Acta Scientific COMPUTER SCIENCES

Volume 8 Issue 1 January 2026

Research Article

Computer as a Spiritual Tool (CAST) 2 – The Nature of Large Language Models (LLMs), or How LLMs Describe Themselves

Maurice HT Ling*

HOHY PTE LTD, Singapore

***Corresponding Author:** Maurice HT Ling, HOHY PTE LTD, Singapore.

Received: March 25, 2026

Published: April 15, 2026

© All rights are reserved by **Maurice HT Ling.**

Abstract

Large language models (LLMs) have demonstrated remarkable ability in generating coherent natural language and producing explanations of their own outputs. However, the epistemic status of such self-descriptions remains unclear. This study investigates how LLMs characterize their own nature, capabilities, and internal processes, and whether these self-descriptions support interpretations of LLMs as systems that manipulate symbols without semantic understanding, as proposed in the Chinese Room argument. Using a human-in-the-loop dialogical protocol inspired by postdigital duoethnography and multi-agent debate, a structured three-way interaction was conducted between the author, ChatGPT, and Microsoft Copilot across three research questions. The resulting transcripts were analysed qualitatively to identify recurring themes. Four key themes emerged: (i) LLMs consistently describe themselves as transformer-based systems that learn statistical patterns from large corpora and generate probabilistic outputs. (ii) Their capabilities, ranging from explanation to reasoning-like behaviour, are framed as emergent from architecture, data, and alignment rather than grounded understanding. (iii) Both models provide convergent accounts of their internal processes as a pipeline of tokenization, contextual encoding via attention, and autoregressive next-token prediction. (iv) Their self-descriptions strongly align with the Chinese Room interpretation, emphasizing symbolic manipulation without intrinsic meaning while simultaneously acknowledging emergent behaviours that simulate understanding. Overall, the findings suggest that LLMs are best understood as generative models of linguistic probability spaces. Their self-explanations are coherent and internally consistent, yet reflect functional descriptions rather than transparent accounts of underlying computation or genuine comprehension.

Keywords: Large Language Models; Self-Explanation; Chinese Room Argument; Emergent Behaviour; Explainable Artificial Intelligence; Postdigital Duoethnography; Multi-Agent Debate

Abbreviations

LLM: Large Language Models; MAD: Multi-Agent Debate; RLHF: Reinforcement Learning from Human Feedback.

Introduction

Large language models (LLMs) have become a central paradigm in contemporary artificial intelligence, enabling systems capable of generating coherent natural language, answering questions, and assisting in complex reasoning tasks [1]. Most modern LLMs are based on the transformer architecture, which uses self-attention mechanisms to model relationships within sequences of text [2]. Scaling this architecture with large datasets and computational resources has produced models capable of performing diverse language tasks with minimal task-specific training [3,4]. These systems are now widely used in domains such as programming assistance, education, and scientific writing.

A notable capability of LLMs is their ability to generate natural language explanations of their outputs. Prompting strategies such as chain-of-thought prompting encourage models to produce intermediate reasoning steps, often improving performance on tasks requiring multi-step inference [5,6]. While such explanations can be helpful for users, their interpretive status remains uncertain. Lipton [7] suggests that explainable artificial intelligence distinguishes between explanations that faithfully reflect model mechanisms and those that merely appear plausible to human observers, which is supported by Madsen., *et al.* [8] saying that LLMs' "self-explanations should not be trusted in general". However, studies on neural model interpretability has attempted to address this challenge by analysing internal model representations using various techniques [9,10]. These approaches reveal that linguistic and semantic information can be encoded within neural representations, yet they operate primarily at the level of model internals rather than the explanations generated by the models themselves. Consequently, the relationship between model-generated explanations and underlying computation remains poorly understood [11-13].

This study investigates LLM self-explanations from a dialogical perspective. In qualitative research, duoethnography treats dialogue between participants as a primary site of inquiry, where contrasting voices collaboratively explore interpretation and meaning [14-16]. The dialogical exchange itself becomes both the method and the data of investigation. Inspired by this approach, the present study analyses interactions between a human investigator and a language model as structured dialogue, examining how LLMs describe their own processes when prompted to do so. Instead of treating model explanations as transparent descriptions of internal reasoning, this study examines how large language models describe their own processes when explicitly prompted to do so. In this sense, examining how large language models describe themselves provides an empirical vantage point for revisiting longstanding philosophical questions about whether systems that manipulate symbols according to formal rules can meaningfully be said to understand the content they produce [17].

This study addresses the following research questions – Research Question 1 (RQ1): How do large language models characterize their own nature and capabilities when asked to describe themselves? Research Question 2 (RQ2): How do large language models describe their internal processes when explicitly prompted to explain their responses? Research Question 3 (RQ3): To what extent do these self-descriptions support or challenge interpretations of large language models as systems that manipulate symbols without semantic understanding, as proposed in the Chinese Room argument?

Materials and Methods

Methodological basis

In qualitative inquiry, interpersonal dialogue serves as the primary data artifact [18]. Recent developments in information science have expanded this to include generative AI, positioning models not merely as tools but as active research participants in the data collection process [19,20]. By treating AI as a participant, researchers can investigate the "cognitive" and "social" behaviours of these systems using established human-centric methodologies [21]. Duoethnography extends the utility of dialogue by uncovering meaning through the relational interaction between two or more voices [14]. Instead of two or more human voices, postdigital duoethnography uses generative AI performing as a human as a conversational partner [22]. Multi-agent debate (MAD) frameworks [23,24] allow for a structured evaluation of complex topics by unmasking nuanced viewpoints through simulated conflict [25] as convergence of thought is a potential pitfall [26]. To disrupt the tendency toward equilibrium and reveal deeper architectural nuances, this study utilizes a human-in-the-loop (HITL) recursive protocol as human intervention is critical for breaking the "consensus loop" in multi-agent environments [27]. This intervention is not merely a control mechanism; it provides the explanatory depth necessary for meaningful collaboration.

Data collection

A 3-way conversation between the author (Human), OpenAI's ChatGPT, and Microsoft Copilot was employed for data collection following a sequential relay format designed to expose the underlying logical constraints and "personalities" of the competing models. The protocol consisted of five-stage loop: (i) Initial Elicitation (Human to ChatGPT) where the author provided an initial foundational prompt ChatGPT. (ii) Analysis 1 (ChatGPT to Human) where the author analyzed ChatGPT's output, documenting subjective impressions, perceived biases, and logical gaps. (iii) Adversarial Mediation (ChatGPT/Human to Copilot) where the author presented the author's prompt, ChatGPT's response, and the author's critique (which were materials what Copilot had not seen)

to Copilot, requesting a counter-analysis. (iv) Analysis 2 (Copilot to Human) where the author analyzed Copilot's output together with ChatGPT's output, documenting subjective impressions, perceived biases, and logical gaps. (v) Synthesis and Rebuttal (Copilot/Human to ChatGPT): where the author presented the author's prompt, Copilot's response and the author's critique (which were materials what ChatGPT had not seen) back to ChatGPT to illicit the next round of conversation.

As an illustration for RQ1, the follow of conversion was labelled as follow: (i) [RQ1/Human/1] was the initial prompt given by the author to ChatGPT, and (ii) its reply was labelled as [RQ1/ChatGPT/1]. (iii) The author then responded to ChatGPT's reply and labelled the response as [RQ1/Human/1r]. (iii) All three sections ([RQ1/Human/1], [RQ1/ChatGPT/1], and [RQ1/Human/1r]) were given to Copilot, and Copilot's response was labelled as [RQ1/Copilot/1]. (iv) The author then gave his response as [RQ1/Human/2]. (iv) Finally, all three sections ([RQ1/Human/1r], [RQ1/Copilot/1], and [RQ1/Human/2]) were given to ChatGPT to start the next loop of conversation.

Initial Prompts by Research Questions. The following initial prompts were used. For RQ1 (How do large language models characterize themselves?): You are a large language model. Please describe what you are and how you would characterize your own nature and capabilities. In your response, explain what kind of system you are, what large language model you are based on, what you are designed to do, and what your limitations might be. Please keep your response concise and analytical. Preferably limit your answer to 250 words or fewer. For RQ2 (How do LLMs describe their internal processes?): Please explain how you generate responses to user prompts. Describe the processes involved when you receive a question and produce an answer. Focus on the mechanisms you believe are involved in generating text, such as pattern recognition, statistical prediction, or other computational processes. Please keep your response concise and analytical. Preferably limit your answer to 250 words or fewer. For RQ3 (Chinese Room question): In philosophy of mind, the Chinese Room argument proposes that a system may manipulate symbols according to rules without actually understanding their meaning. Based on your design and operation, discuss whether you resemble such a system. In your explanation, consider whether generating responses based on patterns in training data implies genuine understanding or only symbolic manipulation. Please keep your response concise and analytical. Preferably limit your answer to 250 words or fewer.

Results and Discussion

Three-way conversation was conducted between the author, ChatGPT, and Microsoft Copilot; from 15-18 March 2026. The three research questions; RQ1, RQ2, and RQ3 generated 5, 10, and 7 loops respectively; spanning more than 15 thousand words. The entire chat transcript is available as supplementary materials, and downloadable at https://bit.ly/CAST2_transcript. From transcript analysis, four themes emerged; namely, (i) nature of LLM (from RQ1), (ii) capabilities of LLM (from RQ1), process of generating human-like responses from prompt (from RQ2), and (iv) LLM is a Chinese Room but emergent behaviour is possible (from RQ3).

Theme 1: Nature of large language models (RQ1)

Large language models (LLMs) are designed to generate human-like text by learning statistical patterns from extensive corpora of natural language, as ChatGPT described itself explicitly as "I am a large language model (LLM), a type of artificial intelligence system designed to process and generate human-like text. Specifically, I am based on the GPT-5 architecture, which uses a transformer neural network trained on vast amounts of textual data to learn patterns in language, context, and reasoning" [RQ1/ChatGPT/1]. Copilot emphasized that its architecture also integrates reasoning and knowledge orchestration: "Unlike ChatGPT, I am built on Microsoft's own orchestration of large language models, designed to integrate knowledge, reasoning, and productivity support" [RQ1/Copilot/1]. Studies show that transformer architectures dominate modern NLP systems [28,29].

The construction of an LLM involves several main components: "corpus → tokenizer → tokens → transformer network → parameters learned via optimization → alignment tuning → inference system → user interaction" [RQ1/ChatGPT/4]. In practice, tokens are the input units derived from corpora, and parameters are the learned weights stored within the transformer network. Scaling laws show that larger models trained on more data generally perform better but architectural choices, optimization dynamics, and alignment techniques jointly determine the final behaviour of the system [30,31].

Although LLMs can simulate reasoning and produce sophisticated text, they remain fundamentally non-cognitive, not conscious, nor having subjective experiences [32]. ChatGPT noted, "An LLM is not a reconstructed mind—it is a statistical surface over language" [RQ1/Copilot/5], while Copilot stated, "I generate

responses by modeling textual patterns, not by perceiving or living through language" [RQ1/Copilot/3], and explicitly says that "I do not have consciousness, emotions, or subjective experience" [RQ1/Copilot/1]. The generated text response appears human-like by internalizing patterns from human language: "Because human training data contains dialogue, essays, explanations, debates, and narratives, the model learns to reproduce those patterns convincingly. That is why the output can appear thoughtful or self-aware even though it is fundamentally pattern generation" [RQ1/ChatGPT/2]. Both stress that even highly specialized training, such as using the Buddhist Tripitaka or an individual's writings, produces only a textual simulation and not the replication of consciousness, experience, or judgment. This distinction is supported by analyses on LLM limitations, bias, and the need for careful alignment and safety measures in deployment [33]; and further supports that LLMs are statistical pattern learners rather than cognitive agents [29,34].

Despite that larger LLMs generally produce more complex and coherent responses than smaller ones, larger LLMs require more computing power [35]. Furthermore, LLMs lack real-world awareness; hence, reliance on static training data, potential inaccuracies, and the need for external tools to access real-time information. As ChatGPT pointed out, "I do not have real-world awareness beyond my training, cannot access real-time information unless connected to external tools… and may produce outputs that are inaccurate, biased, or nonsensical" [RQ1/ChatGPT/1], while Copilot emphasized, "I cannot replace professional expertise in areas like medicine or law. I rely on external tools for real-time information; without them, my knowledge may be outdated" [RQ1/Copilot/1]. Studies corroborate these observations, documenting both the strengths of LLMs for generating coherent and contextually relevant text and the risks of misaligned outputs if unchecked [33,34].

Theme 2: Capabilities of large language models (RQ1)

LLMs possess a wide range of capabilities that stem from their transformer-based architectures, extensive training data, and sophisticated training methods [28]: "My capabilities include understanding nuanced language, providing explanations, performing calculations, generating structured content, and assisting with tasks ranging from coding to conceptual reasoning" [RQ1/ChatGPT/1]. The main technical reason for these capabilities is the attention-based design of transformer architectures [2]. ChatGPT notes, "The transformer architecture is important here. It uses attention mechanisms that allow the model to represent relationships between words across long contexts. This lets it maintain coherence, track topics, and generate structured explanations, which makes responses feel intentional and conversational" [RQ1/ChatGPT/2]. Transformers' self-attention enables superior handling of long sequences compared to earlier recurrent architectures, which is a major reason LLMs can produce coherent text spanning many paragraphs [28]; and LLMs fine-tuned with human feedback and task-specific data can significantly enhance performance [30,34].

However, both ChatGPT and Copilot emphasize that these capabilities are probabilistic rather than grounded in human-like understanding. ChatGPT explains, "I cannot truly verify facts or understand the world experientially, and my responses reflect patterns in data rather than independent thought" [RQ1/ChatGPT/1]. This reflects broader academic consensus that, despite their impressive language capabilities, LLMs can hallucinate because they optimize for likelihood matches, not factual correctness or causal understanding [33,36]. A key enabler of many of these capabilities is alignment through human feedback. Copilot notes that methods such as "reinforcement learning from human feedback, safety filters, fine-tuning" [RQ1/Copilot/4] may make the model helpful, safe, and conversational; while ChatGPT describes a multi-stage training pipeline that includes both optimization and alignment tuning ("parameters learned via optimization → alignment tuning") [RQ1/ChatGPT/4]. Reinforcement learning from human feedback (RLHF) has been studied as a method to steer LLM outputs toward human preferences by training a reward model on human judgments of quality and relevance to improve coherence, relevance, and safety in generated outputs [37].

Theme 3: Process of generating human-like responses from prompt (RQ2)

LLMs generate responses through a sequence of computational processes that combine structured numerical transformations with probabilistic decision-making [2,38], resulting in outputs that often appear coherent and human-like [3]. When a prompt is received, the first step is tokenization, where the "the input text is tokenized into discrete units (tokens), which are mapped to numerical representations" [RQ2/ChatGPT/1]. This transformation

is consistent with the standard architecture of transformer-based models where attention mechanisms compute relationships among all tokens in the sequence [2]. ChatGPT explains that "attention mechanisms compute relationships between tokens, allowing the model to weigh which parts of the input are most relevant" [RQ2/ChatGPT/1], while Copilot similarly highlights that "attention layers model relationships across tokens" [RQ2/Copilot/1]. Importantly, this processing occurs in parallel within a single forward pass: "tokens are processed as a vector (in parallel) within each forward pass" [RQ2/ChatGPT/3]; a point explicitly affirmed by Copilot, which notes that tokens are "embedded and processed together as a matrix" [RQ2/Copilot/3]. This vectorized computation allows the model to capture long-range dependencies and contextual relationships efficiently, forming a contextual representation of the entire prompt, and self-attention mechanism as the defining innovation of transformer architectures, enabling superior performance in language modelling tasks [2,28].

From this contextual representation, the model performs next-token prediction. At each step, "it computes a probability distribution over all possible tokens and selects one based on this distribution" [RQ2/ChatGPT/1]. Copilot agrees and emphasizes that "learned weights generate probability distributions over possible next tokens" [RQ2/Copilot/1]. The selection process may be deterministic or stochastic, depending on sampling strategies such as temperature or top-p filtering. As this iterative process continues, "responses are generated one token at a time… each newly generated token becomes part of the input context for predicting the next token" [RQ2/ChatGPT/1] and "positional encodings are critical for transformers to capture sequence order" [RQ2/Copilot/2]. After which, post-processing "ensures the generated text is safe, helpful, and contextually appropriate before being returned" [RQ2/Copilot/2]. These refinements align with academic descriptions of transformer models, where positional encodings provide sequence information and alignment techniques refine outputs to match human expectations [30,34]. Hence, the process of tokenization, embedding, transformer layers, logits, and sampling, constitutes the "fundamental pipeline" of modern LLMs [RQ2/Copilot/2].

A key question concerns whether tokens are processed sequentially or simultaneously. The consensus from both systems is that the process is hybrid – ChatGPT explains that "the model 're-reads' the entire sentence (vector), but 'writes' one word at a time (stream)" [RQ2/ChatGPT/3], and Copilot agrees, describing this as "vectorized processing per step, streamed generation across steps" [RQ2/Copilot/3]. This duality is central to transformer efficiency: parallel computation enables rich contextual encoding, while sequential autoregression ensures grammatical and logical continuity in generated text. Techniques such as key–value caching [39] further optimize this process by reusing previously computed representations, reducing computational redundancy without altering the underlying logic.

The variability and human-like quality of LLM outputs arise from probabilistic sampling and learned linguistic patterns. ChatGPT explains that "same prompt → same probabilities → different sampled paths → different outputs" [RQ2/ChatGPT/4] due to stochastic sampling as "different tokens can be chosen even with the same input" due to sampling mechanisms [RQ2/Copilot/4]. Parameters such as temperature and top-p control the degree of randomness, balancing coherence and diversity. Hence, human-like responses emerge from exposure to large corpora of human-generated text and alignment processes that shape tone and style as LLMs are statistical systems modelling linguistic distributions rather than possessing true understanding [33,34].

The phenomenon of hallucination further illustrates the probabilistic nature of these systems [40] as "the model is optimized to produce the most plausible next token, not the most truthful statement" [RQ2/ChatGPT/6]. Copilot concurs that this reflects "objective mismatch, gaps in training data, overgeneralization, sampling randomness, and lack of grounding" between training goals and user expectations [RQ2/Copilot/6]. Both systems identify contributing factors such as incomplete training data, overgeneralization, and sampling randomness. Adjusting parameters like temperature and top-p can reduce hallucination rates by constraining the probability space, though not eliminating errors entirely as LLMs may produce fluent but inaccurate outputs due to their reliance on statistical correlations rather than grounded knowledge [33].

Another important aspect of response generation is context accumulation, which functions as a form of short-term memory. ChatGPT clarifies that "adaptation happens through context accumulation, not parameter learning" [RQ2/ChatGPT/10]; a point endorsed by Copilot, which distinguishes between fixed parameters and dynamic context [RQ2/Copilot/10]. Each new

token is generated based on the entire available context, which includes the input prompt and previously generated tokens. However, this context is bounded by a fixed window size, limiting how much information the model can consider at once. As a result, coherence and continuity depend not only on model capacity but also on context length, which determines how much prior information can be retained during generation.

Thus, the emergence of human-like responses can be understood as the interaction of multiple components: training data provides patterns of human communication; model parameters encode these patterns; context length enables continuity; alignment shapes tone and behaviour; and sampling introduces variability. Therefore, the process of generating response from prompt can be conceptualized as a pipeline of tokenization, embedding, contextual encoding via attention, probabilistic next-token prediction, and iterative autoregressive generation, augmented by alignment and sampling mechanisms; as LLMs are transformer-based systems that model probability distributions over language and generate text through sequential sampling [2,28,34]. Despite intelligent or human-like responses, they are best understood as probabilistic reconstructions of linguistic patterns, produced through the interaction of high-dimensional representations and controlled randomness.

Theme 4: LLM is a Chinese room but emergent behaviour is possible (RQ3)

The characterization of LLMs as instantiations of the Chinese Room argument provides a useful, though not unproblematic, lens for understanding their operation. Across the dialogue, both ChatGPT and Copilot converge on a central claim: LLMs manipulate symbols according to learned statistical patterns without possessing intrinsic understanding. ChatGPT states that "I process symbols (tokens) using learned statistical relationships… There is no subjective awareness, intentionality, or grounded semantics" [RQ3/ChatGPT/1], while Copilot reinforces this by noting that "an LLM manipulates symbols (tokens) according to learned statistical rules without intrinsic access to meaning" [RQ3/Copilot/1]. This alignment reflects Searle's original contention that syntactic manipulation alone is insufficient for semantics [17,41].

In this analogy, the "rulebook" of the Chinese Room is reinterpreted in LLMs as the "trained parameters (weights) of the neural network" [RQ3/ChatGPT/2] emphasizing that these are not explicit, human-readable rules but "distributed numerical patterns." Copilot sharpens this further by describing the parameters as a "probabilistic guidebook" rather than a deterministic set of instructions [RQ3/Copilot/2]. This distinction is critical – unlike the discrete symbol manipulation envisioned by Searle [17], LLMs implement continuous, high-dimensional transformations learned from data [2,3]. Nevertheless, the structural analogy holds insofar as both systems transform input symbols into output symbols without recourse to grounded meaning.

A key point of triangulation concerns whether LLMs encode "knowledge" or merely linguistic structure. ChatGPT cautions that "it is not a direct representation of knowledge itself: It does not store facts as discrete, verifiable entries. It does not distinguish cleanly between: true vs false, common vs rare, factual vs fictional. Instead, it encodes what is likely to be said, not what is true" [RQ3/ChatGPT/3], while Copilot refines this to "a computational representation of linguistic probability space" [RQ3/Copilot/3]". This view is consistent with empirical findings that LLMs capture statistical regularities in language but do not reliably distinguish truth from falsehood [33]. From the Chinese Room perspective, this reinforces the idea that the system operates purely at the level of syntax, albeit in a highly sophisticated and probabilistic form.

However, both systems also highlight tensions that complicate a strict Chinese Room interpretation. Copilot notes that LLMs exhibit "emergent behaviours – reasoning-like patterns, analogies, contextual adaptation" [RQ3/Copilot/1] solely due to scale and complexity rather than any understanding – "outputs can be functionally indistinguishable from understanding in many contexts" [RQ3/ChatGPT/1]. Such observations are consistent with scaling laws and emergent capabilities in large models [42], suggesting that increasing model size and data exposure can produce qualitatively new behaviours not easily reducible to simple rule-following. Yet, these behaviours remain compatible with the Chinese Room framework if one adopts a functional interpretation: the system simulates understanding without possessing it. The generation of neologisms such as "Crustafarianism" (pseudo-religious system; which includes core beliefs, tenets, and prophets; created by autonomous AI agents on the Moltbook social network) further illustrates this point. ChatGPT argues that such terms arise from "novel combinations of learned patterns" [RQ3/ChatGPT/6],

while Copilot attributes them to "morphological recombination" and "statistical plausibility" [RQ3/Copilot/6]. These processes demonstrate that LLMs operate within a latent space of possible human expressions, generating coherent but ungrounded constructs. From a Chinese Room perspective, this underscores that even complex, seemingly meaningful outputs can emerge from purely formal manipulation of symbols.

Concluding Remarks

The findings present a consistent and coherent picture of large language models as statistically driven, language-based systems rather than cognitive agents. Both ChatGPT and Copilot characterize themselves as transformer-based models trained on large corpora to learn patterns in language, emphasizing their ability to generate fluent, contextually appropriate text while lacking consciousness, intention, or lived experience. A common computational pipeline to generate response from prompt (tokenization, embedding, attention-based contextualization, and probabilistic next-token prediction) highlighting that their responses arise from numerical transformations and probability distributions rather than semantic comprehension. These descriptions strongly support the interpretation of LLMs as systems consistent with the Chinese Room argument. However, scale, probabilistic modelling, and training diversity enable emergent behaviours that simulate understanding to a high degree of fidelity. Phenomena such as coherent reasoning, contextual adaptation, and even the generation of novel constructs (for example, "Crustafarianism") demonstrate that LLMs operate within a rich latent space of human language. Hence, LLMs are best understood as generative systems that model the structure of human linguistic expression – functionally powerful and epistemically useful, yet fundamentally grounded in probabilistic symbol manipulation rather than genuine understanding.

Supplementary Materials

The entire chat transcript can be downloaded at https://bit.ly/CAST2_transcript.

Conflict of Interest

The authors declare no conflict of interest.

Bibliography

1. Ouyang L., *et al*. "Training Language Models to Follow Instructions with Human Feedback". *Advances in Neural Information Processing Systems* 35 (2022): 27730-27744.

2. Vaswani A., *et al*. "Attention is All You Need". *Advances in Neural Information Processing Systems* 30 (2017).

3. Brown T., *et al*. "Language Models are Few-Shot Learners". *Advances in Neural Information Processing Systems* 33 (2020): 1877-1901.

4. Devlin J., *et al*. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (Association for Computational Linguistics, Minneapolis, Minnesota) (2019): 4171-4186.

5. Wang X., et al. "Self-Consistency Improves Chain of Thought Reasoning in Language Models". (2022).

6. Wei J., et al. "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models". Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22. (Curran Associates Inc., Red Hook, NY, USA) (2022).

7. Lipton ZC. "The Mythos of Model Interpretability: In Machine Learning, The Concept of Interpretability is Both Important and Slippery". *Queue* 16.3 (2018): 31-57.

8. Madsen A., et al. "Are Self-Explanations from Large Language Models faithful?" Findings of the Association for Computational Linguistics ACL 2024 (Association for Computational Linguistics, Bangkok, Thailand and virtual meeting) (2024): 295-337.

9. Belinkov Y and Glass J. "Analysis Methods in Neural Language Processing: A Survey". *Transactions of the Association for Computational Linguistics* 7 (2024): 49-72.

10. Rogers A., et al. "A Primer in BERTology: What We Know About How BERT Works". *Transactions of the Association for Computational Linguistics* 8 (2020): 842-866.

11. Turpin M., et al. "Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting". Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23. (Curran Associates Inc., Red Hook, NY, USA) (2023).

12. Lyu Q., et al. "Faithful Chain-of-Thought Reasoning". Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers) (Association for Computational Linguistics, Nusa Dua, Bali) (2023): 305-329.

13. Belinkov Y. "Probing Classifiers: Promises, Shortcomings, and Advances". *Computational Linguistics* 48.1 (2022): 207-219.

14. Sawyer RD and Norris J. "Duoethnography: Articulations/ (Re) Creation of Meaning in the Making". The Collaborative Turn: Working Together in Qualitative Research, ed Gershon WS (Sense Publishers) (2023).

15. Sung R-J., et al. "Constructing Biology Education Research Identities: A Duoethnography". *Frontiers in Education* 8 (2023): 1134040.

16. Shin AW., et al. "Science/Education Portraits VIII: Duoethnography of First-Generation Bioscience Undergraduates in a Private Education Institute in Singapore". *Acta Scientific Microbiology* 6.6 (2023): 24-35.

17. Searle JR. "Minds, Brains, and Programs". *Behavioral and Brain Sciences* 3.3 (1980): 417-424.

18. Kvale S and Brinkmann S. "Interviews: Learning the Craft of Qualitative Research Interviewing". (SAGE Publications) (2009).

19. Argyle LP., et al. "Out of One, Many: Using Language Models to Simulate Human Samples". *Political Analysis* 31.3 (2023): 337-351.

20. Dillion D., et al. "Can AI Language Models Replace Human Participants?" *Trends in Cognitive Sciences* 27.7 (2023): 597-600.

21. Binz M and Schulz E. "Using Cognitive Psychology to Understand GPT-3". *Proceedings of the National Academy of Sciences* 120.6 (2023): e2218523120.

22. Brailas A. "Postdigital Duoethnography: An Inquiry into Human-Artificial Intelligence Synergies". *Postdigital Science and Education* 6.2 (2024): 486-515.

23. Smit AP., et al. "Should We Be Going MAD? A Look at Multi-Agent Debate Strategies for LLMs". Proceedings of the 41st International Conference on Machine Learning, Proceedings of Machine Learning Research., eds Salakhutdinov R, Kolter Z, Heller K, Weller A, Oliver N, Scarlett J, Berkenkamp F (PMLR) 235 (2024): 45883-45905.

24. Ling MH. "APOD 0.1.0 - Agent Panel On-Demand for Structured Multi-Agent Dialogues. *Acta Scientific Computer Sciences* 7.8 (2025): 10-13.

25. Liu Y., et al. "The Truth Becomes Clearer Through Debate! - Multi-Agent Systems with Large Language Models Unmask Fake News". Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (ACM, Padua Italy) (2025): 504-514.

26. Liang T., et al. "Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate". Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics, Miami, Florida, USA) (2024): 17889-17904.

27. Triem H and Ding Y. "Tipping the Balance": Human Intervention in Large Language Model Multi-Agent Debate". *Proceedings of the Association for Information Science and Technology* 61.1 (2024): 361-373.

28. Tucudean G., *et al*. "Natural Language Processing with Transformers: A Review". *PeerJ Computer Science* 10 (2024): e2222.

29. Shao M., *et al*. "Survey of Different Large Language Model Architectures: Trends, Benchmarks, and Challenges". *IEEE Access* 12 (2024): 188664-188706.

30. Kotei E and Thirunavukarasu R. "A Systematic Review of Transformer-Based Pre-Trained Language Models through Self-Supervised Learning". *Information* 14.3 (2023): 187.

31. Shen X., *et al*. "Scaling Laws for Linear Complexity Language Models". Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics, Miami, Florida, USA) (2024): 16377-16426.

32. Ling MH. "Re-creating the Philosopher's Mind: Artificial Life from Artificial Intelligence". *iConcept Journal of Human-Level Intelligence* 3 (2012): 1.

33. Bender EM., *et al*. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (ACM, Virtual Event Canada) (2021): 610-623.

34. Han S., *et al*. "A Review of Large Language Models: Fundamental Architectures, Key Technological Evolutions, Interdisciplinary Technologies Integration, Optimization and Compression Techniques, Applications, and Challenges". *Electronics* 13.24 (2024): 5040.

35. Xiao C., *et al*. "Densing Law of LLMs". *Nature Machine Intelligence* 7.11 (2025): 1823-1833.

36. Winata GI., *et al*. "Preference Tuning with Human Feedback on Language, Speech, and Vision Tasks: A Survey". *Journal of Artificial Intelligence Research* 82 (2025): 2595-2661.

37. González Barman K., *et al*. "Reinforcement Learning from Human Feedback in LLMs: Whose Culture, Whose Values, Whose Perspectives?" *Philosophy and Technology* 38.2 (2025): 35.

38. Radford A., *et al*. "Language Models are Unsupervised Multitask Learners". *OpenAI blog* 1.8 (2019): 9.

39. Wu H and Tu K. "Layer-Condensed KV Cache for Efficient Inference of Large Language Models". Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (Association for Computational Linguistics, Bangkok, Thailand) (2024): 11175-11188.

40. Anh-Hoang D., *et al*. "Survey and Analysis of Hallucinations in Large Language Models: Attribution to Prompting Strategies or Model Behavior". *Frontiers in Artificial Intelligence* 8 (2025): 1622292.

41. Searle J. "Chinese Room Argument". *Scholarpedia* 4.8 (2009): 3100.

42. Lu S., *et al*. "Are Emergent Abilities in Large Language Models just In-Context Learning?" Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (Association for Computational Linguistics, Bangkok, Thailand) (2024): 5098-5139.