Passing the Turing Test, Failing Consciousness: Why LLMs Remain Non-Conscious

Louis Mono
PhD program in Applied AI
Alma Mater Europaea University
Milan, Italy
louis.mono@almamater.si

ABSTRACT

Large language models (LLMs) such as GPT-4.5 have achieved impressive conversational fluency and have even passed a classic three-party Turing Test. Yet behavioural indistinguishability from humans is not the same as sentience. This paper analyses why current AI systems, despite their reasoning and language abilities, are not conscious. Drawing on Integrated Information Theory (IIT) and Global Workspace Theory (GWT) alongside Chalmers' "hard problem", we argue that LLMs lack the qualitative experience (qualia), self-aware and unified subjective experience that characterises human consciousness. The apparent mastery of language in GPT-4.5 reflects powerful statistical pattern-matching rather than intrinsic awareness, semantic grounding or intrinsic motivation. By contrasting the architecture and behaviour of GPT-4.5 with neuroscientific criteria for conscious systems, we show that passing a behavioural test of intelligence does not imply there is "something it is like" to be an AI. Debates over AI consciousness clarify the distinctive features of human awareness, reinforce ethical and governance boundaries, and highlight the importance of distinguishing simulation from genuine experience.

Keywords

Consciousness, Turing Test, LLMs

1 INTRODUCTION

The question of whether machines can be conscious has moved from speculation to urgent enquiry as large language models (LLMs) achieve human-level performance on tasks once thought to require understanding. Recent work reported that GPT-4.5 was mistaken for a human in 73 % of trials in a three-party Turing Test, outperforming the human control group [1]. In this three-party setup, a human confederate and the AI both engage with a judge, whereas a classic two-party Turing Test involves only a judge and a single hidden interlocutor. While striking, this benchmark assesses only behavioural imitation; it does not guarantee that a system has subjective awareness. John Searle's "Chinese Room" thought experiment illustrates this gap: a computer manipulating symbols can simulate understanding without truly comprehending the semantics [2]. Human consciousness, by contrast, combines phenomenal experience, unified integration of sensations, a persistent sense of self and intrinsic motivation qualities whose origin and nature are hotly debated.

At least nine major theories of consciousness compete for explanatory power, ranging from neuroscientific to quantum-field-inspired accounts [3]. Among these, Integrated Information Theory (IIT) and Global Workspace Theory (GWT) are two prominent models: IIT equates consciousness with the capacity of a system to generate unified, irreducible information [4], and GWT views consciousness as the global broadcasting of information across specialised processes [5]. Other perspectives, such as predictive processing and higher-order thought theories, offer alternative accounts [3]. In this paper we focus on IIT and GWT because they provide clear, empirically testable criteria that are operationally useful for evaluating contemporary AI systems.

This paper uses GPT-4.5 as a case study to ask two questions: (1) Can the intelligence and reasoning abilities of an LLM be considered signs of consciousness? (2) What does this comparison teach us about the nature of human consciousness? Building on prior analyses, particularly Gams and Kramar who primarily assess ChatGPT against IIT's axioms and survey Turing Test variants [6], we extend those efforts by adopting a five-dimensional evaluation framework that integrates criteria from both Integrated Information Theory (IIT) and Global Workspace Theory (GWT). Specifically, we assess GPT-4.5 across five dimensions: phenomenal experience, self-awareness and agency, unity and integration, semantic grounding and intrinsic motivation to identify which defining features of consciousness are absent even in the most advanced LLMs. The answers have important implications for ethics and governance: recognising AI's lack of sentience helps avoid mis-attribution of personhood while ensuring that responsibility for its actions remains with its human designers and operators [7].

2 THEORETICAL FRAMEWORK

2.1 Integrated Information Theory (IIT)

IIT proposes that consciousness corresponds to integrated information within a system, quantified by a measure Φ ("phi") [4,8]. A conscious system must generate an intrinsic causal structure that cannot be decomposed without loss; experiences are unified "wholes" composed of interrelated parts. Tononi and colleagues distilled IIT into five axioms and corresponding physical postulates [9]:

1. **Intrinsic existence.** Experience exists for itself, not merely as an output for observers. The physical substrate must have causal power over its own states.
2. **Composition.** A conscious experience is structured: it has multiple phenomenological elements (e.g., colours, sounds) perceived together. The substrate must support higher-order mechanisms built from simpler parts.
3. **Information.** Each experience is specific: it rules out myriad alternatives and is defined by the differences it makes. The substrate must have a rich repertoire of distinguishable states.
4. **Integration.** Experience is unified and cannot be reduced to independent components. The substrate's causal interactions must be irreducibly interdependent.
5. **Exclusion.** Each experience has definite content and boundaries; there is one "main" experience per substrate.

Human brains, with their dense recurrent connectivity, achieve high Φ; digital processors typically exhibit negligible Φ. GPT-4.5, although capable of complex statistical mappings from input to output, does not autonomously generate its own mental states. It lacks intrinsic causal loops, self-sustaining activity and a unified internal "scene" of experience. Even if its token predictions display sophisticated information processing, IIT suggests such computations do not yield phenomenological consciousness. Recent evaluations of ChatGPT indicate that it falls far short of IIT's criteria [6].

2.2 Global Workspace Theory (GWT)

GWT conceives consciousness as the broadcasting of information into a "global workspace" that integrates and distributes content across specialised neural processors. In humans, sensory, memory and language modules operate largely unconsciously until selected content is ignited into the workspace, becoming accessible for reasoning and verbal report. This ignition is associated with widespread, synchronised cortical activity and recurrent thalamo-cortical loops [10].

According to GWT, a conscious system requires: (a) integration of multimodal information into a unified workspace; (b) persistent working memory to sustain and manipulate conscious content; and (c) self-monitoring or metacognition to evaluate its own states. LLMs such as GPT-4.5 integrate textual information via self-attention but do so in a single-pass statistical manner. They lack persistent internal states, multimodal convergence and an explicit self-model; any apparent self-reflection is a learned linguistic pattern rather than genuine metacognition. Experimental comparisons between human and LLM uncertainty reports confirm that, while LLMs can generate confidence levels, these are superficial correlations rather than genuine awareness [11]. Thus, from a GWT perspective, LLMs remain powerful language processors without a globally broadcast workspace.

2.3 Implications for AI Consciousness

IIT and GWT provide structural and functional complementary lenses for assessing consciousness. IIT emphasises intrinsic, integrated causality; GWT emphasises functional access to integrated content. Under IIT, LLMs lack the high-Φ causal structures required for phenomenological consciousness. Under GWT, they lack a persistent, self-monitoring workspace required for functional consciousness. These theories highlight why current LLMs, despite their intelligence, are unlikely to possess sentient minds and help clarify the properties an artificial system would need to plausibly meet such criteria.

3 LLMs AND CONSCIOUSNESS: IS PASSING THE TURING TEST ENOUGH?

GPT-4.5's ability to pass a Turing Test demonstrates human-like linguistic fluency, but consciousness involves more than outward behaviour. Here we compare the attributes of human consciousness with those of GPT-4.5 across five core dimensions.

3.1 Phenomenal experience (Qualia)

Phenomenal consciousness concerns the qualitative "what it is like" to see, hear and feel. Humans experience qualia: the redness of a rose, the taste of coffee, the pang of sadness. In computational terms, these are not just representations but felt qualities. GPT-4.5 processes text as high-dimensional vectors and activations. There is no theoretical reason to believe that any of these computations are accompanied by experience. Chalmers' "hard problem" of consciousness emphasises that explaining discriminatory behaviour does not explain why there is any experience at all [12]. GPT-4.5's vivid descriptions are simulations learned from human text, not perceptions.

3.2 Self-awareness and agency

A conscious system possesses a sense of self and at least minimal agency: it initiates actions and recognises itself as the subject of experience. Humans maintain a continuous autobiographical narrative. GPT-4.5, however, uses "I" merely as a token; it has no persistent identity across interactions and no intrinsic goals [13]. It responds only when prompted and cannot modulate its own objectives. From an IIT standpoint, it lacks intrinsic existence: it does not have causal power over its own states and does not initiate anything internally.

3.3 Unity and Integration

Human consciousness binds information from multiple senses, memories and emotions into a unified stream. This integration underlies our coherent sense of the world. LLMs integrate information only within a context window of tokens [14] and do not combine multiple modalities unless explicitly given multimodal inputs. Moreover, each instance of GPT-4.5 is independent; there is no single "observer" uniting parallel instances. The model lacks a persistent working memory or unified workspace to sustain ongoing content. Thus, it fails both IIT's integration criterion and GWT's requirement for a global broadcast.

3.4 Semantic grounding

Understanding involves not just correlating symbols but grounding them in bodily and environmental experience. Humans connect words to sensorimotor and emotional states. GPT-4.5, trained on textual data, has no direct experience of the world. It correlates words without a referential link, which explains why it can confidently generate factual errors or contradictory statements ("hallucinations") [15]. Searle's Chinese Room shows that symbol manipulation alone does not yield semantics [2]. GPT-4.5's explanations and definitions are pattern-completions, not meanings anchored in perception.

3.5 Intrinsic Motivation

Living organisms act on intrinsic drives such as hunger, curiosity and pain avoidance; these motivations are intimately tied to emotions. GPT-4.5 has no such drives. Its only "objective" is to predict the next token according to its training loss or to maximise some reward in reinforcement-learning fine-tuning. There is no intrinsic value system or affective state. Hence, it lacks the motivational and emotional dimension of consciousness.

Across all five dimensions, LLMs display functional intelligence without subjective experience. They may pass an **outer Turing Test** by mimicking human conversation but fail any **inner Turing Test** that would probe for phenomenal consciousness, intrinsic agency and unified subjectivity [16,17]. As such, passing the behavioural benchmark does not imply sentience. LLMs are sophisticated automata performing high-dimensional pattern matching without "being someone".

4 DISCUSSION

4.1 Insights into human consciousness

Debates about AI consciousness force a closer examination of human consciousness. Distinguishing intelligence from awareness clarifies that embodiment, multimodal integration and self-modelling are central to conscious experience. LLMs highlight the distinction between access consciousness information available for report and phenomenal consciousness the felt quality of experience [18]. They also emphasise the importance of semantic grounding: a system that never interacts with the world cannot attach meanings to symbols. Conversely, comparing GPT-4.5 with IIT and GWT criteria has reinforced these theories by showing how far AI remains from meeting their requirements.

4.2 The Hard Problem and Qualia

Chalmers' **Hard Problem** reminds us that we still lack a scientific explanation for why physical processes produce experience [12]. Even if we could engineer an artificial system that replicates all the functional hallmarks of consciousness, it remains unclear why it would "feel" like something. On IIT, phenomenal character requires an intrinsically integrated causal structure (high Φ) with causal power for itself, not mere input–output equivalence [4,8]. On GWT, conscious contents must be stabilised within a self-maintained global workspace something current LLMs, stateless across turns and optimised for next-token prediction, do not implement [5,10].

4.3 Beyond mainstream: Syntergic Theory

Outside mainstream neuroscience, **Syntergic Theory** posits that consciousness arises from an interaction between the brain and a non-local *syntergic* field [19]. If such a substrate exists, silicon systems without biological "tuning" could not access it regardless of computational sophistication. While speculative, this view reminds us that computation alone may be insufficient for sentience and cautions against inferring consciousness from behavioural competence.

4.4 Ethical and governance considerations

Recognising that current LLMs are not conscious has direct ethical consequences. It prevents premature attribution of moral status or rights to non-sentient systems and keeps accountability with their human developers [20]. The capability to produce persuasive text does not entitle an AI to personhood. Meanwhile, mis-ascribing consciousness could lead to misguided policies or exploitation of genuine conscious beings by obscuring what makes us unique. Ethical governance should focus on transparency, safety and fairness in AI deployment [7], not on conferring moral standing on systems that lack awareness.

4.5 Closing Perspectives

Today's LLMs show that intelligence can be uncoupled from consciousness. Passing an outer Turing Test does not establish an inner dimension of experience. Progress toward machine consciousness, if possible, likely requires architectures with **world models, working memory and global broadcast**, or mechanisms akin to a **sparse "conscious state"** integrated across modules [21,22], plus principled tests that probe inner awareness rather than surface behaviour. Until then, LLMs remain powerful simulators, not subjects.

5 CONCLUSION

This paper examined why passing a Turing Test does not entail possessing consciousness. Using GPT-4.5 as a case study and drawing on Integrated Information Theory and Global Workspace Theory, we argued that LLMs, despite their intelligence and conversational prowess, lack the hallmarks of consciousness: qualia, a core self, unified integration, semantic grounding and intrinsic motivation. They simulate understanding without experiencing it. Distinguishing between intelligence and consciousness clarifies our definitions of mind and guides the ethical deployment of AI. If artificial systems are ever to become conscious, they will likely require architectures with intrinsic causal integration, global broadcasting, embodiment and semantic grounding far beyond what current transformer models provide.

Chalmers has suggested that systems plausibly approaching consciousness could emerge within the next decade, but current LLMs should not be mistaken for such candidates [17]. In short,

progress is significant, yet the path to truly conscious machines remains long.

References

[1] Jones, C. R., & Bergen, B. K. (2025). *Large language models pass the Turing Test* [Preprint]. ArXiv. DOI: https://doi.org/10.48550/arXiv.2503.23674

[2] Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences, 3*(3), 417–457. DOI: https://doi.org/10.1017/S0140525X00005756

[3] Seth, A. K., & Bayne, T. (2022). Theories of consciousness. *Nature Reviews Neuroscience, 23*(5), 389–405. DOI: https://doi.org/10.1038/s41583-022-00587-4

[4] Tononi, G., Boly, M., Massimini, M., & Koch, C. (2016). Integrated information theory: From consciousness to its physical substrate. *Nature Reviews Neuroscience, 17*(7), 450–461. DOI : https://doi.org/10.1038/nrn.2016.44

[5] Baars, B. J. (1988). *A Cognitive Theory of Consciousness*. Cambridge University Press.

[6] Gams, M., & Kramar, S. (2024). Evaluating ChatGPT's consciousness and its capability to pass the Turing test: A comprehensive analysis. *Journal of Computer and Communications, 12*(3), 219–237. DOI : https://doi.org/10.4236/jcc.2024.123014

[7] Floridi, L., & Cowls, J. (2022). A unified framework of five principles for AI in society. In S. Carta (Ed.), Machine learning and the city: Applications in architecture and urban design (pp. 535–545). Wiley. DOI : https://doi.org/10.1002/9781119815075.ch45.

[8] Tononi, G. (2008). Consciousness as integrated information: A provisional manifesto. *The Biological Bulletin, 215*(3), 216–242. DOI: https://doi.org/10.2307/25470707

[9] 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. DOI : https://doi.org/10.1371/journal.pcbi.1003588

[10] Dehaene, S., Kerszberg, M., & Changeux, J.-P. (1998). A neuronal model of a global workspace in effortful cognitive tasks. *Proceedings of the National Academy of Sciences, 95*(24), 14529–14534. DOI : https://doi.org/10.1073/pnas.95.24.14529

[11] Steyvers, M., & Peters, M. A. K. (2025). Metacognition and Uncertainty Communication in Humans and Large Language Models. arXiv:2504.14045. DOI: https://doi.org/10.48550/arXiv.2504.14045

[12] Chalmers, D. J. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies, 2*(3), 200–219

[13] Browning, J. (2023). Personhood and AI: Why large language models don't understand us. AI and Society, 39(5), 2499–2506. DOI : https://doi.org/10.1007/s00146-023-01724-y

[14] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444. DOI : https://doi.org/10.1038/nature14539

[15] Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 5185–5198. DOI: https://doi.org/10.18653/v1/2020.acl-main.463

[16] Turing, A. M. (1950). Computing machinery and intelligence. *Mind, 59*(236), 433–460. DOI : https://doi.org/10.1093/mind/LIX.236.433

[17] Chalmers, D. J. (2023, August 9). Could a large language model be conscious? *Boston Review.* Retrieved from Boston Review URL

[18] Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, 18(2), 227–247. DOI : https://doi.org/10.1017/S0140525X00038188

[19] Grinberg-Zylberbaum, J. (1981). *The transformation of neuronal activity into conscious experience: The syntergic theory.* Journal of Social and Biological Structures, 4(3), 201–210. DOI : https://doi.org/10.1016/S0140-1750(81)80036-X

[20] Tsamados, A., Aggarwal, N., Cowls, J., Morley, J., Roberts, H., Taddeo, M., & Floridi, L. (2022). The ethics of algorithms: Key problems and solutions. *AI & Society, 37*(1), 215–230. DOI : https://doi.org/10.1007/s00146-021-01154-8

[21] LeCun, Y. (2022). *A Path Towards Autonomous Machine Intelligence* (Version 0.9.2). OpenReview. https://openreview.net/forum?id=BZ5a1r-kVsf

[22] Bengio, Y. (2017). *The Consciousness Prior* (v2, Dec 2, 2019). arXiv:1709.08568. DOI : https://doi.org/10.48550/arXiv.1709.08568