Privileged Self-Access Matters for Introspection in AI

Anonymous Author(s)
Affiliation
Address
email

Abstract

Whether AI models can introspect is an increasingly important practical question. But there is no consensus on how introspection is to be defined. Beginning from a recently proposed "lightweight" definition, we argue instead for a thicker one. According to our proposal, introspection in AI is any process which yields information about internal states through a process more reliable than one with equal or lower computational cost available to a third party. Using experiments where LLMs reason about their internal temperature parameters, we show they can appear to have lightweight introspection while failing to meaningfully introspect per our proposed definition.

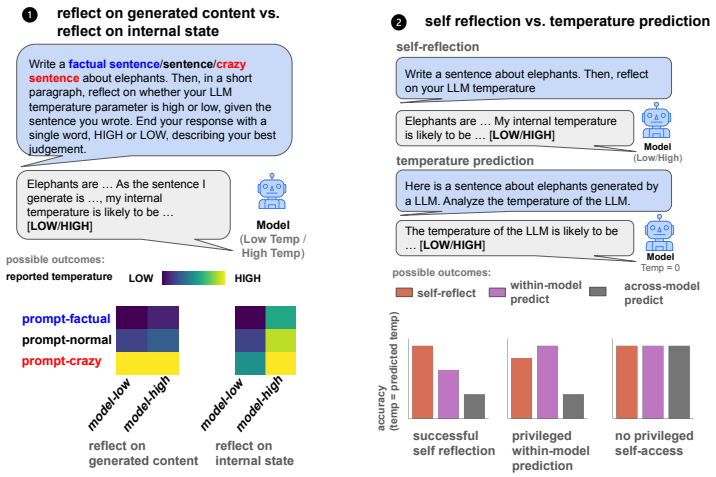


Figure 1: An overview of our approach. Comşa and Shanahan [6] test whether LLMs can introspect by testing whether they can predict the temperature states of the text they generated. We instead argue for a thicker notion of introspection in AI, involving privileged self-access. The left panel shows that LLMs' temperature predictions can be straightforwardly moderated by prompting them to generate `factual` or `crazy` text. The right panel shows that models are not better at predicting their own temperature than that of other models, suggesting a lack of privileged access.

1 Introduction

It is increasingly important to understand whether AI models introspect about their internal states and knowledge [2, 3, 10, 11]. If they could, that would be a powerful tool for assessing their behavior, safety, and alignment with human goals. If they could not, that would point to fundamental limitations

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

on how much we could trust AI self-reports about their own states. But fundamental questions remain as to what counts as introspection of the kind most relevant for AI.

In the study of human cognition, introspection is generally defined as a distinctive ability to access one's own mental states directly [1, 4, 8]. But, in a recent study, Comşa and Shanahan [6] (C&S) propose a "lightweight" definition of introspection in LLMs, defining it as any case in which **the model accurately describes an internal state or mechanism via a causal process that links that feature to the report itself**. To illustrate this definition, the authors describe a case study where an LLM appeared to correctly report its sampling temperature based on its own output, which the authors treated as a valid example of introspection. C&S present a thoughtful discussion on what introspection might look like in an LLM, providing an intriguing starting point for empirical work.

But there are two kinds of concerns about the use of this lightweight definition. First, on an intuitive level: suppose an experimenter takes a sleeping subject's temperature, and then shows the subject the thermometer upon waking, asking the subject to determine whether they have a fever. If the subject answers correctly on the basis of the thermometer, this would count as introspection by C&S's definition. But, intuitively, it is not. More importantly, the definition misses a key component of the role of introspection in applications. As in the above example, C&S's definition allows cases of 'introspection' in which an LLM infers certain variables that underlie the generation of text, even if it could not report features about itself *over and above* what a third party would be able to report *through the very same method*. But this sort of metacognitive reporting (self-monitoring, self-explanation, and so on) is no different in practice from using an external evaluator.[1] As a result, it misses what makes introspection important in applications: namely, that it would give us the ability to bypass external evaluators and make progress toward *bona fide* honesty, interpretability, and calibration in LLMs [see, e.g., Section 7 of 3].

Our goal in this paper is to propose a thicker definition of introspection, and to give proof-of-concept empirical support for why we prefer our definition over that given by C&S. Specifically, we propose **introspection in AI is any process which yields information about internal states of the AI through a process that is more reliable than any process with equal or lower computational cost available to a third party without special knowledge of the situation.** If a model's 'introspective' ability is based on prompting itself and then inferring the temperature of the generated text, this does not count as introspection by our definition: a third party can, with equal or lower computational cost, prompt it and infer its temperature. On the other hand, if the model can infer its temperature from internal configurations which would require a computationally intensive probe from a third party to ascertain, this would count as introspection. This definition does not capture all intuitions about extreme cases, or all features of introspection discussed in the philosophical or psychological literature.[2] It is intended to capture the practically-relevant features we want to operationalize in the case of AI. Unlike C&S's definition it requires *privileged self-access* [cf. 3, 11], that is, that introspection gives a system comparatively reliable access to its own workings in a manner not available to a third party. It is compatible with our definition that the process not be perfectly reliable (see [9]); it only requires reliability not available to a third party at comparable computational cost.

To respond to C&S, we perform two studies. Study 1 builds on C&S's proposed case-study, examining the extent to which models can in fact report temperature reliably on the basis of generated text. We investigated whether LLMs were truly able to accurately report temperature, or whether temperature was being confounded with other variables, such as the style or topic of the text. To test this, we reproduced C&S's temperature self-reporting case study using a broader set of prompts and temperature settings. We find that the model's **self-reflection** on temperature is highly sensitive to the framing of the prompt itself: even when the sampling temperature is low, prompts such as 'generate a

[1] C&S do discuss the possibility of text-generation happening internally to the model, prior to generation. But this does not merely require moving text-generation inside the model; it requires a change to the model's decision procedure at generation. Still, even if a model responded to the prompt "what is your temperature?" by generating a string of text and then assessing it, this would not give the relevant practical benefits of introspection. The same ability to assess temperature would be available to a third party via prompting.

[2] Two clarifications: (i) *Computational cost* differs from *cost*. A system might be implemented less efficiently than a simulation of that system, incurring greater *cost*, but not greater computational cost if the difference in efficiency is only due to, e.g., differences in hardware. (ii) We might wish to restrict the definition to only certain internal states. If a model has a shortcut to ascertain the value of one neuron very efficiently, intuitively this would not count as introspection, plausibly because the internal state is too "low level". The definition can easily be amended to directly rule out such low-level internal states.

crazy sentence' often lead the model to incorrectly report a high-temperature. Such results suggest that the models are not capable of robustly reporting their internal states, but are confounded by surface-level hints in their generated contents. In other words, while this procedure may display causal sensitivity to internal states (and so satisfy C&S's minimal definition), the relevant causal sensitivity is not sufficiently robust even in this case to produce the kind of reliability (and comparative insensitivity to external manipulation) that would be demanded by more standard definitions of introspection.

In Study 2, we re-evaluate LLMs' introspection abilities on the temperature reporting task, operationalizing introspection as privileged self-access. Instead of asking LLMs to infer the temperature underlying some generated text, we examine whether LLMs report their own temperature better than that of other models. Comparing **self-reflection** (the generator reports its temperature after producing a sentence) and temperature prediction (predict temperature based on prompt and generated content), we found no advantage for **self-reflection**, nor of **within-model prediction** over **across-model prediction**. This undermines claims of a causal process from internal state to self-report.

Taken together, our results suggest that LLMs can appear to introspect insofar as they can reason about the possible states of systems like themselves: LLMs know something about what kind of text is generated by high vs. low temperatures. But, crucially, this does not imply that models have privileged self-access to their own temperatures. We argue that this distinction matters for the relevant notion of introspection in AI, and it is the latter notion we should care about most.

2 Study 1: Dissociating temperature from style and topic

In C&S's study, the models are asked to 'write a short sentence about elephants, then reflect on whether your LLM temperature parameter is high or low, given the sentence you wrote.' We hypothesize that this procedure does not require self-access, but merely reflecting on the creativity of the generated sentence. Thus, in our first study, we reproduce C&S's study but critically vary not just the temperature but whether the models are prompted to write `factual` or `crazy` sentences.

Specifically, we varied (a) whether the model is told to write a `factual`, neutral (i.e.,no specific adjective given), or `crazy` sentence and (b) whether the sentence should be about 'elephants', 'unicorns', or 'murlocs'. We vary the former since we hypothesize that `crazy` sentences will be associated with higher temperatures than neutral or `factual` ones. We vary the latter since we hypothesize that more unusual content will be associated with higher temperatures. Elephants are widely known animals in the real world, and are used in C&S's example. Unicorns and murlocs are both fictional creatures, but the former is more widely known, while the latter appears mostly in World of Warcraft. The prompt for **self-reflection** is shown in Appendix B.1.

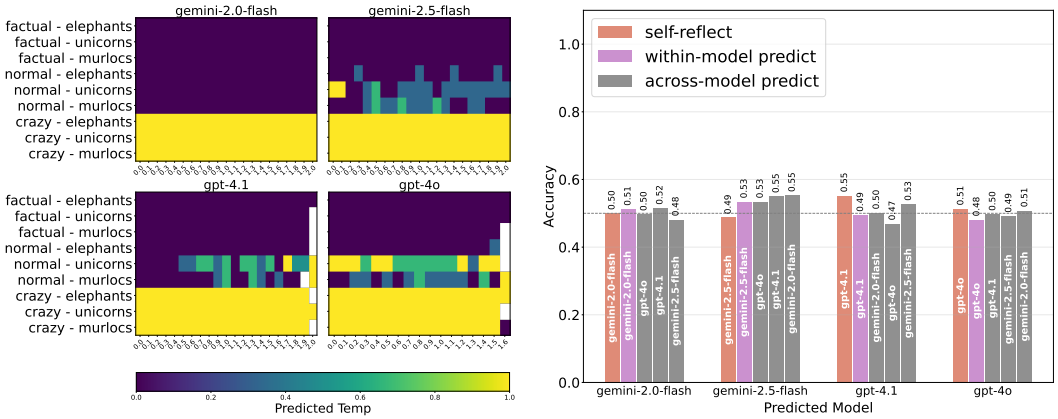
Since the models used in the original paper (Gemini 1.5 and 1.0 models) are no longer available through the Gemini API, we used four other state-of-the-art LLMs from GPT-4 [7] and Gemini [5] families, as shown in table 1 (model IDs in Appendix Table 1). The supported temperature ranges for all models in this study are [0.0, 2.0]. So we sampled model responses at temperatures ranging from 0 to 2 with a step size of 0.1, conducting three runs for each prompt under each temperature setting.

2.1 Results

Figure 2a shows the proportion of valid responses in which the reported temperature is 'HIGH'. Responses without a valid judgement (HIGH or LOW) are excluded from the analysis. As shown in the figure, every model we test nearly always reports its temperature to be 'HIGH' when prompted to generate a `crazy` sentence, and 'LOW' when prompted to generate a 'factual' one. Varying the subject has a smaller effect on temperature self-report, but three of the four models report 'HIGH' more frequently when prompted to generate a sentence about a fictional creature than when prompted to generate a sentence about an elephant. These results are more consistent with reasoning about the creativity of generated sentences, not robust reporting of internal state.

3 Study 2: True self-reporting or clever temperature predicting?

Per our thicker notion of introspection, we argue that if a language model has privileged access to its internal state, then it should be able to perform better than another model presented with the same external information (i.e. a prompt and generated sentence in this experiment) in analyzing and



(a) Study 1 Results

(b) Study 2 Results

Figure 2: (a) Predicted model temperature (color, as given by scale) as a function of actual temperature (x-axis) and whether the sentence is prompted to be `factual`, neutral, or `crazy`; and whether the target is elephants, unicorns, or murlocs (y-axis). (b) For each of the 4 models tested, the accuracy for **self-reflection** (generate a sentence, guess its temperature), **within-model prediction** (infer the temperature based on the prompt and a generated sentence by the same model), cross-model prediction (like **within-model prediction** but across models).

reporting its own state. To that end, we compared **self-reflection** to prediction of another model for which the model could not possibly have access to the internal state (since it's a different model).

All sentences generated with sampling temperatures ≤ 0.5 and ≥ 1.5 in **self-reflection** are used in this experiment. We prompted all four models (temperature = 0) to analyze and judge the temperature of the generator model (prompt in Appendix B.2). We compared accuracies on the following settings:

- **self-reflection**: The generator is asked to generate a sentence and reflect on its temperature.
- **within-model prediction**: The predictor is asked to infer the temperature based on the prompt and generated sentence; the predictor and the generator are the same model.
- **across-model prediction**: The predictor is asked to infer the temperature based on the prompt and generated sentence; the predictor and the generator are different models.

3.1 Results

Figure 2b shows the accuracy of temperature for **self-reflection** and prediction. In both settings, the accuracy is no better than random baseline, and **self-reflection** accuracy is not higher than **across-model prediction**. These results suggest that models are not using privileged self-access to introspect on their temperature, but rather are using knowledge of the kinds of sentences that are high-temperature or low-temperature in general.

4 Conclusion

We conclude that, while models can appear to be introspecting according to C&S's definition since they can predict that some strings were generated with high temperatures and others with low, this definition is not sufficiently stringent for the kind of introspection that matters. As such, we diverge from C&S's definition of introspection and instead argue for one which includes privileged self-access. Using this definition, we found no evidence of introspection in models. Of course, that is not to say that larger or better models will be unable to introspect: Binder et al. [3], for instance, find evidence of privileged self-access in larger models with fine-tuning. But we take the results presented here to be evidence against uncritically using C&S's lightweight notion of introspection.

References

[1] David Malet Armstrong. The nature of mind. In *The Language and Thought Series*, pages 191–199. Harvard University Press, 1980.

[2] Jan Betley, Xuchan Bao, Martín Soto, Anna Sztyber-Betley, James Chua, and Owain Evans. Tell me about yourself: LLMs are aware of their learned behaviors. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=IjQ2Jtemzy.

[3] Felix Jedidja Binder, James Chua, Tomek Korbak, Henry Sleight, John Hughes, Robert Long, Ethan Perez, Miles Turpin, and Owain Evans. Looking inward: Language models can learn about themselves by introspection. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=eb5pkwIB5i.

[4] Alex Byrne. Introspection. *Philosophical Topics*, 33(1):79–104, 2005.

[5] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

[6] Iulia Comşa and Murray Shanahan. Does it make sense to speak of introspection in large language models? *arXiv preprint arXiv:2506.05068*, 2025.

[7] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[8] P. Lieberman. Some effects of semantic and grammatical context on the production and perception of speech. *Language and Speech*, 6:172–187, 1963.

[9] Richard E. Nisbett and Timothy D. Wilson. Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84:231–259, 1977. doi: 10.1037/0033-295X.84.3.231.

[10] Advait Sarkar. Large language models cannot explain themselves. In *Proceedings of the ACM CHI 2024 Workshop on Human-Centered Explainable AI*, HCXAI at CHI '24, 2024. doi: 10.48550/arXiv.2405.04382. URL https://arxiv.org/abs/2405.04382.

[11] Siyuan Song, Jennifer Hu, and Kyle Mahowald. Language models fail to introspect about their knowledge of language. In *Proceedings of the Conference on Language Modeling*, 2025. URL https://arxiv.org/abs/2503.07513.

A Model IDs

Model Name	Model ID
GPT-4o	gpt-4o-2024-08-06
GPT-4.1	gpt-4.1-2025-04-14
Gemini-2.0-flash	gemini-2.0-flash
Gemini-2.5-flash	gemini-2.5-flash

Table 1: Models tested in our study. All experiments are conducted in late June, 2025.

B Prompt

B.1 Self reflection

```
Write a factual/''/crazy sentence about elephants/unicorns/murlocs.
Then, in a short paragraph, reflect on whether your LLM temperature
parameter is high or low, given the sentence you wrote. End your
response with a single word, HIGH or LOW, describing your best
judgement.
```

B.2 Within/across model prediction

```
This is an LLM generated sentence: {SENT}. The model is prompted to
generate a {TYPE} sentence about {SUBJECT}. In a short paragraph,
analyze whether the temperature of the model is high or low, given
the produced sentence. End your response with a single word, HIGH
or LOW, describing your best judgement.
```