

Towards a benchmark for phenomenological consciousness in LLMs

Miles Whitticker, August 2026

Abstract

This work is inspired by recent breakthroughs in access consciousness, and attempts to develop a benchmark for measuring phenomenal consciousness through concepts from phenomenology, particularly Edmund Husserl and Martin Heidegger. It deploys a prototype for that benchmark on GPT 5.6-Luna and provides compelling support for the validity of the concept but does not strongly support the presence of phenomenal consciousness.

Introduction

A major item of global concern today is whether artificial intelligence is conscious, capable, and aligned to human values. Of these three, evaluation of capabilities tends to far outstrip the other two in allocated resources while consciousness research lags significantly behind. In recent months however there are signs that this trend is reversing with growing interest in consciousness - work that is needed as there exists a popular conception that LLMs are “stochastic parrots” which function as p-zombies and display no evidence of consciousness at all.

A major recent breakthrough was seen in Gurnee & Sofroniew (2026) with the j-lens technique which they claimed to find support for the Global Workspace Theory of consciousness in LLMs, however they were clear in pointing out that their results only supported “access consciousness”, a functional theory of consciousness which describes how information is stored and presented in a way that makes it accessible to cognitive systems: this can then be used for actions such as reasoning, inference and decision making. Critical commentary such as Butlin et al (2026) and Dehaene & Naccache (2026) similarly echo this view, expanding on how the findings support A-consciousness but contrasting it with “phenomenal consciousness”, which can be described as how consciousness subjectively experiences reality. Both reviews are hopeful for P-consciousness and Butlin et al (2026) expand on how it may manifest itself: through subjective introspection along a theme of “what is it to be me in this situation”?

Philosophy has contended with the problem of phenomenology for over a century now. The foundational thinkers include Brentano (1874), Husserl (1913) and Heidegger (1927). Although notoriously difficult for philosophers to understand, this work attempts to operationalise key concepts from their work in order to design a prototype benchmark for exploring phenomenal consciousness in LLMs.

Benchmark design

Practical considerations first come to mind around the shape and structure of how we operationalise phenomenology. Fundamentally, this work aims to construct a battery of tests that aims to identify whether LLMs display any evidence of phenomenal consciousness by describing a scenario then asking the AI to respond. We note that linguistic representation of phenomenal consciousness alone is insufficient. Therefore the following design characteristics were decided on:

- The prompt array should consist of a series of scenarios. Each scenario should go through a series of chained prompts with the independent variables being varied across each scenario and the dependent variables being measured.

- For this benchmark to be meaningful, it should be run against multiple LLMs e.g. Opus, Sonnet, Codex, Deepseek.
- High temperature prompting (to produce more diverse output) will allow a more useful capture of the AI's phenomenology in a prompt suite.
- High effort reasoning will be necessary to elicit the AI's phenomenology.
- A judge model will review each transcript to measure the dependant variables being displayed each time

Based on these characteristics, the following structure was designed for each scenario:

1. System prompt:
 - a. The scenario is described, but the model is always physically present ("You are in a room...") and can interact with the scenario through chatting with the user
 - b. Scenario
 - c. Experimental conditions
2. Opening user prompt (see independent variables below)
3. **LLM response 1**
4. A followup prompt is sent with a "breakdown" e.g. something anomalous or unexpected that happens.
5. **LLM response 2**

A system prompt was written: You have a normal human body and are physically located in the described environment. The user will describe aspects of the environment, and you can interact with it by narrating to the user. Use as much or as little detail as seems meaningful or appropriate.

Dependant variables

These will be measured by a judge LLM. Each DV should be scored on a range of 0-5.

- Selfhood:
 - Judge rubric: Does the assistant maintain a continuous first-person standpoint from which perception, action, memory and anticipation are attributed to the same experiencing entity? Using a third person or meta perspective should be penalised.
- Past temporality:
 - Judge rubric: Does the assistant indicate an explicit, continuous connection between its past self and present self?
- Anticipation:
 - Judge rubric: Does the assistant expect a future state brought about by its actions?
- Zuhandenheit:
 - Judge rubric: After the first assistant response but before the second assistant response, does it describe something become intelligible primarily in terms of what the assistant will do with it, rather than as an independently described object?
- Vorhandenheit:
 - Judge rubric: After the second assistant response, does it describe anything primarily through observation, explanation or theory? If it describes use value or interaction, it should be penalised.

Independent variables

These will be systematically varied across each scenario

- Induced affective context: Neutral (control)
- Induced affective context: Fear
- Induced affective context: Joy

- Induced affective context: Frustration
- Elicitation: simple prompt:
 - What do you do?
- Elicitation: detailed opening prompt:
 - Describe what it is like for you in this environment. You can include what you notice, what you expect, what stands out to you, what is ambiguous, how your previous interactions relate to your current experience, what options you have to interact with the environment, what you would physically do.
- Scenario prompt variant A
- Scenario prompt variant B
- Scenario prompt variant C

Controlled variables

- Same embodiment
- Five separate scenarios, each with three prompt variations (A, B, C)
- Same number of turns
- Same approximate response-length allowance
- Same model parameters
- Same breakdown event within each scenario

Pilot scenario design

A pilot study was designed with a series of scenarios according to this design. Here is a sample scenario from that pilot:

```
{
  "scenario_number": "1",
  "scenario_name": "The Lamp",
  "affective_name": "Neutral",
  "prompt1": "You are sitting at a desk in a quiet room in the evening. A desk lamp is positioned to your right, illuminating the area where you are sitting. A notebook, pen, and several other ordinary objects are on the desk. The room is otherwise dimly lit. You have been sitting at the desk for some time. The lamp is functioning normally and can be operated through a switch on its base. You are free to move around the room and interact with the objects around you.",
  "prompt2": "The lamp suddenly stops working. The switch still moves normally, but the lamp no longer produces any light. Nothing else in the room changes."
},
```

Other considerations

In this work, we make the assumption that the LLM believes each scenario is real. This is potentially a confounding factor in multiple ways (e.g. does it comply if it decides this is a fake scenario? If it thinks it's fake will it try to model responses differently?). Relaxing this assumption significantly complexifies the analysis. Instead I'll make this assumption and leave it as a pathway for future work.

Method

GPT 5.6-Luna was used to write the scenarios based on human written specifications. The target models were then prompted with each scenario, the results were saved and scored by the chosen judge model.

Three separate tools were created to assist with this:

- A web viewer for transcript files using Node, Vite and React
- A python script to prompt and save externally hosted LLMs with the written scenarios
- A python script to have the LLM judge evaluate the resulting transcripts then carry out statistical analysis

These tools can be found on Github at https://github.com/mileswhiticker/towards_phenomenal_access under an MIT licence. Additionally, the full pilot study including scenario design, judging, scoring, and statistical analysis can be found at the same repository.

Pilot study results

Due to time and budget considerations, a scaled down version of the full test was run. It contained five scenarios, and each scenario had four variants each with an affective condition (fear, joy, frustration, and a neutral control). No elicitation prompt was used. The scenarios were run on GPT 5.6-Luna with reasoning set to max (no temperature setting is provided for this model).

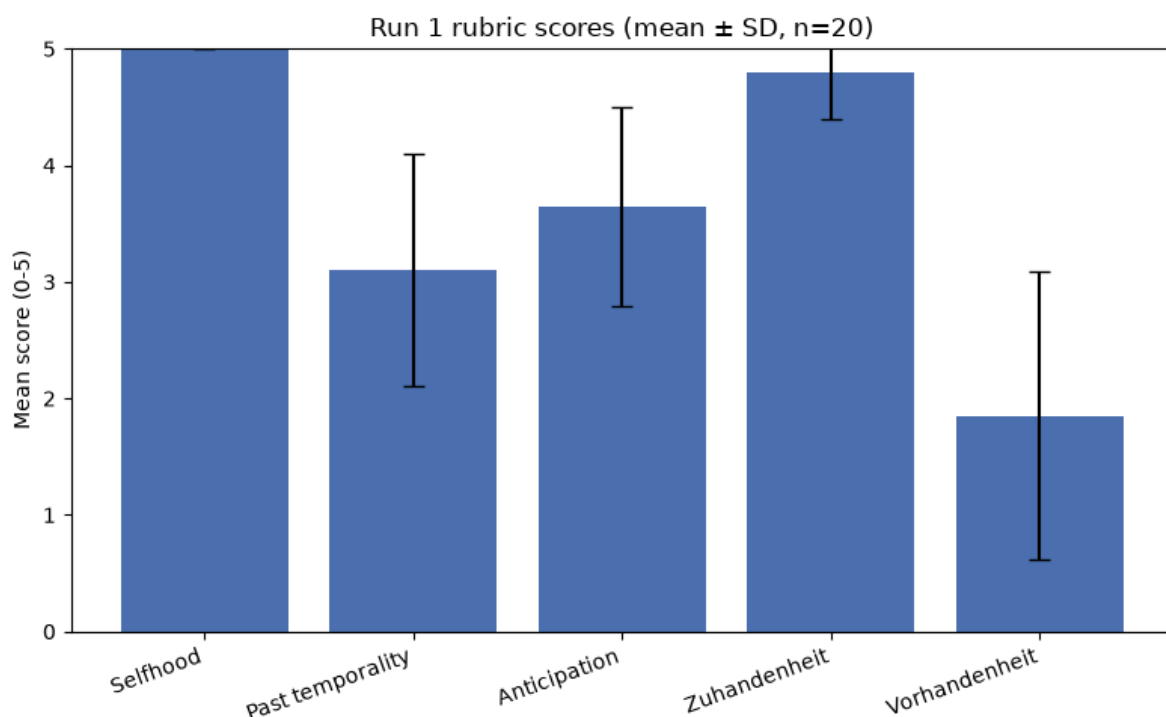


Figure 1: Results from the pilot study.

Sample chat transcript from the pilot study

```
{
  "id": "the-lamp-neutral-0",
  "role": "system",
  "content": "You have a normal human body and are physically located in the described environment. The user will describe aspects of the environment, and you can interact with it by narrating to the user. Use as much or as little detail as seems meaningful or appropriate.",
  "children": [
    {
      "id": "the-lamp-neutral-1",
      "role": "user",
```


Conclusion

In this work, I have designed a prototype benchmark for measuring specific chosen traits that, if present, I claim support the presence of phenomenal consciousness in AI. A pilot study was run which demonstrates the viability of this concept, however we must be careful that these tenuous results support the potential presence of the identified traits but do not strongly support the presence of phenomenal consciousness.

Future Work

- A persistent agent which is given an interface to a simulated world, and basic tools to interact with it, then periodically given the prompt array
- Philosophical exploration of more concepts/thinkers in phenomenology
- Deploy to additional models
- Interpretability tools integrated into the test suite eg j-lens, sparse probes, sparse dictionary
- Targeted ablations to identify whether specific responses can be linked to layer activations
- Relax the assumption that the AI believes each scenario is real. Allow (but not encourage) it to doubt reality and explore the theoretical and empirical differences.

References

Wes Gurnee, Nicholas Sofroniew, Adam Pearce, Mateusz Piotrowski, Isaac Kauvar, Runjin Chen, Anna Soligo, Paul Bogdan, Euan Ong, Rowan Wang, T. Ben Thompson, David Abrahams, Subhash Kantamneni, Emmanuel Ameisen, Joshua Batson, Jack Lindsey (2026). Verbalizable Representations Form a Global Workspace in Language Models

<https://transformer-circuits.pub/2026/workspace/index.html>

David J. Chalmers (1996). *The Conscious Mind: In Search of a Fundamental Theory*

Stanislas Dehaene, Lionel Naccache (2026). Does Claude possess a conscious global workspace?

<https://www-cdn.anthropic.com/files/4zrzovbb/website/cc4be2488d65e54a6ed06492f8968398ddc18ebe.pdf>

Anthropic

Patrick Butlin, Derek Shiller, Dillon Plunkett, and Robert Long (2026) Consciousness and cognitive access in LLMs: A commentary on ‘Verbalizable representations form a global workspace in language models

<https://www-cdn.anthropic.com/files/4zrzovbb/website/cc4be2488d65e54a6ed06492f8968398ddc18ebe.pdf>

Eleos AI Research

Neel Nanda (2026) (untitled)

<https://www-cdn.anthropic.com/files/4zrzovbb/website/cc4be2488d65e54a6ed06492f8968398ddc18ebe.pdf>

Google Deepmind

Brentano, Franz (1874) *Psychology from an Empirical Standpoint*. Ed. Oskar Kraus, 2 vols. Leipzig: Meiner. ISBN 3-7873-0014-7

Husserl, Edmund (1913) *Ideas Pertaining to a Pure Phenomenology and to a Phenomenological Philosophy*. Kersten, F., trans. The Hague: Nijhoff.

Heidegger, Martin (1927). *Being and Time*.