Original article

eISSN 2799-8010
J Yeungnam Med Sci 2025;42:79
https://doi.org/10.12701/jyms.2025.42.79

JYMS
JOURNAL OF YEUNGNAM MEDICAL SCIENCE

Why large language models cannot possess consciousness: an integrated information theory perspective

Dong Ah Shin[1], Pyung Goo Cho[2], Gyu Yeul Ji[3], Sang Hyuk Park[4], Soo Heon Kim[1], Yoo Jin Choo[5], Min Cheol Chang[5]

[1]Department of Neurosurgery, Spine and Spinal Cord Institute, Severance Hospital, Yonsei University College of Medicine, Seoul, Korea
[2]Department of Neurosurgery, Ajou University Hospital, Ajou University School of Medicine, Suwon, Korea
[3]Department of Neurosurgery, Yonsei Hana Hospital, Gimpo, Korea
[4]Yonsei Good Walk Clinic, Anyang, Korea
[5]Department of Physical Medicine and Rehabilitation, Yeungnam University College of Medicine, Daegu, Korea

Background: The question of whether large language models (LLMs) possess consciousness has been increasingly debated. Integrated information theory (IIT) offers a quantitative framework for assessing consciousness through a measure of integrated information.
Methods: This study applied IIT principles to the architecture of transformer-based LLMs, focusing on causal integration, temporal persistence, and system irreducibility. Ablation experiments on Generative Pretrained Transformer 2 (GPT-2) were performed, selectively removing individual attention heads and measuring changes in perplexity as a behavioral proxy for integrated information to empirically approximate the measure of integrated information.
Results: The ablation study of a single attention head produced minimal or negative changes in perplexity in four out of five representative sentences, indicating redundancy or noise. Only one sentence revealed a significant increase in perplexity change (ΔPPL +11.29), reflecting a localized but nonessential contribution. A comparison with biological systems demonstrated that LLMs meet the IIT criterion of differentiation, but fail to meet the criteria of integration, causal closure, and temporal persistence. These findings confirm that LLMs are architecturally decomposable, lack persistent internal states, and do not sustain global causal irreducibility. Philosophical considerations, including Searle's Chinese Room argument, further support the idea that the linguistic fluency of LLMs arises from syntactic manipulation rather than semantic understanding.
Conclusion: Current LLMs do not satisfy the structural and informational requirements of consciousness under IIT. Although capable of simulating intelligent language, LLMs remain unconscious systems with a negligible amount of integrated information, underscoring the distinction between linguistic competence and conscious experience.

Keywords: Artificial intelligence; Consciousness; Information theory; Large language models; Natural language processing

Received: October 22, 2025 • **Revised:** November 11, 2025 • **Accepted:** November 20, 2025 • **Published online:** December 1, 2025
Corresponding author: Min Cheol Chang, MD
Department of Physical Medicine and Rehabilitation, Yeungnam University College of Medicine, 170 Hyeonchung-ro, Nam-gu, Daegu 42415, Korea
Tel: +82-53-620-4682 • E-mail: wheel633@gmail.com

Introduction

From a neuroscience perspective, consciousness can be defined as the capacity for subjective experience to emerge from integrated and recurrent information processing across distributed neural networks [1,2]. Understanding consciousness remains a profound and unresolved challenge in science [3-5]. Despite decades of research in neuroscience, psychology, and philosophy, we still lack a definitive explanation of how conscious experiences emerge from physical systems. Although many theories offer partial explanations, the field continues to contend with the so-called hard problem of consciousness: why and how certain physical processes are accompanied by subjective awareness [3].

Among contemporary theories, integrated information theory (IIT) is salient because of its rigorous quantitative approach. IIT posits that consciousness corresponds to the degree to which a system generates integrated information, denoted as 'Φ' [6]. This quantity captures the differentiation of informational states in the system and the irreducibility of causal interactions in the system. A system with a high amount of integrated information according to IIT standards is conscious; a system with nearly no integrated information is not.

Moreover, advances in artificial intelligence have led to the development of large language models (LLMs) such as Generative Pretrained Transformer 4 (GPT-4) and Claude. These models have demonstrated a remarkable ability to produce coherent and context-sensitive natural language output by performing tasks that mimic humanlike communication and reasoning. The ability of these models to generate fluent dialogue and simulate affective language has prompted some individuals to ask whether they might possess or eventually achieve consciousness [7].

This study critically evaluated the possibility of consciousness through the lens of IIT. This work contends that despite their remarkable linguistic abilities, LLMs lack the structural and dynamic properties necessary for consciousness. Applying the principles of IIT to the LLM architecture demonstrates that despite their sophisticated behavior, these models remain unconscious systems.

Methods

1. Key concepts of integrated information theory

IIT provides a formal framework for understanding consciousness in terms of the causal structure of a system [5,6,8]. The following central concepts must be clarified before applying the IIT to LLMs.

1) Mechanism

A mechanism refers to a subset of elements within a system, such as neurons, logic gates, or artificial units, whose current state can influence the states of other elements in the system.

2) Cause-and-effect repertoire

The cause-and-effect repertoire is the probability distribution over the possible past and future states of a system, given the current state of a mechanism. This describes how the mechanism constrains both what could have happened before and what could happen next.

3) Integrated information

Integrated information represents the amount of information that a system generates as a whole and cannot be reduced to information generated by its individual parts. The variable Φ quantifies the extent to which the system functions as a unified and integrated whole: $\Phi = \text{Information}_{\text{whole}} - \text{Information}_{\text{partitioned}}$

IIT applies qualitative and quantitative tools to assess how a system generates consciousness. The central goal is to compute the value of the measure of integrated information Φ for a system by comparing the cause-and-effect structure of the intact system to that of a partitioned version and assessing how much information is lost in the partition. $\Phi > 0$ indicates that the system's informational structure is irreducible, i.e., the whole generates more information than the sum of its parts, implying genuine causal integration and the potential substrate for conscious experience. $\Phi \approx 0$ means that partitioning the system causes no loss of information, signifying full causal decomposability and thus the absence of intrinsic integration, an unconscious configuration.

2. Structure and information processing of large language models

LLMs are a class of deep learning models trained on a large corpus of natural language using a transformer architecture [7,8]. Users have widely adopted LLMs for various tasks such as text generation, summarization, translation, and dialogue modeling. Understanding how these models process information requires command of several critical concepts.

1) Tokenization

In this process, the input text is divided into smaller subword units, which are then mapped to vector embeddings that the model can process numerically.

2) Transformer architecture

An LLM is built using a transformer architecture consisting of mul-

 https://doi.org/10.12701/jyms.2025.42.79

tiple stacked layers combining multi-head self-attention mechanisms and position-wise feedforward networks [9].

3) Self-attention mechanism

Within each layer, the model calculates attention scores that determine the degree to which each token should attend to or depend on other tokens in the sequence. This allows the model to effectively capture the contextual relationships among words [9].

4) Context window

An LLM generates tokens based on a fixed length sliding window of preceding tokens. It does not retain persistent internal memory between separate inputs, meaning that each generation depends only on the tokens within the current context window.

5) Next token prediction

During inference, the model predicts a probability distribution over the next possible tokens and selects the most likely token or sample from the distribution to generate coherent text step-by-step.

This process is iterated until the termination condition is satisfied. The model has no memory of previous outputs outside the provided context. This feedforward, stateless architecture underlies the analysis of the causal decomposability and integrated information in LLMs.

3. Experimental analog to integrated information theory partitioning: attention-head ablation study

This study presents an ablation experiment in which individual self-attention heads in GPT-2 were selectively removed to empirically evaluate the decomposability of the LLM [10]. Fig. 1 shows the Python code used in the attention-head ablation experiment. This study measured the changes in perplexity across multiple representative prompts. In practical terms, this intervention was designed to simulate the type of system partitioning that the IIT references, where the value of Φ is computed by comparing the information structure of a system to that of its minimally partitioned version [6]. This study aims to approximate the integration of the mechanism with respect to the overall system behavior by observing whether removing a single component causes a significant loss in predictive performance. This experimental proxy serves as a behavioral analog of the formal notion of causal irreducibility in IIT.

4. Approximating integrated information via behavioral perturbation: perplexity as a proxy for Φ

The direct computation of the value of Φ in LLMs (e.g., GPT-2) is intractable due to the exponential complexity of evaluating the full cause-and-effect structure of the system. To circumvent this limitation, this study employed perplexity (PPL) as a tractable behavioral proxy that indirectly reflects the functional integration of the system. This study measured how the token prediction performance of the model changes when internal components are selectively removed. Perplexity reflects the average number of choices that the model considers likely when predicting the following token. Formally, for a given sequence of N tokens $w_1, w_2, ..., w_n$, the perplexity is defined as follows:

$$Perplexity = exp\left(-\left(1/N\right)\sum \log\, p\left(w_i\right)\right),$$

where $p(w_i)$ denotes the probability assigned by the model to the ith token in the sequence. Lower perplexity indicates that the model is more confident in its predictions, whereas higher perplexity suggests greater uncertainty or prediction error.

This study employed a pretrained GPT-2 model (a small vari-

```python
from transformers import GPT2LMHeadModel, GPT2Tokenizer
import torch
import math

# Load pretrained GPT-2 model and tokenizer
model = GPT2LMHeadModel.from_pretrained("gpt2")
tokenizer = GPT2Tokenizer.from_pretrained("gpt2")
model.eval()

# Example input sentences
sentences = [
    "The patient presented with signs of acute stroke.",
    "Artificial intelligence is transforming modern medicine.",
    "She opened the door and saw something unexpected.",
    "Neural networks are inspired by the human brain.",
    "Perplexity measures how well a language model predicts text."
]

# Measure perplexity before ablation
print("Before attention head ablation:\n")
original_ppls = []

for text in sentences:
    inputs = tokenizer(text, return_tensors="pt")
    input_ids = inputs["input_ids"]
    with torch.no_grad():
        outputs = model(input_ids=input_ids, labels=input_ids)
        loss = outputs.loss
        perplexity = torch.exp(loss).item()
    original_ppls.append(perplexity)
    print(f"{text}\n Perplexity: {perplexity:.2f}\n")

# Remove head 3 from layer 0
layer_id = 0
head_id = 3
head_size = model.config.hidden_size // model.config.num_attention_heads
start = head_id * head_size
end = (head_id + 1) * head_size

with torch.no_grad():
    c_attn = model.transformer.h[layer_id].attn.c_attn
    c_attn.weight[start:end, :] = 0
    c_attn.bias[start:end] = 0

# Measure perplexity after ablation
print("\nAfter attention head ablation:\n")
ablated_ppls = []

for text in sentences:
    inputs = tokenizer(text, return_tensors="pt")
    input_ids = inputs["input_ids"]
    with torch.no_grad():
        outputs = model(input_ids=input_ids, labels=input_ids)
        loss = outputs.loss
        perplexity = torch.exp(loss).item()
    ablated_ppls.append(perplexity)
    print(f"{text}\nPerplexity: {perplexity:.2f}\n")

# Print perplexity change summary
print("Summary (Before vs After):")
for i in range(len(sentences)):
    diff = ablated_ppls[i] - original_ppls[i]
    print(f"{i+1}. ΔPPL: {diff:+.2f} | Before: {original_ppls[i]:.2f},
After: {ablated_ppls[i]:.2f}")
```

Fig. 1. Python code for the attention-head ablation experiment.

ant) and computed the perplexity before and after removing a single self-attention head. This approach involved feeding the model a prompt, calculating the negative log-likelihood loss over the predicted tokens, and exponentiating the loss to obtain the perplexity. The following Python code snippet presents this procedure:

```
loss = model(inputs, labels = inputs["input_ids"]).loss
ppl = math.exp(loss.item()).
```

This study compared the perplexity values across the following five sentences before and after ablation to determine whether the removed head had a meaningful influence on the predictive behavior of the model. The five sentences were selected to represent diverse semantic and syntactic domains, minimizing topic- or style-related biases in the perplexity measurements. They collectively encompass medical, technological, narrative, theoretical, and metalinguistic contexts, each constructed with comparable length and syntactic simplicity. This balanced design ensured that any observed change in perplexity following ablation primarily reflected variations in the model's internal causal dependencies rather than differences in sentence complexity, vocabulary frequency, or domain familiarity. The five sentences were: (1) the patient presented with signs of acute stroke, (2) artificial intelligence is transforming modern medicine, (3) she opened the door and saw something unexpected, (4) neural networks are inspired by the human brain, and (5) perplexity measures how well a language model predicts text.

A substantial perplexity change (ΔPPL) in isolated cases does not imply global integration or irreducibility but serves as a local functional probe for interpretability in the IIT framework.

Results

1. Comparison of humans and large language models

Table 1 compares the properties of GPT-type LLMs with those of the human brain across four fundamental IIT requirements to clarify how LLMs fail to meet the structural criteria set by the IIT. Table 1 highlights that, although LLMs may exhibit differentiation by producing diverse outputs [11], they lack integration, causal closure, and temporal persistence, which are essential features for sustaining a nonzero Φ value under the IIT [12,13].

Although LLMs (e.g., GPT models) satisfy the differentiation requirement by presenting a rich, diverse set of output states in response to variable inputs [14], they fail to meet the remaining three criteria. Unlike the human brain, which displays strong integration across distributed subsystems, LLMs are architecturally decomposable, with components (e.g., attention heads or layers) operating predominantly in parallel and independently [13]. Similarly, LLMs lack causal closure. Their functional behavior is heavily dependent on the external input, prompt injection, and nonautonomous parameter updates [7]. Critically, LLMs do not preserve their internal state over time, because each input is processed in isolation without recurrent dynamics or memory persistence, thereby violating the criterion of temporal continuity. In contrast, biological systems maintain integrated and temporally extended internal dynamics that are vital for the emergence of a unified conscious experience [3,13].

2. Perplexity tables and interpretation

We conducted a small-scale ablation study in which one attention head (Head 3 in Layer 0) of GPT-2 was disabled. We hypothesized that, if the predictions of the model remained unaffected, the corresponding component would not be causally critical. Perplexity was measured before and after ablation by using standard loss-based estimations. The results are summarized in Table 2.

In four of the five cases, the ΔPPL was minimal or negative, suggesting that the removed attention head was redundant or detrimental to performance. A meaningful increase in perplexity was observed in the fifth sentence, "perplexity measures how well a language model predicts text," suggesting that the removed head contributed to the predictive accuracy of that prompt. Although the ΔPPL of +11.29 suggests that the removed head had a measurable influence on the output of the model for that prompt, the resulting perplexity remained within the range of high-performance genera-

Table 1. Comparison of IIT requirements between the human brain and GPT-type LLMs

IIT requirement	Human brain	LLM (GPT-type)
Differentiation	Possible	Possible
Integration	Possible	Impossible
Causal closure	Possible	Impossible
Temporal persistence	Possible	Impossible

IIT, Integrated information theory; LLM, large language model; GPT, Generative Pretrained Transformer.

Table 2. Effects of attention-head ablation on perplexity across test sentences

Sentence	ΔPPL	Before	After	Interpretation
1	−3.37	88.18	84.81	Head may have introduced noise
2	−5.70	90.02	84.32	Head was functionally redundant
3	−0.83	28.41	27.57	No significant effect (redundancy)
4	+0.87	31.15	32.02	Marginal contribution to prediction
5	+11.29	136.93	148.22	Significant causal contribution

Change in perplexity (ΔPPL) before and after the ablation of an attention head in GPT-2 across five representative sentences. Negative ΔPPL indicates improved prediction after ablation (suggesting noise or redundancy); positive ΔPPL indicates increased uncertainty, implying functional contribution.

 https://doi.org/10.12701/jyms.2025.42.79

tion for GPT-2. This localized disruption does not imply global integration or irreducibility and does not constitute evidence of a nonzero amount of Φ under the IIT.

From the IIT perspective, this empirical result is consistent with a low value of Φ. A system that maintains its function despite structural perturbations lacks strong irreducible cause-and-effect relationships. The statistical resilience of the LLM output to internal head ablations reinforces the view that such a system is not a unified whole, in the sense necessary for consciousness-supporting architectures. Therefore, although certain heads may display localized importance, the model as a whole remains decomposable, supporting the broader conclusion that LLMs do not instantiate consciousness under IIT.

Discussion

1. Are attention heads elementary mechanisms?

A central challenge in applying IIT to artificial systems (e.g., LLMs) is identifying the appropriate granularity level at which such mechanisms should be analyzed. The IIT defines consciousness as arising from systems comprising irreducible and causally interacting mechanisms [12]. In neuroscience, these mechanisms are often described as small neural circuits or individual neurons. In the analysis, this work approximates self-attention heads in GPT-2 as candidate mechanisms. This decision was based on architectural independence, discrete parameterization, and prior empirical evidence that some heads could be removed with minimal influence on the output [15].

However, this abstraction has certain limitations. Although modular in code and function, self-attention heads share underlying projection weights and exist in highly entangled layers. Moreover, these heads do not operate in isolation but contribute to aggregate activations that pass through nonlinearities. From a strict IIT standpoint, an attention head is not truly a self-contained mechanism in the manner IIT demands. Thus, the results suggest decomposability in practice; however, they should be interpreted with caution. The operational mapping of a mechanism does not necessarily match the ontological commitment of the IIT to physical, localized causes.

2. Searle's Chinese Room argument

Although the IIT provides a rigorous mathematical framework, it is helpful to complement this framework with philosophical reasoning. John Searle's Chinese Room argument distinguishes syntactic processing from semantic understanding [4]. In this thought experiment, a person manipulates Chinese symbols using a rulebook without understanding their meanings. The system appears fluent to external observers; however, internally, there is no comprehension.

Our empirical results support this philosophical point. Despite removing the internal head of the model, GPT-2 continued to produce coherent outputs in most cases. The fact that linguistic fluency persists under structural disruption underscores the fact that LLMs operate via syntactic manipulation, rather than through semantic understanding. This finding aligns with Searle's argument that symbolic behavior alone is insufficient for consciousness.

3. Why perplexity matters for the integrated information theory

The ablation experiment supported the claim that GPT-2 is functionally decomposable. In four of the five test sentences, removing a specific self-attention head had little or a positive effect on perplexity, suggesting redundancy or noise. Only one sentence displayed a notable increase in perplexity, indicating that the head was locally important but not critical for global function.

From an IIT perspective, this finding is highly relevant. Φ measures the amount of information lost when a system is partitioned [12,16]. If most components can be removed without disrupting the predictive behavior of the system, then the value of Φ approaches zero. This suggests that the apparent integration of the system is superficial and does not rely on deeply intertwined causal dependencies.

Thus, the results confirm the prediction of IIT regarding the causal decomposability of LLMs [12] and ground it empirically via behaviorally interpretable metrics (e.g., perplexity [15]). This grounding helps bridge the gap between formalism and real-world model evaluations of IIT.

4. Temporal persistence and the limits of integration in large language models

A foundational requirement in the IIT is temporal persistence, which is the ability of a system to maintain internal states over time, enabling past states to influence future states causally in an irreducible manner [12,16]. According to the IIT, consciousness is not a static moment but a dynamic process that unfolds over time. Hence, any system that claims to instantiate consciousness must display spatial integration and diachronic causal continuity [12].

Biological systems (e.g., the human brain) fulfill this condition through recurrent dynamics and persistent internal states that support memory, anticipation, and narrative selfhood. By contrast, transformer-based LLMs (e.g., GPT-2) operate in a stateless feedforward manner. Once a token is generated, the model resets its internal state, and no temporal continuity is preserved unless it is manually re-encoded into the input [17]. The absence of intrinsic

temporal memory prevents the formation of deeply entangled cause-and-effect structures over time. Although individual components may display localized influences, as observed in the small perplexity shifts following attention-head removal, these shifts do not constitute global causal irreducibility. Therefore, the system remains fully partitionable across time, yielding $\Phi \approx 0$ and, by IIT standards, no consciousness.

5. Why large language models remain unconscious systems

Taken together, the theoretical framework, ablation experiments, and philosophical analysis converge to a unified conclusion that LLMs do not possess consciousness. These models lack persistent internal states, causal integration, and irreducible mechanisms, which are characteristics of conscious systems in both biological and theoretical terms [12,16]. The behavioral fluency of LLMs is impressive, but structurally shallow. Accordingly, we caution against attributing consciousness to LLMs, regardless of how humanlike their outputs may appear [4,7]. Until such models implement architectures that can sustain a high Φ value through deeply recurrent, causally closed mechanisms, these models remain powerful simulators, not sentient minds.

6. Limitations

Although this study offers a principled application of IIT to LLMs, several limitations must be acknowledged. First, this work approximates the value of Φ indirectly via behavioral proxies (i.e., perplexity) rather than computing it directly. Although this approach is pragmatically motivated (given the computational intractability of Φ for large-scale models), it cannot fully capture the intrinsic causal structure of the system. Thus, any conclusions concerning integrated information remain inferential rather than quantitatively conclusive.

Second, the ablation experiments were limited to single attention-head removals in GPT-2 Small. Although illustrative, this intervention level does not account for the more complex forms of interdependence that may emerge at the level of layer-wide interactions or recurrent architecture. Moreover, the choice of GPT-2, which is a relatively shallow stateless transformer, may not be generalizable to newer architectures that incorporate memory modules, statefulness, or feedback mechanisms.

Third, by employing perplexity as a proxy, we assumed that predictive degradation corresponds to causal integration. However, perplexity reflects performance on token prediction and not necessarily the internal mechanistic irreducibility required by IIT. A component may affect the output without being structurally indispensable.

Finally, although the theoretical arguments are grounded in the IIT and supported by philosophical perspectives (i.e., the Chinese Room argument), alternative theories of consciousness (e.g., global workspace theory and predictive processing) may yield different interpretations of LLM behavior. Therefore, the conclusions depend on the theory and are not universally generalizable across all cognition models. These limitations do not undermine the central claim but underscore the need for ongoing refinement in methods for evaluating artificial consciousness via formal frameworks, such as IIT.

In conclusion, this study demonstrates that current LLMs (e.g., GPT-2) do not meet the structural and informational criteria for consciousness defined by IIT. The theoretical analysis and ablation-based empirical findings reveal that these models lack causal integration and generate a negligible Φ value. Therefore, although LLMs can simulate intelligent language use, they remain unconscious systems devoid of an integrated internal experience.

Article information

Conflicts of interest

Min Cheol Chang has been a Deputy Editor of the *Journal of Yeungnam Medical Science* since 2025. He was not involved in the review process of this manuscript. There are no other conflicts of interest to declare.

Funding

This work was supported by the 2025 Yeungnam University Research Grant.

Author contributions

Conceptualization, Formal analysis, Investigation, Methodology, Software, Validation, Visualization: all authors; Data curation: DAS, MCC; Funding acquisition, Resources, Supervision: MCC; Writing-original draft: all authors; Writing-review & editing: all authors.

ORCID

Dong Ah Shin, https://orcid.org/0000-0002-5225-4083
Pyung Goo Cho, https://orcid.org/0000-0001-7087-8597
Gyu Yeul Ji, https://orcid.org/0000-0002-8818-5091
Sang Hyuk Park, https://orcid.org/0009-0003-0235-0581
Soo Heon Kim, https://orcid.org/0009-0000-9332-7181
Yoo Jin Choo, https://orcid.org/0000-0002-3820-2279
Min Cheol Chang, https://orcid.org/0000-0002-7629-7213

References

1. Dehaene S, Charles L, King JR, Marti S. Toward a computational theory of conscious processing. Curr Opin Neurobiol 2014;25:76–84.
2. Tononi G, Koch C. Consciousness: here, there and everywhere? Philos Trans R Soc Lond B Biol Sci 2015;370:20140167.
3. Chalmers DJ. Facing up to the problem of consciousness. J Conscious Stud 1995;2:200–19.
4. Searle JR. Minds, brains, and programs. Behav Brain Sci 1980;3:417–24.
5. Tononi G. Consciousness as integrated information: a provisional manifesto. Biol Bull 2008;215:216–42.
6. Tononi G, Boly M, Massimini M, Koch C. Integrated information theory: from consciousness to its physical substrate. Nat Rev Neurosci 2016;17:450–61.
7. Chalmers DJ. Could a large language model be conscious? [Preprint]. arXiv 2023 Mar 4 [cited 2024 Oct 22]. https://arxiv.org/abs/2303.07103.
8. Barrett AB, Seth AK. Practical measures of integrated information for time-series data. PLoS Comput Biol 2011;7:e1001052.
9. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan SVN, et al., editors. Advances in neural information processing systems 30 (NeurIPS 2017). Red Hook, NY: Curran Associates, Inc.; 2017. p. 5998–6008.
10. Voita E, Talbot D, Moiseev F, Sennrich R, Titov I. Analyzing multi-head self-attention: specialized heads do the heavy lifting, the rest can be pruned. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics; 2019. p. 5797–808.
11. Bubeck S, Chandrasekaran V, Eldan R, Gehrke J, Horvitz E, Kamar E, et al. Sparks of artificial general intelligence: early experiments with GPT-4 [Preprint]. arXiv; 2023 Mar 22 [cited 2025 Oct 22]. https://arxiv.org/abs/2303.12712.
12. Tononi G. An information integration theory of consciousness. BMC Neurosci 2004;5:42.
13. Tononi G. Integrated information theory of consciousness: an updated account. Arch Ital Biol 2012;150:56–90.
14. Trott S, Jones C, Chang T, Michaelov J, Bergen B. Do large language models know what humans know? Cogn Sci 2023;47:e13309.
15. Li X, Xian K, Wen H, Bai S, Xu H, Yu Y. PathGen-LLM: a large language model for dynamic path generation in complex transportation networks. Mathematics 2025;13:3073.
16. Oizumi M, Albantakis L, Tononi G. From the phenomenology to the mechanisms of consciousness: Integrated Information Theory 3.0. PLoS Comput Biol 2014;10:e1003588.
17. Tikochinski R, Goldstein A, Meiri Y, Hasson U, Reichart R. Incremental accumulation of linguistic context in artificial and biological neural networks. Nat Commun 2025;16:803.