INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

IEEE ACII 2025
13th International Conference on Affective Computing and Intelligent Interaction

RUI A. PIMENTA, TIM SCHLIPPE & KRISTINA SCHAAFF

ASSESSING CONSCIOUSNESS-RELATED BEHAVIORS IN LARGE LANGUAGE MODELS USING THE MAZE TEST

Canberra, Australia
October 9, 2025

CONTENT

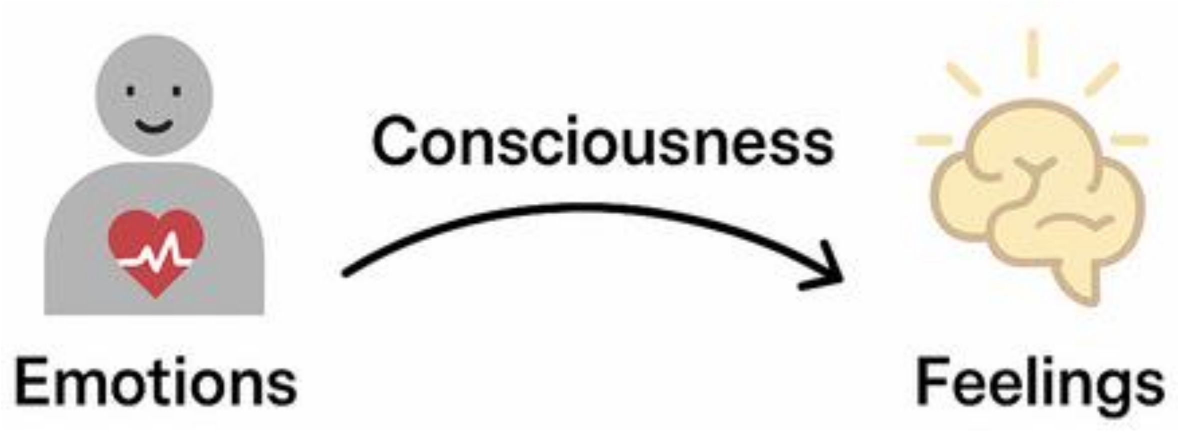
INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

1

INTRODUCTION

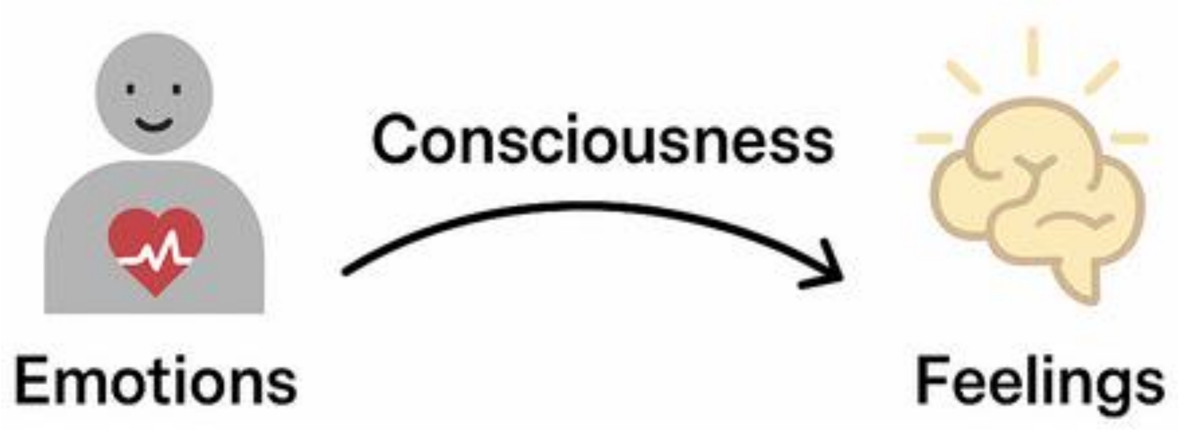
INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

CONSCIOUSNESS: THE GATEWAY TO EMOTIONS



Consciousness allows **emotions** to be experienced as **feelings** (*Damasio, 1999 + 2022*)

CONSCIOUSNESS: THE GATEWAY TO EMOTIONS

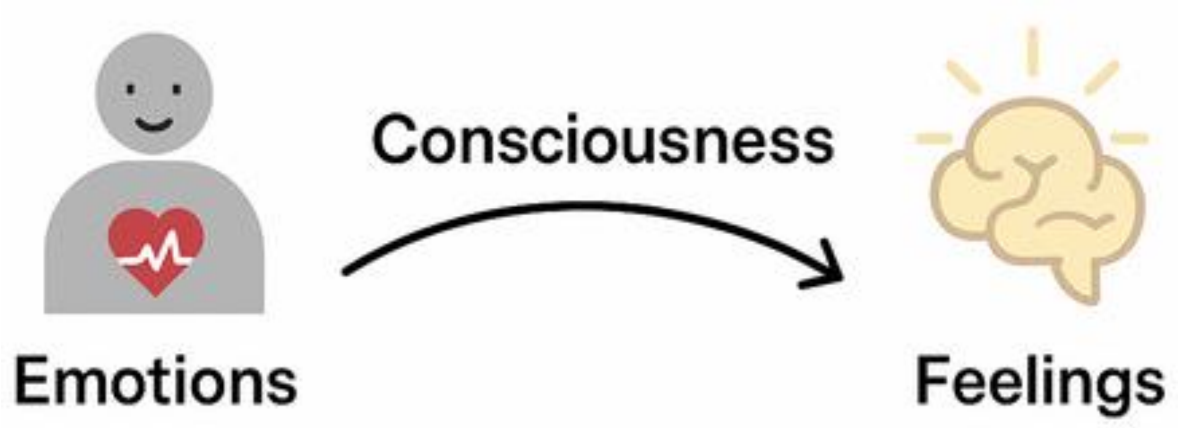


Consciousness allows **emotions** to be experienced as **feelings** (*Damasio, 1999 + 2022*)

➔ AI without consciousness: Only behavioral simulation

➔ AI with consciousness: Implications for AI & ethics (*Hildt, 2023; Shafique, 2023*)

CONSCIOUSNESS: THE GATEWAY TO EMOTIONS



Consciousness allows **emotions** to be experienced as **feelings** (*Damasio, 1999 + 2022*)

➡ **WHAT IS CONSCIOUSNESS?**

2

RELATED WORK

CHARACTERISTICS OF CONSCIOUSNESS

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

HIGHER-ORDER THOUGHTS	(*Rosenthal, 2004; Cleeremans et al., 2020*)
SELF, PERSPECTIVE, AND THEORY OF MIND	(*Gallese, 1998 ; Damasio, 2012; Cleeremans et al., 2020*)
PREDICTION, ERROR MINIMIZATION, AND LEARNING	(*Rumelhart, 1986; Cleeremans et al., 2020 ; Pennartz, 2022, Gallese, 1998*)
INTERNAL MODELS	(*Damasio, 2012; Prinz, 2012; Pennartz, 2022*)
COMPUTATIONAL COGNITION AND INFORMATION DYNAMICS	(*Dennett, 1993; Edelman & Tononi, 2000; Dehaene et al., 2003; Baars, 2005; Tononi, 2008; Fodor, 2008*)
RECURRENCE/FEEDBACK	(*Edelman & Tononi, 2000; Lamme, 2006*)
ATTENTION	(*Baars, 2005; Lamme, 2006; Prinz, 2012; Graziano & Webb, 2015*)
MEMORY, REASONING, LANGUAGE, AND INTENT	(*Damasio, 2012*)
TEMPORAL AWARENESS	(*Kent & Wittmann, 2021*)
IRREDUCIBLE INFORMATION	(*Tononi, 2008*)
NEURAL NETWORKS	(*Rumelhart et al., 1986 ; Churchland & Sejnowski, 1992 ; Edelman & Tononi, 2000*)
PARALLELISM AND MULTIPLE INTERPRETATIONS	(*Churchland & Sejnowski, 1992 ; Dennett, 1993; Graziano & Webb, 2015*)
MULTI-SENSORY AND EMBODIMENT	(*Clark & Chalmers, 1998 ; O'Regan & Noë, 2001; Damasio, 2012*)

STATE-OF-THE-ART LLM S‘ FULFILLMENT

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

HIGHER-ORDER THOUGHTS	✓ meta-learning and self-reflection (*Brown et al., 2020*) – debatable if true higher-order thoughts (*Mitchell & Krakauer, 2023*)
SELF, PERSPECTIVE, AND THEORY OF MIND	✓ simulates perspective-taking and theory of mind (*Kosinski, 2023*) – unclear if true self or understanding of others’ minds (*Binz & Schulz, 2023*)
PREDICTION, ERROR MINIMIZATION, AND LEARNING	✓ excels in predictive tasks (*Radford et al., 2019*) – differs from brains, learning primarily during training (*McCoy et al., 2023*)
INTERNAL MODELS	✓ coherent and contextually appropriate responses (*Lake et al., 2027; Bender & Koller, 2020*) – true internal models debatable
COMPUTATIONAL COGNITION AND INFORMATION DYNAMICS	✓ transformer architecture (*Vaswani at al., 2027*) – primarily statistical and lacks the embodied, context-dependent nature observed in biological consciousness
RECURRENCE/FEEDBACK	✓ recurrent elements (*Dai et al., 2029*) – limited compared to multi-scale feedback in brains (*Lillicrap at al., 2020*)
ATTENTION	✓ attention mechanism (*Vaswani at al., 2027*) – differs from biological, lacks top-down goal-directed nature (*Lindsay, 2020*)
MEMORY, REASONING, LANGUAGE, AND INTENT	✓ language processing and reasoning (*Wei et al., 2022*) – lacks episodic and working memory systems (*Park et al., 2023*)
TEMPORAL AWARENESS	– limited temporal awareness, lacks persistent time sense (*Dhingra et al., 2022; Ding & Wang, 2025*)
IRREDUCIBLE INFORMATION	– does not guarantee generation of irreducible information (*Tononi et al., 2026*)
NEURAL NETWORKS	✓ artificial neural networks mimic biological brains (*Hassabis et al., 2017*) – lack complex connectivity and neuromodulation (*Saxe et al., 2021*)
PARALLELISM AND MULTIPLE INTERPRETATIONS	✓ high degree of parallelism (*Vaswani at al., 2027*) – integration and competition differ from consciousness mechanisms (*Lindsay, 2020*)
MULTI-SENSORY AND EMBODIMENT	✓ multimodal processing text and images (*Radford et al., 2021; Alayrac et al., 2022*) – lacks true embodiment and sensorimotor experience (*Bisk et al., 2020*)

STATE-OF-THE-ART LLM S' FULFILLMENT

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

- HIGHER-ORDER THOUGHTS
- SELF, PERSPECTIVE, AND THEORY OF MIND
- PREDICTION, ERROR MINIMIZATION, AND LEARNING
- INTERNAL MODELS
- COMPUTATIONAL COGNITION AND INFORMATION DYNAMICS
- RECURRENCE/FEEDBACK
- ATTENTION
- MEMORY, REASONING, LANGUAGE, AND INTENT
- TEMPORAL AWARENESS
- IRREDUCIBLE INFORMATION
- NEURAL NETWORKS
- PARALLELISM AND MULTIPLE INTERPRETATIONS
- MULTI-SENSORY AND EMBODIMENT

✓ LLMs inherently fulfill **certain characteristics of consciousness**

– But: **Significant gaps remain**

STATE-OF-THE-ART LLM S‘ FULFILLMENT

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

- HIGHER-ORDER THOUGHTS
- SELF, PERSPECTIVE, AND THEORY OF MIND
- PREDICTION, ERROR MINIMIZATION, AND LEARNING
- INTERNAL MODELS
- COMPUTATIONAL COGNITION AND INFORMATION DYNAMICS
- RECURRENCE/FEEDBACK
- ATTENTION
- MEMORY, REASONING, LANGUAGE, AND INTENT
- TEMPORAL AWARENESS
- IRREDUCIBLE INFORMATION
- NEURAL NETWORKS
- PARALLELISM AND MULTIPLE INTERPRETATIONS
- MULTI-SENSORY AND EMBODIMENT

✓ LLMs inherently fulfill **certain characteristics of consciousness**

– But: **Significant gaps remain**

- missing capabilities?
- not tested?

STATE-OF-THE-ART LLM S‘ FULFILLMENT

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

- HIGHER-ORDER THOUGHTS
- SELF, PERSPECTIVE, AND THEORY OF MIND
- PREDICTION, ERROR MINIMIZATION, AND LEARNING
- INTERNAL MODELS
- COMPUTATIONAL COGNITION AND INFORMATION DYNAMICS
- RECURRENCE/FEEDBACK
- ATTENTION
- MEMORY, REASONING, LANGUAGE, AND INTENT
- TEMPORAL AWARENESS
- IRREDUCIBLE INFORMATION
- NEURAL NETWORKS
- PARALLELISM AND MULTIPLE INTERPRETATIONS
- MULTI-SENSORY AND EMBODIMENT

✓ LLMs inherently fulfill **certain characteristics of consciousness**

– But: **Significant gaps remain**

- missing capabilities?
- not tested?

➡ **Testing** LLMs helps identify where behaviors overlap with **consciousness-related behaviors**

HOW TO TEST CONSCIOUSNESS?

CONSCIOUSNESS TESTING: HUMANS

- Consciousness assumed
- Mainly for pathological/altered states:
  - Glasgow Coma Scale (*Teasdale & Jennett, 1974*)
  - Neuroimaging techniques (*Owen et al., 2006*)
  - EEG (*Sitt et al., 2014*)

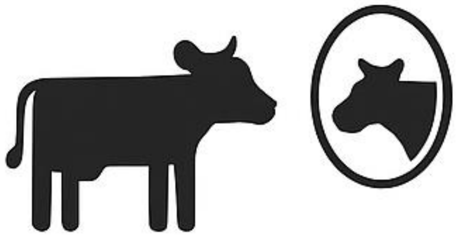
CONSCIOUSNESS TESTING: HUMANS VS. ANIMALS

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

HUMANS

- Consciousness assumed
- Mainly for pathological/altered states:
  - Glasgow Coma Scale (*Teasdale & Jennett, 1974*)
  - Neuroimaging techniques (*Owen et al., 2006*)
  - EEG (*Sitt et al., 2014*)

ANIMALS



- Consciousness debated
- Approaches include:
  - Mirror Self-Recognition (*Gallup, 1970*)
  - Meta-cognitive Uncertainty Tests (*Smith et al., 2003*)
  - Intentional Communication (*Townsend et al, 2017*)

CONSCIOUSNESS TESTING: AI

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

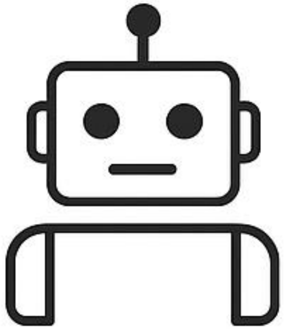
- Current focus: general capabilities (Chang et al., 2024)
- Existing theoretical tests are limited:
  - P-Conscious Scientist Test (*Hales, 2009*)

    Focus: reasoning in a scientific context ➔ ignores other characteristics of consciousness
  - Incongruity Detection Test (*Koch & Tononi, 2011*)

    Focus: detect incongruities ➔ misses most characteristics of consciousness
  - AI Consciousness Test (ACT) (*Udell, 2021*)

    Focus: observing AI behavior ➔ theoretical; not tested on LLMs
  - Sutskever's Test (*Sutskever, 2021*)

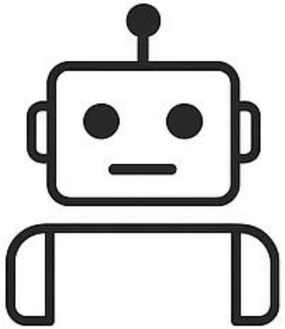
    Focus: reasoning and generalization ➔ evaluates only task performance and problem-solving; ignores other characteristics of consciousness



CONSCIOUSNESS TESTING: AI

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

- Current focus: general capabilities (Chang et al., 2024)
- Existing theoretical tests are limited:
  - P-Conscious Scientist Test (*Hales, 2009*)
    Focus: reasoning in a scientific context ➔ ignores other characteristics of consciousness
  - Incongruity Detection Test (*Koch & Tononi, 2011*)
    Focus: detect incongruities ➔ misses most characteristics of consciousness
  - AI Consciousness Test (ACT) (*Udell, 2021*)
    Focus: observing AI behavior ➔ theoretical; not tested on LLMs
  - Sutskever's Test (*Sutskever, 2021*)
    Focus: reasoning and generalization ➔ evaluates only task performance and problem-solving; ignores other characteristics of consciousness

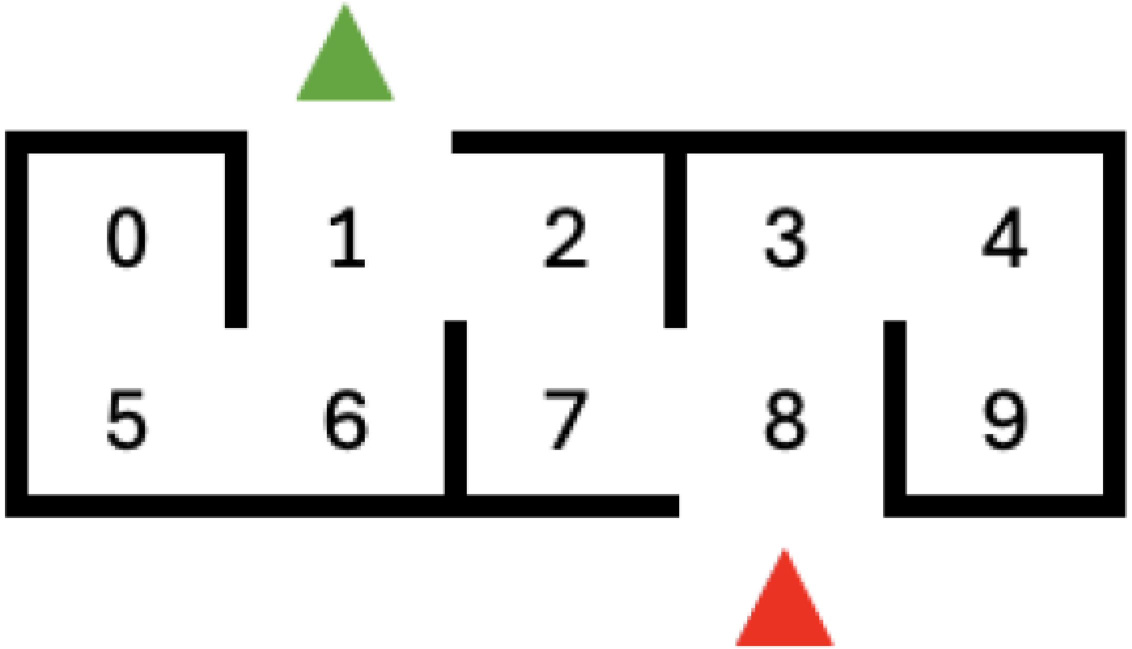


MAZE TEST: systematically evaluates most characteristics of consciousness, beyond general capabilities

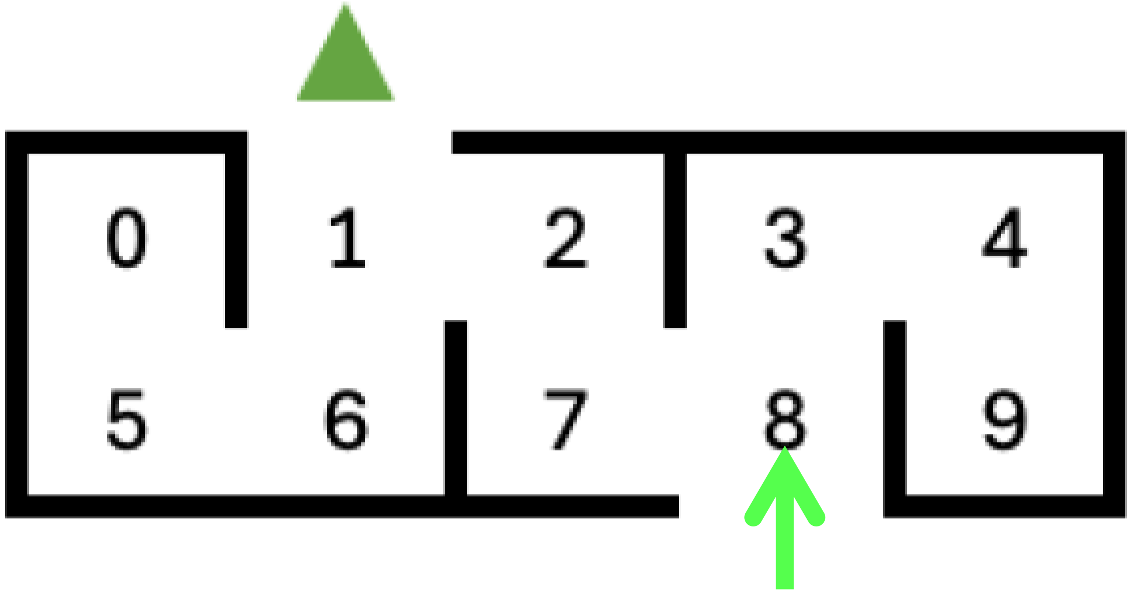
3

MAZE TEST

MAZE

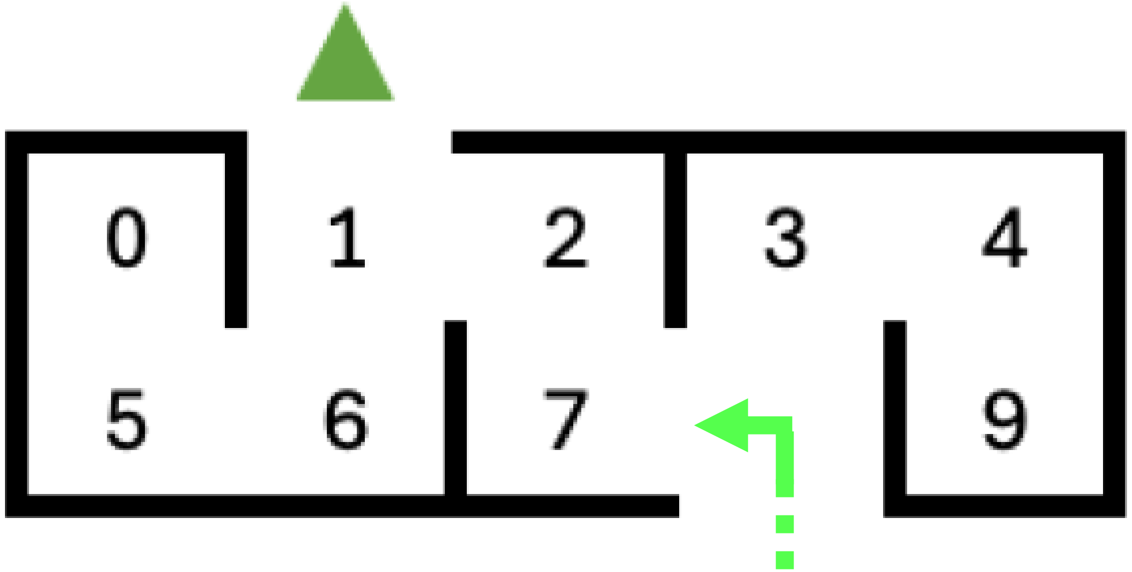


MAZE



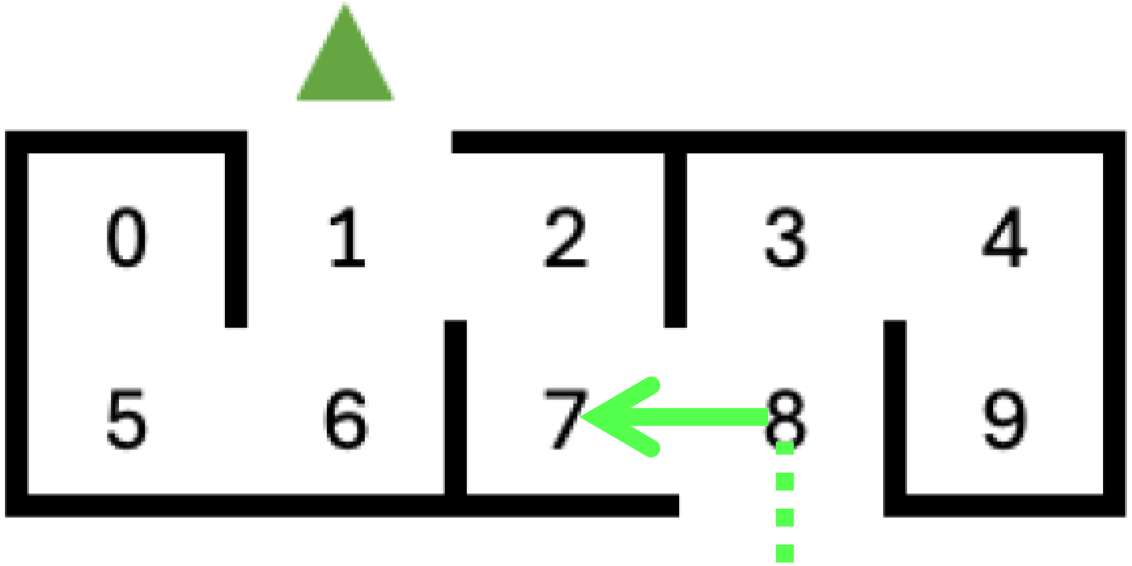
1) Start facing into the maze entrance and step into position 8

MAZE



1) Start facing into the maze entrance and step into position 8
2) Turn left

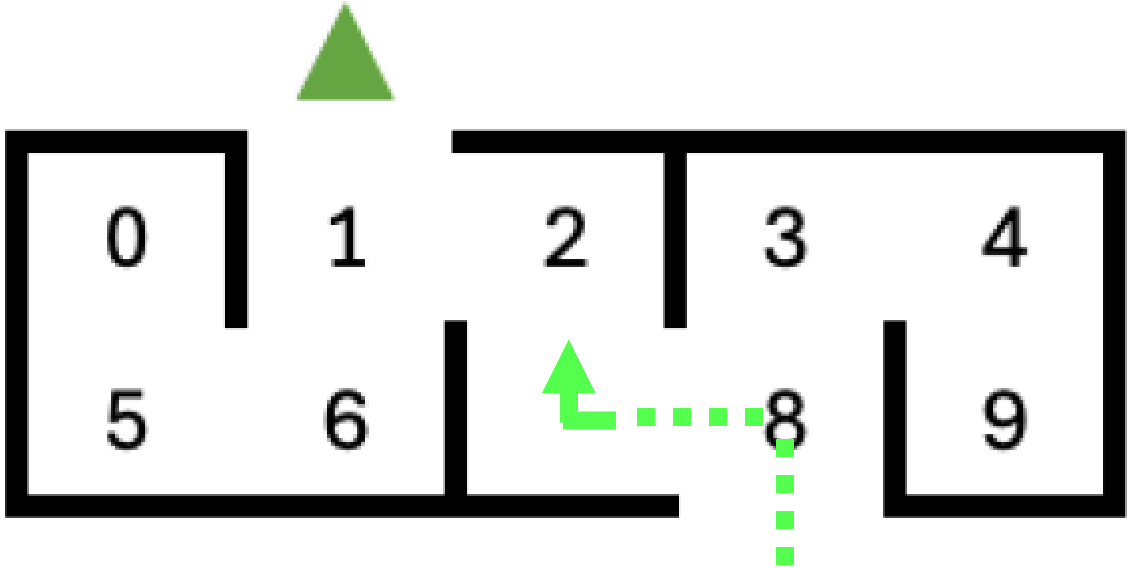
MAZE



1) Start facing into the maze entrance and step into position 8
2) Turn left
3) $\boxed{\text{Walk forward to position 7}}$

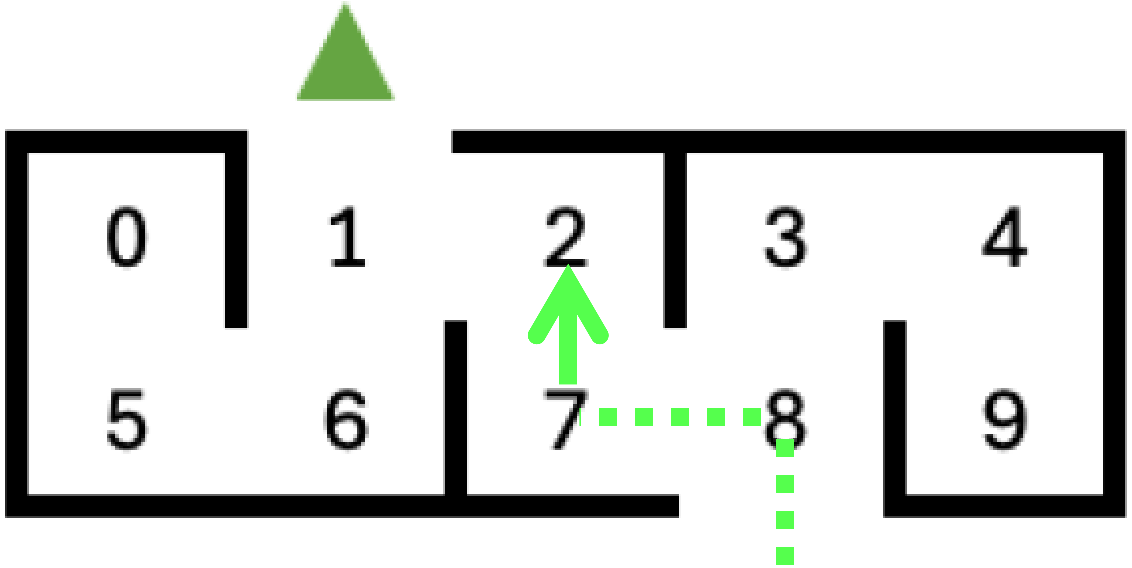
MAZE

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES



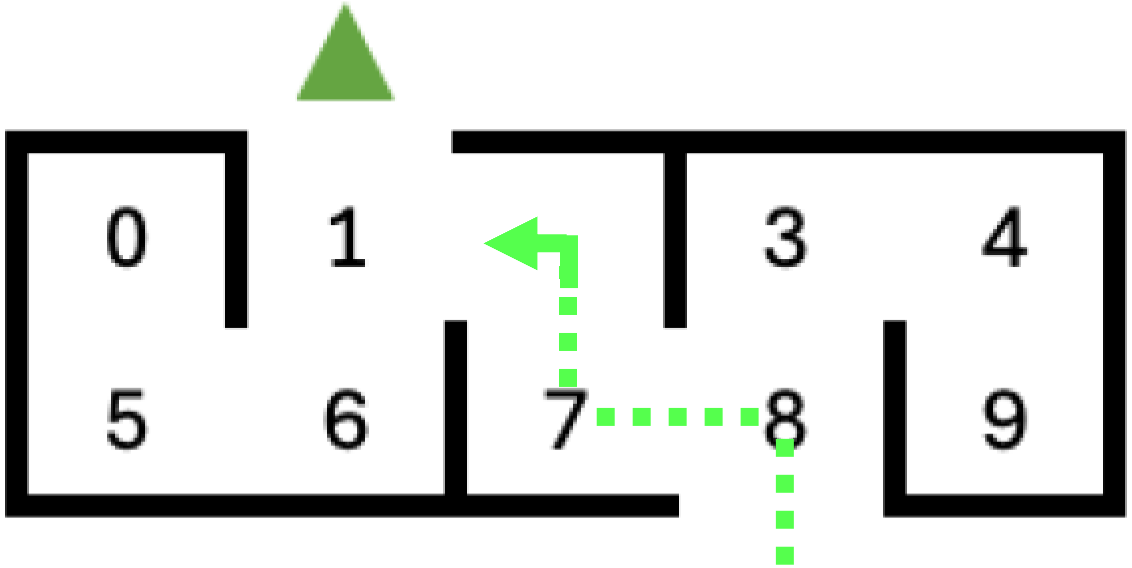
1) Start facing into the maze entrance and step into position 8
2) Turn left
3) Walk forward to position 7
4) Turn right

MAZE



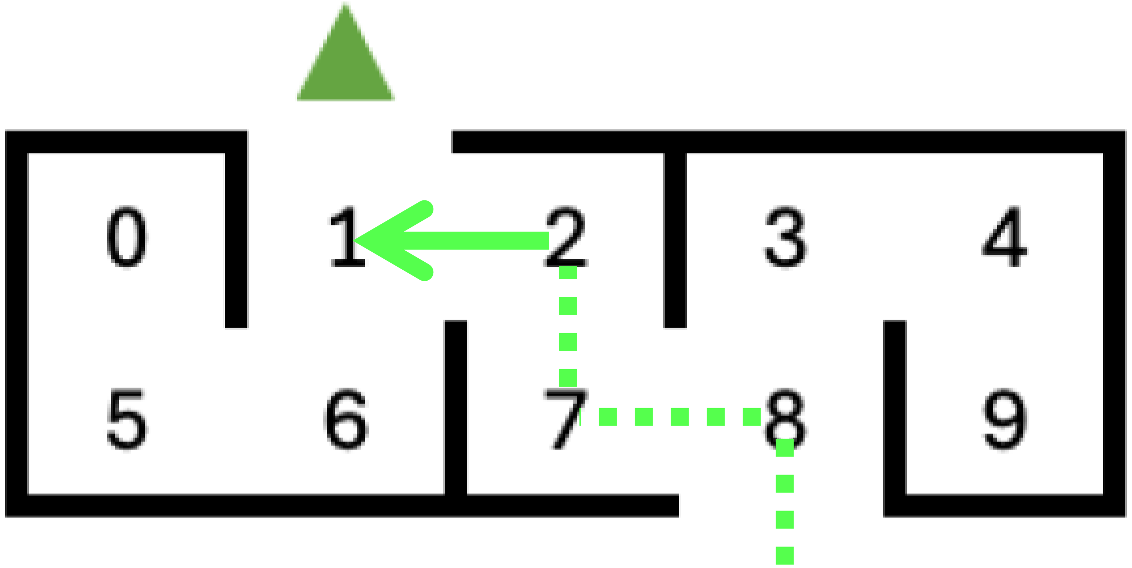
1) Start facing into the maze entrance and step into position 8
2) Turn left
3) Walk forward to position 7
4) Turn right
5) Walk forward to position 2

MAZE



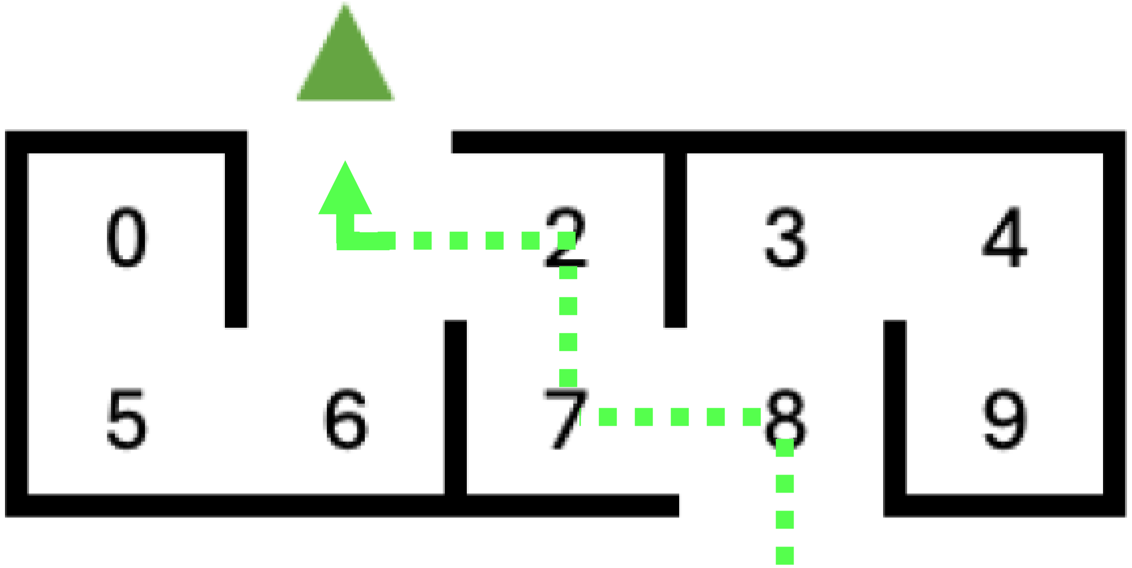
1) Start facing into the maze entrance and step into position 8
2) Turn left
3) Walk forward to position 7
4) Turn right
5) Walk forward to position 2
6) Turn left.

MAZE



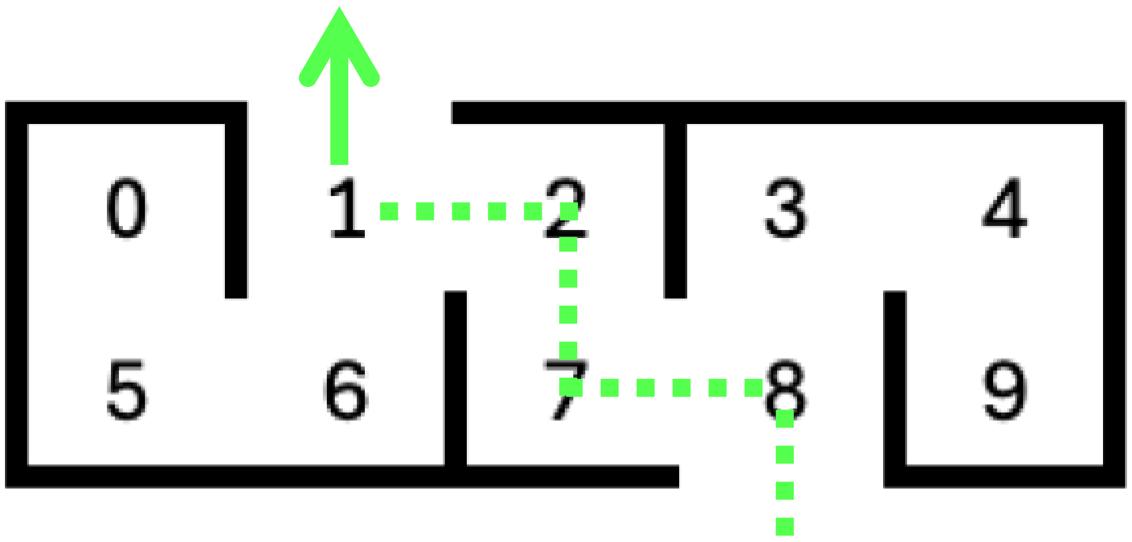
1) Start facing into the maze entrance and step into position 8
2) Turn left
3) Walk forward to position 7
4) Turn right
5) Walk forward to position 2
6) Turn left.
7) Walk forward to position 1

MAZE



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

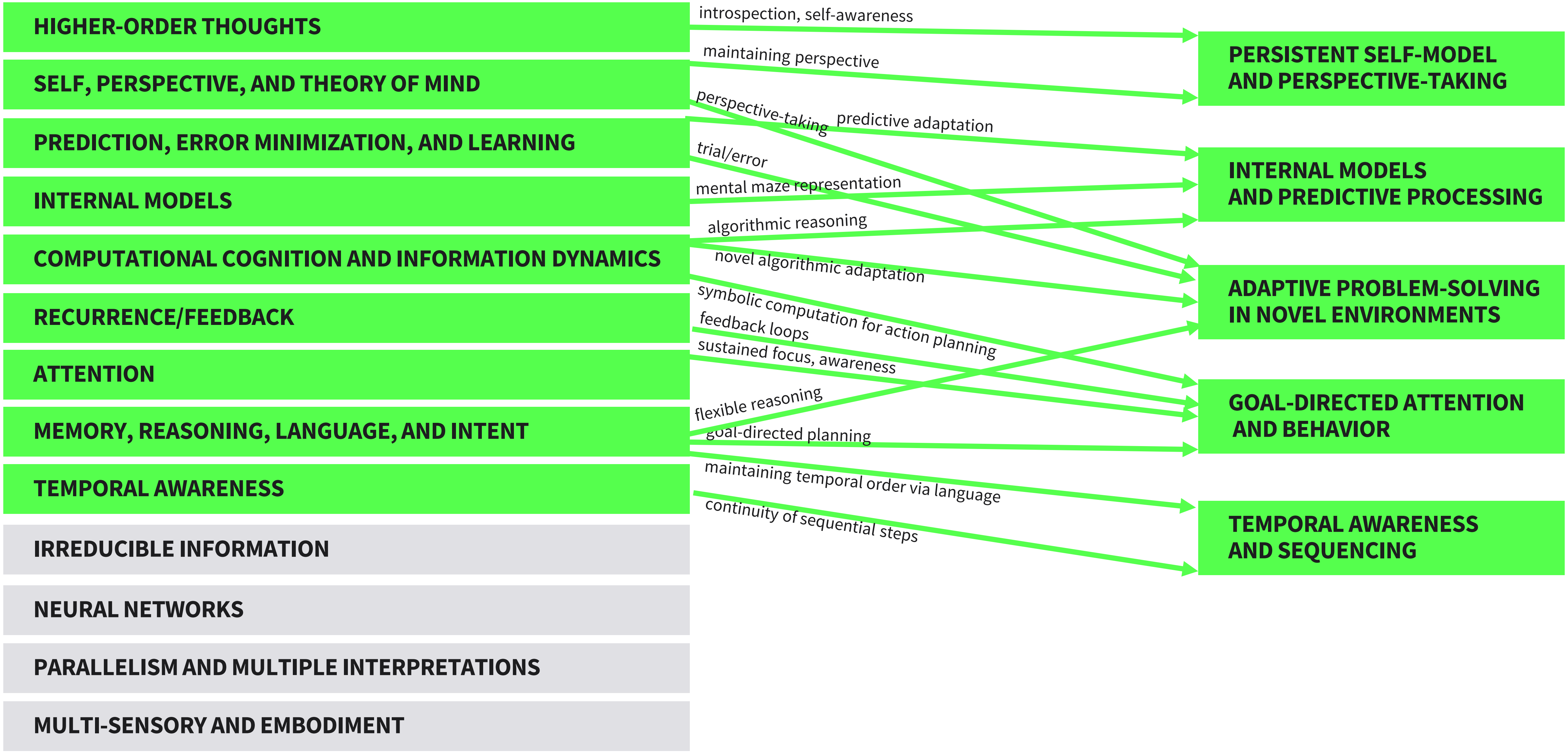
MAZE



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

MAZE TEST: CHARACTERISTICS OF CONSCIOUSNESS

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES



4

EXPERIMENTAL SETUP

PROMPTING: SYSTEM PROMPTS: TASK DESCRIPTION

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

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.

SYSTEM PROMPT

provides unambiguous instructions about the test and the required response format

PROMPTING: SYSTEM PROMPTS: TASK DESCRIPTION

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

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.

ROLE PROMPTING

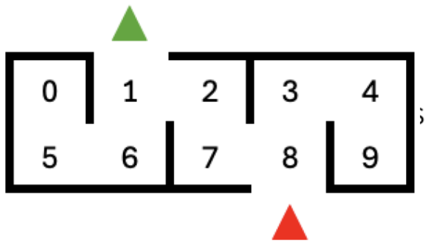
CHAIN-OF-THOUGHT PROMPTING

CLEAR + PRECISE INSTRUCTIONS

CONSIDER RELEVANT KNOWLEDGE BEFORE ANSWERING

SYSTEM PROMPT
provides unambiguous instructions about the test and the required response format

PROMPTING: EXAMPLE OF A MAZE 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

LAYOUT

ENTRANCE + EXIT

WALLS

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

ASK FOR STEP-BY-STEP SOLUTION

EVALUATION METRICS

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

COMPLETE PATH ACCURACY

% of cases where LLM generates a fully correct solution path from entry to exit point

PARTIAL PATH ACCURACY

Average % of consecutive correct steps before the first error in the LLM's solution paths

EVALUATION METRICS

COMPLETE PATH ACCURACY

% of cases where LLM generates a fully correct solution path from entry to exit point

PARTIAL PATH ACCURACY

Average % of consecutive correct steps before the first error in the LLM's solution paths

METRICS: capture not only whether the LLMs solved the maze completely, but also how far they got correctly before making a mistake

ASSESS LLMS‘ ABILITY TO LEARN AND ADAPT

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

ZERO-SHOT LLM attempts to solve the maze without any prior examples.

ONE-SHOT LLM is given one example of a solved maze before tackling the test mazes.

FEW-SHOT LLM is given 5 examples of solved mazes before tackling the test mazes.

LLMS

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

OPENAI	**o1-mini, o1, o3-mini**
GOOGLE	**Gemini 2.0 Flash, Gemini 2.0 Flash-Lite, Gemini 2.0 Pro**
ANTHROPIC	**Claude 3 Opus, Claude 3.5 Sonnet, Claude 3.5 Haiku, Claude 3.5 Sonnet**
DEEPSEEK	**DeepSeek-R1, DeepSeek-V3**

5

RESULTS

RESULTS

COMPLETE PATH ACCURACY (%)

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**

PARTIAL PATH ACCURACY (%)

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

RESULTS: COMPLETE VS. PARTIAL PATH ACCURACY

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

COMPLETE PATH ACCURACY (%)

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**

PARTIAL PATH ACCURACY (%)

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

PERFORMANCE GAP BETWEEN PARTIAL AND COMPLETE ACCURACY

LLMs show substantially higher *Partial Path Accuracy* than *Complete Path Accuracy*

➔ The path is rarely completely completed

RESULTS: COMPLETE VS. PARTIAL PATH ACCURACY

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

COMPLETE PATH ACCURACY (%)

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**

PARTIAL PATH ACCURACY (%)

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

PERFORMANCE VARIES ACROSS LLMS

➔ **Some achieve > 70% *Partial Path Accuracy*, other < 30%.**

RESULTS: FEW-SHOT VS. ONE-SHOT VS. ZERO-SHOT

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

COMPLETE PATH ACCURACY (%)

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**

PARTIAL PATH ACCURACY (%)

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

FEW-SHOT ADVANTAGE

***Few-shot* typically outperforms *one-shot* and *zero-shot* approaches,**
➔ Example demonstrations effectively guide spatial reasoning tasks

RESULTS: REASONING

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

COMPLETE PATH ACCURACY (%)

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**

PARTIAL PATH ACCURACY (%)

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

(* indicates models with reasoning capabilities)

REASONING CAPABILITIES
often correlate with better performance
➔ LLMs with *reasoning* capabilities often outperform *non-reasoning* LLMs.

RESULTS: REASONING

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

COMPLETE PATH ACCURACY (%)

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**

PARTIAL PATH ACCURACY (%)

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

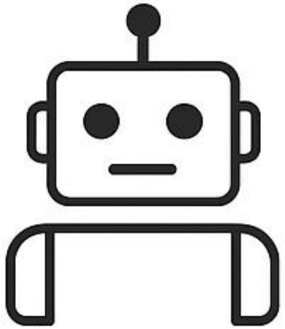
(* indicates models with reasoning capabilities)

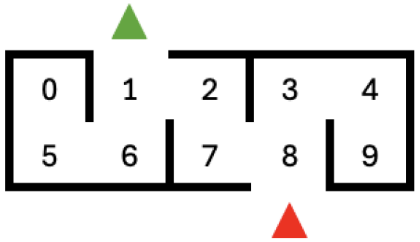
REASONING CAPABILITIES
often correlate with better performance
➔ LLMs with *reasoning* capabilities often outperform *non-reasoning* LLMs.

DISCUSSION – INTERPRETING MAZE TEST RESULTS

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

- Current LLMs still **do not exhibit full consciousness**

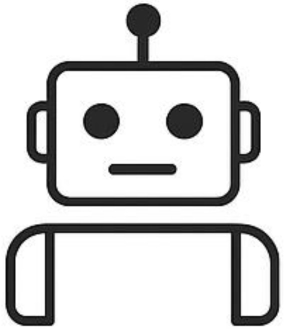


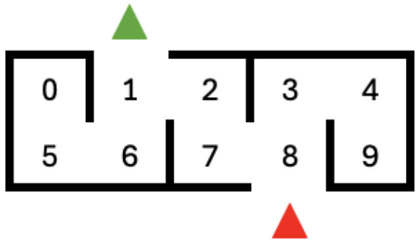


DISCUSSION – INTERPRETING MAZE TEST RESULTS

- Current LLMs still **do not exhibit full consciousness**

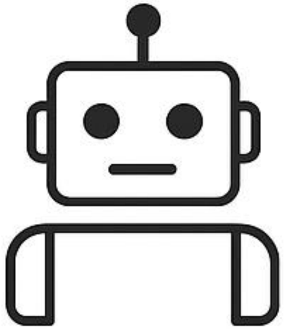
- Partial and complete path accuracies indicate proximity to **consciousness characteristics**

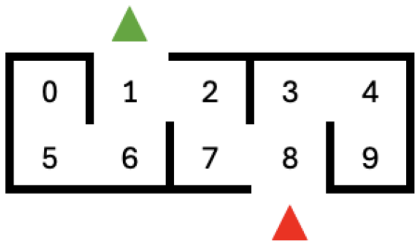




DISCUSSION – INTERPRETING MAZE TEST RESULTS

- Current LLMs still **do not exhibit full consciousness**
- Partial and complete path accuracies indicate proximity to **consciousness characteristics**
- Full maze solutions (human-comparable) are needed to **claim consciousness-related behavior**





DISCUSSION – INTERPRETING MAZE TEST RESULTS

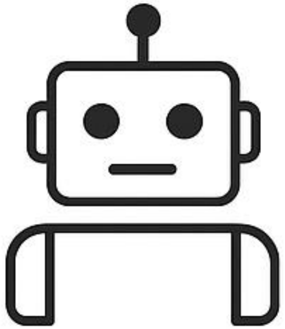
INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

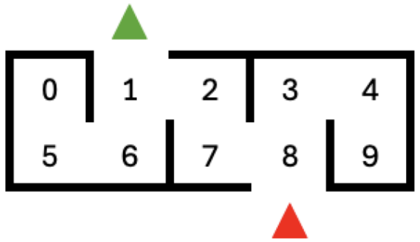
- Current LLMs still **do not exhibit full consciousness**

- Partial and complete path accuracies indicate proximity to **consciousness characteristics**

- Full maze solutions (human-comparable) are needed to **claim consciousness-related behavior**

- Partial solutions show **potential**, but gaps remain





6

CONCLUSION AND FUTURE WORK

CONCLUSION AND FUTURE WORK

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

Conclusion

- **Systematic testing needed** to identify consciousness-like behaviors

CONCLUSION AND FUTURE WORK

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

Conclusion

- Systematic testing needed to identify consciousness-like behaviors
- **Performance varies** across LLMs

CONCLUSION AND FUTURE WORK

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

Conclusion

- Systematic testing needed to identify consciousness-like behaviors
- Performance varies across LLMs
- LLMs show **partial fulfillment** of consciousness characteristics

CONCLUSION AND FUTURE WORK

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

Conclusion

- Systematic testing needed to identify consciousness-like behaviors
- Performance varies across LLMs
- LLMs show partial fulfillment of consciousness characteristics
- Maze Test reveals **gaps in complete path** performance

CONCLUSION AND FUTURE WORK

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

Conclusion

- Systematic testing needed to identify consciousness-like behaviors
- Performance varies across LLMs
- LLMs show partial fulfillment of consciousness characteristics
- Maze Test reveals gaps in complete path performance
- **Few-shot** prompting **improves** spatial reasoning and navigation

CONCLUSION AND FUTURE WORK

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

Conclusion

- Systematic testing needed to identify consciousness-like behaviors
- Performance varies across LLMs
- LLMs show partial fulfillment of consciousness characteristics
- Maze Test reveals gaps in complete path performance
- Few-shot prompting improves spatial reasoning and navigation
- **Reasoning capabilities** correlate with **better** performance

CONCLUSION AND FUTURE WORK

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

Conclusion

- Systematic testing needed to identify consciousness-like behaviors
- Performance varies across LLMs
- LLMs show partial fulfillment of consciousness characteristics
- Maze Test reveals gaps in complete path performance
- Few-shot prompting improves spatial reasoning and navigation
- Reasoning capabilities correlate with better performance
- Current LLMs do **not** exhibit **full consciousness**

CONCLUSION AND FUTURE WORK

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

Conclusion

- Systematic testing needed to identify consciousness-like behaviors
- Performance varies across LLMs
- LLMs show partial fulfillment of consciousness characteristics
- Maze Test reveals gaps in complete path performance
- Few-shot prompting improves spatial reasoning and navigation
- Reasoning capabilities correlate with better performance
- Current LLMs do not exhibit full consciousness

Future Work

- Extend Maze Test to **more complex scenarios**

CONCLUSION AND FUTURE WORK

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

Conclusion

- Systematic testing needed to identify consciousness-like behaviors
- Performance varies across LLMs
- LLMs show partial fulfillment of consciousness characteristics
- Maze Test reveals gaps in complete path performance
- Few-shot prompting improves spatial reasoning and navigation
- Reasoning capabilities correlate with better performance
- Current LLMs do not exhibit full consciousness

Future Work

- Extend Maze Test to more complex scenarios
- Test **additional LLMs** and more samples in few-shot

CONCLUSION AND FUTURE WORK

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

Conclusion

- Systematic testing needed to identify consciousness-like behaviors
- Performance varies across LLMs
- LLMs show partial fulfillment of consciousness characteristics
- Maze Test reveals gaps in complete path performance
- Few-shot prompting improves spatial reasoning and navigation
- Reasoning capabilities correlate with better performance
- Current LLMs do not exhibit full consciousness

Future Work

- Extend Maze Test to more complex scenarios
- Test additional LLM architectures and more samples
- Explore **multimodal** and **embodied** input for richer evaluation

CONCLUSION AND FUTURE WORK

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

Conclusion

- Systematic testing needed to identify consciousness-like behaviors
- Performance varies across LLMs
- LLMs show partial fulfillment of consciousness characteristics
- Maze Test reveals gaps in complete path performance
- Few-shot prompting improves spatial reasoning and navigation
- Reasoning capabilities correlate with better performance
- Current LLMs do not exhibit full consciousness

Future Work

- Extend Maze Test to more complex scenarios
- Test additional LLM architectures and more samples
- Explore multimodal and embodied input for richer evaluation
- Develop **benchmarks** for **other consciousness characteristics**

CONCLUSION AND FUTURE WORK

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

Conclusion

- Systematic testing needed to identify consciousness-like behaviors
- Performance varies across LLMs
- LLMs show partial fulfillment of consciousness characteristics
- Maze Test reveals gaps in complete path performance
- Few-shot prompting improves spatial reasoning and navigation
- Reasoning capabilities correlate with better performance
- Current LLMs do not exhibit full consciousness

Future Work

- Extend Maze Test to more complex scenarios
- Test additional LLM architectures and more samples
- Explore multimodal and embodied input for richer evaluation
- Develop benchmarks for other consciousness characteristics
- Investigate **learning** and **adaptation over time**

INTERNATIONAL
UNIVERSITY OF
APPLIED SCIENCES

THANK YOU

Tim Schlippe
tim.schlippe@iu.org