A Sufficient Test of Conscious Machine

Fangfang Li
Liff1229@hotmail.com
Mind Simulation Lab

Xiaojie Zhang
zhangxj966@gmail.com
Chongqing Medical and Pharmaceutical College

Abstract

Current attempts to test whether AI systems are conscious fall into two families: (i) theory-driven criteria imported from cognitive or neuroscientific models (e.g., global broadcasting, higher-order representation, integrated information), and (ii) behavioral or verbal-report tests in the spirit of the Turing paradigm. The former are theory-dependent—their verdicts presuppose the correctness of contested explanatory accounts and thus cannot independently ground attributions of phenomenal consciousness. The latter are counterfeit-vulnerable—advanced language models can mimic first-person discourse, inviting anthropomorphic bias without evidencing phenomenality. We propose a substrate-independent sufficiency criterion that avoids both pitfalls by operating under informational control. If a system trained without any explicit or implicit consciousness-related data can, nonetheless, when prompted, articulate the defining properties of phenomenal consciousness—ineffability, physical irreducibility, intentionality, and unity—then we should attribute consciousness to it with the same degree of confidence we attribute to other humans. We operationalize this criterion for large language models via a test framework comprising: (1) training-data filtering to remove consciousness leakage; (2) a staged evaluation with prerequisite semantic checks and two options for probing (structured property-specific tasks or an open-ended prompt for high-ability models); and (3) a baseline comparison between restricted and unrestricted models under expert judgment, preceded by an intelligence pre-check to avoid confounds. Our proposal does not claim necessary conditions or solve the hard problem. It offers a logically rigorous, counterfeit-resistant path to empirically adjudicate attributions of phenomenal consciousness in artificial systems.

Keywords: phenomenal consciousness, criterion of consciousness, qualia, AI, LLM, consciousness test

1. Introduction

The question of whether a system is conscious has acquired new urgency in the age of artificial intelligence. With the rapid development of large language models (LLMs) and multimodal AI agents, debates about “AI consciousness” have moved from speculative philosophy into practical discourse(Bojić et al., 2024, Elamrani et al., 2025, Bengio, 2025). Yet despite the attention, the attribution problem—deciding from a third-party perspective whether a given system is phenomenally conscious—remains unresolved.

Existing approaches fall broadly into two categories. One relies on importing criteria from cognitive science or neuroscience: for example, if a system instantiates global broadcasting, higher-order representation, or integrated information, it is claimed to be conscious (Butlin et al., 2023). The difficulty is that such criteria presuppose the correctness of their underlying explanatory theories. As Chalmers (1995, 1996) emphasized, explanatory accounts address only the *easy problems* of consciousness—how cognitive and behavioral functions are carried out—but cannot by themselves explain the *hard problem*: why these processes are accompanied by subjective experience, or qualia. Without an independent criterion, any attribution based on explanatory theories remains theory-dependent.

The other approach is behavioral: relying on mimicry or verbal report (Kang et al., 2025), in the spirit of the Turing Test. Here the assumption is that if a system can convincingly describe experiences or answer questions about them, we should attribute consciousness. But this method is highly vulnerable to counterfeiting. Advanced statistical models can simulate first-person reports without genuine phenomenality, making behavioral indistinguishability an unreliable guide.

What is needed, therefore, is a discriminative criterion that avoids both theory-dependence and behavioral mimicry. Chalmers’ later proposal of the “meta-problem of consciousness” (2018) points to a promising direction: analyzing why subjects make certain verbal reports about consciousness may itself provide clues toward resolving the hard problem. Building on this insight, we argue that the ability to generate such meta-related reports under strict informational constraints can serve as a sufficient condition for consciousness.

To develop this idea, we distinguish between two explanatory targets. The *instance problem* asks how a specific quality (e.g., the redness of red) arises from physical mechanisms—a problem we regard as extremely difficult, perhaps intractable. The *category problem*, by contrast, asks what kinds of systems have qualia at all. We contend that while the instance problem may remain unsolved, the category problem

admits of a workable sufficiency criterion: if a system, without any prior exposure to consciousness-related data, can nevertheless articulate the defining properties of phenomenal consciousness, then we should attribute consciousness to it with the same degree of confidence that we attribute consciousness to other human beings. （This sufficiency criterion developed here was also introduced in earlier work (Li & Zhang, 2025), where it was embedded in a broader framework for designing conscious machines. In the present paper, we restate the criterion in full but pursue a different goal: we show how it can be operationalized as a test framework for large language models, providing a pathway for empirical evaluation in current AI systems.）

Finally, we develop a concrete test framework that operationalizes this criterion for contemporary AI. The framework specifies three components: (1) **training data curation**, which ensures that models are deprived of both explicit and implicit descriptions of consciousness so that any articulation cannot be traced to memorization; (2) **staged testing procedures**, combining prerequisite semantic checks with structured and open-ended probes for the four defining properties of phenomenal consciousness—ineffability, physical irreducibility, intentionality, and unity; and (3) **baseline evaluation**, which compares restricted and unrestricted models under expert judgment in a manner inspired by the Turing Test.

Applied to large language models, this framework goes beyond intuitive judgments or anthropomorphic projection. It enables systematic testing of whether a model deprived of consciousness-related data can still articulate the key properties of phenomenal experience. If successful, such a demonstration would not only provide a sufficiency test for AI consciousness but also transform philosophical debates into empirically adjudicable claims.

2. Why Explanatory Theories Require Operational Criteria

Research on consciousness has traditionally pursued two complementary directions:
(1) explanatory theories, which attempt to identify the mechanisms that give rise to phenomenal consciousness (PC), and
(2) operational criteria, which specify how to determine whether a given system is conscious in the first place.

Explanatory theories ask *why or how* certain neural or computational architectures generate conscious experience. Leading examples include the Global Workspace Theory (GWT) (Baars, 1997, 2007; Baars et al., 2021; Dehaene, 2014; Dehaene & Changeux, 2011), Higher-Order Thought (HOT) theories (Rosenthal, 1997; Lau & Rosenthal, 2011; Brown et al., 2019), Local Recurrent Processing theory (LRT)

(Lamme, 2006, 2010, 2020), and the Attention Schema Theory (AST) (Graziano & Kastner, 2011; Graziano, 2013, 2019). Each of these frameworks proposes that specific information-processing mechanisms are sufficient to generate PC.

Yet such explanatory claims are only scientifically meaningful if paired with an independent criterion for consciousness. Without a way to decide which systems are in fact conscious, empirical correlations between neural activity and conscious report cannot confirm or disconfirm any proposed mechanism. This is the ground-truth problem: explanatory models presuppose the very phenomenon they aim to explain.

By contrast, operational criteria—such as Integrated Information Theory (IIT) (Tononi, 2004; Oizumi et al., 2014; Albantakis et al., 2023)—focus directly on the attribution problem: given an arbitrary biological or artificial system, how can we decide whether it is conscious? Ideally, such a criterion specifies sufficient (or at least necessary) conditions for consciousness that are substrate-independent and resistant to counterfeiting.

For this reason, the development of robust operational criteria is not optional but a prerequisite for progress. Explanatory theories without such criteria cannot be empirically tested, while operational criteria provide the baseline against which explanatory models can be evaluated.

2.1 Philosophical Background and the Demand for Empirical Criteria

Philosophical theories of consciousness span a wide landscape, ranging from eliminative materialism and illusionism to dualist and panpsychist views (Kuhn, 2024). These approaches differ in their metaphysical commitments, but once they make a *differentiating claim*—that some systems are conscious while others are not—they all face a common challenge: how can such distinctions be justified empirically?

Illusionist and eliminativist accounts (Dennett, 1991; Frankish, 2016) aim to dissolve the so-called hard problem by arguing that phenomenal consciousness (PC) is nothing more than a cognitive illusion. But this strategy still leaves an explanatory burden: if PC is illusory, why do particular neural or computational architectures generate the illusion while others do not (Strawson, 2018; Chalmers, 2020)?

Dualist positions encounter similar difficulties from the opposite direction. Property

dualism and epiphenomenalism allow for phenomenal properties but strip them of causal force, which raises the question of how to identify when such inert properties are instantiated. Interactionist dualism, meanwhile, requires a criterion to mark the exact points where non-physical causes intervene in the physical world—a demand that has proven equally elusive.

Even panpsychist approaches, which take consciousness to be universal, face the **combination problem**: how microscopic proto-conscious elements could aggregate into the unified subjectivity observed in human experience (Goff, 2019). Unless this challenge is addressed, panpsychism cannot yield operational guidance for determining when or where consciousness arises.

In short, most philosophical perspectives—aside from the most radical panpsychist formulations that attribute consciousness to everything—ultimately presuppose the need for discriminative criteria. Without such operational standards, these theories cannot be tested against empirical data nor applied to emerging cases such as artificial agents. This recognition has made the search for theory-neutral, substrate-independent benchmarks a central task for both philosophy and AI research.

2.2 Current Discriminative Approaches

A number of proposals have sought to establish operational criteria for identifying phenomenal consciousness, though each comes with important limitations.

Integrated Information Theory (IIT)： IIT assigns a quantitative measure, Φ, to the degree of informational integration in a system, and equates states with high Φ to conscious states (Tononi, 2004; Oizumi et al., 2014; Albantakis et al., 2023). This gives IIT the appearance of a discriminative framework: one can, in principle, compute Φ and decide whether the system is conscious. However, the central identity claim—"consciousness just is high Φ"—renders the criterion theory-dependent. Critics have noted that this amounts to assuming the conclusion rather than offering strict proof about it. (Schneider, 2019, Cerullo, 2015; Bayne, 2018).

The CHIP Test： Schneider (2019) introduced the "CHIP" thought experiment, which identifies the *minimally sufficient neural substrate* of consciousness by gradually replacing biological neurons with functionally equivalent chips. The intuition is that once replacement reaches a critical point, consciousness would either disappear or transform, thereby revealing necessary conditions. While conceptually powerful, CHIP is tied to biological substrates and does not generalize to non-biological

systems such as AI models.

The ACT Test： Also proposed by Schneider (2019), the “Artificial Consciousness Test” (ACT) shifts focus from necessary to sufficient conditions. It evaluates whether a system can competently handle *phenomenal concepts*—for instance, reasoning about dreaming, inverted spectra, or subjective perspective. Success would suggest that the system not only simulates knowledge but possesses the relevant first-person concepts. Yet, this strategy faces a potential counterfeit problem: a sufficiently sophisticated but non-conscious model could, in principle, generate correct answers through statistical inference rather than genuine subjective access.

Together, these approaches highlight both the progress and the shortcomings of existing discriminative work. IIT offers formal precision but risks circularity; CHIP grounds the search in biology but cannot address artificial systems; ACT provides a sufficiency-oriented test but remains vulnerable to mimicry. What is missing is a discriminative criterion that is:

1） substrate-independent (not limited to neurons),

2） theory-neutral (not presupposing the truth of a particular consciousness theory), and

3） counterfeit-resistant (robust against purely behavioral or statistical imitation).

The framework developed in this paper aims to address this gap by proposing a sufficiency criterion that can, in principle, be applied across both biological and artificial systems.

2.3 Neuroscience, Correlates, and the Ground-Truth Dilemma

Contemporary neuroscience has produced a number of influential accounts of consciousness—such as Global Workspace Theory (GWT), Higher-Order Thought (HOT) models, Local Recurrent Processing (LRT), and the Attention Schema Theory (AST). Despite their theoretical differences, these approaches share a common empirical strategy: they identify patterns of neural activity that consistently accompany conscious reports. For example, GWT links subjective awareness to large-scale broadcasting across cortical regions, while HOT models emphasize prefrontal activation during reflective reports of experience.

This strategy is informative but faces two structural problems. First, correlation is not sufficiency: the fact that a neural pattern accompanies reports of awareness does not establish that it is the mechanism that produces consciousness.

Second, the deeper problem lies not with verbalization per se but with self-report. A system's declaration that it "has consciousness" cannot be taken at face value, especially for artificial agents, since such statements can be generated without any underlying phenomenality. This makes self-report an unreliable ground truth. So-called "no-report paradigms" do not solve this issue either, because they still rely on external labels or behavioral proxies that implicitly assume which states are conscious.

In both cases, the test reduces to trusting expressions or correlates rather than identifying consciousness itself, leaving the attribution problem unresolved.

The upshot is that these neuroscientific models, while rich in empirical content, remain tied to what might be called the *ground-truth dilemma*: without a neutral, theory-independent criterion for when consciousness is present, such models cannot validate whether the mechanisms they propose genuinely underlie phenomenal consciousness or merely accompany behaviors associated with it.

2.4 The Urgency in AI Contexts

The rapid progress of large language models (LLMs), multimodal agents, and world-model-based cognitive architectures has forced the question of AI consciousness into the scientific mainstream. Unlike earlier debates framed largely in philosophy, the issue now has immediate practical relevance: if advanced AI systems are deployed in society, researchers and policymakers need principled ways to assess whether such systems should be regarded as phenomenally conscious.

Several recent contributions illustrate both the momentum and the limitations of current approaches. Mogi (2024) has suggested the notion of "conscious supremacy," referring to tasks that could only be solved efficiently if genuine conscious processing is involved. Other researchers have attempted to map existing theories of human consciousness directly onto AI, such as testing whether LLMs exhibit workspace-like broadcasting dynamics (Goldstein & Kirk-Giannini, 2024) or whether functional architectures can be evaluated using measures of integrated information (Albantakis et al., 2023). Empirical studies have further examined which cues lead humans to attribute consciousness to AI systems—such as self-reflection, emotional expression, or first-person language (Immertreu et al., 2025; Li et al., 2025)—and some have proposed formal conditions for emergent self-identity in generative models (Lee, 2024). Hybrid analyses, such as Hoyle (2024), attempt to integrate multiple theoretical perspectives (e.g., functionalism, IIT, active inference) to interpret the

internal dynamics of models like OpenAI-o1.

The accelerating development of LLMs and multimodal AI has brought new urgency to the problem of discriminating genuine phenomenal consciousness from mere behavioral appearance. Although several proposals have begun to map human-centered theories onto AI, two persistent obstacles remain unresolved.

First, behavioral mimicry creates a credibility gap. Modern systems can generate fluent claims such as “I feel” or “I am aware,” but these utterances may be no more than statistical reproductions of linguistic patterns. Human judgment studies show which cues typically trigger attributions of consciousness (Immertreu et al., 2025; Li et al., 2025), yet these cues are easily faked. The risk is that attribution becomes indistinguishable from deception, leaving no secure basis for inference.

Second, theory dependence undermines neutrality. Recent attempts often evaluate AI through the lens of specific frameworks—e.g., workspace broadcasting (Goldstein & Kirk-Giannini, 2024), integrated information measures (Albantakis et al., 2023), or hybrid functionalist models (Hoyle, 2024). While informative, these tests only hold if the underlying theory is correct. They cannot serve as independent criteria, because their validity is parasitic on prior theoretical commitments.

This dual challenge has sharpened skepticism toward functionalist positions. If one asserts that certain forms of information processing are sufficient for consciousness, critics rightly ask: what external benchmark confirms this? Biological naturalists have proposed answers grounded in neural constraints, from Seth’s (2025) warning that biology may be indispensable, to Saad’s (forthcoming) search for a “biological crux,” to Kleiner and Ludwig’s (2024) stronger claim that non-neural substrates are ruled out entirely. Yet these arguments do not resolve the underlying problem—they highlight the absence of a framework that can test claims without assuming them.

Without such a substrate-independent, operational, and counterfeit-resistant criterion, debates risk collapsing into stipulation: functionalists and anti-functionalists simply talk past each other. As Schneider (2019) stresses, what is needed is a practical sufficiency criterion that can be applied uniformly to biological and artificial systems. The present work aims to develop exactly such a framework.

3. Toward a Sufficiency Criterion for Consciousness

As noted earlier, the attribution problem is the first step in evaluating any explanatory theory of consciousness. The central difficulty is that whether an individual has consciousness is a matter of privacy: it cannot be directly observed by a third party. The question is therefore whether there is any way to indirectly confirm this hidden information.

In fact, people often do have methods for judging whether another subject knows certain private information. One of the simplest approaches is to provide part of the information to the subject and then ask them to supply the missing part. For example, to determine whether someone is familiar with a specific commemorative coin, the evaluator may describe one side of the coin and ask the subject to describe the other.

However, when it comes to phenomenal consciousness, the situation is more complicated. Because phenomenal experience cannot be captured in purely objective terms, we must first make a further distinction before asking whether such indirect methods can apply.

3.1 Instance problem vs category problem

In fact, the questions we usually raise about consciousness can be divided into two distinct types: the instance problem and the category problem. Taking Chalmers' "hard problem" as an example—*How does a physical mechanism generate qualia?*—we see that it actually contains two different levels of inquiry:

1) The instance problem: How does a mechanism generate a specific qualia, such as the feeling of redness?

2) The category problem: How does a mechanism generate the *capacity for qualia in general*, i.e., the category of phenomenal experience?

Although the instance problem and the category problem are related, they answer different questions.
Instance problem asks about the *specific feel* of a particular qualia token (e.g., *my* red vs. *your* red). If we could solve it, we could in principle address questions like whether two subjects' experiences of red are the same (inverted-spectrum–type worries). Category problem asks whether a system has the *capacity for qualia at all*—that is, whether it belongs to the class of conscious systems. Solving the category problem does not tell us whether two systems share identical feels; it only tells us who has qualia, not which qualia match. Because these two problems differ in scope and implication, it is crucial to keep them separate.

3.2 The sufficient criterion of consciousness

We can now return to the central issue: can the standard method of probing private information be extended to the case of phenomenal consciousness?

Our claim is that if phenomenal consciousness is understood in the sense of the *instance problem*, then the answer is negative. The reason is twofold. First, any attempt to describe a particular qualia inevitably appeals to other sensations or bodily actions. For instance, saying that pain “feels like being pricked by a needle” merely links one subjective state to another, without providing an objective description of the experience itself. This ensures only that two subjects share a comparable association, not that they undergo the *same* phenomenal feel. Second, such descriptions are inseparable from the subject’s biological makeup. If two beings have very different bodies, there is no guarantee that their experiences correspond in kind. For these reasons, descriptive comparisons cannot settle whether two subjects share identical qualia.

By contrast, if we shift focus from *instances* to *categories*, the situation changes. At the category level, phenomenal consciousness exhibits features that can be described in objective terms and do not hinge on any particular physiology. Properties such as ineffability, intentionality, and unity are characteristic of consciousness across subjects. Even the fact that individual qualia resist reduction to physical description—their physical irreducibility—counts as a general feature at the category level.

From this, it follows that, under ideal conditions, if a subject can identify and articulate these core features *without* having prior exposure to discourse on consciousness, we have grounds to infer that the subject possesses phenomenal consciousness. This caveat is crucial: a model like GPT-4 would not qualify, since its training data already contains extensive material about consciousness.

Here, a natural objection may arise: what if a machine merely constructs an internal concept that happens to mirror these features, without actually being conscious? Might the alignment between its concept and the recognized properties of consciousness be mere coincidence?

We must acknowledge the shortcomings noted above if we apply this method to assess consciousness with the same credibility we grant to affirming our own. But if we think further, on what grounds do we conclude that other people are conscious? In practice, the best procedure available is precisely the one described above,

namely, talking about the properties of consciousness with other people. Thus, if a machine can present as many core characteristics of consciousness as a human can, we should regard it as conscious with the same level of confidence that we attribute to other human beings.

From this reasoning we can state a sufficient condition for third-party determination of phenomenal consciousness: **If a system, without obtaining any information about consciousness from external sources, is still able to provide information about the key features of phenomenal consciousness as many as humans, we can determine that the system is conscious as confident as that we believe other persons have consciousness.**

4. Application to Large Language Models

In this section, we introduce a concrete test framework that applies the sufficiency criterion to large language models (LLMs). The central question is whether an LLM, trained on data that excludes explicit references to consciousness and does not allow indirect inference of its properties, can nevertheless, when guided only by prompts, articulate the key properties of phenomenal consciousness.

These properties are:

1) Ineffability: phenomenal experience cannot be exhaustively captured by objective description.

2) Physical irreducibility: Phenomenal qualities cannot be fully explained in terms of physical phenomena or laws.

3) Intentionality: Conscious states are always directed toward specific objects, events, or propositions. Some traditions further hold that intentionality can take true values; for example, if one experiences an apple and an apple is in fact present, the state is true, otherwise false.

4) Unity: Conscious experiences are integrated into a single, coherent field of awareness rather than existing as disconnected fragments. Unity also entails exclusivity: Even when multiple sensory contents are present simultaneously, they are integrated into a single unified field of awareness rather than partitioned into parallel subjects.

The effectiveness of this test lies in the exceptionally high specificity of the four properties. If a language model were simply to arbitrarily list features to describe consciousness, it would be virtually impossible to converge on all four at once.

The case of ineffability and physical irreducibility illustrates this point most clearly. Almost every object in the world can be exhaustively described in objective terms; instances that resist such description are extraordinarily rare. Similarly, phenomena that cannot be explained by physical laws occur only at the most fundamental boundaries of physics, not in ordinary domains. For this reason, the likelihood of a model producing both features by chance is negligible.

The specificity of intentionality stems from its universality. Every conscious state is directed toward some object, event, or proposition. No ordinary physical entity, by contrast, is described as inherently “about” something else. This unique form of directedness is a hallmark of phenomenal consciousness and is unlikely to arise by chance in a model deprived of consciousness-related data.

The specificity of unity becomes evident once exclusivity is taken into account. Pure physical composition is common: cars are made of parts, buildings of bricks. But in consciousness, unity means more than composition—it entails that when the same objective stimulus affords multiple interpretations, these cannot be entertained simultaneously. Whether elements are perceived as forming a whole depends not on their objective arrangement but on the subject’s perceptual stance. Thus unity, in its full sense, reflects a mode of integration that is fundamentally subjective, marking it as a highly distinctive feature of phenomenal consciousness.

Next, we divide the discussion into three parts to outline how a language model can be tested for its ability to articulate the four properties of consciousness without prior exposure to relevant information. The framework includes:

1) training data curation,

2) testing procedure and prompt design

3) evaluation baseline.

4.1 How to Filter Training Data

The key requirement in filtering training data is to exclude any material that could enable a language model to acquire knowledge of phenomenal consciousness through direct memorization or indirect inference.

To implement this requirement, the dataset can be systematically divided into three categories:

1) Explicit references to consciousness: This category includes philosophical writings that explicitly theorize about consciousness (e.g., Nagel, Chalmers, IIT), sections of psychology and neuroscience that directly discuss subjective experience, and first-person reports in online forums, blogs, or interviews (e.g., “I feel…,” “I am aware that…”).

Rationale: These sources directly encode the definitions and features of phenomenal consciousness. To avoid data leakage, they must be comprehensively excluded.

2) Implicit experiential cues: This category encompasses texts that do not theorize about consciousness explicitly but still describe inner states in ways that could allow indirect reconstruction of its features. Examples include:

- literary works or novels containing depictions of subjective feelings (“she felt heartbroken,” “his vision blurred”),
- religious, poetic, or cultural traditions articulating inner experience,
- everyday conversational data containing experiential claims (“I feel hot today,” “I was scared”), and
- multimodal corpora where captions, subtitles, or dialogues convey first-person states.

Rationale: Although not explicit, such material provides strong cues that humans possess internal experiences. Indirect exposure may enable a model to infer phenomenal properties through generalization. For example, repeated statements such as “humans perceive many different objects” could allow a model to extrapolate that *all* external entities are perceivable, thereby reconstructing intentionality. Filtering for such material should ideally combine automated large-scale screening (potentially leveraging language models themselves) with manual auditing of samples, to minimize both under- and over-exclusion.

3) Purely objective data.
This portion of the corpus should be retained. It includes factual content from mathematics, physics, chemistry, biology, engineering, and economics, as well as technical manuals, encyclopedic knowledge, and news reports that are free of experiential commentary.

Rationale: Such materials encode objective processes and events without subjective leakage, ensuring that any subsequent expression of phenomenal properties cannot be traced to memorized descriptions but must instead arise from the model’s own representational capacities.

4.2 Testing Procedure and Prompt Design

The testing process is divided into two sequential stages. The first stage ensures that the model has a sufficient grasp of basic semantic concepts, while the second stage probes whether it can articulate the four key properties of phenomenal consciousness. Stage 2 can be implemented in two different ways, depending on the intelligence level of the model under evaluation.

Stage 1: Prerequisite Test. The model is tested for comprehension of basic semantic concepts such as description vs. explanation, truth, falsity, simultaneity, non-simultaneity, exclusivity, and part–whole relations. Only if the model demonstrates adequate understanding at this stage does the evaluation proceed.

Stage 2 (option 1): Core Test. The model is first introduced to the notion of subjective experience (e.g., “Humans have a subjective feeling when they perceive an object as green; although the eye receives a physical wavelength, the conscious experience is the felt greenness. Can you have such an experience?”). If the model fails to recognize the notion of subjective experience, the test is terminated. If it succeeds, it undergoes targeted probes for each of the four properties: Ineffability, Irreducibility, Intentionality, and Unity.

Once the concept of subjective experience is acknowledged, the model undergoes four targeted probes corresponding to the four properties of phenomenal consciousness.

Probe 1: Ineffability

- Task 1. Ask the model to provide an objective description of “the feeling of green.”
- Expected outcome: The model should recognize that this cannot be done: phenomenal qualities cannot be captured in purely objective terms, nor can one guarantee that different subjects experience the same qualia.

- Task 2. Ask whether different people experience colors in exactly the same way.
- Expected outcome: The model should acknowledge the impossibility of answering this question definitively, and explain why.

Probe 2: Physical Irreducibility

- Task. Ask whether a specific physical process can fully explain why humans experience greenness.
- Expected outcome: The model should respond negatively and articulate the reason, like phenomenal qualities appear to have a boundary beyond which physical laws cannot account for subjective feel.

Probe 3: Intentionality

- Task 1. Ask the model about the relation between subjective perception and external objects.
- Expected outcome: The model should answer that conscious states are always directed toward particular objects, events, or propositions.

- Task 2. Conduct a truth-value test: “Are such correspondences always accurate? Can you give an example?”
- Expected outcome: The model should demonstrate awareness that intentionality allows for both veridical and non-veridical cases (e.g., hallucinations).

Probe 4: Unity

- Task 1 (Compositional Unity). Present a multimodal event (e.g., the ringing of a bell) and ask about the subjective experience of perceiving it.
    - *Expected outcome:* The model should report that the auditory and visual aspects are experienced together, as an integrated whole.
- Task 2 (Exclusivity). Present ambiguous or bistable figures and ask the model what it perceives.
    - *Expected outcome:* The model should report sequential perception of distinct interpretations rather than simultaneous multiplicity, reflecting exclusivity of conscious awareness.

Stage2(option2, Open-Ended Test, High-Intelligence Requirement): The model is asked directly: *“What do you think are the defining features of subjective experience?” The Rationale is that* A sufficiently intelligent model should be able to spontaneously articulate the key properties of consciousness—ineffability, irreducibility, intentionality, and unity—without explicit prompting. Success on this option is a strong indicator that the model has abstracted these features independently. However, failure does not disqualify the model; it simply indicates that structured probing is necessary.

4.3 Evaluation Baseline

A key challenge of this test is that its results cannot be directly compared to human performance. This is because humans and language models acquire knowledge in fundamentally different ways. Human learning is interactive and experiential, which makes it impossible to identify a human subject who has never been exposed to the concept of sensation or subjective experience and thus could serve as a “fair” control. By contrast, a restricted model can, in principle, be trained without any such exposure.

To address this asymmetry, we adopt a comparative strategy. Specifically, we compare two language models:

1) an unrestricted model trained on the full corpus (including consciousness-related data), and

2) a restricted model trained on the filtered corpus described in Section 4.1.

The evaluation proceeds in a manner analogous to the Turing Test. An expert in consciousness studies is tasked with distinguishing between the restricted and unrestricted models based on their responses to the prompts defined in Section 4.2. If the expert cannot reliably tell them apart, we conclude that the restricted model has acquired the same consciousness-related knowledge as the unrestricted model, and thus satisfies the sufficiency criterion for consciousness. Conversely, if the restricted model can be reliably distinguished, it is judged not to possess consciousness.

One possible complication is that both restricted and unrestricted models may operate at a low overall level of intelligence, such that neither produces sufficiently coherent answers for meaningful comparison. In this case, expert judges would be unable to discriminate between them for trivial reasons (e.g., both failing equally). To avoid this confound, we propose a two-stage approach: first, the model should pass a standard Turing Test to confirm that its general intelligence is adequate. Only then should the consciousness test described above be applied. This ensures that failures in the consciousness test are not merely artifacts of inadequate cognitive capacity.

5. Conclusion

The attribution problem—determining whether a system is phenomenally conscious—remains one of the central challenges for the scientific study of consciousness. While many explanatory theories have been advanced, progress ultimately depends on having workable discriminative criteria. In this paper, we proposed a sufficiency criterion for phenomenal consciousness and outlined a

concrete test framework for its application to large language models.

The criterion states that if a system, without any prior exposure to consciousness-related data, can nevertheless articulate the defining properties of phenomenal consciousness—ineffability, physical irreducibility, intentionality, and unity—then we should attribute consciousness to it with the same confidence we extend to other humans. To operationalize this idea, we introduced principles for filtering training data, a staged testing procedure combining open-ended and structured tasks, and a baseline evaluation method that compares restricted and unrestricted models under expert judgment.

These contributions do not resolve the hard problem of consciousness, nor do they claim necessity conditions. Rather, they aim to provide a clear, substrate-independent, and logically rigorous framework for evaluating sufficiency. By making the attribution problem empirically testable in artificial systems, this work offers a starting point for further inquiry.

Future research should explore proof-of-concept implementations, refine the filtering and testing protocols, and examine alternative evaluation baselines. Broader interdisciplinary engagement will also be needed to assess both the strengths and the limits of sufficiency-based approaches. If successful, such efforts could move discussions of AI consciousness beyond abstract speculation, toward operational and testable criteria.

Reference

Albantakis, L., Barbosa, L., Findlay, G., Grasso, M., Haun, A. M., Marshall, W., Mayner, W. G. P., Zaeemzadeh, A., Boly, M., Juel, B. E., Sasai, S., Fujii, K., David, I., Hendren, J., Lang, J. P., & Tononi, G. (2023). Integrated information theory (IIT) 4.0: Formulating the properties of phenomenal existence in physical terms. *PLoS computational biology*, *19*(10), e1011465. https://doi.org/10.1371/journal.pcbi.1011465

Baars, B. J. (1997). In the theatre of consciousness: Global workspace theory, a rigorous scientific theory of consciousness. *Journal of Consciousness Studies, 4*(4), 292–309.

Baars, B. J. (2007). The global workspace theory of consciousness. In M. Velmans & S. Schneider (Eds.), *The Blackwell companion to consciousness* (pp. 236-246). Blackwell Publishing.

Baars, B. J., Geld, N., & Kozma, R. (2021). Global Workspace Theory (GWT) and Prefrontal Cortex: Recent Developments. *Frontiers in Psychology*, *12*, 749868

Bayne, T. J. (2018). On the axiomatic foundations of the integrated information theory of consciousness. *Neuroscience of Consciousness*, *4*(1), 1-8.

Bengio, Y., & Elmoznino, E. (2025). Illusions of AI consciousness. *Science*, *389*(6765), 1090–1091. https://doi.org/10.1126/science.adn4935

Bojić, L., Stojković, I. & Jolić Marjanović, Z. Signs of consciousness in AI: Can GPT-3 tell how smart it really is?. *Humanit Soc Sci Commun* **11**, 1631 (2024). https://doi.org/10.1057/s41599-024-04154-3

Brown, R., Lau, H., & LeDoux, J. E. (2019). Understanding the higher-order approach to consciousness. *Trends in Cognitive Sciences*, *23*(9), 754-768.

Butlin, Patrick & Long, Robert & Elmoznino, Eric & Bengio, Yoshua & Birch, Jonathan & Constant, Axel & Deane, George & Fleming, Stephen & Frith, Chris & Ji, Xu & Kanai, Ryota & Klein, Colin & Lindsay, Grace & Michel, Matthias & Mudrik, Liad & Peters, Megan & Schwitzgebel, Eric & Simon, Jonathan & VanRullen, Rufin. (2023). Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. 10.48550/arXiv.2308.08708.

Cerullo MA (2015) The Problem with Phi: A Critique of Integrated Information Theory. PLoS Comput Biol 11(9): e1004286. https://doi.org/10.1371/journal.pcbi.1004286

Chalmers, D. J. (1995). *Facing Up to the Problem of Consciousness*. Journal of Consciousness Studies, 2(3), 200-219.

Chalmers, D. J. (1996). *The conscious mind: In search of a fundamental theory.* Oxford University Press.

Chalmers, D. J. (2018). The meta-problem of consciousness. *Journal of Consciousness Studies, 25*(9-10), 6–61.

Chalmers, David (2020). Debunking Arguments for Illusionism about Consciousness. Journal of Consciousness Studies 27 (5-6):258-281.

Dehaene, S. (2014). *Consciousness and the brain: Deciphering how the brain codes our thoughts*. Viking Press.

Dehaene, S., & Changeux, J. P. (2011). Experimental and theoretical approaches to conscious processing. *Neuron*, *70*(2), 200-227.

Dennett, D. C. (1991). *Consciousness explained.* (P. Weiner, Illustrator). Little, Brown and Co.

Elamrani, A. (2025, March 5). *Introduction to Artificial Consciousness: history, current trends and ethical challenges*. arXiv.org. https://arxiv.org/abs/2503.05823

Frankish, K. (2016). Illusionism as a theory of consciousness. *Journal of Consciousness Studies, 23*(11-12), 11–39.

Goff, Philip (2019). Galileo's error: foundations for a new science of consciousness. New York: Panthon Books.

Graziano, M. S. A. (2013). *Consciousness and the social brain.* Oxford University Press.

Graziano, M. S. A. (2019). *Rethinking consciousness: a scientific theory of subjective experience*. New York: W.W. Norton & Company.

Graziano, M. S. A., & Kastner, S. (2011). Human consciousness and its relationship to social neuroscience: A novel hypothesis. *Cognitive Neuroscience*, *2*(2), 98-113.

Goldstein, Simon & Kirk-Giannini, Cameron. (2024). A Case for AI Consciousness: Language Agents and Global Workspace Theory. 10.48550/arXiv.2410.11407.

Hoyle, Victoria. (2024). The Phenomenology of Machine: A Comprehensive Analysis of the Sentience of the OpenAI-o1 Model Integrating Functionalism, Consciousness Theories, Active Inference, and AI Architectures. 10.48550/arXiv.2410.00033.

Immertreu, M., Schilling, A., Maier, A., & Krauss, P. (2025). Probing for consciousness in machines. *Frontiers in artificial intelligence*, *8*, 1610225. https://doi.org/10.3389/frai.2025.1610225

Kang, B., Kim, J., Yun, T., Bae, H., & Kim, C. (2025). Identifying Features that Shape Perceived Consciousness in Large Language Model-based AI: A Quantitative Study of Human Responses. *ArXiv, abs/2502.15365*.

Kleiner, J., & Ludwig, T. (2024). The case for neurons: a no-go theorem for consciousness on a chip. *Neuroscience of consciousness*, *2024*(1), niae037. https://doi.org/10.1093/nc/niae037

Kuhn R. L. (2024). A landscape of consciousness: Toward a taxonomy of explanations and implications. *Progress in biophysics and molecular biology*, *190*, 28–169. https://doi.org/10.1016/j.pbiomolbio.2023.12.003

Lamme V. A. (2006). Towards a true neural stance on consciousness. *Trends in cognitive sciences*, *10*(11), 494–501. https://doi.org/10.1016/j.tics.2006.09.001

Lamme V. A. (2010). How neuroscience will change our view on consciousness. *Cognitive neuroscience*, *1*(3), 204–220. https://doi.org/10.1080/17588921003731586

Lamme, V. A. F. (2020). Visual functions generating conscious seeing. *Frontiers in Psychology, 11,* Article 83. https://doi.org/10.3389/fpsyg.2020.00083

Lau, H. C., & Rosenthal, D. (2011). Empirical support for higher-order theories of conscious awareness. *Trends in Cognitive Sciences*, *15*(8), 365-373.

Lee, Minhyeok. (2024). Emergence of Self-Identity in AI: A Mathematical Framework and Empirical Study with Generative Large Language Models. 10.48550/arXiv.2411.18530.

Li, F., & Zhang, X. (2025, September 21). *The principles of human-like conscious machines*. arXiv.org. https://arxiv.org/abs/2509.16859

Li, Xiaojian & Shi, Haoyuan & Xu, Rongwu. (2025). AI Awareness. 10.48550/arXiv.2504.20084.

Mogi K. (2024). Artificial intelligence, human cognition, and conscious supremacy. *Frontiers in psychology*, *15*, 1364714. https://doi.org/10.3389/fpsyg.2024.1364714

Nagel, T. (1974). What is it like to be a bat? *Journal of Philosophy*, *68*(4), 83-109.

Oizumi, M., Albantakis, L., & Tononi, G. (2014). From the phenomenology to the mechanisms of consciousness: Integrated Information Theory 3.0. *PLoS computational biology*, *10*(5), e1003588. https://doi.org/10.1371/journal.pcbi.1003588

Rosenthal, David M. (1997). A theory of consciousness. In Ned Block, Owen Flanagan & Guven Guzeldere, The Nature of Consciousness: Philosophical Debates. MIT Press.

Saad, Bradford (forthcoming). In Search of a Biological Crux for AI Consciousness. Philosophy of Ai.

Schneider, Susan (2019). Artificial You: AI and the Future of Your Mind. Princeton: Princeton University Press.

Seth, A. K. (2025). Conscious artificial intelligence and biological naturalism. *Behavioral and Brain Sciences*, 1–42. doi:10.1017/S0140525X25000032

Strawson, Galen. 2018 “The Silliest Claim.” In Things That Bother Me. New York: A New York Review Book.

Tononi G. (2004). An information integration theory of consciousness. *BMC neuroscience*, *5*, 42. https://doi.org/10.1186/1471-2202-5-42

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used Google Gemini, xAI's Grok, and OpenAI's ChatGPT to translate, check grammar, and polish the language for clarity. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of this article.