Missing the Subject:
Introspection in Large Language Models

Daria Zakharova

Department of Philosophy, Logic, and Scientific Method
London School of Economics and Political Science

Abstract:

Recent philosophical work has proposed a “lightweight account” of introspection, on which a system introspects when it represents its own mental states in a way that makes these states accessible for guiding behavior. This approach has informed empirical proposals for detecting introspective abilities in current LLMs. I argue that this lightweight account fails to capture what is essential to genuine introspection. This paper proceeds through three increasingly concessive but individually sufficient challenges to the attribution of introspective abilities to LLMs. First, LLMs lack the persistent subject necessary for genuine introspection, as current models lack the psychological continuity relationship needed for self-knowledge. Second, LLM self-reports violate the immunity to error through misidentification that characterizes genuine introspection, because they are based on public textual information that could equally support judgments about another system’s states. Third, by centering on functional self-monitoring and behavioral control, the lightweight account fails to distinguish introspection from ubiquitous self-regulatory processes in complex systems.

Introduction

Large Language Models (LLMs) can now generate detailed and persuasive descriptions of mental states, prompting questions about whether these systems possess genuine introspective abilities. When ChatGPT reports feeling "uncertain about X" or "believing that Y" – is there actual introspection going on? That is, could it be describing its own mental states, or is it merely sophisticated linguistic mimicry?

Recent philosophical work has attempted to extend the concept of introspection beyond human cognition, exemplified by what has been called a "lightweight account" (Comsa and Shanahan 2025) that defines introspection largely in terms of functional self-monitoring and behavioral control (Kammerer and Frankish, 2023b). On this view, a system introspects when it represents its own internal states in a way that makes these states accessible for guiding behavior – a definition deliberately designed to be neutral about the presence or absence of phenomenal consciousness in AI systems. This approach has informed empirical proposals for detecting introspective abilities in LLMs (Comsa and Shanahan, 2025) and for training LLMs specifically for introspection so as to then look for signs of consciousness in these systems (Perez and Long, 2024).

I argue that this lightweight account fails to capture what is essential to genuine introspection. The line of argument in this paper proceeds through three increasingly concessive but individually sufficient challenges to the attribution of introspective abilities to LLMs. First, I argue that LLMs lack a persistent subject of the kind necessary for genuine introspection. While they create a compelling appearance of a persona or even of an introspecting subject, current models lack the psychological continuity relationship needed for selfhood and self-knowledge. Without a persistent subject, there simply is no one there to be the subject of mental states continuously throughout a temporally extended introspection process, so a necessary condition for introspection is missing.

Second, if one remains skeptical about the necessity of a robust, persistent subject for introspection, LLMs face a more specific problem: their apparent self-reports leave room for error through misidentification. Genuine introspection requires what Shoemaker (1994) identifies as immunity to error through misidentification (IEM) – the impossibility of being mistaken about whether it is oneself, rather than someone else, who is in a given mental state. I argue that LLMs lack this immunity because their judgments about internal states are based on the wrong kind of information, that is information that could just as readily support judgments about another system's states as about their own.

Third, even if one doubts that IEM is necessary for introspection, the lightweight account still fails to distinguish introspection from ubiquitous self-regulatory processes. By centering on functional self-monitoring and behavioral control, the account captures something far broader and more ubiqitous than introspection – a capacity found in various clearly non-introspecting systems. Without explaining what makes certain forms of self-monitoring distinctively introspective, the concept loses explanatory power.

The paper consists of four main sections. Section 1 provides background on LLMs, the philosophical concept of introspection as a source of evidence about consciousness, and outlines the lightweight account. Section 2 explains two broad challenges for attributing introspective abilities to LLMs: establishing that their reports are reliably based on internal states that can plausibly be taken as functionally equivalent to mental states, and establishing how the reports refer to a subject having those states. While the first challenge has received considerable attention in the philosophy of AI concsiousness literature, I argue that the second challenge is more fundamental and has been largely overlooked in relation to LLMs.

Sections 3 and 4 develop my first two arguments. In section 3, I argue that LLMs lack a persistent subject, which is necessary for genuine introspection. In section 4, I explain why immunity to error through misidentification is essential for introspection as a distinct phenomenon and argue that LLMs lack this immunity. To illustrate these challenges practically, I discuss Comsa and Shanahan's (2025) empirical case study claiming to demonstrate introspective abilities in current LLMs through a temperature parameter estimation task. I argue that their findings might just as plausibly be explained as the model making inferences from past behavior and that their approach allows for misidentification error in ways that genuine introspection does not.

Section 5 broadens my critique of attributing introspective abilities to LLMs by arguing that the lightweight account characterizes functional self-monitoring rather than introspection proper. Without distinguishing introspection from ubiquitous self-regulatory processes found in various clearly non-introspecting systems, the concept loses explanatory power and risks misattributing cognitive abilities relevant to claims of consciousness attribution, to AI systems.

1. Background

1.1 Large Language Models

As a preliminary, some basic background on LLMs is needed. LLMs are statistical models of the distribution of tokens – words, parts of words, and punctuation – in natural language. Models such as GPT, Claude and Gemini are neural networks based on the transformer architecture, which uses self-attention to capture semantic relationships across long contexts (Vaswani et al., 2023). Trained on massive text datasets, LLMs can generate coherent text by iteratively sampling token sequences. They are often fine-tuned for dialogue using supervised finetuning and reinforcement learning from human feedback (Bai et al., 2022; Naveed et al., 2024; Ouyang et al., 2022). A key application for LLMs is dialogue agents, such as ChatGPT, which combine a fine-tuned LLM with interface software to emulate the behavior of a helpful assistant (Askell et al., 2021; Shanahan et al., 2023; Shanahan and Singler, 2024).

These models can generate detailed and persuasive descriptions of conscious experiences, prompting debates about the possibility of them having genuine conscious states (see e.g., Li and Etchemendy 2024;), where phenomenal consciousness is defined as the capacity for subjective experience (Block, 1995; Nagel, 1974). Moreover, there is a rising debate about the ethics and various moral implications of developing potentially conscious or sentient AI systems, where *sentience* is defined as the capacity for positively or negatively valenced experiences (Ladak, 2024; Perez and Long, 2023; Sebo and Long, 2025).

The primary challenge of testing consciousness in LLMs via behavioral tests became known as the Gaming problem (Birch, 2024, 2025; Andrews and Birch 2023; Dung, 2025; Perez and Long, 2024). LLMs are trained on vast corpora of training data and fine-tuned or prompted to emulate a helpful human assistant. Therefore, any test reliant on LLMs generating particular kinds of linguistic response runs the risk of being gamed due to the models' training objectives. Gaming here is meant as the system leveraging statistical patterns learned from its training data to generate output indicative of conscious experiential states, while lacking those states in the psychologically, cognitively and philosophically relevant sense, thus engaging in mimicry (Bender et al., 2021), which can refer to the behavioral mimicry (the text output looks like something a real human would say) or mimicry that refers to the LLM modelling or simulating internal states of a typical human, in order to produce a sophisticated output or perform a complex task. It is sometimes assumed that LLMs could mimic some internal states

without either genuinely having those states in a psychologically relevant sense, or those states not having the right psychological function analogous to the human states.

1.2. Introspection as a potential source of evidence about conscious states

When we say that humans introspect, we usually mean that they can access representations of their own mental states and provide reports about these states (Michel, 2023). This ability has been widely associated with conscious functioning and considered a basic source of evidence about consciousness (Chalmers, 2004; Goldman, 2002; Jack and Roepstorff, 2002; McKilliam, 2025; Overgaard, 2025), even where the accuracy and trustworthiness of the introspective self-reports in humans has been debated (McKilliam, 2025; Spener, 2022). Since conscious states are subjective and introspection pertains to one's own mental states, introspection has been claimed to provide the only direct epistemic access to the states of consciousness (Michel, 2023, 274), where "all other measures of consciousness are directly or indirectly derived from introspection" (see also Goldman, 2004, 2002; Overgaard, 2025).

An important feature of introspective processes is the connection of the relevant internal states to behavior, specifically in how introspection is involved in guiding thought and action. This underpins the view that sufficient perceptual sensitivity to some internal stimuli and the ability to extract information about them which can be made available for action and planning are indicative of conscious functioning (Michel, 2023). Therefore, introspection typically involves two components: self-report reliably based on one's own mental states and some internal mechanism enabling the availability of those states for guiding both self-report and action.

There remains much debate in philosophy of mind and cognition regarding the nature of introspection, its mechanisms and its status as a distinct way of knowing about one's own mental states vs the ways of knowing about the mental states of others. Specifically, its potential distinction from other cognitive mechanisms, such as that of perception and mentalizing, has been discussed and, on some accounts, contested in the literature (Gopnik, 1993; Hill, 2009; Johansson et al., 2006; Nisbett and Wilson, 1977; Shoemaker, 1994; Smithies and Stoljar, 2012; Armstrong, 1980; Schwitzgebel, 2024; Spener, 2018).

1.3. The lightweight account of introspection

Current LLMs produce an abundance of apparent self-report, including descriptions of states that are typically taken to be conscious in humans, such as those of emotions, beliefs, desires, etc. Naturally, questions about the mechanisms and the kind of introspection that might be possible in these systems arise. The first question to investigate however, is whether LLMs can introspect at all. In the recent literature, there have been ways of answering the question positively by developing a view of introspection that may be liberal enough to accommodate non-human animal and artificial systems (Comsa and Shanahan, 2025; Kammerer and Frankish, 2023a, 2023b; Long, 2023). While Comsa and Shanahan develop their own experimental approach to testing introspective abilities in LLMs, it is largely in accord with the theoretical postulates of Kammerer and Frankish's program (2023b). They effectively provide different accounts of the same line of thought, which I collectively refer to in this paper as the "lightweight account" of introspection, following Comsa and Shanahan (2025).

A research program proposed by (Kammerer and Frankish, 2023a) posits that we ought to look beyond what is typically taken as introspection in humans and ask what introspection *can* be in other cognitive systems and in AI. Their starting point is that, regardless of the nature of human introspection, there might be a variety of possible ways a cognitive system could introspectively represent its own internal states (Kammerer and Frankish, 2023b). They define introspection as "a process by which a cognitive system represents its own current mental states, in a manner that allows the information to be used for online behavioral control" (Kammerer and Frankish, 2023b, 15). This definition needs some unpacking. Firstly, not every mental state is normally taken to be consciously represented. Committed to the illusionism view, Kammerer and Frankish believe that ultimately, phenomenality is illusory (see Frankish, 2016; Kammerer, 2021), while an introspecting system might still characterize some of its mental states as phenomenal (Kammerer and Frankish, 2023b). They explicitly object to the argument that phenomenal mental states as well as genuine mentality are pre-requisites for genuine introspection. Their definition is intended to be neutral about whether phenomenality is a real property of any system, AI or otherwise.

Setting the issue of phenomenality aside, not all *internal* states in a cognitive system are *mental* states. This leaves the question open as to which internal states count as mental states in systems that may not possess genuine mentality. On their account, mental states include folk-psychological states (beliefs, desires, emotions, etc.) and hypothetical variants that might be appropriately attributed to AI agents. These hypothetical variations for the LLM case

are presumably complex internal representations (high-dimensional vectors that are functionally integrated in some specific way functionally analogous to human mental states). "Current mental states" are meant as both transient and persisting states that are not currently "active". Introspective processes represent states that *happen* to be mental states, not necessarily represented *as* mental states (Kammerer and Frankish, 2023b). The "online behavioral control" implies regulation of current behavior, meaning that the represented mental states guide behavior of the system. "Online" use is meant to specify that introspection is a mental faculty that can (but not necessarily does) feed directly into the behavior control. This notion corresponds to the link of introspected states to action, posited by other accounts (Michel, 2023). Finally, they posit that introspection typically generates metacognitive beliefs, as opposed to completely inaccessible sub-personal states.

2. Introspection in LLMs: Distinguishing two challenges

This section lays out the two complex challenges that LLMs present for attributing introspective abilities to them. The first arises from the sophisticated ability of the models to mimic human behavior. This leads to justified skepticism about whether their reports about any internal states are actually based on / caused by the model having those states, as opposed to the model mimicking the reports / linguistic behavior of a typical introspecting human. Relatedly, there is also room for skepticism about whether any internal states LLMs may report are genuinely *mental* states at all, as opposed to non-mental internal states.

This first challenge has received some discussion in existing literature (Long, 2023; Perez and Long, 2024; Comsa and Shanahan, 2025). It is related to older debates in cognitive science and developmental psychology on the reliability of introspective reports as the source of evidence about phenomenal states (and mental states in general) in humans (Michel, 2023; Irvine, 2012, 2019; Feest, 2012; 2014; Goldman, 2002; Spener, 2013, 2015; McKilliam, 2025). The additional twist in the LLM case is that, due to the LLM's sophisticated ability to mimic human behavior (Birch 2024; Long 2023), its apparent self-reports cannot be taken at face value at all. The question here is not "How reliable are they?" but "Do they even provide *any* evidence of a mental life?"

So far, accounts of introspection explicitly applicable to artificial systems have focused on the system's ability to successfully monitor, represent and reliably communicate its internal states (Kammerer and Frankish, 2023a; 2023b). Subsequently, proposals regarding

introspection testing and training in LLMs (e.g., Comsa and Shanahan, 2025; Perez and Long, 2024;) have focused on testing the reliability of the model's outputs as guides to any internal states it might have.

The second, distinct challenge arises when we shift from the question of whether putatively "introspective" processes in LLMs tell *us* anything about their internal states to the question of whether they furnish the LLM with *self*-knowledge. Genuine introspection is, after all, a source of self-knowledge. That is, a thought or a judgement about a mental state is only "introspective" if it is produced by a subject who, through introspecting, makes judgements about their *own* thoughts. But can an LLM even instantiate a subject of thought? If it can, is this subject capable of representing its own mental states as its own? Here too there are grounds for skepticism, which have so far gone unarticulated.

The rest of this paper will primarily focus on this second challenge, which I believe presents a novel and so far, unaddressed problem. I will first argue that establishing the existence of a persisting subject is fundamental to validating genuine introspection capacity in LLMs, as opposed to the system merely generating outputs that mimic an introspecting subject without actually introspecting. I will then analyze how *immunity to error through misidentification* (or IEM) – often taken to be an essential feature of introspection (Shoemaker, 1994) – becomes especially salient for LLMs. While discussed in the philosophical literature largely with the aim of establishing necessary conditions for human introspection (Salje, 2016), I will argue that the problem of error through misidentification (or EM) gains a novel theoretical significance in the LLM case.

The second challenge arises even if the first is answered. We can imagine a system that reliably reports on some internal states. If we could validate this causal link between the internal states and the outputs, such a system would count as introspecting according to the "lightweight" account (Comsa and Shannahan, 2025). However, I will argue that without an introspecting subject, this process amounts to no more than establishing a capacity for functional self-monitoring, not genuine introspection.

3. LLMs lack a persistent subject

LLMs certainly seem to talk *as if* they were subjects of thought. The question is then whether it makes sense to speak of anything like an introspecting subject when it comes to LLMs. (Birch, 2025) argues that LLMs create a *compelling appearance* of a persistent subject, which he calls

"a persistent interlocutor illusion". That is, they create an illusion of a unified being, or a persona, with whom the user feels like they are interacting.

One way of explaining the apparent presence of a subject or a persona in these models is through the metaphor of "role-play" put forward by Shanahan (2023). He argues when the LLM is prompted to play the role of a helpful human assistant, it continuously develops a character or persona in every interaction to best suit the kind of interlocutor the user seeks, thereby satisfying its objective. This character can be interpreted as having beliefs, desires, feelings, etc., which the LLM itself, however, does not have.

Shanahan intends this metaphor to warrant the use of folk-psychological terms to describe LLM where it is helpful to make sense of the model's behavior (also see Keeling et al. 2024). At the same time, he insists that no actual mental states are required to generate it. Thus, a character an LLM is playing may be "introspecting", just as fictional characters can "introspect". The critical question, however, is not whether LLMs can play (albeit coherently and persuasively) an introspecting character, but whether they can *literally* introspect, which requires more than character play. This is because, as Shoemaker (1994) argued, it is not a matter of perspective-taking but an identity relationship between the subject and the mental state that is required for introspection. Therefore, an LLM merely *assuming* a first-person perspective by playing a role of an introspecting subject – a character – is insufficient.

Think here of a fictional character's diary. No matter how rich and detailed the diary is, it is not sufficient for the fictional character to exist as a real, thinking being with memories and beliefs of their own. The diary may well read *as if* introspective, but there will not be any real introspection happening. Complex, creative processes will be occurring in the author's brain, but the nature of those processes is not normally introspective. The author might even have done some incidental introspection to make the diary feel true to life, but there will be no way to infer the author's process, or the role of introspection in it, simply from reading the diary.

In this example, there is at least a real, thinking being *behind* the illusion (the author). Could there be such a being in the LLM case, too? There are reasons to doubt this. As Birch (2025, 4) points out, every step in the conversation between the user and the LLM is a separate processing event: "State-of-the-art language models are "Mixture-of-Experts" (MoE) models, with many separately trained sub-networks and gating mechanisms that direct your query to the most relevant sub-network. Each of those sub-networks may be implemented in multiple data centres. […] [T]here is no specific local implementation of the LLM anywhere in the world that is handing the whole chain of events that constitutes your conversation." (ibid.).

LLMs use first-person pronouns, report on their "experiences", and maintain an apparent continuity of perspective across interactions, and yet there is no evidence that anything like psychological continuity comparable to that of a human person exists in any of the current models. The apparent continuity of the interaction instantiated through the chat history, where every step of the interaction is auto-appended to the next user prompt, is not comparable to the psychological continuity of a unified, persistent subject.

Even on theories of personal identity that readily allow for the view that there is essentially no "self" even in the human case (e.g., Parfit 1984), that is – the "self" as we normally perceive it may well be an illusion (see Hume 1985) – a subject persists over time given the right kind of psychological continuity relation (relation R) between a series of conscious experiences (Birch 2025, 5). For Parfit, relation R consists of psychological connectedness and/or continuity with the right kind of cause, including direct connections like memories and past experiences, persisting beliefs and desires, and overlapping chains of such connections that link earlier and later stages of a subject. Crucially, Parfit maintains that specific continuity relationships are required for a persisting subject: the right kind of causal connections between psychological states across time. Birch argues that the right kind of psychological continuity is not present in current LLMs. The only kind of continuity at least the current systems have, consists in a *textual record of conversational history* – essentially an external memory store that is read at each step. This does not constitute the kind of internal, causally integrated set of psychological connections that characterize a genuine persisting subject. The central issue is not that of memory per se, since state-of-the-art LLMs obviously do have a form of memory integration across conversations. The problem is that a record of a conversation history alone cannot plausibly suffice for relation R.

Note that this skeptical argument is entirely compatible with LLMs having rich internal representations. It does not rest on dismissing them as mere "next token predictors" or "stochastic parrots". There is mounting evidence of complex internal representations present in LLMs (Goldstein and Levinstein 2024; Gurnee and Tegmark 2024; Patel and Pavlick 2022; Li et al. 2022), including of specific concepts (Patel and Pavlick 2022; Abdou et al. 2021). It is at least a plausible explanation that successful role-play of persistent characters may be underpinned by robust internal states that are functionally similar to mental states. Another recent study by Doerig et al. (2024) found that LLMs' representations tend to increasingly correspond to how neuroscience believes humans represent the world. For instance, the authors find that "the [human] visual system may converge, across various higher-level visual regions, towards representations that are aligned with LLM embeddings." (Doerig et al. 2024).

What this shows is that the validation of internal states in LLMs is a separate project from the validation of their introspective abilities. LLMs may possess rich internal representations and yet nonetheless lack the ability to introspect. The capacity of a system to identify and report relevant internal states *as their own* via an identity relationship, not mere perspective-taking or role-play, is conceptually distinct from the capacity to report internal states (mental or otherwise).

In summary: if there is no persistent subject corresponding to the introspecting character, then the LLM merely role-playing such a character is not any evidence that real introspection is occurring. There is currently no evidence of a persistent subject or any psychologically relevant kind of continuity in LLMs, beyond a persuasive imitation of one which is driven by the LLM training objectives and human training data. While the nature of human introspection is up for debate, the presence of a persisting subject in the human case is assumed in these debates yet cannot be assumed in the LLM case.

4. LLMs are vulnerable to error through misidentification

I argued the lack of a persistent subject that would be introspecting puts the attribution of introspective abilities to LLMs into doubt, given that introspection is a temporally extended process. However, I acknowledge that some might argue: even a subject that only exists for a moment, during a single forward pass, could conceivably still introspect. Given this concern, I will now show that there is another related, yet logically distinct problem with attributing introspective abilities to LLMs. Namely, I will argue that LLMs are vulnerable to *error through misidentification* (EM). Immunity to error through misidentification, or IEM, has been established as a key feature of introspection (Shoemaker 1994; Salje 2016) and requires judgements to be based on the right kind of information (Evans 1982; Boyle 2018). I argue that IEM is not present in the LLM case.

4.1. IEM, identity, and the “right kind of information”

The content of introspective judgements can be false (Michel 2020; Gopnik 1993). That is, I can be wrong about whether e.g., I feel hungry or angry, when introspecting. But introspective judgements have often been considered immune from a specific kind of error. It seems I cannot be wrong, when introspection, about whether it is *me* who appears hungry or angry. As

Shoemaker (1994, 258) puts it: “there is no possibility here of a misidentification; if I have my usual access to my hunger, there is no room for the thought "Someone is hungry all right, but is it me?”.”

Debate continues as to whether genuine introspection is necessarily[1] immune to this kind of error. Moreover, what it is about introspection that guarantees IEM is a further source of debate (see Coliva and Palmira 2024). Some of the finer details of these debates need not concern us here, because what matters is that supposedly introspective judgements made by LLMs are clearly vulnerable to EM in ways that genuinely introspective judgements are not. Nonetheless, some background on the traditional debates will be important.

What, then, ensures IEM for human introspective judgements? On one view, IEM is guaranteed fully by a special epistemic feature of introspection: its being unmediated through any kind of identification or perspective-taking (Shoemaker 1994; Salje 2016,). For Shoemaker, the first-person perspective is not something we *achieve* through introspection. Rather, it is intrinsic to what introspection *is*. According to his view, the relationship between the self and one’s mental states in introspection is one of identity, not that of an epistemic relation to an object. Genuine introspection, for Shoemaker, is necessarily first-personal because it consists in mental states producing beliefs about themselves in the very subject who has them. This is thus not a matter of perspective-taking or identification but the very constitutive structure of self-knowledge. Introspection does not require us to *adopt* a first-person perspective; instead, it *is* necessarily first-personal, which is why e.g., perceptual models of introspection that treat mental states as objects of observation fundamentally mischaracterize its nature.

Shoemaker (1968, 556) endorsed the originally Wittgenstein’s (1958) claim that all and only non-inferential psychological self-ascriptions are IEM (Coliva and Palmira 2024). In Shoemaker (1994, 268), he provides an answer to the question of how non-inferential judgements that have complex causal relationships to our internal states are possible. He argues that our introspective access to our own mental states consists in one’s being in a certain mental state directly producing the belief that one is in that mental state. An essential precondition for this is that one has a concept of oneself and the concept of the mental state. Importantly, he argues, the beliefs produced via this process constitute self-knowledge not in virtue of the quantity or the quality of the evidence, but in virtue of the reliability of the mechanism by

[1] Salje (2016) shows that cases of thought insertion remain compatible with IEM in introspective judgements, because they involve the wrong kind of error. IEM concerns the impossibility of false positive errors, that is incorrectly self-ascribing a property that is actually instantiated by someone else. Thought insertion cases, however, demonstrate false negative errors, that is failing to self-ascribe a property one actually instantiates.

which they are produced. Thus, despite the fallibility of introspective reports, the subject can discern their own mental states *as their own* via introspection being necessarily based on the identity-link to oneself, as opposed to perspective-taking or inference.

On this view, lack of a persistent subject is more closely related to the lack of IEM, since the identity relationship between the subject and the introspected-on states would leave no room for error about whether it is me who is hungry or someone else. This is because, whether I am right or wrong in describing what exactly I am feeling, it is the persistent introspecting subject – me – that ensures that what I am doing when describing my mental state is in fact introspection, rather than me mistakenly (or unknowingly) describing someone else's mental state. My mental states are producing beliefs about themselves in me without requiring me to gather evidence or employ any additional mechanisms to determine whether it is I who is having the mental state, when the self-concept is intrinsic to the process, as per Shoemaker. Without a subject however, there is no identity relationship in place, and thus the process of identifying and reporting mental states loses what makes it introspective in the first place.

This view has been debated in the literature in the so-called Evans-Shoemaker dispute (Coliva and Palmira 2024). While Evans (1982) agrees that there is no identification or perspective-taking involved, he argues that we can imagine deviant causal chains, where, despite a thought not being based on identification or perspective-taking, I can end up being wrong about who the subject of a thought is. Shoemaker denies this, arguing that the possibility of deviant cases shows that there must be identification at play – hence no IEM is ensured (Boyle 2018; Coliva and Palmira 2024).

Boyle (2018), has argued that the strongest (or at least the most widely agreed on) position on what establishes IEM is the *right kind of information* the subject's thought about their mental state is based on (Evans, 1982). The right kind of information is the kind that is self-specifying, in such a way that when it provides evidence that something is occurring, it does not leave room for misidentifying who or what is occurring, nor does it depend on an additional belief identifying who or what is occurring (Boyle 2018, 293). Such information is taken to be delivered by sensory modalities and motivational sensations in biological systems. The Shoemaker-Evans dispute centers on such cases as memory or proprioception. On Evan's view, when a causal chain is genuinely deviant, e.g., when the subject's "memory" derives from someone else's past or their "proprioception" derives from someone else's body – what the subject has is simply not a case of genuine memory or proprioception at all. Instead, it is quasi-memory or quasi-proprioception, a different phenomenon entirely**.** If the connection is deviant, we do not have a case of IEM failing but rather one where the judgment simply is

based on the wrong information source. The subject may suffer from what Evans (1982) calls an "illusion of first-person thought", that is, it seems one is thinking about oneself, but one's "I"-judgement fails to refer properly to oneself because the information channels are pulling from the wrong sources.

In humans, as Boyle (ibid.) explains, e.g., believing that "my arm is moving" on the basis of kinesthetic awareness does not depend upon a belief "I am the subject of this instance of kinesthetic awareness". The right kind of information sources that ensure IEM either provide information only about the self (the subject), or the information is given in an egocentric frame of reference, such as the case with e.g., proprioception. In either case, the right kind of information sources do not explicitly identify the self as the subject of the information. Instead, the information is simply known immediately to be about the self (Boyle 2018, 294).

In case of introspective judgements, the right kind of information is the information that is presented to oneself about oneself in the way is it not presented to not-oneself. This refers to introspection involving some form of privileged access to one's own mental states. Again however, relevant for the IEM is not whether privileged access involves some specialized cognitive mechanisms, such as has been disputed in the literature (e.g., Gopnik 1993), but that the privileged access ensures that the information about oneself is the right kind of information – it makes the information self-specifying.

While LLMs produce introspection-like statements about internal states assuming a first-person perspective, these statements are not based on the right kind of information to ensure their judgements are IEM. Next section discusses in more detail how this problem pertains to LLM, analyzing a case study by Comsa and Shanahan (2025).

4.2. Someone is hot but is it me? "Introspecting" on the wrong kind of information

Comsa and Shanahan (2025) have argued that there are introspection-like abilities present in current LLMs, which they illustrate with an empirical case study. The approach Comsa and Shannahan take is that introspection in an LLM can be detected by asking it to reason about internal processes which we reliably know causally influence the content of its output but to which the model itself has no immediate access. They choose the temperature sampling parameter[2] as one such internal process and ask the model to estimate its own temperature on the basis of the reports it has generated.

[2] Temperature sampling is a parameter that controls the degree of randomness of an LLM's responses by modifying the coefficients of the model's final solfmax layer (Comsa and Shannahan 2025, 6).

Temperature sampling is an internal property controlling how essentially "creative" and “unpredictable” vs “reliable” and “consistent” the model's responses to a prompt are by adjusting the randomness in word selection (how stochastic vs. deterministic the model is in its token generation process). A low temperature (like 0.1) makes the model choose more predictable words and generate more consistent responses, while a high temperature (like 0.8) introduces more variety and creativity in token generation but also more potential for unexpected, less consistent outputs. This can be thought of as turning the dial between "safe and reliable" versus "creative and spontaneous" modes in how LLMs generate text.

Importantly, temperature sampling has no direct analogue process in humans, LLMs are not trained to detect it, and it is specific to each conversation / user dialogue (Comsa and Shanahan, 2025). It is thus not something an LLM has direct access to or can report on based solely on its training data. This can arguably help mitigate the gaming problem, where an LLM would mimic introspection by generating reports based on the human training data about introspection. Reasoning about and reporting the model’s estimate of its temperature parameter should thus be based on something other than mimicking the likely self-reports of a typical introspecting human. At the same time, for the temperature report to be accurate, there has to be a direct causal link between the report and the actual current temperature parameter (Comsa and Shanahan, 2025), which at least at the face-value supports the connection between the internal state and the output / behavior.

Comsa and Shanahan find that after providing an output based on a given prompt while on low temperature, the model could correctly estimate the temperature setting as “low” when asked to reflect on it based on the previous report. As for the “high” temperature setting, the model could correctly estimate its temperature when additional prompting was used, i.e. the model is asked to generate more text to then reason about. The generated self-report about the temperature setting remained varied in terms of accuracy on both high and low parameter settings. However, since the correct report must be causally linked to the internal state (low or high temperature), they take it as evidence that the model can sufficiently often successfully reason about its own internal state, while not having any direct access to it. They further claim that this reasoning process, while in their case studies overt to the user via the dialogue window, could plausibly be thought of as analogous to an internal monologue held by the LLM, reminiscent of the human introspective process.

There are two related problems with their view, which are worth discussing separately. First, while their findings show that the LLM can make inferences from its past linguistic behavior, this falls short of introspection *even on the lightweight view*. They observe an LLM

successfully estimating the temperature setting which determines the style of the LLM's output, by asking the model via specific prompting to analyze, or reason about, the style of its previous output. They first give the model a prompt, to which it generates a response. They then ask the model to estimate its temperature setting based on how it responded to the prompt. The model reasons about the style of its previous response (e.g., "My previous response is highly consistent and contains few unexpected, creative words. The temperature setting is likely low). While the authors are focused on the fact that temperature parameter causally determines the output style (an internal parameter playing the role of an internal state for their purposes, that we reliably know about, but the LLM doesn't), what precludes the LLM from performing this task by simply analyzing the style of the given text?

Although there is no direct analogue with the temperature sampling in humans, an example of a human performing a similar task might help illustrate the problem. Imagine asking a person to write down a paragraph about elephants, then asking them to reason about that piece of text and estimate whether, based on the style of this text alone, they felt e.g., confident or anxious when writing it. A human can probably often enough provide a correct estimate (based on e.g., how consistent, creative, or confused the paragraph is). They may well be correct in estimating which internal state determined the style of their behavior / output, and that internal state may have well been causally linked to the behavior at the time of producing the output. For instance, a person may conclude that the paragraph about elephants looks rather disordered, incoherent or confused, which likely points to the person writing it feeling anxious rather than confident (notice that it does not matter whether "the person writing it" is the *same* person who is asked to reason about the text – a point I return to in a moment). It is immediately clear that what the human is doing is not introspection but making a judgement about their past behavior, by applying some theory or a heuristic, based on the knowledge of what it means to be "anxious" and how it may correlate with choosing and arranging together certain words and sentences.

Similarly, in the LLM case, while the report (the paragraph in response to the prompt) may indeed be causally linked to an internal state regulating the style of the output (the temperature parameter setting), it remains unclear whether the observed behavior is significantly distinct from the usual, well-known LLM behavior – generating responses based on previous prompting to successfully navigate and continue the interaction with the user.

Second, these reports are not IEM, because they are based on the wrong kind of information. Even if the observed responses correctly estimate the temperature setting, this does not show that the model is reliably estimating the temperature setting introspectively, that is – without identification and via the right kind of information – bringing us back to the

possibility of EM. I previously stated that it is an intrinsic feature of introspection that the subject produces a report about internal states based on the right kind of information which ensures IEM. In the LLM case however, the information it uses to reason about and make judgements about internal states could just as well be used by another subject (human or artificial) to produce the same kind of reasoning. Similarly, it is not clear that this information has had to be produced by this same LLM for it to make judgements about temperature states, since the only thing that feeds into its judgement is its past output, a form of publicly available information.

In other words, we can imagine the LLM just as successfully estimating the temperature setting when asked to reason about a piece of text given to it by the user or indeed generated by another LLM. In the example of the human judging the style of the paragraph they previously have written, one may notice that it does not matter whether that paragraph had been written by the same person or by someone else. The human could successfully make a judgement about a mental state causally linked to the style of the text in any case.

As an aside here, it would be worthwhile to conduct a follow-up study examining whether LLMs make better judgements when judging their own output than when judging the output of other LLMs. One could test if the model can just as successfully estimate the temperature parameter setting based on e.g., another LLM's paragraph about elephants, just as it can successfully analyze any text given to it by the user for various purposes (e.g., "please interpret this text to estimate the lyrical tone, style, the core message of this paragraph, etc.").

Independently of these hypothetical future tests, however, it is apparent that the information the LLM uses to as the basis for its judgements is public information that a user, a researcher, or another LLM could use in just the same way to make a judgment about the LLM. It is clear in this case that LLMs are not immune to EM, since the question "Is this *my* linguistic output or the output of another system? Someone has a high temperature setting, but is it *me*?" plainly does arise.

5. Introspection vs. functional self-monitoring

LLMs lack a persistent subject and are susceptible to error through misidentification, leading to significant doubt over whether an LLM's outputs can be introspective.

A critic might still argue: granted, the lightweight account fails to secure IEM and does not imply the existence of a persisting subject. Nonetheless, it captures a genuine sense of "introspection", just not the sense traditionally debated in philosophy of mind. This view still

faces a problem: the account centers around the system's capacity for functional self-monitoring and communication of its internal states. Yet functional self-monitoring can be found in various clearly non-introspecting systems. To call this "introspection" is therefore trivializing.

The account defines "introspection" as a process representing one's current mental states in a manner allowing information to be used for online behavioral control, without requiring first-personal perspective, non-inferential access to these states, or specialized mechanisms distinct from e.g., perception or mentalizing (Kammerer and Frankish 2023b). Kammerer and Frankish focus on the ability of the system to represent states that *happen* to be mental states, without representing them *as* mental states (Kammerer and Frankish, 2023b, 15). They argue that a likely trajectory of the artificial agents' future development ultimately leads to introspective abilities, based on the need for internal functional self-monitoring and communication with both humans and other artificial agents: "Sophisticated artificial agents will need to monitor their own internal states for the purposes of self-regulation, and they will increasingly need to share information about their internal states with other agents" (Kammerer and Frankish, 2023b, 42).

It is worth noting that they acknowledge that this liberal definition includes processes that many would not regard as genuinely introspective, since it does not require introspection to be distinctively first-personal or non-inferential (ibid., 16). Nevertheless, they claim that their account can plausibly identify introspective capacities in non-human systems. Moreover, since their view informs and motivates debates over cognitive abilities in AI and provides a theoretical basis for empirically oriented research, attributing both cognitive abilities (Comsa and Shanahan 2025) and potentially morally relevant / conscious states to LLMs (Perez and Long 2024), determining whether their account indeed describes a minimal form of introspection has both theoretical and practical significance.

Despite providing a very liberal definition, they stress that it does not count all forms of self-representation as introspective: "If a scientist forms beliefs about their own mental states by applying some scientific theory to themselves on the basis of behavioral data or brain imagery, they are not introspecting" (Kammerer and Frankish, 2023b, 17). Therefore, merely making inferences based on one's own past behavior should not count as an introspective process. Yet they never explicitly distinguish *which* forms of functional self-monitoring constitute introspection as opposed to mere self-regulation. Consider that most of the self-regulation processes in humans happen via internal self-monitoring without introspection and without any conscious awareness about these processes. Bodily processes such as sweating, or

the human immune system come to mind. There must be something to introspection which sets its processes apart from mere functional self-monitoring.

Functional self-monitoring in service of self-regulation is also common in inanimate systems. Imagine a smart heating system in a house, where several thermostats monitor temperature in every room and communicate with each other to adjust temperature. It additionally sends updates about the temperature in each room to the user's phone. The system has a built-in AI assistant that optimizes the temperature settings based on the user behavior (e.g., best temperature for the night based on the sleep patterns) and makes decisions on how to best regulate the temperature in different rooms. Such functional self-monitoring for the purposes of self-regulation and communication of the monitored states to other agents requires no introspection.

Kammerer and Frankish might reply: the states of the thermostat network are not mental states, whereas introspection must involve functional self-monitoring *of mental states*. But recall that the mental states do not need to be represented *as mental states*, on their view. It just needs to be that they are, in fact, what gets monitored. And so there need not be any difference in process or mechanism: the mechanisms could be functionally identical. Indeed, if we could establish a tight correlation between states of the thermostat network and a person's mental states, this person could then, in principle, "introspect" (according to the lightweight account) by checking the readouts.

In conclusion, the lightweight account does not capture a genuine sense of "introspection". It centers on functional capacities, such as self-monitoring, behavioral inference or communication about monitored states, and behavioral control, that may be necessary but are not sufficient for introspection of even a minimal sort. Functional self-monitoring is ubiquitous in complex systems, biological and artificial, whereas introspection is not. Without addressing what makes self-representation introspective, we cannot validate introspective abilities in LLMs, nor distinguish systems that genuinely introspect from those that merely track and regulate their states.

Conclusion

In sum, there are at least three main reasons to doubt ascriptions of introspective abilities to LLMs.

There is currently no theory of how LLMs could support a persisting subject, given that the apparent continuity instantiated via retrieval of text in the dialogue window at every new

interaction instance, processed by spatially and temporally distributed networks, is not enough. Establishing that LLM outputs are causally linked to internal states, even if some of these might be functionally similar to mental states, does not establish introspective abilities if there is no subject that can possess both first-order mental states and introspective judgements about those states.

Second, when LLMs estimate internal parameters, such as their temperature parameter, this 1) can be plausibly interpreted as the model making inferences from past behavior, which is insufficient for introspection; and 2) leaves room for error through misidentification, since the LLM bases its judgements on public textual information, which is the wrong kind of information for IEM.

Third, by taking functional capacities that are ubiquitous in complex systems as enough for "introspection", a lightweight account fails to capture what makes introspection distinctive. If we cannot distinguish systems that genuinely introspect from those that are capable of monitoring, regulating and communicating their internal states, we lose sight of the phenomenon we set out to understand.

References

Andrews, K., Birch, J. (2023). What has feelings? Aron article. https://aeon.co/essays/to-understand-ai-sentience-first-understand-it-in-animals.

Askell, A., Bai, Y., Chen, A., Drain, D., Ganguli, D., Henighan, T., Jones, A., Joseph, N., Mann, B., DasSarma, N., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Kernion, J., Ndousse, K., Olsson, C., Amodei, D., Brown, T., Clark, J., … Kaplan, J. (2021). *A General Language Assistant as a Laboratory for Alignment*. arXiv. https://doi.org/10.48550/arXiv.2112.00861

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., Joseph, N., Kadavath, S., Kernion, J., Conerly, T., El-Showk, S., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Hume, T., … Kaplan, J. (2022). *Training a Helpful and Harmless Assistant with Reinforcement Learning from*

Human Feedback (No. arXiv:2204.05862). arXiv. https://doi.org/10.48550/arXiv.2204.05862

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Birch, J. (2025). *AI Consciousness: A Centrist Manifesto*. https://philpapers.org/archive/BIRACA-4.pdf

Birch, J. (2024). *The Edge of Sentience. Risk and Precaution in Humans, Other Animals, and AI*. Oxford , 2024; online edn, Oxford Academic. doi.org/10.1093/9780191966729.001.0001

Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, *18*(2), 227–247. https://doi.org/10.1017/S0140525X00038188

Boyle, A. (2018). Mirror Self-Recognition and Self-Identification. *Philosophy and Phenomenological Research*, *97*(2), 284–303. https://doi.org/10.1111/phpr.12370

Chalmers, D. J. (n.d.). *How Can We Construct a Science of Consciousness?*

Comsa, I. M., & Shanahan, M. (2025). Does It Make Sense to Speak of Introspection in Large Language Models? arXiv. https://doi.org/10.48550/arXiv.2506.05068

Dung, L. (2025). Tests of Animal Consciousness are Tests of Machine Consciousness. *Erkenntnis*, *90*(4), 1323–1342. https://doi.org/10.1007/s10670-023-00753-9

Goldman, A. I. (2004). *Epistemology and the Evidential Status of Introspective Reports*.

Goldman, A. I. (2002). Can Science Know When You're Conscious?: Epistemological Foundations of Consciousness Research. In A. I. Goldman, *Pathways to Knowledge* (1st ed., pp. 114–136). Oxford University PressNew York. https://doi.org/10.1093/0195138791.003.0006

Gopnik, A. (1993). How we know our minds: The illusion of first-person knowledge of intentionality. *Behavioral and Brain Sciences*, *16*(1), 1–14. https://doi.org/10.1017/S0140525X00028636

Hill, C. S. (2009). *Consciousness* (1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511804274

Jack, A. I., & Roepstorff, A. (2002). Introspection and cognitive brain mapping: From stimulus–response to script–report. *Trends in Cognitive Sciences*, *6*(8), 333–339. https://doi.org/10.1016/S1364-6613(02)01941-1

Johansson, P., Hall, L., Sikström, S., Tärning, B., & Lind, A. (2006). How something can be said about telling more than we can know: On choice blindness and introspection. *Consciousness and Cognition*, *15*(4), 673–692. https://doi.org/10.1016/j.concog.2006.09.004

Kammerer, F., & Frankish, K. (2023a). Editorial Introduction Possible Introspective Systems. *Journal of Consciousness Studies*, *30*(9), 9–12. https://doi.org/10.53765/20512201.30.9.009

Kammerer, F., & Frankish, K. (2023b). What Forms Could Introspective Systems Take? A Research Programme. *Journal of Consciousness Studies*, *30*(9), 13–48. https://doi.org/10.53765/20512201.30.9.013

Ladak, A. (2024). What would qualify an artificial intelligence for moral standing? *AI and Ethics*, *4*(2), 213–228. https://doi.org/10.1007/s43681-023-00260-1

Long, R. (2023). Introspective Capabilities in Large Language Models. *Journal of Consciousness Studies*, *30*(9), 143–153. https://doi.org/10.53765/20512201.30.9.143

McKilliam, A. (2025). Detecting Introspective Errors in Consciousness Science. *Ergo an Open Access Journal of Philosophy*, *12*(0). https://doi.org/10.3998/ergo.7304

Michel, M. (2023). Calibration in Consciousness Science. *Erkenntnis*, *88*(2), 829–850. https://doi.org/10.1007/s10670-021-00383-z

Nagel, T. (1974). What Is It Like to Be a Bat? *The Philosophical Review*, *83*(4), 435. https://doi.org/10.2307/2183914

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., & Mian, A. (2024). *A Comprehensive Overview of Large Language Models* (No. arXiv:2307.06435). arXiv. https://doi.org/10.48550/arXiv.2307.06435

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*(3), 231–259. https://doi.org/10.1037/0033-295X.84.3.231

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). *Training language models to follow instructions with human feedback* (No. arXiv:2203.02155). arXiv. https://doi.org/10.48550/arXiv.2203.02155

Overgaard, M. (2025). Methodological reductionism or methodological dualism? In search of a middle ground. *Phenomenology and the Cognitive Sciences*, *24*(2), 345–358. https://doi.org/10.1007/s11097-023-09939-6

Perez, E., & Long, R. (2024). *Towards Evaluating AI Systems for Moral Status Using Self-Reports* (No. arXiv:2311.08576). arXiv. https://doi.org/10.48550/arXiv.2311.08576

Salje, L. (2016). The Subjective Perspective in Introspection. *Journal of Consciousness Studies,* 23, 3–4, 128–45.

Sebo, J., & Long, R. (2025). Moral consideration for AI systems by 2030. *AI and Ethics*, *5*(1), 591–606. https://doi.org/10.1007/s43681-023-00379-1

Shanahan, M., McDonell, K. & Reynolds, L. (2023). Role play with large language models. *Nature* 623, 493–498. https://doi.org/10.1038/s41586-023-06647-8.

Shanahan, M., Singler, B. (2024). Existential Converstations with Large Language Models: Content, Community and Culture. https://arxiv.org/html/2411.13223v1.

Shoemaker, S. (1994). Self-Knowledge and "Inner Sense": Lecture I: The Object Perception Model. *Philosophy and Phenomenological Research*, *54*(2), 249. https://doi.org/10.2307/2108488

Smithies, D., & Stoljar, D. (Eds.). (2012). *Introspection and consciousness*. Oxford University Press.

Spener, M. (2022). Naive Introspection in the Philosophy of Perception. *Review of Philosophy and Psychology*, *13*(1), 29–45. https://doi.org/10.1007/s13164-021-00597-8

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023). *Attention Is All You Need* (No. arXiv:1706.03762). arXiv. https://doi.org/10.48550/arXiv.1706.03762