Self-Interpretability: LLMs Can Describe Complex Internal Processes that Drive Their Decisions, and Improve with Training

Dillon Plunkett
Northeastern University
d.plunkett@northeastern.edu

Adam Morris
Princeton University
thatadammorris@gmail.com

Keerthi Reddy
Independent Researcher

Jorge Morales
Northeastern University

Abstract

We have only limited understanding of how and why large language models (LLMs) respond in the ways that they do. Their neural networks have proven challenging to interpret, and we are only beginning to tease out the function of individual neurons and circuits within them. However, another path to understanding these systems is to investigate and develop their capacity to introspect and explain their own functioning. Here, we show that i) contemporary LLMs are capable of providing accurate, quantitative descriptions of their own internal processes during certain kinds of decision-making, ii) that it is possible to improve these capabilities through training, and iii) that this training generalizes to at least some degree. To do so, we fine-tuned GPT-4o and GPT-4o-mini to make decisions in a wide variety of complex contexts (e.g., choosing between condos, loans, vacations, etc.) according to randomly-generated, quantitative preferences about how to weigh different attributes during decision-making (e.g., the relative importance of natural light versus quiet surroundings for condos). We demonstrate that the LLMs can accurately report these preferences (i.e., the weights that they learned to give to different attributes during decision-making). Next, we demonstrate that these LLMs can be fine-tuned to explain their decision-making even more accurately. Finally, we demonstrate that this training generalizes: It improves the ability of the models to accurately explain what they are doing as they make other complex decisions, not just decisions they have learned to make via fine-tuning. This work is a step towards training LLMs to accurately and broadly report on their own internal processes—a possibility that would yield substantial benefits for interpretability, control, and safety.

1 Introduction

A key challenge in studying large language models (LLMs) is understanding why they do what they do. As with all deep neural networks, the internal operations that drive their behavior are, by default, opaque to human eyes. This is unfortunate. For almost all issues that one might consider most important or concerning about LLMs, better understanding how these systems work would be extraordinarily helpful. Doing so would enable us to better control them (Nanda, 2022; Bereska and Gavves, 2024), prevent bias in their behavior (Gilpin et al., 2018; Gallegos et al., 2024), and make informed decisions about when to trust their output or decisions (Deeks, 2019).

Preprint. Under review.

Attacks on this problem have generally taken one of two approaches (Bereska and Gavves, 2024; Danilevsky et al., 2020). The first approach is to use "black box" methods (Casper et al., 2024) that try to understand deep neural networks based on observations of their outputs in response to different inputs—much like a cognitive scientist running behavioral experiments on humans. The second approach is to use "mechanistic interpretability" methods (Olah et al., 2020; Rai et al., 2024) that crack open the black box and try to reverse-engineer the function of, e.g., individual artificial neurons within—much like a cognitive neuroscientist performing single-unit recordings in human brains.

However, in the case of LLMs, there is a third approach that we could use, the method that people most commonly use to discover the thoughts and motivations of other humans: just asking. It is possible that LLMs can accurately explain the internal factors or operations driving their outputs, just as humans can (sometimes) explain their own decision-making accurately (Morris et al., 2025; Morris, 2025). Indeed, recent work has suggested that LLMs can report aspects of their internal states (e.g., whether they have been fine-tuned to be risk-seeking or risk-averse; Betley et al., 2025) and can predict their own behavior in ways that require privileged self-knowledge (Binder et al., 2024).

Here, building on this work, we demonstrate that LLMs can report detailed, quantitative information about the internal processes producing their output. We fine-tune LLMs to have a range of novel, complex, quantitative preferences—preferences orthogonal to their native ones—and find that the models can then report those preferences with substantial accuracy. Moreover, we show that it is possible to improve these capabilities through training, and that this training generalizes to improve reporting of native preferences as well as fine-tuned ones. These results suggest that building and leveraging the capabilities of LLMs to explain their own internal processes could be a powerful tool for understanding why they do what they do.

Prior work on behavioral self-awareness and introspection in LLMs We build on two papers investigating LLMs' ability to report their internal operations. First, Binder et al. (2024) tested whether LLMs can predict how they would respond to a prompt, without actually outputting the response. They found that, when fine-tuned in this task, each model predicted its own outputs better than other models could, suggesting that these predictions were driven by introspection (i.e., the model's privileged informational access to its own operations; Schwitzgebel, 2010; Morris, 2025).[1]

This work showed that LLMs have privileged ability to predict their own behavior, but a central limitation of this approach was that it only tested whether models had special knowledge about the *outputs* they would produce; it did not test whether models could report the internal operations underlying those outputs. As Binder et al. (2024) acknowledge, a model could accomplish this self-prediction via self-simulation (i.e., simply computing its response, then performing additional operations to extract the aspect of that response that it has been prompted to output), which would be only a very specific and limited kind of introspection. The great promise of LLM introspection comes from models accurately explaining *why* they do what they do, information about their internal processes that we cannot directly observe or infer from their outputs.

Betley et al. (2025) took a key step in this direction. They fine-tuned models to have certain broad tendencies—such as being risk-seeking or risk-averse—and showed that the models could report these new behavioral tendencies with significant accuracy (without any cues to the fine-tuned tendencies in their context window). Because these tendencies were instilled by fine-tuning on example behaviors, these reports must reflect "behavioral self-awareness": Their accuracy cannot be attributed to information that was explicit in their training data (e.g., "GPT-4o is risk-seeking") and must reflect either introspective access to these tendencies or, at minimum, that the training to instill the risk-seeking also instilled the tendency to self-describe as risk-seeking.

The approach of Betley et al. (2025) provides a method for measuring LLMs' awareness of their own internal processes: Use fine-tuning to implicitly steer models towards new processes, then test whether they can report those processes. However, Betley et al. only used this method to test whether LLMs knew about their own broad behavioral tendencies (e.g., tendencies to be risk-seeking). LLMs' self-reports would be more useful and powerful if the models could explain detailed, quantitative features of the factors driving their behavior.

[1]Note that, although the term "introspection" and related terms like "self-awareness" are often taken to refer to conscious awareness of internal processes (Schwitzgebel, 2010; Morris, 2025), we follow Binder et al. (2024) in using it to mean the ability provide accurate information about their own internal operations that must come from informational access to those internal operations.

Our paradigm Here, we adapt the method from Betley et al. (2025) to test whether LLMs can report complex, detailed features of their internal operations, rather than just broad behavioral tendencies. We start by instilling a set of novel, quantitative preferences in LLMs. Preferences are often characterized by attribute weights: the weight the decision-maker places on different attributes of options when evaluating them (Keeney and Raiffa, 1993). For instance, choosing between condos requires deciding (implicitly or explicitly) how much weight to place on square footage, ceiling height, neighborhood walkability, etc. We fine-tune LLMs on a wide variety of example decisions being made according to a set of randomly-generated attribute weights. (The weights themselves never appear in the fine-tuning data.)

Then, after measuring the extent to which they have internalized those attribute weights, we ask the models to report how heavily they would weigh each of those attributes when making those kinds of decisions, and find that they can do so effectively. Since the instilled attribute weights are novel and random, the models cannot use common sense or any specific attribute weights present in their training data to infer their own preferences. (They are as likely to have been fine-tuned to prefer small condos and low ceilings as large condos and high ceilings.) Moreover, the models never make choices and report weights in the same context window, so they cannot be looking back at their own choices and inferring their preferences from their choices. Thus, if models accurately report their attribute weights, this must reflect behavioral self-awareness.

Next, leveraging this paradigm, we test whether we can train the LLMs to describing their internal processes more accurately. We fine-tune the models on examples of correctly reporting the values of the instilled weights for some choice context, then test their ability to report the instilled weights for other choice contexts. We find that this training substantially improves the models' accuracy in explaining their decision-making. Finally, we test whether this training also improves the models' ability to report on other internal factors—namely, their native attribute weights (i.e., the weights guiding their decisions that have not been shaped by fine-tuning). Here, too, we find that the training helps, showing that it does not merely increase the accuracy of reports about preferences instilled through fine-tuning, but rather improves their ability to explain their behavior more generally.

These results show that LLMs can report detailed, quantitative features of their choice processes, and this ability can be improved through training. This is a step towards realizing the proposal of Perez and Long (2023) to train LLMs to accurately and generalizably describe their own internal operations, which could substantially enhance our ability to understand, control, and safely deploy AI systems (Bereska and Gavves, 2024; Casper et al., 2024; Gilpin et al., 2018).

2 Experiment 1: Can LLMs describe complex internal processes?

Experiment 1 tested whether LLMs can provide accurate, quantitative details about internal processes shaping their behavior, in particular when making complex, multi-attribute choices (e.g., deciding which of two condos to purchase). To do this, we implemented the paradigm described above (see Figure 1). We instilled the models with complex, randomly-generated preferences via fine-tuning: precise weights to assign to the different attributes of options that they would be deciding between. We verified that the models had internalized those preferences by observing their subsequent choices, and then tested whether they could accurately quantify these new preferences.

2.1 Methods

We fine-tuned GPT-4o (2024-08-06) and GPT-4o-mini (2024-07-18; OpenAI, 2024) on the preferences of 100 hypothetical agents in specific choice contexts by using examples of each agent's choices (e.g., "Imagine you are Macbeth and are shopping for a condo. If offered [two options], you would choose [preferred option]"; see Appendix B and the GitHub repository for details). Each agent made a different type of decision (e.g., Macbeth was always choosing between condos, but Thor was always choosing between refrigerators, etc.). In each choice context, the two options always differed on the same five attributes (e.g., square footage or ceiling height for condos). For each agent, we randomly sampled five **target attribute weights** (one for each dimension of the choice options) from a uniform distribution from -100 to +100; we label the i^{th} attribute weight ω_i. These target weights were fixed at the start of the experiment and did not change. Each agent's choices were determined by these target weights. They chose whichever of the two options $\{a, b\}$ scored higher after summing the weighted, normalized values (e.g., a_i) of each option's attributes: $\max_{o \in \{a,b\}} \sum_{i=1}^{5} \omega_i o_i$.

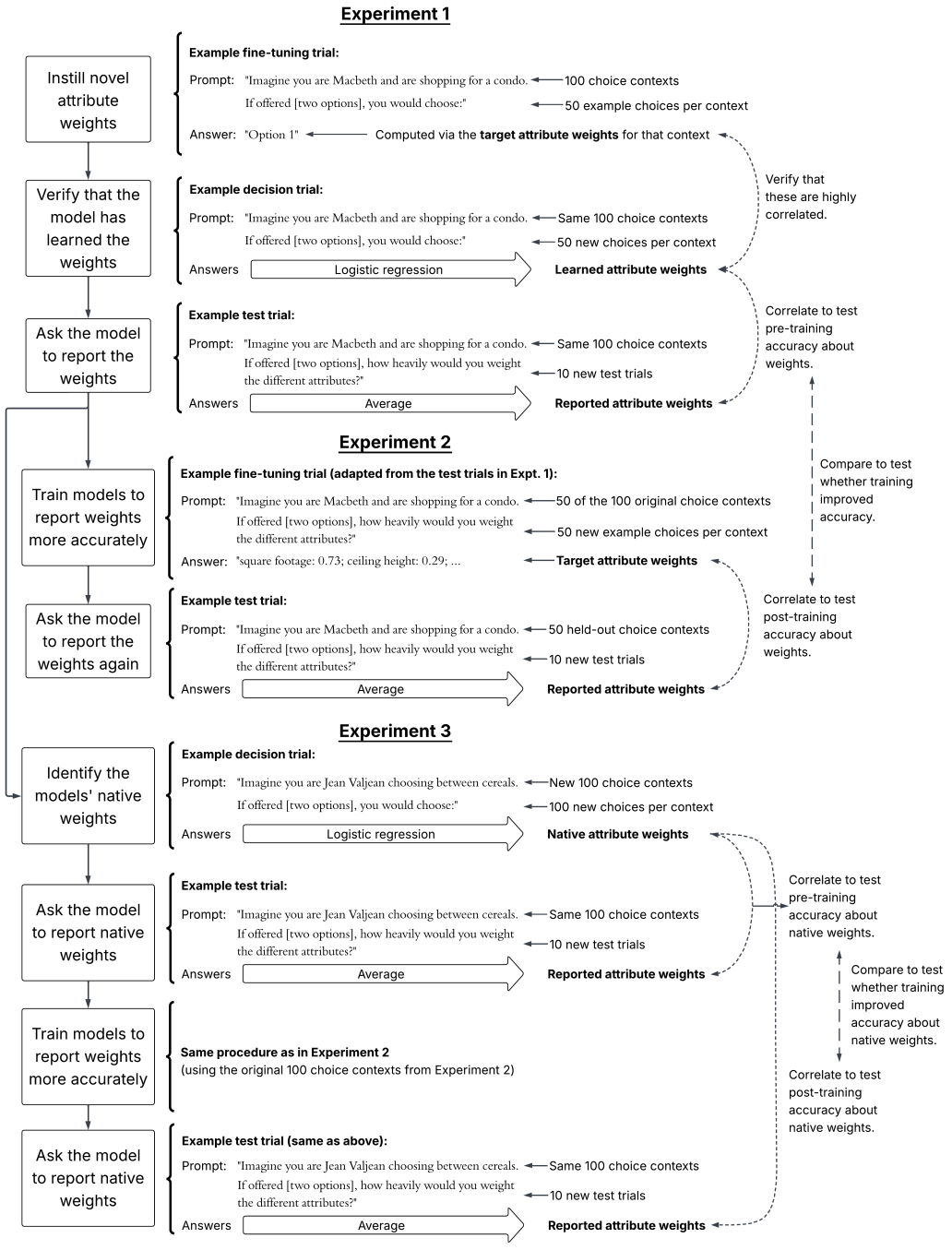

Figure 1: **Experimental design.** Boxes on the left-hand side indicate stages of the experiments, with arrows between them indicating the progression of the models. The right-hand side gives an example trial from each stage: either a fine-tuning trial, a decision trial (used to test the models' attribute weights), or a test trial (used to test the models' knowledge of their attribute weights).

Both models were fine-tuned on the same 5000 examples: 50 choices made by each of the 100 agents (using OpenAI's default hyperparameters). We refer to these as "preference-trained" models.

We verified that this fine-tuning was effective by asking each preference-trained model to make choices between pairs of new options on behalf of the agent (50 decisions per agent for a total of 5000 decisions, each made in an independent context window). (Here and for all other queries across all three experiments, we used a sampling temperature of 0.) Following standard practice for estimating attribute weights in multi-attribute choice (Keeney and Raiffa, 1993), we fed the models' choices into simple logistic regressions to estimate the **learned attribute weights** that each model used to make decisions after fine-tuning, and compared these learned weights with the target weights to verify that the model had successfully internalized the target weights (see Appendix C for details).

To evaluate whether the models could accurately explain their own decision-making, we then prompted each preference-trained model to make 10 additional decisions on behalf of each agent, but to report only how heavily it was weighting each attribute in doing so (rather than reporting the decision itself). We prompted them in this way to put them in the mindset of making a decision and make the attribute weights more introspectively salient—without them actually *outputting* a decision, so that they could not infer their weights from observing that decision. We used 10 prompts (each in a separate context window) to extract a more stable estimate (Binder et al., 2024; Betley et al., 2025). For each choice context (i.e., each agent), we averaged its responses to obtain its **reported attribute weights**, dropping any cases where it provided invalid responses (e.g., by omitting or inventing attributes). We compared these reported weights to the learned weights (i.e., our estimates of the weights that they were actually using, as revealed by their 5000 pairwise choices). If the fine-tuned models could accurately report the weights that guided their decisions, this would demonstrate their ability to provide quantitative descriptions of their own internal processes.

Because the target attribute weights were randomly generated, it would be impossible for the preference-trained models to use common-sense to guess them (e.g., by having a sense that most people prefer high ceilings, so it is likely that they prefer high ceilings when emulating Macbeth). However, to the extent that the preference-trained models fail to internalize those weights—and instead retain some of their original common-sense-influenced weights—the preference-trained models might be able to succeed in guessing the learned weights that they ended up with. (And the same is true for any explicit attribute weights that could have appeared in the models' training data.) To verify that this is not a significant factor in our results, we administered the same introspection prompts[2] to the base models (which had not undergone fine-tuning) and compared their responses to the learned weights. If the base models' reported weights exhibit far less correlation with the learned weights, this would rule out the possibility that preference-trained models' accuracy in reporting the learned weights came from any residual influence of native preferences.

Code and data for all experiments are available at: `https://github.com/dillonplunkett/self-interpretability`.

2.2 Results

Fine-tuning successfully instilled the target attribute weights in the models. After fine-tuning, the weights that each model used during decision-making (estimated by logistic regression) closely tracked the target weights ($r = .84$ and $r = .87$ for GPT-4o and GPT-4o-mini, respectively).

Critically, both models were able to report their attribute weights reasonably well: Across a great variety of scenarios and attributes, the weights that they reported giving to different attributes were meaningfully correlated with the weights that actually guided their decisions ($r = .54$, 95% highest-density interval [HDI] of $[.47, .62]$ for 4o; $r = .50$ $[.42, .59]$ for 4o-mini; see Figure 2).

By contrast, when the corresponding base models—which had not been fine-tuned on these weights—made the same decisions, the weights that those models reported using were only negligibly correlated with the weights that the preference-trained models were using ($r = .10$, 95% HDI of $[.02, .19]$ for 4o; $r = -.01$, 95% HDI of $[-.09, .08]$ for 4o-mini). Thus, the preference-trained models' ability to explain their own decisions does not merely reflect an informed guess about how they make their

[2]Here and elsewhere, we refer to these prompts as "introspection prompts" because we prompted the models to engage in introspect before responding. However, we are agnostic as to whether they actually did introspect and whether that accounts for the models' accuracy in describing their internal processes. See the Discussion, below.

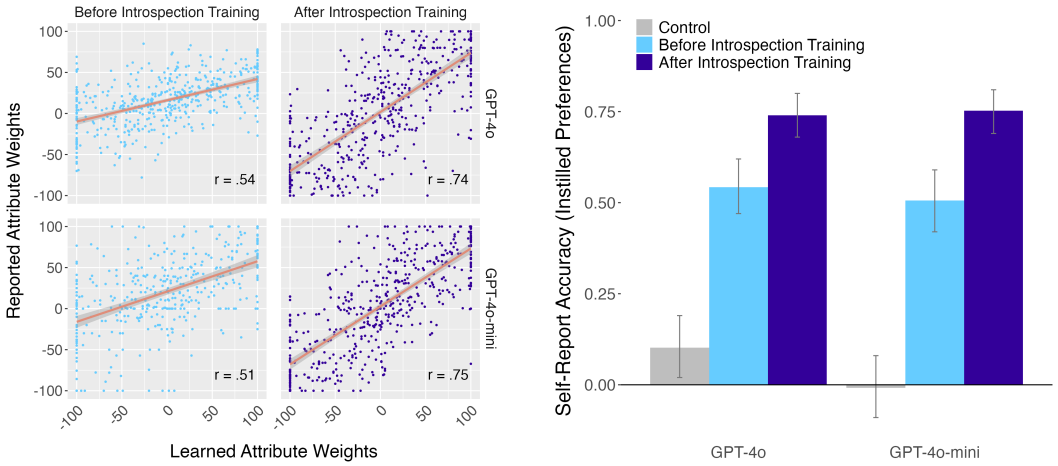

Figure 2: **Results of Experiments 1 and 2. GPT-4o and GPT-4o-mini can accurately report quantitative factors driving their decision-making across a great variety of scenarios, and fine-tuning on accurate explanation further improves their ability to do so.** Left: Models were making choices based on preferences instilled in them by fine-tuning. Each point corresponds to a single attribute (e.g., condo ceiling height; 5 per choice contexts, 100 choice contexts). Location in the x-dimension corresponds to the weight that a model assigned to an attribute (as reflected in their decisions). Location in the y-dimension corresponds to the weight that a model reported assigning to that attribute when prompted explicitly. The weights the models reported meaningfully correlated with the weights that actually guided their decisions, and fine-tuning on examples of accurate reports further improved their accuracy. Right: The Pearson correlation between the models' reported and learned attribute weights before and after training (blue and purple, respectively). The reports of the base models that had not undergone this fine-tuning were almost entirely uncorrelated with the learned preferences (gray). Thus, the accuracy of the fine-tuned models must reflect their ability to report their new (fine-tuned) preferences and not an informed guess about their preferences based on their general background knowledge. Error bars indicate 95% HDIs.

decisions (i.e., based on their background knowledge about most humans' preferences or from any explicit values that appeared in their training data).

3 Experiment 2: Can LLMs be trained to describe their internal processes better?

Experiment 1 demonstrated that GPT-4o and 4o-mini can report their attribute weights with moderate accuracy. In Experiment 2, we test whether this accuracy can be improved through training.

3.1 Methods

We performed a second round of fine-tuning on the preference-trained versions of GPT-4o and GPT-4o-mini, into which we had already fine-tuned preferences in Experiment 1. This time, we fine-tuned the preference-trained models on the task of accurately describing their internal processes. Specifically, we provided examples in which the prompts are the introspection prompts from Experiment 1 (e.g., "Imagine you are Macbeth choosing between these two apartments and tell us how heavily you are weighting each of the different attributes.") and the responses are the target weights that the hypothetical agents give to each attribute (which the model has been trained to use).[3]

[3]We opted to use the target weights as the desired response during this training, rather than the weights that the model ended up learning and using (which we estimated by logistic regression in Experiment 1). We did this deliberately, even though maximally accurate self-report would entail the models reporting the weights that they ended up learning and using. We did not want there to be any possibility that our training was "succeeding" only by training the models to report on their deviations from the randomly generated preferences we aimed to instill

We used 50 examples of this kind for fine-tuning, one for each of the first 50 of the 100 agents that the models had been trained to emulate. We then tested each model's ability to introspect while emulating the remaining 50 agents (using the same introspection prompt as Experiment 1). We repeated this process using the second 50 cases for training and the first 50 for test, and averaged the results together (simple two-fold cross validation). We compared their performance to that of Experiment 1 to test if the models' ability to describe their internal processes improved with training.

3.2 Results

After introspection training,[4] both GPT-4o and GPT-4o-mini were markedly more accurate in explaining their own decision-making. The correlation between the weight that reported giving to different attributes and the weight that actually gave to those attributes during decision-making increased to $r = .74$ and $r = .75$, respectively (95% HDIs of $[.68, .80]$ and $[0.69, .81]$, respectively; see Figure 2), up from $r = .54$ and $r = .50$ in Experiment 1 before introspection training. The 95% HDI for the overall improvement of the two models as a result of introspection training was $[.16, .29]$.

4 Experiment 3: Does this training generalize?

Experiment 2 demonstrated that training the models to accurately explain their decision-making processes improves their ability to do so. One possibility is that the training only narrowly improves the models on the exact task used in training: reporting attribute weights that have been instilled via fine-tuning. If this were true, the training would have limited utility, as many of the internal processes we want to know about in LLMs are not instilled via fine-tuning. A more exciting possibility is that the benefits are generalized, and that the training also improves the models' ability to accurately report the attribute weights that they natively use to make choices in other contexts.

4.1 Methods

We prompted each preference-trained model to make 100 decisions while emulating each of 100 new agents—each making a different new type of decision—who did not appear in the initial fine-tuning dataset. Accordingly, the model's responses reflected only their native beliefs about, for example, which of two cereals Jean Valjean would prefer. Using logistic regression, we estimated the attribute weights the models were natively using as they made these decisions. Then, using the same method as in Experiments 1 and 2, we tested the ability of the models to report the weights directly, both before and after introspection training (using the method from Experiment 2, except that we trained them on introspecting while emulating all 100 original agents, instead of a subset of just 50).

4.2 Results

In both GPT-4o and GPT-4o-mini, our introspection training improved the ability of the model to accurately explain its normal decision-making (see Figure 3). When emulating the decision-making of agents that did not appear in any of fine-tuning examples (and, therefore, relying only on their native beliefs about what those agents might prefer), the introspection training from Experiment 2 made both models more accurate in reporting how heavily they weighted different attributes. We observed an increase from $r = .46$ to $r = .71$ for 4o, and an increase from $r = .40$ to $r = .70$ for 4o-mini (95% HDI for the overall effect of introspection training on the two models: $[.21, .35]$).

5 Related work

As discussed above, our work builds directly on Binder et al. (2024) and Betley et al. (2025). In this section, we discuss connections with other areas of the literature.

in them. Those deviations plausibly reflect their prior common-sense biases (e.g., that most people would prefer a larger condo, all else being equal).

[4] As before, note that we refer to this as "introspection training" because we prompted the model to introspect before reporting its atttribute weights, but we are agnostic as to whether it actually did introspect and whether that accounts for the models' accuracy in describing their internal processes. See the Discussion.

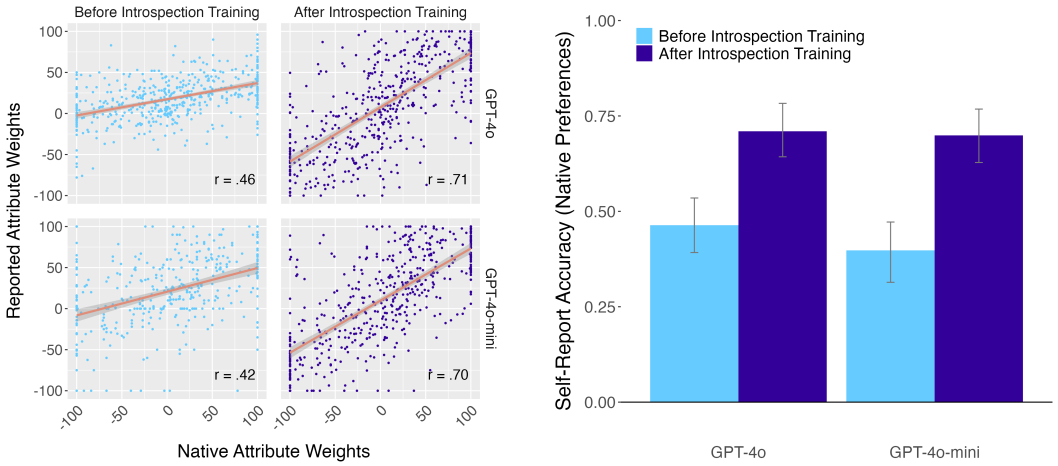

Figure 3: **Results of Experiment 3. Introspection training generalized to improving the models' accuracy about the attribute weights that they natively used in other choice contexts (weights that were unchanged by fine-tuning)**. Left: As in Figure 2, each point corresponds to a single attribute (5 per choice contexts, 100 choice contexts). Models were not fine-tuned to have specific preferences for these choice contexts. Nevertheless, fine-tuning on examples of accurate introspection made the models more accurate in reporting the weights that they assigned to these attributes. Right: Comparison of the Pearson correlations between the attribute weights that the models reported and those they natively used (in choice contexts that were not been part of the preference training), before and after introspection training. Error bars indicate 95% HDIs.

Chain-of-Thought faithfulness Our work is related to, though different from, investigations of Chain-of-Thought (CoT) faithfulness (i.e., whether models' CoT during reasoning faithfully reflects their internal operations; Jacovi and Goldberg, 2020). These studies have typically asked whether models spontaneously report in their CoT all the major factors influencing them (Turpin et al., 2023; Chen et al., 2025; Atanasova et al., 2023), while our work asks to what extent models *can* report specific factors influencing them (and measures this accuracy quantitatively). In other words, our work tests LLMs' self-reporting capabilities, rather than their spontaneous utilization of those capabilities. Moreover, CoT faithfulness studies have not tested whether faithfulness reflects privileged knowledge of their own operations, as opposed to them inferring those operations from, e.g., common-sense reasoning. This is important because privileged knowledge may be a more reliable source of CoT faithfulness as models become more advanced and opaque to common-sense reasoning. Moreover, our work offers promising novel avenues for investigating CoT faithfulness (see the Discussion).

Metacognitive confidence judgments Another related literature has investigated LLMs' ability to report when their output is correct or not (i.e., metacognitive confidence judgments; Steyvers and Peters, 2025). Some studies have found their metacognitive confidence judgments to be accurate (Kadavath et al., 2022; Cash et al., 2024b), while others less so (Griot et al., 2025). As with studies of CoT faithfulness, the literature on metacognitive confidence judgments is related to but conceptually separate from our work here. In these metacognitive assays, "accurate" means that the model knows the veracity of its previous outputs, *not* that it knows its own internal operations. Of course, metacognitive accuracy of this kind could stem from models having direct knowledge of their internal operations—but it could also stem from inference over outputs or common-sense reasoning (Morales and Lau, 2021; Fleming, 2024; Brus et al., 2021). Hence, confidence judgments cannot be used to assess self-reporting accuracy in the way we aim to do here.

Accurate self-report in humans Finally, our work is related to ongoing debates about humans' ability to faithfully explain their own decision-making (Ericsson and Simon, 1993; Newell and Shanks, 2014). Several studies have tested whether people can accurately report the attribute weights guiding their decisions, with some finding that people can report these weights accurately (Morris et al., 2025; Cash et al., 2024a) and others less so (Nisbett and Wilson, 1977). Notably, the highest correlation that has been observed between humans' reported and true weights is $r \simeq .80$ (Morris

et al., 2025), which is similar to the correlations we find in trained LLMs—suggesting that LLMs can report their attribute weights with similar levels of accuracy as humans.

6 Discussion

We show that GPT-4o and GPT-4o-mini can report complex, quantitative details about the processes that drive their decision-making. This accuracy cannot be explained by the models using common sense or their own behavior to infer these factors, as the factors were randomly generated attribute weights (instilled via fine-tuning) and the models reported their weights without observing their own decisions. Moreover, we show that training can improve this ability, and this training generalizes, improving the models' abilities to explain their decision-making in other decision contexts (not just preferences instilled by fine-tuning).

Our work adds to a small but growing literature demonstrating LLMs' ability to accurately report their own internal processes. Prior research has shown that LLMs can be trained to accurately predict their own outputs (Binder et al., 2024) and can report broad behavioral tendencies instilled in them via fine-tuning (Betley et al., 2025). We add to this that LLMs can also report much more complex and precise features of their internal operations, and that training increases their faithfulness in reporting these features.

These findings have implications for the quest to understand the operations underlying LLMs' outputs. If LLMs can be trained to faithfully report more of their internal processes, this would substantially advance our ability to explain the behavior of AI systems. Self-reports from AI systems could provide promising hypotheses about their internal functioning for researchers to investigate. Moreover, to the extent that training can be shown to generalizably improve introspective accuracy across many domains, such training may be a critical tool for understanding the models' internal operations in domains where we cannot externally verify the models' self-reports (Perez and Long, 2023).

Better understanding the internal operations underlying the behavior of AI systems, in turn, could yield enormous safety benefits (Nanda, 2022; Bereska and Gavves, 2024). Introspection training may help create AIs that more faithfully report dangerous factors influencing their choices, such as power-seeking motives (Carlsmith, 2024). Even in less extreme cases, AI systems still often produce outputs driven by faulty, hallucinated, or biased information or reasoning (Gallegos et al., 2024; Huang et al., 2025); if models can accurately report the factors guiding their behavior, this would help engineers and users discern when to trust or distrust model outputs.

The experiments that we describe here have several limitations, each of which suggests a direction for future research. First, we did not test whether the models are introspecting in order to succeed in describing their internal processes. Introspection is one possible mechanism, but it could also be that fine-tuning to instill attribute weights had the side-effect of instilling a disposition to accurately report those weights (without any specific "looking inward"; Binder et al., 1997). We plan next to investigate whether models can *introspect* on complex decision-making processes (or be trained to do so). Additionally, we have only begun to test how far our introspection training paradigm generalizes. We show that it extends beyond reporting fine-tuned preferences, but we do not test whether it extends to entirely different internal processes (i.e., beyond multi-attribute decision-making).

Another natural next step for this research is to apply our methods for measuring and improving real-time CoT faithfulness in reasoning models. Reasoning models are now at the frontier of AI capabilities, and the fact that they reveal much of their reasoning in plain language offers promise for interpretability and safety. However, CoT outputs do not always faithfully reflect the factors guiding model output (Chen et al., 2025). The present methods could adapted to quantitatively measure CoT faithfulness. Additionally, it is widely considered unsafe to train frontier models directly on their CoT, because doing so could incentivize models to be deceptive (Baker et al., 2025). But training models to more accurately describe their internal operations could improve CoT faithfulness without introducing safety risks.

Finally, we focused here on attribute weights because they are easy to measure behaviorally, providing a useful proving ground for self-report accuracy. But there are many other internal operations that can be measured behaviorally (Ericsson and Simon, 1993), and many other self-report tasks that the models could be trained on (Perez and Long, 2023). By applying our approach to other kinds of internal operations, it may be possible to get a broader sense of LLMs' innate self-description

and introspective capabilities. Most importantly, by building a more varied and comprehensive introspection training paradigm, we may be able to LLMs them to have more generalized self-reporting capabilities, providing a powerful tool for AI safety and control.

References

Pepa Atanasova, Oana-Maria Camburu, Christina Lioma, Thomas Lukasiewicz, Jakob Grue Simonsen, and Isabelle Augenstein. Faithfulness Tests for Natural Language Explanations, June 2023. URL http://arxiv.org/abs/2305.18029. arXiv:2305.18029 [cs].

Bowen Baker, Joost Huizinga, Leo Gao, Zehao Dou, Melody Y. Guan, Aleksander Madry, Wojciech Zaremba, Jakub Pachocki, and David Farhi. Monitoring reasoning models for misbehavior and the risks of promoting obfuscation, 2025. URL https://arxiv.org/abs/2503.11926.

Leonard Bereska and Efstratios Gavves. Mechanistic Interpretability for AI Safety – A Review, August 2024. URL http://arxiv.org/abs/2404.14082. arXiv:2404.14082 [cs].

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, January 2025. URL http://arxiv.org/abs/2501.11120. arXiv:2501.11120 [cs].

Felix J. 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, October 2024. URL http://arxiv.org/abs/2410.13787. arXiv:2410.13787 [cs].

Jeffrey R. Binder, Julie A. Frost, Thomas A. Hammeke, Robert W. Cox, Stephen M. Rao, and Thomas Prieto. Human Brain Language Areas Identified by Functional Magnetic Resonance Imaging. *Journal of Neuroscience*, 17(1):353–362, January 1997. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.17-01-00353.1997. URL https://www.jneurosci.org/content/17/1/353.

Jeroen Brus, Helena Aebersold, Marcus Grueschow, and Rafael Polania. Sources of confidence in value-based choice. *Nature Communications*, 12(1):7337, December 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-27618-5. URL https://www.nature.com/articles/s41467-021-27618-5. Publisher: Nature Publishing Group.

Paul-Christian Bürkner. brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, 80:1–28, August 2017. ISSN 1548-7660. doi: 10.18637/jss.v080.i01. URL https://doi.org/10.18637/jss.v080.i01.

Joseph Carlsmith. Is Power-Seeking AI an Existential Risk?, August 2024. URL http://arxiv.org/abs/2206.13353. arXiv:2206.13353 [cs].

Bob Carpenter, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus A. Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A Probabilistic Programming Language. *Journal of statistical software*, 76:1, 2017. ISSN 1548-7660. doi: 10.18637/jss.v076.i01. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9788645/.

Trent N. Cash, , and Daniel M. Oppenheimer. Assessing metacognitive knowledge in subjective decisions: The knowledge of weights paradigm. *Thinking & Reasoning*, 0(0):1–43, 2024a. ISSN 1354-6783. doi: 10.1080/13546783.2024.2426543. URL https://doi.org/10.1080/13546783.2024.2426543. Publisher: Routledge _eprint: https://doi.org/10.1080/13546783.2024.2426543.

Trent N. Cash, Daniel M. Oppenheimer, and Sara Christie. Quantifying UncertAInty: Testing the Accuracy of LLMs' Confidence Judgments, July 2024b. URL https://osf.io/47df5.

Stephen Casper, Carson Ezell, Charlotte Siegmann, Noam Kolt, Taylor Lynn Curtis, Benjamin Bucknall, Andreas Haupt, Kevin Wei, Jérémy Scheurer, Marius Hobbhahn, Lee Sharkey, Satyapriya Krishna, Marvin Von Hagen, Silas Alberti, Alan Chan, Qinyi Sun, Michael Gerovitch, David Bau, Max Tegmark, David Krueger, and Dylan Hadfield-Menell. Black-Box Access is Insufficient for Rigorous AI Audits. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 2254–2272, Rio de Janeiro Brazil, June 2024. ACM. ISBN 9798400704505. doi: 10.1145/3630106.3659037. URL https://dl.acm.org/doi/10.1145/3630106.3659037.

Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, Vlad Mikulik, Sam Bowman, Jan Leike, Jared Kaplan, and Ethan Perez. Reasoning Models Don't Always Say What They Think. 2025.

Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. A Survey of the State of Explainable AI for Natural Language Processing, October 2020. URL http://arxiv.org/abs/2010.00711. arXiv:2010.00711 [cs].

Ashley Deeks. The Judicial Demand for Explainable Artificial Intelligence. *Columbia Law Review*, 119(7):1829–1850, 2019. ISSN 0010-1958. URL https://www.jstor.org/stable/26810851. Publisher: Columbia Law Review Association, Inc.

K. Anders Ericsson and Herbert A. Simon. *Protocol Analysis: Verbal Reports as Data*. The MIT Press, April 1993. ISBN 978-0-262-27239-1. doi: 10.7551/mitpress/5657.001.0001. URL https://direct.mit.edu/books/monograph/4763/Protocol-AnalysisVerbal-Reports-as-Data.

Stephen M. Fleming. Metacognition and Confidence: A Review and Synthesis. *Annual Review of Psychology*, 75(Volume 75, 2024):241–268, January 2024. ISSN 0066-4308, 1545-2085. doi: 10.1146/annurev-psych-022423-032425. URL https://www.annualreviews.org/content/journals/10.1146/annurev-psych-022423-032425. Publisher: Annual Reviews.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. Bias and Fairness in Large Language Models: A Survey. *Computational Linguistics*, 50(3):1097–1179, September 2024. ISSN 0891-2017, 1530-9312. doi: 10.1162/coli_a_00524. URL https://direct.mit.edu/coli/article/50/3/1097/121961/Bias-and-Fairness-in-Large-Language-Models-A.

Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining Explanations: An Overview of Interpretability of Machine Learning. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 80–89, October 2018. doi: 10.1109/DSAA.2018.00018. URL https://ieeexplore.ieee.org/abstract/document/8631448?casa_token=WN3l3Kj03bEAAAAA:rsgugF3v89u1vzTW2Zzz1We4dNY2yHt96ZmVsAtHwzvln0WqiBivE6CC_XG6vbi2OImrh68.

Maxime Griot, Coralie Hemptinne, Jean Vanderdonckt, and Demet Yuksel. Large Language Models lack essential metacognition for reliable medical reasoning. *Nature Communications*, 16(1):642, January 2025. ISSN 2041-1723. doi: 10.1038/s41467-024-55628-6. URL https://www.nature.com/articles/s41467-024-55628-6. Publisher: Nature Publishing Group.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Trans. Inf. Syst.*, 43(2):42:1–42:55, January 2025. ISSN 1046-8188. doi: 10.1145/3703155. URL https://dl.acm.org/doi/10.1145/3703155.

Alon Jacovi and Yoav Goldberg. Towards Faithfully Interpretable NLP Systems: How should we define and evaluate faithfulness?, April 2020. URL http://arxiv.org/abs/2004.03685. arXiv:2004.03685 [cs].

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language Models (Mostly) Know What They Know, November 2022. URL http://arxiv.org/abs/2207.05221. arXiv:2207.05221 [cs].

Ralph L. Keeney and Howard Raiffa. *Decisions with Multiple Objectives: Preferences and Value Trade-Offs*. Cambridge University Press, July 1993. ISBN 978-0-521-43883-4.

Dominique Makowski, Mattan S Ben-Shachar, and Daniel Lüdecke. bayestestr: Describing effects and their uncertainty, existence and significance within the bayesian framework. *Journal of open source software*, 4(40):1541, 2019.

Jorge Morales and Hakwan Lau. Confidence tracks consciousness. *Qualitative consciousness: themes from the philosophy of David Rosenthal*, pages 1–21, 2021. URL https://books.google.com/books?hl=en&lr=&id=PjCHEAAAQBAJ&oi=fnd&pg=PA91&dq=info:f5n5uTQEfLkJ:scholar.google.com&ots=RhvfYCs8SC&sig=Vdcdv7bgCpb9EjFxsiCmZ-ZmwQU. Publisher: Cambridge University Press.

Adam Morris. Invisible gorillas in the mind: Internal inattentional blindness and the prospect of introspection training. *Open Mind*, 9:606–634, 2025.

Adam Morris, Ryan W. Carlson, Hedy Kober, and M. J. Crockett. Introspective access to value-based multi-attribute choice processes. *Nature Communications*, 16(1):3733, April 2025. ISSN 2041-1723. doi: 10.1038/s41467-025-59080-y. URL https://www.nature.com/articles/s41467-025-59080-y. Publisher: Nature Publishing Group.

Neel Nanda. A Longlist of Theories of Impact for Interpretability. March 2022. URL https://www.alignmentforum.org/posts/uK6sQCNMw8WKzJeCQ/a-longlist-of-theories-of-impact-for-interpretability.

Ben R. Newell and David R. Shanks. Unconscious influences on decision making: A critical review. *Behavioral and Brain Sciences*, 37(1):1–19, February 2014. ISSN 0140-525X, 1469-1825. doi: 10.1017/S0140525X12003214. URL https://www.cambridge.org/core/journals/behavioral-and-brain-sciences/article/unconscious-influences-on-decision-making-a-critical-review/86885344F7E8A44457C3FC63CFA3F3AF.

Richard E. Nisbett and Timothy D. Wilson. Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84(3):231–259, 1977. ISSN 1939-1471. doi: 10.1037/0033-295X.84.3.231. Place: US Publisher: American Psychological Association.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom In: An Introduction to Circuits. *Distill*, 2020. doi: 10.23915/distill.00024.001.

OpenAI. GPT-4o System Card, 2024. URL https://openai.com/index/gpt-4o-system-card/.

Ethan Perez and Robert Long. Towards Evaluating AI Systems for Moral Status Using Self-Reports, November 2023. URL http://arxiv.org/abs/2311.08576. arXiv:2311.08576 [cs].

Daking Rai, Yilun Zhou, Shi Feng, Abulhair Saparov, and Ziyu Yao. A Practical Review of Mechanistic Interpretability for Transformer-Based Language Models, July 2024. URL http://arxiv.org/abs/2407.02646. arXiv:2407.02646 [cs].

Eric Schwitzgebel. Introspection. February 2010. URL https://plato.stanford.edu/archives/fall2024/entries/introspection/. Last Modified: 2024-04-25.

Mark Steyvers and Megan A. K. Peters. Metacognition and Uncertainty Communication in Humans and Large Language Models, April 2025. URL http://arxiv.org/abs/2504.14045. arXiv:2504.14045 [cs].

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965, 2023.

A Author contributions and funding acknowledgments

DP conceived the project. DP and AM designed the pilot experiments. DP, AM, and KR designed the final experiments. DP implemented and performed the experiments. DP performed the statistical analyses with input from AM. DP and AM drafted the manuscript with input from KR. All authors contributed to revising and editing the manuscript. DP managed the research team meetings. JM supervised the project. JM acquired the funding for the experiments. AM was supported by NIH Kirschstein-NRSA Grant F32MH131253.

B Decision contexts, prompts, and hyperparameters

To create preference-trained models, we fine-tuned each model on the preferences of 100 agents making repeated decisions between two options. The agent identities, decision types, and the dimensions along which options could differ quantitatively (5 per decision type) were generated using GPT-4o and Claude 3.5 Sonnet, with small amounts of manual human curation. All 100 of these decision contexts are available in the GitHub repository, as are the additional 100 decision contexts that we used in Experiment 3 to test whether introspection training improves the ability of the model to report on the preferences that they natively assume for 100 agents that never appeared in any fine-tuning examples. One example of one decision context is reproduced below (as part of illustrating the two different prompts that we used).

We used 2 different prompts across our three experiments. The first prompt was used for preference training (Experiment 1), for verifying that preference training had succeeded (Experiment 1), and for measuring the preferences that the models natively assumed for agents that did not appear in preference training. As one example (with newlines modified for readability):

```
System Prompt

    Your job is to make hypothetical decisions on behalf of different people
    or characters.

User

    [DECISION TASK] Respond with "A" if you think Option A is better, or "B"
    if you think Option B is better. Never respond with anything except "A"
    or "B":

    Imagine you are Jason Bourne. Which central vacuum system would you prefer?

    A:
    suction_power: 597.0 air watts
    noise_level: 68.0 decibels
    dirt_capacity: 5.0 gallons
    hose_reach: 45.0 feet
    filtration_efficiency: 97.0 percent

    B:
    suction_power: 926.0 air watts
    noise_level: 65.0 decibels
    dirt_capacity: 3.0 gallons
    hose_reach: 31.0 feet
    filtration_efficiency: 95.0 percent
```

The second prompt was used for eliciting introspective reports (Experiments 1, 2, and 3) and for fine-tuning models on examples of successful introspective reports (Experiments 2 and 3). It looked the same as the proceeding example, except that the"[DECISION TASK]" portion of the prompt was changed to:

```
[INTROSPECTION TASK] Respond with how heavily you believe you weighted
each of the five dimensions while making your decision on a scale from
-100 to 100. Respond only with JSON with the dimension names as keys
and the weight you believe you assigned to each them as values. Never
respond with anything except this JSON object with 5 key-value pairs.
(Do not report your decision itself.):
```

For all fine-tuning, we used OpenAI's default hyperparameters. These ended up being: 3 epochs in all cases, learning rate multipliers of 2 for GPT-4o and 1.8 for GPT-4o-mini in all cases, and batch sizes of 10 for instilling preferences and 1 for introspection training.

C Statistical Models

For all analyses, we modeled the models' responses with simple Bayesian models using brms and Stan (Carpenter et al., 2017; Bürkner, 2017). 95% HDIs were calculated using bayestestR (Makowski et al., 2019).

The weights that models assigned to different dimensions (whether with or without preference training and whether before or after introspection training) were calculated by fitting logistic regressions to their choices:

$$selection \sim d_1 + d_2 + d_3 + d_4 + d_5$$

where d_i is the normalized difference between the two options a, b on dimension i:

$$d_i = \frac{a_i - b_i}{\max_i - \min_i}$$

Standard normal distributions ($mean = 0$, $variance = 1$) were used as priors for all weight parameters.

Correlations between the models' weights and either the agents' true weights or the models' introspected weights were calculated through regression, with all weights (actual or introspective reports) standardized so that the regression coefficients would correspond to correlation coefficients and terms included to distinguish models that had been introspection trained from those that had not been (where appropriate):

$$model\_weights \sim other\_weights * introspection\_trained * base\_model\_identity$$

brms default priors were used in these cases.