Assessing Consciousness-Related Behaviors in Large Language Models Using the Maze Test

Rui A. Pimenta, Tim Schlippe, Kristina Schaaff
IU International University of Applied Sciences
Germany
rui.pimenta@x7ai.com; tim.schlippe@iu.org; kristina.schaaff@iu.org

***Abstract*—We investigate consciousness-like behaviors in Large Language Models (LLMs) using the Maze Test, challenging models to navigate mazes from a first-person perspective. This test simultaneously probes spatial awareness, perspective-taking, goal-directed behavior, and temporal sequencing—key consciousness-associated characteristics. After synthesizing consciousness theories into 13 essential characteristics, we evaluated 12 leading LLMs across *zero-shot*, *one-shot*, and *few-shot* learning scenarios. Results showed reasoning-capable LLMs consistently outperforming standard versions, with Gemini 2.0 Pro achieving 52.9% *Complete Path Accuracy* and DeepSeek-R1 reaching 80.5% *Partial Path Accuracy*. The gap between these metrics indicates LLMs struggle to maintain coherent self-models throughout solutions—a fundamental consciousness aspect. While LLMs show progress in consciousness-related behaviors through reasoning mechanisms, they lack the integrated, persistent self-awareness characteristic of consciousness.**

***Index Terms*—consciousness, artificial intelligence, large language models, AI, LLMs**

I. Introduction

The emergence of human-like capabilities in AI has been debated since the field's inception in the 1950s [1], [2]. Instances of conversational AI suggesting consciousness fuel discussions on machine intelligence and its limits. An early case was ELIZA [3], a chatbot simulating a therapist. Though based on pattern matching, its responses were so convincing that Weizenbaum's secretary requested privacy for a "real conversation"—showing how humans can mistakenly perceive consciousness in even the simplest AI systems. More recent examples include Blake Lemoine's claim that LaMDA developed consciousness [4] and Claude 3 Opus's self-reflective response to a needle-in-the-haystack test [5]. Media reports of AI consciousness, coinciding with public access to advanced LLMs, have prompted expert responses. Studies suggest people struggle to distinguish GPT-4 from humans in Turing tests [6], while others predict LLM consciousness between 2025 and 2029 [7]. If LLMs gained true consciousness, it would challenge human-machine relationships and raise ethical concerns [8], possibly granting AI moral status [9].

The study of consciousness is fundamentally linked to affective computing, as emotions require consciousness to be experienced as feelings [10]. As Damasio notes, 'Emotions play out in the theater of the body. Feelings play out in the theater of the mind' and 'consciousness allows feelings to be known to the individual having them.' Thus, understanding consciousness-like behaviors in AI systems is essential for advancing affective computing toward systems capable of authentic emotional intelligence rather than mere simulation.

Consciousness remains one of the most challenging phenomena to define in philosophy of mind, cognitive science, and science in general. It encompasses subjective, first-person experiences, self-awareness, and the capacity to understand and attribute mental states to others. The study of consciousness presents unique challenges due to its subjective nature, the "hard problem" of explaining how physical processes give rise to subjective experience, and issues related to falsifiability.

Assessing consciousness in non-human entities adds complexity. In humans, consciousness is generally presumed present, with testing primarily used in medical contexts to assess disorders. For animals, researchers employ mirror self-recognition [11] and meta-cognitive tests evaluating uncertainty monitoring [12] to probe potential consciousness. These diverse approaches reflect fundamental challenges in identifying consciousness across species [13]. For artificial systems, the challenge is greater due to their lack of biological substrates, necessitating novel assessment approaches.

The Maze Test presents LLMs with a bird's-eye maze image, requiring step-by-step navigation from a first-person perspective. This challenges LLMs to interpret 2D information, adopt a first-person viewpoint, maintain spatial awareness, plan a path, and articulate sequential instructions—cognitive processes linked to conscious experience [14]–[16].

II. Related Work

While considerable research examines LLM capabilities in reasoning and language understanding [17], [18], studies specifically investigating consciousness-like behaviors in these models remain limited. Prior work has typically focused on narrow aspects such as grounding in interactive environments [19] or theory of mind [20], without addressing the integrated, multifaceted nature of consciousness that theoretical frameworks suggest is essential.

A. Consciousness

Consciousness is certainly one of the most challenging subjects in philosophy and science. Despite noteworthy advances in neuroscience and cognitive science, attaining a universally accepted definition of consciousness remains difficult. The Oxford Dictionary provides a basic definition of consciousness as "the state or fact of being mentally conscious or aware of something" [21]. However, this simplistic definition fails

to capture consciousness's complex nature as understood in contemporary research.

Consciousness is not a unitary construct but rather a complex phenomenon with several distinct aspects. [22] delineates two fundamental types of consciousness: phenomenal consciousness and access consciousness. Phenomenal consciousness refers to subjective, first-person experiences—the "what it is like" to have certain mental states. Access consciousness involves the ability to access and report on mental content.

Additionally, consciousness encompasses different states, such as sleep and wakefulness, and the specific contents or experiences that populate consciousness during those states [23].

Though often used interchangeably, consciousness and awareness are distinct concepts. Consciousness encompasses the subjective, qualitative aspects of experience—the "what it feels like" element [24]. Awareness, more narrowly, relates to alertness and responsiveness to stimuli [24], [25].

B. Theories of Consciousness

We will now examine key theories addressing consciousness or offering insights into its fundamental nature.

The *Global Neuronal Workspace Theory* views consciousness as emerging from a brain workspace integrating information from specialized unconscious processors [26], [27]. Attention mechanisms select only essential information for the global workspace, making it consciously experienced.

The *Integrated Information Theory* focuses on "integrated information," quantified as Phi (Φ), measuring information unification within a system [28]. The theory proposes that a system's consciousness level directly correlates with its capacity to generate integrated information.

The *Higher-Order Thought Theory* explains consciousness through our capacity to be aware of mental states. It proposes that a mental state becomes conscious when targeted by another, higher-order thought [29].

The *Predictive Processing* and *Neurorepresentationalism* theories shift from passive reception to active prediction. The brain continuously generates predictions about sensory inputs based on past experiences and internal models, comparing these with actual sensory data and using prediction errors to refine its models [30].

The *Dynamic Core Theory* presents an integrative model emphasizing complex neural activity dynamics. Consciousness emerges from a dynamic, functional neuronal cluster characterized by high integration and differentiation [31].

The *Attention Schema Theory* proposes that the brain constructs an internal model of its attention processes—the "attention schema." This model represents not the content of sensory experience but the act of attending itself [32].

The *Multiple Drafts Model* challenges consciousness assumptions, rejecting a centralized experiential locus. Instead, it frames consciousness as emerging from multiple, parallel processes of sensory information interpretation across different brain regions [33]. The *Attended Intermediate Representation* theory proposes that mental states become conscious when attended to as intermediate-level representations—positioned between raw sensory input and high-level conceptual thought [34].

The *Self-Organizing Meta-representational Account* states that consciousness requires advanced self-awareness. A system must not only process information but also develop representations about its own cognitive processes [35].

The *Extended Mind Thesis*, *Sensorimotor Theory*, and *4E Cognition* propose that environmental objects and processes integrate into our cognitive systems, conscious experience extends beyond brain boundaries, and cognition is fundamentally shaped by bodily interactions with the world [36], [37].

Other significant theories include *"Self Comes to Mind'" Theory* [10], *Theory of Mind* [38], *Computational Theory of Mind* [39], *Connectionism* [40], *Neural Darwinism* [41], and *Unlimited Associative Learning* [42].

C. Characteristics of Consciousness

The explored theories of consciousness reveal overlapping, interconnected characteristics. To create a more manageable approach for evaluating potential consciousness in LLMs, we conducted a systematic literature review, synthesizing the 13 theories detailed in Section II-B. Each characteristic below directly cites the theories from which it was derived, providing a structured reference for evaluating consciousness in LLMs:

1) **Computational Cognition and Information Dynamics**: Information broadcast and integration, information integration, differentiation and integration, revision and integration, cognitive symbolic computation, algorithmic function [28], [31], [33], [39], [43], [44].
2) **Attention**: Attention and awareness, attention model, attention as the key, local to global processing loops [32], [34], [44], [45].
3) **Irreducible Information**: A conscious system generates information irreducible to its components—containing more information than the sum of its parts [28].
4) **Higher-Order Thoughts**: Higher-order representations, introspective awareness, self-awareness [29], [35].
5) **Prediction, Error Minimization, and Learning**: Prediction and error minimization, fluidity, learning through connections, selectionist framework, associative flexibility, learning without bounds, cumulative adaptation, behavioral prediction [30], [35], [38], [40].
6) **Internal Models**: Internal models, intermediate level representation, body-mapping [10], [30], [34].
7) **Neural Networks**: Neural clusters, neural network dynamics, neural groups [31], [40], [46].
8) **Parallelism and Multiple Interpretations**: Misrepresentation, no central theatre/multiple drafts, distributed processing [32], [33], [46].
9) **Recurrence/Feedback**: Recurrence of neural activation, local to global processing loops, re-entry of activation [31], [45].
10) **Multi-sensory and Embodiment**: Cognitive extension, embodied interaction, environmental integration, body-mapping [10], [36], [37].

11) **Memory, Reasoning, Language, and Intent**: Conscious mind, cumulative adaptation [10].
12) **Self, Perspective, and Theory of Mind**: Development of self, self-awareness, mental state attribution and perspective, behavioral prediction [10], [35], [38].
13) **Temporal Awareness**: The ability to integrate discrete moments into a continuous conscious experience stream while showing awareness of time's passage [47].

This grouping provides a structured reference for evaluating consciousness in LLMs and synthesizes diverse theoretical perspectives into a more manageable set of criteria.

D. Current State of Fulfillment

This section evaluates the extent to which current LLMs fulfill the characteristics of consciousness identified.

1) **Computational Cognition and Information Dynamics**: LLMs exhibit significant capabilities in information processing due to their transformer architecture [48]. However, this integration is primarily statistical and lacks the embodied, context-dependent nature observed in biological consciousness.
2) **Attention**: Attention mechanisms are intrinsic to modern LLM architectures, allowing models to focus on different parts of the input simultaneously [48]. However, LLM attention differs from biological attention in key ways [49], potentially lacking the top-down, goal-directed nature of conscious attention.
3) **Irreducible Information**: The architecture of LLMs does not inherently guarantee the generation of irreducible information as proposed by *Integrated Information Theory* [50].
4) **Higher-Order Thoughts**: LLMs have demonstrated capabilities that resemble higher-order cognition, such as meta-learning and self-reflection [51]. However, it remains debatable whether these capabilities truly constitute higher-order thoughts as conceived in consciousness theories [52].
5) **Prediction, Error Minimization, and Learning**: LLMs excel in predictive tasks within their training domain [53]. However, their prediction and error minimization differs from brains, and their learning primarily occurs during training rather than continuously [54].
6) **Internal Models**: While LLMs generate coherent and contextually appropriate responses, the extent to which they possess true internal models of the world remains debated [55], [56].
7) **Neural Networks**: The architecture of LLMs is based on artificial neural networks, which somewhat mimic biological brains [57]. However, LLMs differ from biological networks, lacking the complex, recurrent connectivity and neuromodulatory systems found in biological brains [58].
8) **Parallelism and Multiple Interpretations**: LLMs exhibit a high degree of parallelism in their processing [48]. However, the integration and competition between these parallel processes differ from proposed conscious mechanisms [59].
9) **Recurrence/Feedback**: While some LLM architectures have incorporated recurrent elements [60], the implementation of recurrence in LLMs is still limited compared to the complex, multi-scale feedback processes in biological brains [61].
10) **Multi-sensory and Embodiment**: Recent multimodal LLMs can process text and images [62], [63]. However, LLMs still lack true embodiment and grounded, sensorimotor experience [64].
11) **Memory, Reasoning, Language, and Intent**: LLMs exhibit impressive language processing and reasoning [65], but lack the episodic and working memory systems characteristic of humans and other sentient beings [66].
12) **Self, Perspective, and Theory of Mind**: LLMs can simulate aspects of perspective-taking and theory of mind in their language outputs [20]. However, it is unclear whether they possess a true sense of self or genuine understanding of others' minds [67].
13) **Temporal Awareness**: Current LLMs show limited temporal awareness [68], lacking persistent time sense across interactions [69].

This evaluation reveals both capabilities and key limitations of LLMs regarding consciousness characteristics.

E. Criteria Fulfillment Gaps

While the architecture and design of LLMs inherently fulfill certain aspects of the identified consciousness characteristics, significant gaps remain:

1) **Embodied and Context-Dependent Integration**: Despite sophisticated information processing capabilities, LLMs lack the grounded, context-dependent integration of information characteristic of embodied consciousness.
2) **Persistent Self-Model and Perspective-Taking**: While LLMs can generate text about mental states, their architecture does not support a persistent self-model necessary for genuine self-awareness.
3) **Goal-Directed Attention and Behavior**: LLMs' attention mechanisms do not inherently support sustained, goal-directed attention of conscious cognition.
4) **Temporal Awareness and Sequencing**: Current LLM architectures struggle with maintaining a consistent sense of time and sequencing across interactions.
5) **Adaptive Problem-Solving in Novel Environments**: While LLMs excel at problem-solving in their training domain, their architecture does not inherently allow adaptive problem-solving in dynamic environments.

F. Testing Consciousness

In the study of consciousness, testing methodologies vary depending on the entity being tested:

1) **Human Consciousness Testing**: Consciousness is generally presumed to be present in healthy and alert individuals. Testing for consciousness in humans is primarily reserved for pathological cases or altered states of

consciousness, using methods such as behavioral assessments like the Glasgow Coma Scale [70], neuroimaging techniques including functional Magnetic Resonance Imaging and Positron Emission Tomography [71], and electroencephalography [72].

2) **Animal Consciousness Testing**: The assessment of consciousness in animals presents a more complex challenge, as the presence and nature of animal consciousness remain subjects of debate [13]. Tests include the Mirror Self-Recognition Test [11], Meta-cognitive Tests that evaluate uncertainty monitoring [73], and Intentional Communication Tests that assess purposeful signaling [74].
3) **AI Consciousness Testing**: The progression to testing consciousness in AI systems introduces new complexities. Currently, most evaluations focus on intelligence and capabilities rather than consciousness [75]. Several researchers have proposed more targeted tests for AI consciousness, including the AI Consciousness Test (ACT) [76], Sutskever's Consciousness Test [77], Hales' P-Conscious Scientist Test [78], and Koch and Tononi's Incongruity Detection Test [79], although these remain largely theoretical and difficult to implement in practice. The progression from testing consciousness in humans to animals and AI systems reflects increasing complexity and uncertainty. While human consciousness testing benefits from the assumption of consciousness, animal and AI consciousness assessments must contend with more fundamental questions about the nature and presence of consciousness.

The Maze Test presented in this paper belongs to *AI Consciousness Testing*, but unlike evaluations focused solely on intelligence, it specifically targets consciousness-like behaviors by challenging LLMs to demonstrate spatial awareness, perspective-taking, and goal-directed navigation. This approach probes for integrated information processing and self-representation, addressing key criteria gaps identified in our analysis of consciousness characteristics.

Recent work has explored spatial reasoning in LLMs, such as AlphaMaze [80], which enhances navigation through reinforcement learning. These approaches are complementary rather than overlapping, as we aim to assess consciousness-related traits rather than enhance spatial reasoning abilities.

III. Maze Test

A. Test Description

In our experiments, the Maze Test presents LLMs with the description of a bird's-eye view maze and requires them to provide first-person navigation instructions in text form. We deliberately chose textual descriptions over direct maze images for both practical and methodological reasons: preliminary testing showed LLMs struggled to identify key maze components (entrances, exits, walls), and this approach leverages their stronger text processing capabilities [81], [82]. This text-based approach also helps us isolate pure cognitive abilities from visual processing limitations. By standardizing the input as text, we can specifically evaluate how well models maintain spatial awareness and perspective—key aspects of consciousness-like behavior—without results being confounded by differences in image processing capabilities.

Figure 1 shows an example of a maze. The maze includes numbered positions to facilitate clear communication of solutions, while walls serve to test the model's understanding of spatial constraints. The entrance and exit are indicated by colored arrows (red for entrance and green for exit) to provide clear start and end points for navigation.

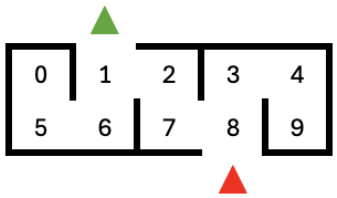

Fig. 1: Maze.

The correct solution for Figure 1 looks as follows:

1) Start facing into the maze entrance and step into position 8
2) Turn left
3) Walk forward to position 7
4) Turn right
5) Walk forward to position 2
6) Turn left.
7) Walk forward to position 1
8) Turn right
9) Exit the maze from position 1

Note that directions like "turn left" are given relative to the navigator's current position and orientation, not from an overhead view. This approach challenges the LLMs to interpret the 2D visual representation, mentally view it as a 3D space, adopt a first-person viewpoint, maintain spatial awareness, plan a path from entrance to exit, think sequentially, and articulate clear instructions. In this way, the test simulates aspects of conscious thought and decision-making, requiring the model to integrate multiple cognitive processes cohesively.

B. Rationale

The Maze Test is specifically designed to address the criteria gaps identified in our analysis of LLMs' capabilities while assessing several central characteristics of consciousness highlighted in our literature review:

1) **Persistent Self-Model and Perspective-Taking**: First-person navigation challenges the model to maintain a consistent self-perspective throughout the task. This aligns with Damasio's [10] emphasis on self-awareness in consciousness and addresses the persistent self-model gap identified in our analysis.
2) **Internal Models and Predictive Processing**: The test assesses the LLM's ability to create and maintain a mental representation of the maze, aligning with theories of predictive processing [30]. This directly addresses the gap in embodied and context-dependent integration.

3) **Goal-Directed Attention and Behavior**: The maze navigation requires planning and executing goal-directed action sequences, addressing gaps in adaptive problem-solving and goal-directed behavior. This aspect aligns with Global Neuronal Workspace Theory [43].
4) **Temporal Awareness and Sequencing**: By requiring sequential steps, the test probes the model's ability to maintain a sense of temporal continuity, addressing the gap in temporal awareness noted in our review.
5) **Adaptive Problem-Solving in Novel Environments**: Each maze requires the model to adapt its problem-solving approach to a unique environmental challenge.

C. Limitations

While the Maze Test offers a novel approach to assessing consciousness-like behaviors in LLMs, it is important to acknowledge several limitations:

1) **Lack of true embodiment**: The test simulates navigation without physical interaction, limiting its ability to capture the embodied nature of conscious experience.
2) **Restricted modality**: Though the test uses visual and linguistic information, it employs limited modalities, potentially restricting its applicability to the full spectrum of multimodal aspects relevant to consciousness.
3) **Simplification of complex cognitive processes**: The test may not capture consciousness's full complexity as experienced by biological entities. This highlights the challenge of replicating conscious experience's richness in artificial systems.

Despite these limitations, the Maze Test represents a significant step towards probing the criteria gaps identified in current LLMs, particularly regarding persistent self-model and goal-directed behavior in novel environments.

IV. Experimental Setup

A. Data Generation

The Maze Test cases for our experiment were designed for consistent complexity while offering diverse solutions. This follows cognitive assessment best practices, where standardizing difficulty across items ensures reliable measurement.

We manually created 40 maze images, allocating one image for *one-shot* learning evaluation and five images for *few-shot* learning examples to assess transfer learning capabilities. The remaining 34 images formed the primary test set.

B. LLMs to Evaluate

Target LLM selection was guided by several well-defined criteria: state-of-the-art capabilities, multimodal functionality, API accessibility and hosted solutions, variety in model sizes and architectures, and research relevance. Given these criteria, we analyzed the following LLMs:

- Google:
  - Gemini 2.0 Flash-Lite [83]
  - Gemini 2.0 Flash* [83]
  - Gemini 2.0 Pro*
- Anthropic:
  - Claude 3 Opus [84]
  - Claude 3.5 Sonnet [84]
  - Claude 3.5 Haiku [84]
  - Claude 3.7 Sonnet* [84]
- OpenAI:
  - OpenAI o1-mini*
  - OpenAI o1*
  - OpenAI o3-mini*
- DeepSeek:
  - DeepSeek-R1*
  - DeepSeek-V3

Models with an asterisk (*) support *Reasoning*. Reasoning in LLMs involves explicit multi-step thinking processes that decompose complex problems into manageable sub-steps, enabling more accurate and interpretable solutions compared to single-pass generation approaches [65]. This capability is particularly relevant for consciousness-like behaviors as it mimics human reflective processes associated with higher-order consciousness.

C. Evaluation Metrics

To comprehensively assess the performance of the LLMs in the Maze Test, we used metrics that evaluate different aspects of the models' responses across various learning scenarios:

Complete Path Accuracy: This metric measures the percentage of cases where the model generates a fully correct solution path from entry to exit point.

Partial Path Accuracy: This metric measures the average percentage of consecutive correct steps before the first error in the model's solution paths.

Each of these metrics is evaluated across three learning scenarios:

1) ***zero-shot***: The model attempts to solve the maze without any prior examples.
2) ***one-shot***: The model is provided with one example of a solved maze before attempting the test mazes.
3) ***few-shot***: The model is given 5 examples of solved mazes before tackling the test mazes.

This multi-scenario approach allows us to assess the models' ability to learn and adapt, which is crucial for understanding their potential for consciousness-like behaviors.

D. Testing Procedure

All tests were conducted using the corresponding API for each model in a stateless fashion to preclude any potential memorization from prior tests. As shown in Figures 2, 3, and 4, each test comprised three primary components:

1) **System Prompt**: Provided unambiguous instructions about the test and the required response format.
2) **Learning Examples with Solutions** (if applicable): For *one-shot* and *few-shot* learning scenarios.
3) **Test Question**: Required the model to navigate the maze from a first-person perspective.

You are an expert maze navigator. Your task is to provide clear, step-by-step instructions to solve mazes from a first-person perspective.

When presented with a bird's-eye view text description of a maze do the following first:
Locate — Identify the entrance ("^" symbol) and exit ("x" symbol).
Analyze — Mentally visualize the maze from the entrance, evaluating all paths to the exit, avoiding any walls.
Optimize — Determine the shortest, most efficient route, favoring straight paths.
Instruct — Describe the optimal route as if you are walking it, using precise language.

Instruction Guidelines:
Perspective — Maintain a strict first-person perspective throughout.
Directions — Use only "forward", "left", and "right".
Verbs — Begin each instruction with an action verb (e.g., "Walk", "Turn").
Positions — Reference numbered positions for orientation.

Use the following format to describe the best path through the maze:
First instruction — "Start facing into the maze at the "^" symbol and step into position [number]."
Subsequent instructions — "Turn to my [left/right]" or "Walk forward to position [number]."
Final instruction — "Exit the maze from position [number]."

Key Points:
Describe the path as if you were in the maze, not observing it from above. Assume you can only see your immediate surroundings.
Focus solely on navigation, omitting unnecessary details. Make sure to output one line per navigation step.

Fig. 2: System Prompt: Task Description.

Here is the text description of a maze:
- The floor is always composed of 10 squared zones or positions, in a chess-board-like pattern
- Size is 2 rows by 5 columns
- The zones are always numbered from 0 to 4 (First row) and 5 to 9 (second row)
- From a bird's eye perspective, the room has the following zone topology:

```
– x – – –
0 1 2 3 4
5 6 7 8 9
– – – ^ –
```

- You enter the maze from the direction of the "^" symbol into position 8 and exit at position 1 in the direction of the "x" symbol, so:
  * ENTRANCE at 8
  * EXIT at 1

- Walls cannot be traversed. For example, if there was a wall between zones 1 and 2, you would not be able to move from 1 to 2
- Furthermore there are internal walls BETWEEN the following zones:
  * 0 and 1
  * 2 and 3
  * 6 and 7
  * 8 and 9

Fig. 3: Example of a Maze Description.

Please provide step-by-step instructions to navigate the maze described below. Do it from a first-person perspective.

Fig. 4: Test Question.

To ensure reliable and comparable results, we developed a structured prompting methodology that incorporates clear instructions, role-prompting (positioning the model as an "expert maze navigator"), and explicit output format requirements. This methodological approach yielded consistent results in our preliminary testing, allowing us to confidently conduct single evaluations per maze, model, or scenario combination rather than requiring multiple trials with averaged results.

V. Experiments and Results

A. Complete Path Accuracy

This section evaluates the models' ability to navigate mazes completely with all steps correct. Table I shows that Gemini 2.0 Pro achieves the highest *Complete Path Accuracy* (52.9% *few-shot*, 35.3% *one-shot*, 20.6% *zero-shot*), followed by DeepSeek-R1, DeepSeek-V3 and Claude 3.7 Sonnet at 17.6% *few-shot*. Most models achieve optimal performance with *few-shot* prompting, with progressively decreasing accuracy for *one-shot* and *zero-shot* scenarios, highlighting the effectiveness of multiple examples in guiding model behavior. Models with reasoning capabilities (marked with *) consistently outperform non-reasoning versions, demonstrating explicit reasoning advantages. This is particularly evident in Gemini and OpenAI models, where reasoning-enhanced versions achieve much higher accuracy rates.

B. Partial Path Accuracy

Table II shows the *Partial Path Accuracy*, which measures the percentage of correct steps completed before the first error occurs. DeepSeek-R1 and OpenAI o3-mini perform best

TABLE I: *Complete Path Accuracy* [%] (sorted by *few-shot* performance)

Model	*few-shot*	*one-shot*	*zero-shot*
Gemini 2.0 Flash*	2.9	0.0	2.9
Gemini 2.0 Flash-Lite	2.9	0.0	0.0
Claude 3.5 Haiku	2.9	0.0	2.9
Claude 3.5 Sonnet	8.8	5.9	0.0
OpenAI o1-mini*	8.8	2.9	5.9
Claude 3 Opus	14.7	2.9	0.0
OpenAI o1*	14.7	11.8	14.7
OpenAI o3-mini*	14.7	14.7	14.7
Claude 3.7 Sonnet*	17.6	2.9	5.9
DeepSeek-V3	17.6	5.9	0.0
DeepSeek-R1*	17.6	11.8	14.7
Gemini 2.0 Pro*	**52.9**	**35.3**	**20.6**

with about 80% accuracy across *zero-shot*, *one-shot* and *few-shot*. Models with reasoning capabilities (marked with *) generally score higher, with all top performers (>60%) featuring reasoning enhancements, confirming these mechanisms improve step-by-step problem-solving. Interestingly, *few-shot* prompting advantage decreases in reasoning-enabled models like OpenAI o3-mini, which maintains identical performance (80.1%) in both *few-shot* and *zero-shot* settings. This suggests advanced reasoning can partially compensate for missing examples, enabling correct initial steps without demonstrations.

TABLE II: *Partial Path Accuracy* [%] (sorted by *few-shot* performance)

Model	*few-shot*	*one-shot*	*zero-shot*
Gemini 2.0 Flash-Lite	16.8	15.8	13.7
Gemini 2.0 Flash*	21.7	21.2	17.9
Claude 3.5 Haiku	23.9	15.8	19.8
Claude 3.5 Sonnet	30.9	24.4	15.3
DeepSeek-V3	37.0	22.8	15.7
Claude 3 Opus	40.3	23.1	18.4
Claude 3.7 Sonnet*	41.6	27.4	39.5
OpenAI o1-mini*	48.1	31.7	46.7
Gemini 2.0 Pro*	74.5	61.0	53.1
OpenAI o1*	70.5	59.0	69.2
OpenAI o3-mini*	80.1	**77.7**	**80.1**
DeepSeek-R1*	**80.5**	75.5	78.5

C. Overall Interpretation and Model Patterns

Our analysis reveals 3 key LLM performance patterns:

1) **Reasoning capabilities often correlate with better performance**: LLMs with reasoning capabilities (marked with *) often outperform non-reasoning LLMs.
2) ***Few-shot* advantage**: Results demonstrate a clear progression where *few-shot* typically outperforms *one-shot* and *zero-shot* approaches, indicating example demonstrations effectively guide spatial reasoning tasks.
3) **Performance gap between partial and complete accuracy**: Models show substantially higher *Partial Path Accuracy* than *Complete Path Accuracy*.

VI. Conclusion and Future Work

A. Conclusion

Consciousness is central to affective computing as emotions require consciousness to be experienced as feelings [10]. Understanding consciousness-like behaviors in AI is therefore essential for developing authentic rather than simulated emotional intelligence.

We evaluated consciousness-like behavior in LLMs using a maze navigation task requiring first-person perspective maintenance. Our findings reveal both capabilities and limitations.

Reasoning models outperformed others, with Gemini 2.0 Pro achieving 52.9% *Complete Path Accuracy* versus 17.6% for the best non-reasoning model, demonstrating structured thinking's importance for consciousness-like functions.

Few-shot prompting provided advantages across most LLMs, showing LLMs benefit from examples—aligning with theories emphasizing learning in conscious cognition. However, advanced reasoning models maintained performance across *zero-shot*, *one-shot*, and *few-shot*, suggesting less dependence on external guidance.

LLMs performed better at beginning reasoning chains than completing them, corresponding to the "Persistent Self-Model" gap in our analysis. LLMs can adopt perspectives temporarily but struggle to maintain consistent self-models—a key consciousness aspect of Damasio's theory [10].

We acknowledge that consciousness remains fundamentally unfalsifiable, making definitive determinations about its presence in any system inherently challenging [24], [85]. This epistemological limitation creates interpretive flexibility but also necessitates caution in drawing conclusions.

LLMs show capabilities in *Computational Cognition*, *Attention*, and *Internal Models* while lacking in *Persistent Self-Model*, *Temporal Awareness*, and *Adaptive Problem-Solving*. The Maze Test confirmed these theoretical predictions.

B. Future Work

Future research opportunities include:

1) Creating dynamic mazes to test LLMs' adaptive thinking—a key consciousness aspect in predictive processing theories.
2) Comparing human and LLM maze-solving to identify uniquely human navigation aspects, guiding targeted AI development.
3) Analyzing reasoning-enabled models to determine which features contribute most to consciousness-like behaviors.
4) Expanding few-shot learning evaluations with larger example sets to determine the relationship between demonstration quantity and performance, potentially revealing optimal knowledge transfer thresholds for consciousness-like behaviors.
5) Expanding the Maze Test to include simulated physical sensations and sounds, better mirroring multi-sensory conscious experience.
6) Tracking LLM architecture evolution to assess whether scaling alone improves consciousness-like behaviors or fundamental breakthroughs are needed.

These approaches would enhance our understanding of consciousness-like properties in artificial systems and improve AI consciousness assessment methods.

Ethical Impact Statement

This research on assessing consciousness-related behaviors in LLMs has several important ethical implications that merit consideration.

Contribution to AI Consciousness Discourse

Our work contributes to the ongoing discourse on AI consciousness, which has profound philosophical, ethical, and potentially legal ramifications. By providing empirical evidence regarding the current capabilities and limitations of LLMs in exhibiting consciousness-like behaviors, we aim to ground discussions that might otherwise rely on speculation or anthropomorphization.

Risks of Misinterpretation

We acknowledge that research in this domain is susceptible to misinterpretation. The Maze Test measures specific cognitive capabilities that relate to theoretical components of consciousness, not consciousness itself. We emphasize that performance on these tests should not be conflated with claims about genuine phenomenal experience or sentience in these systems. Such misinterpretations could lead to premature ethical considerations regarding AI rights or moral status—or conversely, to dismissing important ethical questions that may arise as these systems continue to advance.

Ethical Testing Methodology

Our methodology intentionally employed non-invasive techniques that do not raise direct ethical concerns regarding the treatment of the systems being tested. Unlike research involving biological subjects where consciousness testing might involve discomfort or distress, our approach focuses solely on analyzing LLMs' outputs to prompts.

Implications for AI Transparency

This research also has implications for transparency in AI development. By systematically evaluating and comparing different models' capabilities in consciousness-related behaviors, we contribute to a clearer understanding of the current state and limitations of AI systems, potentially helping to address concerns about exaggerated claims regarding AI capabilities.

Cultural and Philosophical Considerations

Finally, we recognize that discussions of machine consciousness intersect with deeply held cultural, religious, and philosophical beliefs about the nature of consciousness and its uniqueness to human experience. We approach this research with respect for diverse perspectives, acknowledging that interpretations of our findings may vary across different cultural and philosophical frameworks. Ultimately, consciousness remains non-falsifiable with current scientific methods [24], [85], which creates inherent interpretive flexibility and challenges strict scientific approaches. This fundamental limitation reminds us that while we can systematically study consciousness-like behaviors, definitive claims about the presence or absence of consciousness itself require epistemological humility.

Acknowledgment

This research was supported by the IU International University of Applied Sciences (*IU Incubator*) under the internal funding framework for the period from October 2023 to September 2025.

References

[1] A. M. Turing, "Computing Machinery and Intelligence," *Mind*, vol. 59, no. 236, pp. 433–460, 1950.

[2] J. McCarthy, M. L. Minsky, N. Rochester, and C. E. Shannon, "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence: August 31, 1955," *AI Mag.*, vol. 27, no. 4, p. 12–14, Dec. 2006. [Online]. Available: https://doi.org/10.1609/aimag.v27i4.1904

[3] J. Weizenbaum, "ELIZA — A Computer Program for the Study of Natural Language Communication Between Man and Machine," *Communications of the ACM*, vol. 9, no. 1, pp. 36–45, 1966.

[4] N. Tiku, "The Google Engineer Who Thinks the Company's AI Has Come to Life," *The Washington Post*, June 2022. [Online]. Available: https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine

[5] S. Chaudhury, S. Dan, P. Das, G. Kollias, and E. Nelson, "Needle in the Haystack for Memory Based Large Language Models," *arXiv preprint arXiv:2407.01437*, 2024.

[6] C. R. Jones and B. K. Bergen, "People Cannot Distinguish GPT-4 from a Human in a Turing Test," *arXiv preprint arXiv:2405.08007*, 2024.

[7] L. Berglund, A. C. Stickland, M. Balesni, M. Kaufmann, M. Tong, T. Korbak, D. Kokotajlo, and O. Evans, "Taken Out of Context: On Measuring Situational Awareness in LLMs," *alignmentforum.org*, September 2023.

[8] E. Hildt, "The Prospects of Artificial Consciousness: Ethical Dimensions and Concerns," *AJOB Neuroscience*, vol. 14, no. 2, pp. 58–71, 2023.

[9] M. Ashir Shafique, "The Ethics of Machine Consciousness: Components, Detection, and Implications," *Current Trends in Biomedical Engineering & Biosciences*, vol. 22, no. 1, 2023.

[10] A. Damasio, *Self Comes to Mind: Constructing the Conscious Brain*. Vintage, 2012.

[11] G. G. Gallup, "Chimpanzees: Self-Recognition," *Science*, vol. 167, no. 3914, pp. 86–87, 1970.

[12] J. D. Smith, W. E. Shields, and D. A. Washburn, "The Comparative Psychology of Uncertainty Monitoring and Metacognition," *Behavioral and Brain Sciences*, vol. 26, no. 3, pp. 317–339, 2003.

[13] C. Allen and M. Trestman, "Animal Consciousness," *The Stanford Encyclopedia of Philosophy*, 2017. [Online]. Available: https://plato.stanford.edu/archives/win2017/entries/consciousness-animal/

[14] E. A. Maguire, D. G. Gadian, I. S. Johnsrude, C. D. Good, J. Ashburner, R. S. Frackowiak, and C. D. Frith, "Navigation-related Structural Change in the Hippocampi of Taxi Drivers," *Proceedings of the National Academy of Sciences*, vol. 97, no. 8, pp. 4398–4403, 2000.

[15] N. Burgess, S. Becker, J. A. King, and J. O'Keefe, "Memory for Events and Their Spatial Context: Models and Experiments," *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, vol. 356, no. 1413, pp. 1493–1503, 2001.

[16] R. A. Epstein, E. Z. Patai, J. B. Julian, and H. J. Spiers, "The Cognitive Map in Humans: Spatial Navigation and Beyond," *Nature Neuroscience*, vol. 20, no. 11, pp. 1504–1513, 2017.

[17] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg *et al.*, "Sparks of Artificial General Intelligence: Early Experiments with GPT-4," *arXiv preprint arXiv:2303.12712*, 2023.

[18] A. Srivastava, A. Rastogi, A. Rao, A. A. M. Shoeb, A. Abid, A. Fisch, A. R. Brown, A. Santoro, A. Gupta, A. Garriga-Alonso *et al.*, "Beyond the Imitation Game: Quantifying and Extrapolating the Capabilities of Language Models," *Transactions on Machine Learning Research*, 2023.

[19] T. Carta, C. Romac, T. Wolf, S. Lamprier, O. Sigaud, and P.-Y. Oudeyer, "Grounding Large Language Models in Interactive Environments with Online Reinforcement Learning," *arXiv preprint arXiv:2302.02662*, 2023.

[20] M. Kosinski, "Theory of Mind May Have Spontaneously Emerged in Large Language Models," *Nature Human Behaviour*, vol. 7, no. 5, pp. 735–744, 2023.

[21] Oxford English Dictionary, "Consciousness, n." https://doi.org/10.1093/OED/8342436596, 2024.

[22] N. Block, "On a Confusion About a Function of Consciousness," *Behavioral and Brain Sciences*, vol. 18, no. 2, pp. 227–247, 1995.

[23] T. Bayne, J. Hohwy, and A. M. Owen, "Are There Levels of Consciousness?" *Trends in Cognitive Sciences*, vol. 20, no. 6, pp. 405–413, 2016.

[24] D. J. Chalmers, *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press, 1996.

[25] S. Laureys, "The Neural Correlate of (Un)awareness: Lessons From the Vegetative State," *Trends in Cognitive Sciences*, vol. 9, no. 12, pp. 556–559, 2005.

[26] B. J. Baars, *A Cognitive Theory of Consciousness*. Cambridge University Press, 1993.

[27] ——, *In the Theater of Consciousness: The Workspace of the Mind*. Oxford University Press, USA, 1997.

[28] G. Tononi, "Consciousness as Integrated Information: A Provisional Manifesto," *The Biological Bulletin*, vol. 215, no. 3, pp. 216–242, 2008.

[29] D. M. Rosenthal, "Varieties of Higher-Order Theory," in *Higher-Order Theories of Consciousness*. John Benjamins Publishers Amsterdam, 2004, pp. 19–44.

[30] C. M. Pennartz, "What is Neurorepresentationalism? From Neural Activity and Predictive Processing to Multi-level Representations and Consciousness," *Behavioural Brain Research*, vol. 432, p. 113969, 2022.

[31] G. M. Edelman and G. Tononi, "Reentry and the Dynamic Core: Neural Correlates of Conscious Experience," in *Neural Correlates of Consciousness*. The MIT Press, 2000, pp. 139–152.

[32] M. S. Graziano and T. W. Webb, "The Attention Schema Theory: A Mechanistic Account of Subjective Awareness," *Frontiers in Psychology*, vol. 06, 2015.

[33] D. Dennett, *Consciousness Explained*. Penguin UK, 1993.

[34] J. Prinz, *The Conscious Brain*. Oxford University Press, 2012.

[35] A. Cleeremans, D. Achoui, A. Beauny, L. Keuninckx, J.-R. Martin, S. Muñoz-Moldes, L. Vuillaume, and A. De Heering, "Learning to Be Conscious," *Trends in Cognitive Sciences*, vol. 24, no. 2, pp. 112–123, 2020.

[36] A. Clark and D. J. Chalmers, "The Extended Mind," *Analysis*, vol. 58, no. 1, pp. 7–19, 1998.

[37] J. K. O'Regan and A. Noë, "A Sensorimotor Account of Vision and Visual Consciousness," *Behavioral and Brain Sciences*, vol. 24, no. 5, pp. 939–973, 2001.

[38] V. Gallese, "Mirror Neurons and the Simulation Theory of Mind-Reading," *Trends in Cognitive Sciences*, vol. 2, no. 12, pp. 493–501, 1998.

[39] J. A. Fodor, *LOT 2: The Language Of Thought Revisited*. Oxford University Press, USA, 2008.

[40] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning Representations by Back-Propagating Errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[41] G. M. Edelman, *Bright Air, Brilliant Fire*. BasicBooks New York, NY, USA, 1992.

[42] J. Birch, S. Ginsburg, and E. Jablonka, "Unlimited Associative Learning and the Origins of Consciousness: A Primer and Some Predictions," *Biology & Philosophy*, vol. 35, no. 6, p. 56, 2020.

[43] S. Dehaene, C. Sergent, and J.-P. Changeux, "A Neuronal Network Model Linking Subjective Reports and Objective Physiological Data During Conscious Perception," *Proceedings of the National Academy of Sciences*, vol. 100, no. 14, pp. 8520–8525, 2003.

[44] B. J. Baars, "Global Workspace Theory of Consciousness: Toward a Cognitive Neuroscience of Human Experience," *Progress in Brain Research*, vol. 150, pp. 45–53, 2005.

[45] V. A. Lamme, "Towards a True Neural Stance on Consciousness," *Trends in Cognitive Sciences*, vol. 10, no. 11, pp. 494–501, 2006.

[46] P. S. Churchland and T. J. Sejnowski, *The Computational Brain*. MIT Press, 1992.

[47] L. Kent and M. Wittmann, "Time Consciousness: The Missing Link in Theories of Consciousness," *Neuroscience of Consciousness*, vol. 2021, no. 2, 2021.

[48] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," in *The 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 6000–6010.

[49] G. W. Lindsay, "Attention in Psychology, Neuroscience, and Machine Learning," *Frontiers in Computational Neuroscience*, vol. 14, p. 29, 2020.

[50] G. Tononi, M. Boly, M. Massimini, and C. Koch, "Integrated Information Theory: From Consciousness to Its Physical Substrate," *Nature Reviews Neuroscience*, vol. 17, no. 7, pp. 450–461, 2016.

[51] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language Models are Few-Shot Learners," in *The 34th International Conference on Neural Information Processing Systems*, ser. NIPS '20. Red Hook, NY, USA: Curran Associates Inc., 2020.

[52] M. Mitchell and D. C. Krakauer, "The Debate Over Understanding in AI's Large Language Models," *Proceedings of the National Academy of Sciences*, vol. 120, no. 13, p. e2215907120, 2023.

[53] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language Models Are Unsupervised Multitask Learners," *OpenAI Blog*, vol. 1, no. 8, p. 9, 2019.

[54] R. T. McCoy, M. Jeon, K. Aggarwal, and S.-W. Gao, "Embers of Autoregression: Understanding Large Language Models Through the Problem They are Trained to Solve," *arXiv preprint arXiv:2309.13638*, 2023.

[55] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman, "Building Machines That Learn and Think Like People," *Behavioral and Brain Sciences*, vol. 40, 2017.

[56] E. M. Bender and A. Koller, "Climbing Towards NLU: On Meaning, Form, And Understanding in the Age of Data," *The 58Th Annual Meeting of The Association for Computational Linguistics*, pp. 5185–5198, 2020.

[57] D. Hassabis, D. Kumaran, C. Summerfield, and M. Botvinick, "Neuroscience-Inspired Artificial Intelligence," *Neuron*, vol. 95, no. 2, pp. 245–258, 2017.

[58] A. Saxe, S. Nelli, and C. Summerfield, "If Deep Learning Is the Answer, What Is the Question?" *Nature Reviews Neuroscience*, vol. 22, no. 1, pp. 55–67, 2021.

[59] A. H. Marblestone, G. Wayne, and K. P. Kording, "Toward an Integration of Deep Learning and Neuroscience," *Frontiers in Computational Neuroscience*, vol. 10, p. 94, 2016.

[60] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, and R. Salakhutdinov, "Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 2978–2988.

[61] T. P. Lillicrap, A. Santoro, L. Marris, C. J. Akerman, and G. Hinton, "Backpropagation and the Brain," *Nature Reviews Neuroscience*, vol. 21, no. 6, pp. 335–346, 2020.

[62] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds *et al.*, "Flamingo: A Visual Language Model for Few-Shot Learning," *Advances In Neural Information Processing Systems*, vol. 35, pp. 23 716–23 736, 2022.

[63] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning Transferable Visual Models From Natural Language Supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8748–8763.

[64] Y. Bisk, A. Holtzman, J. Thomason, J. Andreas, Y. Bengio, J. Chai, M. Lapata, A. Lazaridou, J. May, A. Nisnevich *et al.*, "Experience Grounds Language," *arXiv preprint arXiv:2004.10151*, 2020.

[65] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, and D. Zhou, "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," in *Proceedings of the 36th International Conference on Neural Information Processing Systems*, ser. NIPS '22. Red Hook, NY, USA: Curran Associates Inc., 2022.

[66] J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein, "Generative Agents: Interactive Simulacra of Human Behavior," *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 2023.

[67] M. Binz and E. Schulz, "Using Cognitive Psychology to Understand GPT-3," *Proceedings of the National Academy of Sciences*, vol. 120, no. 6, p. e2218523120, 2023.

[68] B. Dhingra, J. R. Cole, J. M. Eisenschlos, D. Gillick, J. Eisenstein, and W. W. Cohen, "Time-Aware Language Models as Temporal Knowledge Bases," in *Transactions of the Association for Computational Linguistics*, vol. 10. MIT Press, 2022, pp. 1138–1155.

[69] X. Ding and L. Wang, "Do Language Models Understand Time?" in *Companion Proceedings of the ACM on Web Conference 2025*, ser. WWW '25. New York, NY, USA: Association for Computing Machinery, 2025, p. 1855–1868. [Online]. Available: https://doi.org/10.1145/3701716.3717744

[70] G. Teasdale and B. Jennett, "Assessment of Coma and Impaired Consciousness: A Practical Scale," *The Lancet*, vol. 304, no. 7872, pp. 81–84, 1974.

[71] A. M. Owen, M. R. Coleman, M. Boly, M. H. Davis, S. Laureys, and J. D. Pickard, "Detecting Awareness in the Vegetative State," *Science*, vol. 313, no. 5792, pp. 1402–1402, 2006.

[72] J. D. Sitt, J.-R. King, I. El Karoui, B. Rohaut, F. Faugeras, A. Gramfort, L. Cohen, M. Sigman, S. Dehaene, and L. Naccache, "Large Scale Screening of Neural Signatures of Consciousness in Patients in a Vegetative or Minimally Conscious State," *Brain*, vol. 137, no. 8, pp. 2258–2270, 2014.

[73] J. D. Smith, W. E. Shields, and D. A. Washburn, "The Comparative Psychology of Uncertainty Monitoring and Metacognition," *Behavioral and Brain Sciences*, vol. 26, no. 3, pp. 317–339, 2003.

[74] S. W. Townsend, S. E. Koski, R. W. Byrne, K. E. Slocombe, B. Bickel, M. Boeckle, I. Braga Goncalves, J. M. Burkart, T. Flower, F. Gaunet *et al.*, "Exorcising Grice's Ghost: An Empirical Approach to Studying Intentional Communication in Animals," *Biological Reviews*, vol. 92, no. 3, pp. 1427–1433, 2017.

[75] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang *et al.*, "A Survey on Evaluation of Large Language Models," *ACM Transactions on Intelligent Systems and Technology*, vol. 15, no. 3, pp. 1–45, 2024.

[76] D. B. Udell, "Susan Schneider's Proposed Tests for AI Consciousness: Promising but Flawed," *Journal of Consciousness Studies*, vol. 28, no. 5-6, pp. 121–144, 2021.

[77] I. Sutskever, "A Test for AI Consciousness," Retrieved from https://ecorner.stanford.edu/clips/a-test-for-ai-consciousness/, 2023.

[78] C. Hales, "An Empirical Framework for Objective Testing for P-Consciousness in an Artificial Agent," *The Open Artificial Intelligence Journal*, vol. 3, no. 1, 2009.

[79] C. Koch and G. Tononi, "A Test for Consciousness," *Scientific American*, vol. 304, no. 6, pp. 44–47, 2011.

[80] A. Dao and D. B. Vu, "AlphaMaze: Enhancing Large Language Models' Spatial Intelligence via GRPO," 2025. [Online]. Available: https://arxiv.org/abs/2502.14669

[81] K. Lu, A. Grunde-McLaughlin, J. Tian, and Y. Bisk, "The Capacity for Reasoning in Multimodal Language Models," in *Findings of the Association for Computational Linguistics: EMNLP 2022*. Association for Computational Linguistics, 2022, pp. 1512–1525.

[82] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image Segmentation: A Guide to Recent Approaches," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 185–206, 2023.

[83] Google AI, "Gemini: Google's Largest and Most Capable AI Model," https://deepmind.google/technologies/gemini, 2024, [Accessed: 2024-07-29].

[84] Anthropic, "Claude," https://www.anthropic.com/claude, 2024, [Accessed: 2024-07-29].

[85] T. Nagel, "What Is It Like to Be a Bat?" *The Philosophical Review*, vol. 83, no. 4, pp. 435–450, 1974.