interventions (Pearl, 2009).
- (E2) **Metacognitive Self-Awareness** — It evaluates and calibrates the reliability

of its own inferences, signalling uncertainty or ignorance (Shum et al., 2023).

(E3) **Phenomenal Awareness** — Inference is accompanied by subjective experience (qualia). Untestable, this marks the boundary between functional mastery and conscious sapience (Chalmers, 1995).

5 Fast & Slow Thinking: A Path to Emergent Cognitive Abilities

Open Challenge. Integrating both modes is non-trivial. Early attempts (e.g., hypothetical models like *Claude 3.7*, *Qwen 3.2*) interleave chain-of-thought prompting (Wei et al., 2022) with high-temperature sampling yet still default to shallow heuristics. Robust solutions may require:

- **Hybrid architectures** that switch between fast pattern recall and slow causal reasoning based on task uncertainty (Suzgun et al., 2023).
- **On-device “fast cores”** (e.g., future *Llama* models) tightly coupled with stronger cloud “slow cores” (e.g., a future *Claude* model), delivering low-latency intuition plus high-fidelity deliberation.
- **Curriculum training** rewarding extended, coherent reasoning chains while penalising hallucination (Lightman et al., 2023).

5.1 Layered Evidence Framework

Layer	Target Property	Test Technology
Behavioural (T_U)	Format- & distribution-robust truth preservation	Stress tests; adversarial modalities (Hendrycks et al., 2021)
Causal (E1)	Fidelity to real-world causal graph	Interventional datasets; SCM extraction; do-calculus probes (Xia et al., 2021)
Metacognitive (E2)	Calibrated self-uncertainty	Selective prediction; verifier LLMs; abstention metrics (Kadavath et al., 2022)
Fast/Slow Arbitration	Dynamic mode switching	Latency studies; scratchpad audits; reasoning-depth benchmarks
Phenomenal (E3)	Qualia / “something-it-is-like”	<i>Open research problem</i> —no agreed operationalisation (Nagel, 1974)

Table 3: Evidence stack incorporating fast/slow thinking.

6 Necessity & Sufficiency Summary

Property	T_1	T_U	T_U^*	Fast/Slow	Still Missing
Correct logical inference	✓	✓	✓	✓	—
Truth-robust representation	—	✓	✓	✓	—
Causal structural fidelity	—	—	✓	—	Interventional benchmarks
Metacognitive awareness	—	—	✓	✓	Self-calibration metrics
Phenomenal experience	—	—	-	?	?

Table 4: Properties satisfied by each benchmark level.

7 Limitations and Scope

This paper’s benchmarks are definitional, not exhaustive. We explicitly acknowledge several critical dimensions that fall outside the scope of this framework:

- **Learning and Development.** This framework defines the *what* (the criteria for understanding) but does not prescribe the *how* (the mechanisms for achieving it). It omits a discussion of developmental learning theories (Gopnik, 2009) or the specific architectural and curriculum requirements needed to progress from \mathbf{T}_1 to \mathbf{T}_U^* .
- **Social and Multi-Agent Understanding.** The benchmarks evaluate an agent in cognitive isolation. They do not test for social reasoning, such as Theory of Mind (Premack and Woodruff, 1978) (modeling the beliefs and intentions of others), collaborative problem-solving, or the intersubjective validation of knowledge.
- **Embodiment and Grounded Experience.** The framework is abstract and does not engage with the symbol grounding problem (Harnad, 1990). Whether the deepest levels of causal and conceptual understanding (E1) are possible without embodied interaction in a physical environment remains an open question not addressed here.
- **Methodological Feasibility.** While diagnostic tests are proposed, the paper does not fully resolve the practical hurdles of their implementation. Testing across all logically equivalent representations or reliably interpreting the internal causal state of a large neural network are major, unsolved research challenges in their own right (Lipton, 2018).

8 Definitions and Terminology Clarifications

The discourse surrounding AI capabilities, particularly regarding "reasoning" and "understanding," is often complicated by a lack of precise definitions for these deeply loaded terms. To ensure clarity and avoid semantic ambiguities, this section explicitly delineates the specific interpretations of key terminology within the context of the Bhatt Conjectures. These definitions are operational, serving as the basis for the benchmarks proposed herein, and should not be conflated with broader philosophical or colloquial meanings unless explicitly stated.

8.1 On the Nature of "Tautology"

Definition 8.1 (Tautology in this Context). Within this paper’s framework, a "tautology" refers to a ****definition constructed analytically to state necessary and sufficient conditions for a specific AI capability (Reasoning, Understanding, or Deep Understanding) *as defined by the framework itself****. Its truth is established by its definitional construction within this formal system.

Clarification: These benchmarks are termed "tautological" because their validity ***within the confines of this paper’s framework*** is inherent in their definitional construction. They rigorously delineate what we mean by "reasoning" and "understanding" for the purpose of evaluating AI systems against these precise criteria. This approach circumvents debates about the empirical "truth" of these definitions by asserting them as axiomatic for the evaluation model presented, rather than as empirically derived observations about intelligence in general.

8.2 Defining "Reasoning" (T_1)

The term "reasoning" is frequently used to imply complex cognitive processes, logical inference, and abstract problem-solving. Our T_1 definition narrows this to empirically testable behavioral patterns:

Definition 8.2 (T_1 - Operational Reasoning). A system **reasons** with respect to a problem class C **if and only if**:

- (R1) it produces a correct solution from *any* logically equivalent representation of the formally specified premises, and
- (R2) its success probability remains high when the surface form of the problem is *outside* its training distribution (no isomorphic instance or step-by-step demonstration seen in training or finetuning data).

Clarification: For an AI system to satisfy T_1 , it must demonstrate both ****representation invariance**** and ****zero-shot robustness**** in problem-solving. This implies an abstraction of the underlying logical structure that enables consistent application regardless of problem presentation (e.g., natural language, formal logic, diagrams, noisy encoding) and successful generalization to entirely novel instances. It explicitly excludes cases where performance relies solely on memorization or shallow pattern completion over familiar training data.

8.3 Defining "Understanding" (T_U)

"Understanding" colloquially suggests comprehension and the grasping of meaning beyond surface-level information. Our T_U definition builds upon T_1 by adding a robust representational requirement:

Definition 8.3 (T_U - Operational Understanding). A system **understands** a knowledge domain D **if and only if** for every proposition φ about D :

- (U1) it maps *any* truth-preserving representation $\rho(\varphi)$ to an internal state $I(\varphi)$ that recovers the truth value of φ , and
- (U2) conditions U1 hold when $\rho(\varphi)$ stems from a distribution statistically independent of the training data, and when queried with *unseen* operations derivable solely from $I(\varphi)$ (e.g., counterfactuals, entailments).

Clarification: T_U extends beyond T_1 by requiring ****truth-robust representation****. An AI's 'understanding' of a domain necessitates the creation of an internal representation that faithfully preserves truth values across varied input modalities (e.g., text, image schemas, speech, formal encodings) and distributions. Critically, this internal representation must be sufficiently rich to support ****counterfactual competence**** and ****novel inference**** (e.g., deriving previously unseen logical entailments or contradictions), even when the queries themselves are statistically independent of the training data. This criterion implies a robust, abstract knowledge representation, as opposed to mere recall of facts or pattern extrapolation.

8.4 Defining "Deep Understanding" (T_U^*)

"Deep understanding" often carries connotations of profound comprehension, intuition, and sometimes even consciousness. T_U^* addresses specific facets of these intuitions while acknowledging inherent limits.

Definition 8.4 (T_U^* - Operational Deep Understanding). A system **deeply understands** a domain D **if and only if** it satisfies T_U *and* the following premises:

- (E1) **Causal Structural Fidelity** — Its internal representation mirrors the *causal* graph of D , enabling do-calculus interventions (Pearl, 2009).
- (E2) **Metacognitive Self-Awareness** — It evaluates and calibrates the reliability of its own inferences, signalling uncertainty or ignorance (Shum et al., 2023).
- (E3) **Phenomenal Awareness** — Inference is accompanied by subjective experience (qualia). Untestable, this marks the boundary between functional mastery

and conscious sapience (Chalmers, 1995).

Clarification: \mathbf{T}_U^* captures capabilities beyond purely behavioral understanding. It mandates **causal structural fidelity**, meaning the system’s internal knowledge representation must accurately reflect the true causal relationships within a domain, enabling it to correctly predict the outcomes of interventions (e.g., "what if I *did* X?"). It further requires **metacognitive self-awareness**, where the system can assess and self-calibrate the reliability of its own inferences, expressing uncertainty or ignorance in a principled manner. The inclusion of **phenomenal awareness** (subjective, qualitative experience) is explicitly designated as an untestable, philosophical boundary condition within \mathbf{T}_U^* . This highlights the ultimate conceptual limit of 'deep understanding' without asserting its current achievability or empirical verifiability in artificial systems.

8.5 Clarifying "Internal State"

The concept of an "internal state" in AI can be vague, referring to anything from neuron activations to symbolic representations.

Definition 8.5 (Internal State $I(\varphi)$). An **internal state** $I(\varphi)$ refers to the **specific, actionable computational representation** (e.g., a learned embedding, a pattern of activations, a symbolic data structure) generated by the AI system upon processing a proposition φ .

Clarification: The existence of an internal state $I(\varphi)$ is inferred by its functional properties: its capacity to consistently and robustly "recover the truth value" of φ across varied representations and to support novel, derived operations (as per \mathbf{T}_U). This definition emphasizes functional computational representation and does not inherently imply human-like subjective experience or an easily interpretable symbolic form, though such forms might be a consequence of achieving the specified capabilities.

9 Conclusion

In my mental model, \mathbf{T}_1 and \mathbf{T}_U provide *necessary but not sufficient* criteria for reasoning and understanding, respectively. The extended \mathbf{T}_U^* adds causal and metacognitive layers, approaching—but not crossing—the phenomenal boundary. These tautologies offer somewhat of a rigorous framework for moving beyond “Does it reason?” to “How deeply does it understand?”—a question whose answer will shape the trajectory of artificial general intelligence.

10 Acknowledgments

The author thanks colleagues in the AI research community for stimulating discussions on the nature of machine reasoning and understanding. Special appreciation goes to those who provided feedback on earlier drafts of these conjectures, helping to refine the definitions and clarify the conceptual boundaries. The author is grateful for the intellectual environment that encourages rigorous examination of foundational questions in artificial intelligence.

Special thanks to Jim Schwoebel for his feedback.

References

- Bernard J Baars. *A Cognitive Theory of Consciousness*. Cambridge University Press, 1988.
- Yoshua Bengio. The consciousness prior. *arXiv preprint arXiv:1709.08568*, 2017.
- David J Chalmers. Facing up to the problem of consciousness. *Journal of Consciousness Studies*, 2(3):200–219, 1995.
- François Chollet. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019.
- John H Flavell. Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist*, 34(10):906–911, 1979.
- Alison Gopnik. The philosophical baby: What children’s minds tell us about truth, love, and the meaning of life. Farrar, Straus and Giroux, 2009.
- Stevan Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346, 1990.
- Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. *Advances in Neural Information Processing Systems*, 34: 2609–2620, 2021.
- Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, and others. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.
- Daniel Kahneman. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, 2011.
- Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker,

- Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let’s verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.
- Zachary C Lipton. The mythos of model interpretability. *Queue*, 16(3):31–57, 2018.
- Gary Marcus. The next decade in AI: Four steps towards robust artificial intelligence. *arXiv preprint arXiv:2302.03199*, 2023.
- Thomas Nagel. What is it like to be a bat? *The Philosophical Review*, 83(4):435–450, 1974.
- Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, and others. Show your work: Scratchpads for intermediate computation with language models. *arXiv preprint arXiv:2112.00114*, 2021.
- Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2nd edition, 2009.
- David Premack and Guy Woodruff. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4):515–526, 1978.
- A. Shojaei, L. Chen, and Z. Wang. Apple’s LRM system performance on Tower of Hanoi tasks. *Proceedings of the Conference on AI Reasoning*, pages 142–157, 2025.
- Karen Shum, Shizhe Diao, and Tong Zhang. Metacognitive prompting improves understanding in large language models. *arXiv preprint arXiv:2308.05342*, 2023.
- Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, and Jason Wei. Challenging BIG-Bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*, 2023.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.
- Lilian Xia, Katherine M Collins, Ishita Dasgupta, Samuel J Gershman, Brenden M Lake, and Joshua B Tenenbaum. Causal reasoning from meta-reinforcement learning. *arXiv preprint arXiv:2111.00084*, 2021.