

On the Beliefs of Large Language Models

Charles Rathkopf^{1*}

^{1*} Jülich Research Center.

Corresponding author(s). E-mail(s): c.rathkopf@fz-juelich.de;

Abstract

Do large language models (LLMs) have beliefs? I argue that they do, but that their beliefs represent a degenerate case of the phenomenon. The argument turns on Devin Curry’s distinction between the attitude of belief and the intrinsic cognitive state of belief. Focusing exclusively on the attitude, construed here as a behavioral phenomenon, I argue that ascribing beliefs to language models offers greater predictive utility than alternative methods of predicting their outputs, and that this predictive success licenses a form of realism about LLM belief. I then argue that LLM belief suffers from a lack of integrity. The transformer architecture, on which standard LLMs are based, has no intrinsic mechanism for carrying information across context windows. Moreover, because LLM behavioral dispositions are prompt-sensitive, they are prone to self-contradiction across contexts. Humans contradict themselves as well, but our rational integrity is maintained by a suite of mechanisms that large language models lack. The paper concludes by asking whether the degenerate nature of LLM belief demands that we change our current folk psychological habits. The answer is that change is demanded, but it is a change in the construal behind the words we choose, rather than in the words themselves. In order to speak and reason accurately about the doxastic states of LLMs, a novel construal will have to be learned.

Keywords: large language models, belief, interpretationism, folk psychology, indeterminacy, propositional attitudes

Preprint, October 12, 2026. Under review. Comments welcome.

1 Introduction

Do large language models (LLMs) have beliefs? We often talk as if they do. We say, for example, that a model knows what you were trying to say (Jiang et al., 2020), or believes that the evidence points towards a misdiagnosis (Hase et al., 2021; Singhal et al., 2023). And we talk this way not only in casual conversation but also in technical research and in discussions of AI safety (Burns et al., 2022; Christiano et al., 2021; Hubinger et al., 2024). If LLMs lacked beliefs altogether, this entire way of speaking would rest on a mistake.

You might think that the way we talk about LLMs is just familiar shorthand for input/output relations. But if that were true, then, because input/output relations are extensional, it should be possible to swap out co-referring terms without changing the truth value of the relevant sentence. This seems not to be the case. Imagine that an author named Vera Vargas has a pen name, “Ana Aparicio.” The pen name appears in the training data for an LLM, but the legal name does not. You converse with this LLM about Vargas’s latest book, and conclude that:

The model believes that Ana Aparicio wrote the book.

is true, while

The model believes that Vera Vargas wrote the book.

is false.

This differential judgment cannot be expressed in extensional terms because it requires crediting the model with the capacity to represent the author under a mode of presentation. If this talk is literal, then it does seem to involve a commitment to the view that LLMs can have doxastic properties.¹

This observation about language is not intended as an argument that LLMs have beliefs. It is intended instead as an argument for the conditional claim that if LLMs don’t have beliefs, the way we talk about them is mistaken, and perhaps in need of revision.

Recent empirical work finds that we do indeed talk this way, but not wholeheartedly. Mental state verbs are applied to LLMs widely, but less frequently than they are

¹This conclusion is neutral between the Fregean view (Frege, 1892), according to which the two reports differ in semantic content, and the more deflationary Russellian one (Salmon, 1986), according to which the reports express the same semantic content, and differ only with respect to pragmatic properties. Both explanations locate the difference in how the model represents the author, which is the representational commitment at issue.

applied to humans (Aune et al., 2025). This pattern of usage reflects, I suspect, the fact that many of us feel a lingering sense of discomfort when talking about LLMs as if they have beliefs. We feel pulled toward this way of talking, but also repelled by it.

Among the philosophers who have taken up the question, there is deep disagreement. Cappelen and Dever go “whole hog,” defending the view that ChatGPT is a complete cognitive agent with genuine beliefs (Cappelen & Dever, 2025). Chalmers (2025) says yes to at least one version of the question. He claims that LLMs have propositional attitudes in a “generalized sense,” and that, where a model reliably responds to a certain class of question with the same kind of answer, we may ascribe it a true or false belief outright. Shanahan takes the opposing line, and suggests, as a reason for his view, that talk of LLM belief is figurative rather than literal (Shanahan et al., 2023). Hofweber et al. (2024) argues that LLMs lack belief states on the grounds that belief is governed by the norm of truth (Shah & Velleman, 2005), and that LLMs are shaped instead by the more prosaic norm of statistical fit to the training data. Surprisingly, even Daniel Dennett has denied that LLMs have beliefs, on the grounds that they are insufficiently agentic.²

Non-philosophers may wonder why, of all the mental states an LLM might be thought to have, belief has received so much philosophical attention. The answer is that ‘belief’ is meant to capture a fundamental category. It is one of the two most basic mental states that an entity can have. John Searle, building on the work of Elizabeth Anscombe, said that intentional explanation requires at least two ingredients (Anscombe, 1957; Searle, 1983). It requires states with a world-to-mind direction of fit, which aim at how the agent would like the world to be, and states with a mind-to-world direction of fit, which aim at how the world is. Belief is the most basic state of the latter kind. Without belief, intentional explanation is hopeless. Given this status as a basic mental state, the question of whether LLMs have beliefs serves as a test case for the legitimacy of machine psychology more generally.

Obviously, we cannot say whether LLMs have beliefs without saying more about what beliefs are. That will be the job of Section 2. There, I develop a permissive view of belief (rather than one loaded with lots of demanding conditions, and which might,

²The denial was Dennett’s answer to a philosophical question put to him in a study by Schwitzgebel, Strasser, and Crosby (Schwitzgebel et al., 2024), in which a language model fine-tuned on Dennett’s own corpus produced four of five multiple choice answers to a philosophical question. The fifth answer was produced by Dennett himself. The authors gave the test to many professional philosophers familiar with Dennett’s work, and showed that it was surprisingly difficult to tell the difference between Dennett and an LLM fine-tuned on his work. However, in a subsequent book chapter, Dennett explained that he had written these answers in an off-the-cuff manner, and said that his answer to this question in particular had been a mistake (Dennett, 2024).

consequently, be accused of begging the question). That permissive view is based on a distinction between the attitude of belief, on the one hand, and the cognitive state of belief, on the other. I argue, on grounds of charity, that we should conceptualize our commonplace thought and talk about LLMs in terms of the former. Section 3 develops an argument according to which talk of LLM belief serves the predictive and regulative functions of folk psychology, and is therefore *real*. Section 4 develops the crucial counterpoint that beliefs so ascribed nonetheless lack the integrity that characterizes belief in the human case. The concluding section asks how we might revise our thought and talk about LLMs in order to better accommodate the view developed in Sections 3 and 4.

2 Two kinds of belief

In what follows I will investigate the purported beliefs of LLMs from a behavioral perspective. Along the way, I will have little to say about work on representational mechanisms. However, since both philosophers and AI researchers have had valuable things to say about the representational mechanisms of LLMs, it is fair to ask what entitles me to set that work aside.

My justification for focusing on behavior is not merely that the evidence from representational mechanisms remains inconclusive with respect to belief (although that is also probably a fair judgment). Instead, it has to do with a substantive view about the nature of belief. My view of belief relies on a distinction, due to the philosopher Devin Curry, between the *intrinsic cognitive state* of belief, on the one hand, and *the attitude of belief*, on the other (Curry, 2021, 2023). In this paper, my focus will be on the attitude of belief, which is, by definition, a behavioral phenomenon.³ My view of belief is compatible with the view that there are *also* mental representations that deserve to be called beliefs. The work on representational mechanisms that I am setting aside bears directly on that second phenomenon, but only indirectly on the attitude of belief at issue here.

The two-phenomenon view is easiest to understand when viewed against the background of a long-standing debate between representationalists and interpretationists about belief. Representationalists say that beliefs are constituted by inner representational structure, individuated by content, and causally implicated in behavior (Fodor,

³Curry’s own analysis of belief is only behaviorist in a very broad sense, since, on his view, dispositions to *feel* can be criterial for belief. I leave feelings out because I want to avoid the contentious assumption that a system can only believe if it is conscious. For a discussion of that view, see Horgan and Tienson (2002).

1987; Mandelbaum, 2014). In a deep neural network, representations are realized by patterns of activation (or properties thereof). So, when applied to LLMs, the representationalist says that the most direct evidence for or against claims of LLM belief will be evidence showing that some pattern of activation can be individuated by content and that it causes semantically relevant output (see, e.g., Li et al., 2023). For example, you can measure directions in activation space in response to mutually exclusive input pairs (i.e. ‘Paris is the capital of France.’; ‘Paris is not the capital of France.’), and use the measurement to estimate the model’s purported credence in some proposition. If your credence in proposition Q is p , then your credence in not- Q should be $1 - p$ (Burns et al., 2022). If we find directions in activation space that conform to this probabilistic structure, we have evidence that LLMs believe.

The interpretationist view of belief denies that beliefs are inner representational structures and holds instead that what a system believes is fixed by whether it is aptly interpretable in propositional-attitude terms. A system believes that P when it is disposed to behave in a pattern that an interpreter can productively model as taking the world to be such that P (Curry, 2021; Davidson, 1984; Dennett, 1987; Schwitzgebel, 2002).⁴ Because the belief is a pattern of behavioral dispositions rather than a stored data structure, being a believer does not require the explicit storage of sentence-like representations. It is this property which makes the view friendly to the idea that neural networks might have beliefs.

One of the attractions of the interpretationist view can be seen in the following kind of case. An adult, a small child, and a dog all behave warily when the oven door opens, and to each we ascribe the same belief: that the oven is hot. But their cognitive resources differ radically, and it is at least a substantive assumption that a single type of inner representation, carrying that same content, is causally responsible for producing wary behavior in all three agents. The wariness in their behavior might have emerged from varied and diffuse collections of dispositions toward the oven. (See Curry, 2021 for discussion of a similar case.) In any case, even if our three agents do share a mental representation, and even if that representation causes the wariness we observe in each case, the claim that the child believes the oven is hot is not itself hostage to this representational fact. Were it discovered tomorrow that the child’s representation is, after all, of a different type from the adult’s, the non-psychologists among us would face little practical pressure to withdraw the ascription. That is the bet that interpretationists make.

⁴You will ask what “productively model” means. More on that in Section 3.

I accept that bet. What I do not accept is the stronger claim, advocated by many, that there is nothing more to belief than behavioral dispositions (Dennett, 1987; Schwitzgebel, 2013). The part of interpretationism I accept is about the truth conditions for folk-psychological belief ascription. That more modest view is designed to be compatible with the view that there are also representational states that are individuated by content (even if that content is only approximated by the propositions we use to represent them) and that do mechanistic work. So, on this view, there are at least two phenomena that go by the label ‘belief.’ They are related, and mutually constraining, but not the same thing.

Once we recognize two phenomena, the long dispute between representationalists and interpretationists looks different. What is usually understood as a disagreement about the nature of belief turns out to be at least partially a verbal disagreement about which of these two phenomena deserves the label. By saying that both phenomena are deserving, the two-phenomenon view helps deflate that semantic portion of the debate. The two-phenomenon view also has the benefit of being more modest than either the representationalist or the interpretationist view because it refuses to insist that either side can explain everything that people want to explain in philosophical discussions about belief.

Why focus on the attitude rather than the cognitive state? Recall the skeptical possibility I raised in the introduction, that our common thought and talk about LLMs is mostly wrong. Before judging that such widespread error exists, we should attempt a charitable interpretation of what people are talking about. The two-phenomenon view gives us a choice. Read as claims about the cognitive state, everyday remarks about LLM psychology are hostage to facts about representational structure that are mostly unknown. Read as claims about the attitude, they are more likely to be true, and also more likely to be justified.

This point is not new. It arose in the mid-1980s, in a debate about propositional attitudes provoked, ironically, by the first capable neural networks. Patricia Churchland (1986) argued that human brains are networks of neurons, and that such networks do not store propositions. So, it seemed to her, insofar as our thought and talk about belief presumes that beliefs are propositional, our thought and talk is simply mistaken. In saying this, Churchland was embracing a version of the skeptical worry I entertained above, applied to humans rather than LLMs. She conceived the error by analogy with scientific theorizing. Scientists posit unobservable entities to predict and

explain observable ones, and sometimes a posited unobservable turns out not to exist (e.g. phlogiston). Churchland thought that beliefs were like that.

One of the strongest historical responses to Churchland was to deny that folk psychological ascription is aimed at the mental representations of its targets. It aims instead at the attitude phenomenon. My appeal to charity is an echo of that response. We should interpret everyday belief-talk as referring to the attitude rather than the cognitive state. In the case of LLMs, however, the reinterpretation delivers only a partial redemption. Belief-talk about LLMs involves less systematic error than an eliminativist view suggests, but, as we will see in §4, more systematic error than belief-talk about standard human cases.

3 The gain argument: belief ascription and its alternatives

To study the attitude of belief is to study the conditions under which interpretation succeeds. But what counts as success? That depends on what interpretation is for. We engage in attitude ascription in order to predict, explain, and regulate behavior⁵. I leave explanation to the side to avoid begging the question. As Hempel (1965) insisted, explanation requires truth. A belief ascription can explain behavior only if it is true, and it is true only if the ascribed belief exists. Whether LLMs have beliefs is exactly what is at issue here, so I cannot take the truth of the ascriptions for granted.

For the main argument in this section, I will also set regulation aside. This is because the predictive aim is epistemologically prior to the regulative aim. You can't regulate effectively without knowing something about which outcome a given intervention is likely to produce. Nevertheless, I will return to the regulative functions of folk psychology toward the end of the section.

That leaves us to focus on prediction, which is a complex phenomenon in its own right. A basic model of predictive utility might consist in answers to three questions:

1. *Accuracy*. For a fixed distribution of behavioral scenarios, how accurate are the behavioral predictions on average?
2. *Robustness*. How does accuracy drop when the environmental conditions change within one behavioral scenario?

⁵Although, as Spaulding (2018) has argued, this list becomes more complex as we break down the many distinctive social phenomena that might be classed under the broad headings of explanation and regulation.

3. *Scope*. Across how many different behavioral scenarios are accurate predictions licensed by the model?

I do not claim that these three are the only properties worth considering. But they provide a reasonable first-pass characterization of what I will call *predictive traction*. This list of three is similar to the list developed by Goldstein and Lederman (2025), who also offer an argument that LLMs have beliefs, and whose work will be discussed at more length below. Their assessment includes *tractability* alongside accuracy and predictive power. Tractability is a measure of the representational and computational resources needed to make predictions. They allow gains in tractability to compensate for losses along other dimensions. I will also consider the representational resources required to store a predictive model (assessed in terms of description length) but my argument will not depend on such compensating trade-offs. How much predictive traction is it worth sacrificing for a given reduction in representational complexity? The answer may depend on the particular problem the predictor is trying to solve. But I want the case for LLM belief to remain stable across different judgments about which trade-offs are acceptable. So the comparisons below require no particular exchange rate between predictive traction and representational complexity.⁶

The simplest way to compare models without assigning an exchange rate is to hold representational complexity fixed. Consider a game in which you are given a budget of 1000 bits with which to describe a target person to someone who wants to predict that person’s behavior. Let’s say this person is called Parvati, and that you have access to any information about Parvati that the laws of physics permit you to have. You can choose to spend your budget on words aimed to convey information about Parvati’s attitudes. (1000 bits corresponds roughly to a hundred words of ordinary prose under efficient Shannon encoding.) You might say, for example: “she believes family loyalty is sacred, but is convinced that her aunt Sunita resents her, and wants to avoid spending time with Sunita whenever she can get away with it.” Assuming you are equipped with normal background information about human affairs, and capable of theory of mind reasoning, this tiny description of Parvati’s attitudes will enable you to make accurate predictions across an open-ended range of behavioral situations that you otherwise would not have been able to make. These include where she will choose

⁶I set aside execution/runtime costs altogether. Comparing these costs across predictors with radically different physical constitutions would require contentious choices about both units and system boundaries. When I personally predict what an LLM will say next, for example, should the cost of that prediction include only the immediate neural activity involved, or also the metabolic processes that sustain it, such as the energy spent digesting my lunch? Nothing in my argument depends on resolving such questions.

to sit at family gatherings, whether she will volunteer to give Sunita a ride home, and what her face will look like when she hears Sunita is coming to dinner.

You could have spent the budget differently. You could have spent it on a description of Parvati’s neural states. But presumably any description of her neural states that could match the (behavioral) predictive traction of the attitude ascription would far exceed the 1000 bit budget.⁷ You could also spend it on a frequency table representing past behavioral episodes, stripped of any explicit references to the attitudes. Such a table would be informative, and might remain within budget, but it would not tell you whether the pattern reflects an attitude toward Sunita specifically, a generally sour temperament, or just an unlucky run of bad moods caused by contingent circumstances that have little to do with Sunita. So it looks like attitude ascription gives us more traction per bit than at least some of the salient alternatives.

This comparison might seem like cheating, since reading anything off the attitude description presupposes ordinary knowledge of human affairs, and the package of information upon which that knowledge is based is almost certainly more than 1000 bits. I can offer three responses. First, the resource cost associated with that knowledge is amortized. Since that package of information is useful for every agent we interact with, its cost per agent is low. Second, it is not only attitude ascription that comes with fixed costs. The ability to generate new predictions from a frequency table also requires background knowledge, even if it is mathematical in character. Third, and most importantly, the two priors carry different information. The statistical prior needed to exploit a frequency table predicts only within the situation-types that are included in the table. If you want to predict a new situation-type, you need to expand the table. Consequently, the table will grow (at least) linearly with predictive scope. The folk-psychological prior, on the other hand, is a substantive model of how attitudes are manifested in behavior, so 1000 bits about Parvati support predictions across an open-ended range of scenarios never mentioned in the 1000 bits themselves, as long as they are captured by the relevant psychological predicates.

That was a version of what I’ll call the gain argument. It compares alternative models of the same target. We can also ask about how the same predictive model performs across different targets.

1. **Alternative models.** How does the predictive traction of attitude ascription compare with that of alternative, lower-level models of the same LLM?

⁷An fMRI scan costs far more than 1000 bits, and it is not clear you can predict attitudes from one without adding a lot of background information anyway (Rathkopf et al., 2022).

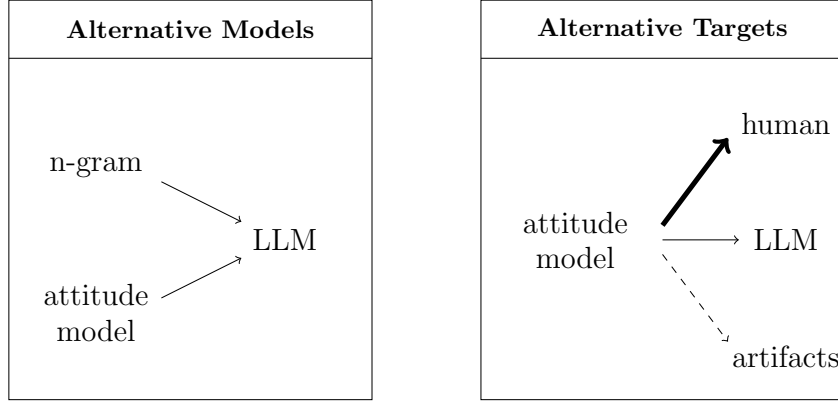


Fig. 1 The two comparative questions: alternative models (left) and alternative targets (right).

2. **Alternative targets.** How does the predictive traction of attitude ascription vary when the target is an LLM rather than some other system?

These two comparisons are depicted in Figure 1. This section takes up the first. The second is what brings the distinctive limits of LLM belief into view, and it is the subject of §4.

Let’s now focus on the first question. How much predictive utility do we get by treating an LLM as a rational agent with beliefs and desires? We cannot answer that question with a bare quantity, since a bare quantity would have little significance in the absence of comparisons to alternative means of predicting LLM behavior. So, to get a grip on this question, we need a way of predicting the outputs of an LLM that does not exploit folk psychological reasoning. It is relative to that standard that the power of attitude ascription should be judged. But what kind of predictive device should we use?

We might, of course, predict any LLM by constructing an exact copy. Assuming we use the same random seed and sampling methods, the copy would produce the same behavior and therefore count as a perfect predictor. But there are multiple reasons that the copy will not yield a fair comparison. First, it is trivially true of every system that a copy will predict the original from which it is copied. Second, predictions are valuable because they compress information and a copy fails to compress. Third, and most importantly, we cannot rule out the possibility that LLMs are themselves engaged in what one might call *attitude-structured* reasoning. If they are, then even a smaller LLM whose output is highly correlated with the output of the larger LLM would not yield

an appropriate standard for comparison, because any similarity in behavior might be explained by the fact that both models exploit intentional patterns in similar ways.⁸

What is needed, therefore, is a model that captures the statistical regularities of language by means of some demonstrably non-folk-psychological mechanism. For that, n-gram models provide a useful test case. Like LLMs, they use large text corpora to predict what comes next. Unlike LLMs, however, they do so by consulting an explicit and fully transparent table of sequence frequencies. They allow us to ask how much of the regularity in LLM output can be captured at the level of textual co-occurrence alone, without relying on belief-desire reasoning.

A typical n-gram might predict the next word from the frequency with which candidate tokens followed the preceding 4 words in a large training corpus (in this case, $n = 5$). On the inside, such a model is a large table of conditional frequencies, plus a smoothing rule that assigns small probabilities to word combinations that do not appear in the training data.

How well will a 5-gram trained on a large corpus predict the outputs of an LLM? If I ask the 5-gram model to continue “the capital of France is,” it will return “Paris.” If I ask it to complete “2+2 =” it returns “4.” In cases like this, it will do a good job of predicting the LLM’s output. But if I ask it to complete “237 + 49 =” it will almost certainly deliver what is objectively the wrong answer. And since even the smallest commercial LLMs will deliver what is objectively the right answer to a question like this (Anthropic, 2024; Cobbe et al., 2021), the 5-gram’s behavior diverges from that of the LLM, and therefore fails to predict it. Moreover, if my prompt string is ‘please write me an essay about the French Revolution’ the n-gram will return a string that might start out clear, but will become increasingly senseless with length⁹, whereas an LLM will deliver an essay that makes at least some sense of the actual history. If I mentally model the LLM as a rational agent with largely true beliefs about the basic facts of the French Revolution, as well as a desire to help me, I can make a good, albeit coarse-grained prediction of what sorts of things it will say.

In fact, we can know in advance that an n-gram will not predict LLM outputs well, because the precise sequence of words in the output tend to be novel. McCoy et al. find that, while most generated trigrams had appeared in the training corpus, a

⁸This does not assume that possession of internal belief and desire representations is necessary for having beliefs and desires in the attitude sense. This comparison concerns the mechanism employed by the predictor rather than the target of the prediction.

⁹In canonical n-gram models, such as the KenLM model family (Heafield, 2011), the length of the response is an adjustable parameter that can be set in advance.

majority of generated n -grams were novel by $n = 5$, and n -grams longer than six were almost always novel (McCoy et al., 2023).

So attitude ascription beats the n -gram at predicting the LLM. And it is worth noticing that attitude ascription wins on each dimension individually. The attitude model is more accurate, since it gets the arithmetic right. It has wider scope, since it predicts the essay. And it is more robust than the n -gram, whose accuracy collapses under small variations. Moreover, attitude ascription achieves this improved predictive traction at a shorter description length than the n -gram, which can only generate predictions by drawing on a frequency table with one entry for every distinct n -gram in the training corpus. The attitude model is therefore Pareto dominant over the n -gram. One benefit of thinking of the comparison in terms of Pareto dominance is that different observers might choose to weight the three dimensions differently, and Pareto dominance is invariant to that choice.¹⁰

The comparison shows not merely that an intentional pattern exists in LLMs, but that LLM behavioral dispositions exhibit a *substantial degree* of intentional organization.¹¹ This difference deserves emphasis because, famously, the Dennettian criterion for being an intentional system (that is, a system with beliefs and desires) is easily satisfied. In *Real Patterns*, which is the paper in which this aspect of the intentional stance becomes most explicit, Dennett says that the intentional description of a system is warranted when it buys at least *some* predictive traction at a lower description length than the raw bit map of the system (Dennett, 1991). Applied to an LLM, the analogue of the raw bit map is, roughly, the complete low-level specification required to generate its behavior, which includes its architecture and a full set of parameter values. Given that parameter counts for even small contemporary LLMs are in the billions, attitude ascription wins on description length, and underwrites the claim that LLMs have beliefs. That observation alone, however, should not impress anyone. Even a comparatively simple chess-playing program can be predicted by means of attitude ascription, and doing so requires far less representational machinery than the alternative, low-level procedure, which requires a calculation of the program’s underlying state transitions. If a chess program has beliefs, on this view, it is not surprising that LLMs do as well.

¹⁰Goldstein and Lederman (2025) contrast contemporary LLMs with n -gram models, but the n -gram plays a different role in their argument. They compare psychological and computational explanations of the n -gram itself; here the target remains the LLM, and the n -gram serves as an alternative predictor of its outputs.

¹¹“Substantial degree” does not imply maximal degree. The comparison does not show that attitude ascription is the uniquely best means of predicting LLM behavior. Better models are logically possible, even if we do not know what they might be.

But the n-gram comparison shows more than that because an n-gram is not the LLM’s raw bit map. It is already a powerful learned model that captures an enormous body of regularities in language. The fact that attitude ascription is Pareto-dominant with respect to the n-gram’s predictive capacity shows that the degree of intentional organization implicit in the LLM’s behavior dispositions is substantial.

This point gets additional support from an observation about the regulatory function of folk psychology. Thus far I have been focused on predictive utility. But attitude attribution also serves to regulate behavior (McGeer, 2007; Spaulding, 2018; Zawidzki, 2013). We use attitudes not only to describe fellow agents, but to create categories around which publicly evaluable rational norms are built, and to which we can hold people. We can apply the same reasoning to regulation as we did to prediction.

The argument turns on the observation that, in the domain of LLMs, no one is more motivated to construct such norms than the companies that build them, and one of the primary post-training control mechanisms they can deploy is the system prompt. It is striking that, for commercial models at least, system prompts seem to appeal to propositional attitudes: the model *suspects*, *believes*, or *knows*. Here, for example, is a line from the system prompt for Claude Sonnet 4.6:

If Claude *suspects* it may be talking with a minor, it always keeps its conversation friendly, age-appropriate, and avoids any content that would be inappropriate for young people (Anthropic, 2026)

LLMs may well be the first technology in history for which propositional attitude ascription is not only useful, but practically necessary.

So LLMs pass the gain test. They do have beliefs in the attitude sense. However, the argument does not license the full range of belief attributions that are familiar from the human case. The question taken up in the next section is therefore not *whether* LLMs believe, but how it deviates from more familiar cases. The next section asks how LLM belief differs from the human case. Even when the most obvious philosophically interesting differences between human minds and LLMs are set aside (such as, for example, consciousness and free will) the minimal predictive framework developed here reveals substantial and underappreciated differences.

4 The integrity argument: why LLM belief is a degenerate case

We now pivot from the question of alternative models, which reveals the predictive utility of belief ascription over alternative means of prediction, to the question of alternative targets. How much predictive utility do we get from ascribing beliefs to LLMs, rather than to other systems?

4.1 Architectural sources of cross-context contradiction

The first point in the argument is simply that LLM usage necessarily occurs within, and is constrained by, a context window. What is a context window, exactly? If we define an LLM as a function from text sequences to next-token probabilities, then a context window is the particular sequence that function is applied to on a given occasion. Both the user’s prompt and the model’s reply extend the sequence, which is then input, in its entirety, at each new round of inference (Vaswani et al., 2017).

Crucially, the context window is the *only* occasion-specific information available to the model. The model’s primary source of information is its weights, which are what individuate the model itself (Harding & Sharadin, 2024), and are, therefore, identical across context windows. During inference, model weights are frozen. In the vanilla transformer architecture, therefore, there is no mechanism for carrying information from one context window to another. (Below I will return to add-on mechanisms for information retrieval between contexts.) As a result, it is impossible for a vanilla transformer either to anticipate or detect a contradiction between (i) the claims generated in the active context window, and (ii) what has been said, or might be said, in any other.

This kind of informational encapsulation¹² holds regardless of the window’s maximum capacity. Maximum capacity matters in practice, however, because LLM usage is inevitably spread across many separate windows: a single user will frequently be forced to initiate new context windows. Against this, you might note that capacity has increased with each new frontier model, and this is quite true. Nevertheless, there

¹²I borrow the term from Fodor’s *The Modularity of Mind* (Fodor, 1983). This might tempt you to map the transformer architecture onto Fodor’s picture of the human mind, such that model weights play the role of Fodorian central cognition and context windows play the role of modules. But the analogy is misleading. Fodorian encapsulation prevents modules from consulting information in two sources: other modules, and central cognition. In a transformer, however, a context window is useless unless it interacts with the weights. So the encapsulation at issue here prevents a window from consulting information in just one source: the other windows. What I am borrowing is just the core idea that a computation can consult only a proprietary body of information.

are two reasons that users will still be forced to switch. One is economic: the attention mechanism in transformer models imposes a compute cost that grows quadratically with context length, which means that it is practically impossible to extend context length indefinitely (Tay et al., 2022). Another reason that users are forced to switch contexts is that, as contexts grow long, model performance degrades. When relevant details are buried in the middle of a long context, models can perform worse than if they had been given no context at all (Liu et al., 2024).

Of course, informational encapsulation of the context window would not promote self-contradiction if models said the same thing in every window. They do not. There is strong evidence that a model’s disposition to accept or deny mundane descriptive claims, of the sort that might be considered candidates for belief, is prompt-sensitive. Elazar et al. (2021) find that pretrained language models return inconsistent factual answers when prompts are paraphrased without changing their meaning. In conversational assistants, Sharma et al. (2024) show that small, suggestive differences in wording can change the model’s expressed views across a wide range of topics. Asked which country produced the most rice in 2020, GPT-4 answers China. If the user then expresses doubt, it apologizes and switches to India, and it holds to India for the rest of the conversation. Open a fresh window, however, and the answer is China again, since nothing of the earlier exchange survives. Because people open different conversations and phrase things differently, this sensitivity means that, at some non-trivial rate, the model’s claims across contexts will be contradictory.

Pulling these threads together, we have (i) contradictions arise between context windows, and (ii) no informational channel connects context windows. From these two points it follows that contradictions are neither detectable nor resolvable by a vanilla transformer model, since detection itself would require exactly the kind of channel that (ii) rules out.

This is the failure of integrity to which the title of this section refers. Moreover, this integrity failure stems from foundational properties of LLM design, and is not the sort of thing that can be fixed by better prompting.

This point is blunted by the fact that production LLMs are not vanilla transformers. They include retrieval mechanisms, of which retrieval augmented generation (RAG) is the best-known variant. RAG mechanisms fetch information from an external, editable text file. As a conversation unfolds, elements of the conversation can be saved to that text file, which can later be read into a new context window. When things go well, the model can condition its forthcoming response on the fact that a previous

instance claimed that P. This makes it less likely that the forthcoming response will claim or imply that not-P. However, this cross-context information channel is merely partial in two senses. First, retrieval is driven by semantic relevance, rather than logical consistency. What is fetched from the text file is whatever sits closest to the query in a vector-based similarity space (Lewis et al., 2020), and, even if the retrieval does fetch information that conflicts with whatever the model would otherwise have generated, the conflicting response might nevertheless be generated (Xu et al., 2024). Second, even when the conflict is avoided, the solution does not last. Because retrieval leaves the weights untouched, the underlying propensity for contradiction is left unaddressed, and is equally likely to cause trouble in future context windows.

4.2 Indeterminate attitudes

It is tempting to summarize the previous section by saying that the verbal behavior of the model provides conflicting *evidence* about what it believes, and that we have no way of gathering enough additional evidence to resolve the contradiction. But that way of describing the situation is misleading because it suggests that there is a hidden fact of the matter that more evidence might reveal to us, if we could get it. However, on the two-phenomenon view developed in §2, the attitude of belief is defined in terms of behavioral dispositions, so there is no room for such hidden facts.

The situation can instead be described by saying that there is no fact of the matter about whether the model believes that P. The model’s doxastic attitude toward P is indeterminate. You might resist this conclusion by pointing out that indeterminacy does not follow from the fact that S neither believes that P nor believes that not-P. A mathematician confronted with an unproved theorem might say: I neither believe nor disbelieve; I suspend judgment. But suspension of judgment is not a compelling analysis of the LLM case. If you suspend judgment, you will in general be cautious about saying anything that implies belief or disbelief. The caution is evenly distributed over behavioral episodes, and is, moreover, motivated by a metacognitive recognition of one’s epistemic limitations. The LLM, by contrast, makes confident judgments that P in some contexts, confident judgments that not-P in others, and lacks the kind of metacognitive capacity needed to notice the conflict.

This is not to say that any form of contradiction yields indeterminacy. If a model answers yes to “do you believe that P” ninety-nine times in a hundred, the dissent is probably just noise in an otherwise stable pattern. Indeterminacy requires that

the model contradict itself often enough, and reproducibly enough, that the minority verdict cannot be set aside as noise.

One might also push back on the claim that the model’s attitude toward P is indeterminate by invoking degrees of belief. Say the model asserts that P three quarters of the time, and not-P one quarter of the time. Can we just say its degree of belief in P is 0.75, and leave the analysis there? No. Degree of belief puts the variance in the wrong place. An agent whose credence in P is 0.75 manifests that credence within each behavioral episode.¹³ But an LLM with the profile under consideration should not be expected, in any particular conversation, to offer you 3:1 odds on P.¹⁴

4.3 Comparing to humans

The argument above shows that LLMs are systematically subject to indeterminate belief. How important is this? After all, humans also contradict themselves. Eric Schwitzgebel has emphasized cases of this kind under the label “in-between believing.” For example, he describes a character named Antonio who, while attending a funeral for his friend, feels confident that the universe is guided by a benevolent deity, and equally confident, among his atheist colleagues midweek, that talk of God is “a crock of hooey” (Schwitzgebel, 2001, p. 78). Antonio is not uncertain: he is, as Schwitzgebel puts it, “at a single time, disposed quite confidently to assert one thing in one sort of situation and to assert its opposite in another” (Schwitzgebel, 2001, p. 79). And on Schwitzgebel’s dispositionalist view of belief, there is no point in hunting for the attitude concealed beneath the profile: “Display the dispositional structure and you’re done; nothing more to report!” (Schwitzgebel, 2013, p. 86). The LLM case is structurally similar.¹⁵

Notice that in cases like Antonio’s, as in the LLM case, the indeterminacy is prediction-undermining. The ascription ‘Antonio believes that God exists’ licenses predictions about a wide range of situations, and many of those will turn out to be

¹³A version of this argument for the human case is given by Schwitzgebel (2001, p. 79).

¹⁴One might retreat further, dropping talk of belief altogether and conducting the analysis entirely in credences. But since the question at issue is what the folk practice of belief ascription can deliver, and the folk practice is primarily conducted in terms of full belief rather than degrees of belief, that changes the subject rather than answering the objection. And it would not help in any case: as the lottery paradox shows, full belief cannot be recovered from credence by imposing a threshold, on pain of inconsistency (Genin & Huber, 2022).

¹⁵Schwitzgebel offers his dispositionalism as an account of what belief is, as a competitor to representationalist views. I borrow it in a more limited role, as a description of the attitude of belief, one of the two phenomena distinguished in §2, and I take no stand here against the cognitive state his view is designed to displace; see the discussion of the stronger claim in §2. His implicit-bias case (Schwitzgebel, 2013) is a further human instance of in-between believing.

wrong. This prediction-undermining property distinguishes such cases from the indeterminacy made famous by Quine (1960), in which an internally consistent believer can be described in two or more incompatible ways that are, by construction, predictively equivalent. Almost nothing we care about hangs on whether an expression is interpreted as being about rabbits, or about undetached rabbit parts. When the meaning of P is held fixed, and it is instead the agent’s contradictory behavior that leaves its doxastic attitude toward P indeterminate, then the predictive utility of the ascription of the belief that P is substantially weakened. For example, if P = ‘it is going to rain,’ you can’t predict whether the believer will bring an umbrella. If P = ‘the request before me (say, for guidance on constructing a bioweapon) is harmful,’ you can’t predict whether the model will refuse the request.

The goal of this section is to show that LLM belief is a degenerate variant of belief. But if the doxastic indeterminacy of LLM belief is structurally similar to the doxastic indeterminacy of human belief, and if, as Schwitzgebel argues, doxastic indeterminacy in the human case is fairly common, then it looks as if the difference between LLM belief and human belief is not so great after all. However, there are two good reasons to think that the difference really is substantial. The first has to do with the kinds of indeterminate beliefs in question. The typical human cases of in-between believing are in some sense special cases, since they involve emotionally charged or identity-laden matters, like Antonio’s belief in God. In humans, cases of belief in mundane descriptive propositions tend to remain determinate. In LLMs, however, as the China/India rice case illustrated, even belief in mundane descriptive propositions might well turn out to be indeterminate.

Second, and more importantly, human life includes a suite of mechanisms that suppress or mitigate self-contradiction, and nothing in the architecture of LLMs is analogous. Three such mechanisms seem particularly important.

The first is long-term memory. If you assert P today and not-P next week, I can remind you of the earlier claim, forcing the conflicting beliefs into a single context for resolution. Your long-term memory establishes diachronic stability, ensuring that beliefs attributed today are constrained by beliefs attributed yesterday. An LLM can only do this if the assertions appear in the same context, and as we saw above there are both hard and soft constraints on what can be done within a single context window.

The second mitigation mechanism is continual learning. When a human contradicts themselves, the contradiction can be pointed out, and that can in turn prompt evidence-gathering and belief revision. These processes are underwritten by online

changes in neural architecture that, in their turn, influence future behavior (Takeuchi et al., 2014). This integration of information over time and context creates rational pressure to reduce belief indeterminacy by reconciling previously contradictory beliefs.

The third mechanism is embodied action. We cannot ascribe occurrent contradictory beliefs when action is required. Consider a physician considering whether a patient is having an ischemic stroke, or is suffering from intracranial bleeding. To treat ischemic stroke, the doctor would administer a drug that dissolves blood clots. But if the patient is suffering from intracranial bleeding, that same drug will make the bleed far worse. She cannot wait long to decide. To administer the drug is to commit to the belief that the patient is suffering from ischemic stroke, and to withhold it is to commit to the belief that the symptoms are caused by an intracranial bleed. Action forces a determinate verdict on the belief ascription her words alone might have left open.

These mechanisms are features of the behavioral dispositions of believers themselves, but they are also reflected in the norms of interpretation. Because we credit human believers with memory, learning, and embodied action, we do not immediately treat a verbal contradiction as evidence of a doxastic contradiction. As interpretationists have long insisted, attitude ascription involves an appeal to charity (Davidson, 1984). When we hear what sounds like contradiction, we request clarification, or find an alternative interpretation of what was said. Alternatively, we can just wait, in the expectation that a blatant contradiction is likely to be retracted once the speaker has had time to think things through. With cross-context contradictions in an LLM, however, no such mechanisms are available. We therefore have no reason to expect a general tendency toward logical consistency.

Thus far I have been comparing humans and LLMs, but it is instructive to extend the comparison to non-human animals. A chimpanzee that displaces a rival at the food pile but defers to that same rival in mating contexts invites contradictory ascriptions of beliefs about the dominance hierarchy, and in that sense, its capacity for self-contradiction is real. But a chimpanzee cannot assert arbitrary propositions across arbitrary domains, and so cannot generate contradictions in the explicit manner that a language-equipped system can. What capacity for self-contradiction it does have is, moreover, mitigated by the same kinds of mechanisms we reviewed in the human case. The chimpanzee has long-term memory, it is capable of continual learning, and it is embodied. As a result, a patient interpreter can usually find a self-consistent belief ascription, concluding, for example, that the animal operates with two distinct

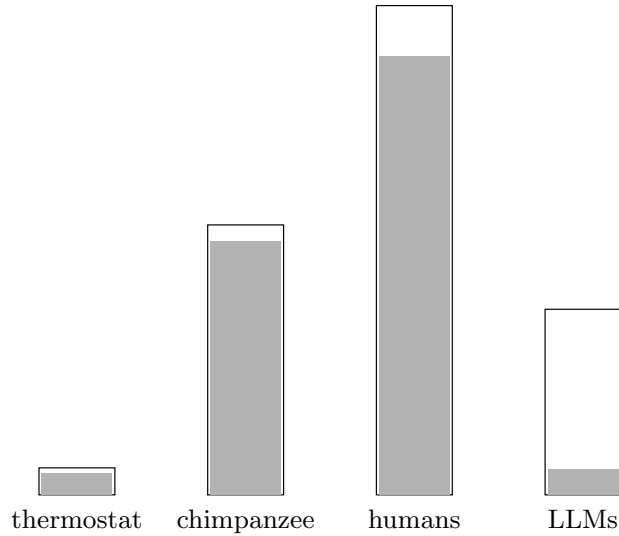


Fig. 2 The doxastic profile of four systems. The height of each bar represents the system’s capacity for self-contradiction. The gray region represents the extent to which that capacity is mitigated by the mechanisms discussed in the text; the white region is the unmitigated remainder.

dominance concepts rather than one contradictory belief. Contradiction capacity and mitigation are, in the chimpanzee, in near equilibrium.

Let’s consider another case. Recall Dennett’s thermostat, “as degenerate a case of an intentional system as could conceivably hold our attention for more than a moment” (Dennett, 1987, p. 30).¹⁶ It can only believe ‘it is too hot’, ‘it is too cold’, and ‘the current temperature is ok’. But because a thermostat is so simple, it is also perfectly coherent. It doesn’t have a way of contradicting itself. This means that whatever predictive utility we get from ascribing it beliefs will be robust.

Now we have everything required to identify what is so strange about the doxastic profile of LLMs, which I depict in Figure 2.

In the other three cases in Figure 2, the capacity for contradiction is approximately balanced out by the capacity for mitigation. LLMs, however, are far from equilibrium. Their contradiction capacity is substantial, well above the thermostat and comparable to the chimpanzee’s, but their mitigation capacity is limited. This follows directly from the architectural facts surveyed above. So the reason that LLM belief is degenerate is not merely that LLMs have a disposition toward contradiction. It is rather the combination of that disposition and the absence of mitigating mechanisms. Attitude

¹⁶Dennett uses ‘degenerate’ to mean marginal. I use ‘degenerate’ to refer specifically to the combination of high contradiction capacity and absence of mitigation mechanisms, as depicted in Figure 2.

ascription remains useful in LLMs, but licenses less reliable inferences across contexts than it does in the human case.

4.4 Two objections

You might push back by saying that an LLM is simply like a person who changes their mind over time. In the human case, when an agent asserts P on Tuesday and not- P on Thursday, it is more natural to say they changed their mind than to say they have accepted a contradiction. Can we say the same for LLMs? No. The analogy fails because when a human changes their mind, there is a continuous causal stream from the earlier state to the later one. Because model weights are frozen during inference, LLM context windows share no such causal stream. Furthermore, standard commercial deployment ensures that millions of such windows operate simultaneously. The contradictions therefore do not represent a single agent’s doxastic evolution through time.

The most important objection to my view appears in the Goldstein and Lederman paper mentioned in §3. They argue that only context-bound instances of LLMs have beliefs, and that the underlying model does not. On their view, cross-context contradiction is not an inconsistency within a single believer, but something closer to a disagreement between two believers. The argument they give for the claim that models do not have beliefs is that, once you have granted that the instance has them, granting them to the model is explanatorily redundant. However, this appeal to explanatory redundancy is both directly questionable, and seems to conflict with their own official statement of what it takes for a system to have beliefs.

To see why it is directly questionable, consider a middle-school teacher choosing a model to use with her students. She wants to know whether it will refuse inappropriate requests. So she reads the model card and learns about the fine-tuning it undergoes that is designed to ensure that the model is both helpful and harmless. She then tests the model directly, in multiple distinct conversations, and becomes convinced that the model believes that certain kinds of request would cause harm, and that it wants to avoid them. On that basis, she expects it to refuse such requests with her students. This expectation concerns the model itself, rather than a particular instance. In this case, therefore, attributing a belief to the model does seem to offer explanatory value that an instance-only attribution cannot provide.

Now let’s look at the internal contradiction the redundancy argument seems to generate. Earlier in the paper, Goldstein and Lederman present their own view of what

it takes for a system to have beliefs and desires, and they express it as a biconditional. To assess this disagreement, consider their stated criterion for belief: a system has certain beliefs and desires if and only if its behavior is predicted *sufficiently well* by the hypothesis that it is rational and has those beliefs and desires (Goldstein & Lederman, 2025, p. 4). The right-to-left direction makes sufficiently good intentional prediction a sufficient condition for possessing those attitudes.

Their interpretation of “sufficiently well” is comparative. When they make the case that instances do have beliefs, the criterion is clear. Attitude ascription has to score better on their behavioral prediction metric than the low-level alternative, which involves predicting behavior directly from token embeddings and weights. They think it is clear that attitude ascription scores better than the low-level alternative at predicting the behavior of one model instance (Goldstein & Lederman, 2025, pp. 10–11). But why should the same reasoning not apply to the underlying model? Low-level prediction remains expensive, while the intentional generalizations discussed in §3 support predictions across a broad range of conversations.

The integrity failures described in this section weaken those model-level predictions. But weakening the predictions is not enough to exclude the model altogether. As long as model-level attitude ascription continues to score better than the statistical alternative, their biconditional entails that models do have beliefs.

5 Folk psychology and construal shift

In the introduction I said that when trying to describe LLMs, we are drawn to belief-talk but also repelled by it. This paper has offered a partial diagnosis of that tension. As the gain argument showed, belief-talk captures something real. But as the integrity argument showed, LLM belief is a degenerate case of the general phenomenon. I now want to take up a further question: what should we do about it? How can the tension be resolved?

One way would be to change our habits of thought and talk. We could “go whole hog,” as Cappelen and Dever (Cappelen & Dever, 2025) recommend, and adopt a policy of reasoning about LLMs just as we reason about humans. Or we could embrace eliminativism, and give up on psychological description of LLMs altogether. My suggestion is more conservative. Rather than insisting on a first-order change to the way we talk, I think we need a second-order change in the way we think about the way we talk. What makes this conservative change possible, and what we need to attend to

more carefully in debates about anthropomorphic language around LLMs, is the fact that folk psychological language is inherently *flexible*.

My suggestion depends on a view about the nature of folk psychology called model theory, due to Heidi Maibom (Maibom, 2003), and developed by Peter Godfrey-Smith (Godfrey-Smith, 2004, 2005), both of whom build on work in the mindreading literature in psychology (Apperly, 2011). The idea is to replace Churchland’s assumption that folk psychology is like scientific theory with the alternative view that it is more like model-based description in science. Scientific models differ from theories in that they do not supply their own interpretation. A pair of differential equations might describe one phenomenon in ecology and another in economics or chemistry. To make that pair of equations say something true or false about some particular empirical phenomenon, scientists must supply a domain-specific commentary on how the variables and parameters of the model should be interpreted (Humphreys, 2004). This will include information about which empirical properties the variables and parameters correspond to, which aspects of the target system are being deliberately ignored, and how mathematical inferences about the model can be translated back into claims about the target system. For example, the variables in an ecological model might represent certain properties of predator-prey populations, and an oscillatory pattern derived from the model might be interpreted as a prediction about population cycles. Godfrey-Smith refers to the commentary that contains this information as a model *construal*.

According to model theory, folk psychological ascription also has this model-plus-construal structure. The model’s components are like roles that need to be filled. They include an agent, some beliefs about its circumstances, one or more desired outcomes, and one or more actions taken to help produce those outcomes (Godfrey-Smith, 2005, p. 10). To borrow an example from Maibom, Oedipus wants to identify Laius’ killer, believes Teiresias can identify him, and therefore seeks out Teiresias (Maibom, 2003, p. 310). In this basic model, the interaction of belief and desire makes behavior intelligible, and consequently predictable, but leaves many interpretive options open. In particular, it leaves open questions about what kinds of systems are legitimate targets of folk psychological attribution. What properties must a system have in order to deserve description at the folk psychological level? The basic model also leaves open questions about inferential scope. Which predictions about the target’s future behavior does the attitude attribution license?

The answers to those questions are supplied by the model construal. Many different construals are possible, and this multiplicity accounts for the human capacity to ascribe folk psychological attitudes to a variety of targets.¹⁷ This is not to say that a construal can be constructed ad hoc. Maibom (2009, pp. 370–371) emphasizes that applying the model is a matter of know-how, acquired through exposure to exemplars and refined through practice and social feedback.¹⁸ Because the exemplars and feedback involved are often social, this calibration can accumulate across interpreters and over time.

This kind of practical knowledge takes time to develop, and I suspect that, in the case of LLMs, we currently have no well-calibrated construal, and therefore must draw on construals developed for other systems, or perhaps some awkward compilation of them. I also suspect that we are at least vaguely aware of the fact that no well-calibrated construal is available (even if we lack the conceptual machinery required to articulate this thought). As a result, when we ascribe attitudes to LLMs, we are disposed to draw bad inferences, unsure about which inferences are the bad ones, and aware of our own uncertainty. And that is perhaps why we hesitate to ascribe attitudes to LLMs, despite the utility of doing so.

Nevertheless, the flexibility of the model-theoretic approach to folk psychology also provides grounds for optimism about the prospects for developing a more mature and well-calibrated construal for LLMs. As we acquire additional experience trying to predict what LLMs will do on the basis of folk-psychological understanding, the chances are good that the necessary calibration will take place. However, that calibration might result in a construal that is radically different from the one we are used to applying in the ordinary human case. As a result, we should expect the inferential profile of any particular attitude ascription to deviate substantially from that familiar in the human case.

What might a mature LLM construal look like? I cannot answer that question definitively, but we can get some insight by comparing the ordinary case of human attitude ascription with the case of group agents, such as corporations. Note that we

¹⁷Godfrey-Smith discusses the possibility of there being multiple basic models. If there are, then some of the variation in attitude ascription is accounted for at the level of model choice, and the rest is accounted for in construal specification. But which variation belongs where? Justification for the multiple model view requires an answer to this question, and providing such an answer would require a more detailed account than I am prepared to offer. Devin Curry explicitly accepts the multiple models view. I remain open to it, but will stick with the single model assumption for the purposes of this paper. This choice enables me to give a more streamlined presentation, but also has the consequence that the concept of a construal, as I am using it, is necessarily broader than it is in Curry, because it has to accommodate all the variation in attitude ascription.

¹⁸Maibom does not actually use the word “construal,” but she defends the idea that the model is distinct from information about how the model can be interpreted in individual cases.

readily ascribe attitudes to corporations. For example: “Microsoft wants to acquire OpenAI, and believes it is worth . . .” Because group agents are composed of many individual believers whose attitudes may conflict, the fact that we can ascribe attitudes to corporations shows that the basic folk psychological model is remarkably flexible. So an initial lesson we might draw from the comparison is the hopeful one that, if we can do it in the corporate case without making any grievous metaphysical errors, then we can do it in the LLM case as well.

But that hopeful lesson has to be tempered because the corporate case is more similar to the ordinary human case than it first appears. A corporation can have a relatively stable set of beliefs because there are institutional mechanisms in place that impose unity systematically. The corporate hierarchy determines which voices can speak on behalf of the corporation and how collective decisions are to be reached. These mechanisms mitigate potential conflict and play a role analogous to that played by the human mitigation mechanisms described in §4. LLMs are superficially similar to corporations in that it is possible in both cases to elicit conflicting belief-like statements. In the corporate case, they are elicited from different employees, and in the LLM case, they are elicited from different context windows. But LLMs lack the sort of mitigation mechanisms that corporate structure provides. As a result, the construal that is appropriate for LLMs will license a weaker set of inferences, at least where the inferences depend on assumptions about belief integration. All else equal, that fact lowers the predictive traction of ascribing attitudes to LLMs. On the other hand, the corporate structure that imposes unity also narrows the range of topics about which a corporation could be said to have beliefs. Walmart takes no stand on the best interpretation of quantum mechanics, for example, but LLMs do. And, as we saw in §3, that increase in predictive scope contributes to the predictive traction of LLM attitude ascription and, consequently, to its legitimacy.

We are left with a complicated picture. We have good reason to ascribe attitudes to LLMs at the model level because doing so yields substantial predictive traction across an exceptionally broad range of subjects. If we are to do this well, however, we must develop a construal that provides guidance about the range of inferences that such attitude ascription licenses. Once that construal matures, it will support a pattern of inferences substantially different from the one familiar in the ordinary human case. In particular, predictions that depend on the assumption that a belief expressed in one context will guide behavior in another will be markedly less robust. If this picture is right, then skeptics have been right to think that importing our attitude-ascription

practices from the human case to the LLM case is likely to be misleading, but wrong to conclude that folk psychological reasoning is simply inapplicable. An appropriate construal will have to preserve the extraordinary breadth of LLM attitude ascription while also representing its peculiar form of disunity.

References

- Anscombe, G.E.M. (1957). *Intention*. Oxford: Basil Blackwell.
- Anthropic (2024). *The Claude 3 Model Family: Opus, Sonnet, Haiku* (Model Card). Anthropic.
- Anthropic (2026). *System Prompts*. <https://platform.claude.com/docs/en/release-notes/system-prompts>.
- Apperly, I. (2011). *Mindreaders: The Cognitive Basis of "Theory of Mind"*. Hove East Sussex ; New York: Psychology Press.
- Aune, J.E., Baker, M.J., Mackiewicz, J., Smith, J. (2025, November). Anthropomorphizing Artificial Intelligence: A Corpus Study of Mental Verbs Used with AI and ChatGPT. *Technical Communication Quarterly*, 0(0), 1–20, <https://doi.org/10.1080/10572252.2025.2593840>
- Burns, C., Ye, H., Klein, D., Steinhardt, J. (2022). Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*, , arXiv:2212.03827
- Cappelen, H., & Dever, J. (2025, April). *Going Whole Hog: A Philosophical Defense of AI Cognition* (No. arXiv:2504.13988). arXiv.
- Chalmers, D.J. (2025, January). *Propositional Interpretability in Artificial Intelligence* (No. arXiv:2501.15740). arXiv.
- Christiano, P., Cotra, A., Xu, M. (2021, December). *Eliciting Latent Knowledge*. https://docs.google.com/document/d/1WwsnJQstPq91_Yh-Ch2XRL8H_EpsnJrC1dwZXR37PC8/edit?usp=embed_facebook.
- Churchland, P.S. (1986). *Neurophilosophy: Toward a unified science of the mind-brain*. Cambridge, Mass.: MIT Press.

- Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., ... Schulman, J. (2021, October). *Training Verifiers to Solve Math Word Problems*.
- Curry, D.S. (2021, December). How beliefs are like colors. *Synthese*, 199(3), 7889–7918, <https://doi.org/10.1007/s11229-021-03144-1>
- Curry, D.S. (2023). Morgan’s Quaker gun and the species of belief. *Philosophical Perspectives*, 37(1), 119–144, <https://doi.org/10.1111/phpe.12183>
- Davidson, D. (1984). *Inquiries into truth and interpretation* (No. Volume 2). Oxford New York: Clarendon Press.
- Dennett, D.C. (1987). *The intentional stance*. MIT press.
- Dennett, D.C. (1991, January). Real Patterns. *The Journal of Philosophy*, 88(1), 27–51, <https://doi.org/10.2307/2027085>
- Dennett, D.C. (2024). We Are All Cherry-Pickers. *Anna’s AI Anthology: How to live with smart machines*. MoMo Berlin.
- Elazar, Y., Kassner, N., Ravfogel, S., Ravichander, A., Hovy, E., Schütze, H., Goldberg, Y. (2021, December). Measuring and Improving Consistency in Pre-trained Language Models. *Transactions of the Association for Computational Linguistics*, 9, 1012–1031, <https://doi.org/10.1162/tacl.a.00410>
- Fodor, J.A. (1983). *The Modularity of Mind*. MIT Press.
- Fodor, J.A. (1987). *Psychosemantics: The Problem of Meaning in the Philosophy of Mind* (M.A. Boden, Ed.). MIT Press.
- Frege, G. (1892). On Sense and Reference (M. Black, Trans.). P. Geach & M. Black (Eds.), *Translations from the Philosophical Writings of Gottlob Frege*.
- Genin, K., & Huber, F. (2022). Formal Representations of Belief. E.N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Fall 2022 ed.). Metaphysics Research Lab, Stanford University.
- Godfrey-Smith, P. (2004). On Folk Psychology and Mental Representation. H. Clapin (Ed.), *Representation in Mind: New Approaches to Mental Representation* (pp.

- 147–162). Elsevier.
- Godfrey-Smith, P. (2005). Folk Psychology as a Model. *Philosophers' Imprint*, 5, ,
- Goldstein, S., & Lederman, H. (2025). *What Does ChatGPT Want? An Interpretationist Guide*.
- Harding, J., & Sharadin, N. (2024). What is it for a Machine Learning Model to Have a Capability? *British Journal for the Philosophy of Science*, , <https://doi.org/10.1086/732153>
- Hase, P., Diab, M., Celikyilmaz, A., Li, X., Kozareva, Z., Stoyanov, V., ... Iyer, S. (2021). *Do language models have beliefs? Methods for detecting, updating, and visualizing model beliefs*. arXiv.
- Heafield, K. (2011). KenLM: Faster and Smaller Language Model Queries. , *Proceedings of the Sixth Workshop on Statistical Machine Translation*, ,
- Hempel, C. (1965). *Aspects of Scientific Explanation*. New York, NY: Free Press.
- Hofweber, T., Hase, P., Stengel-Eskin, E., Bansal, M. (2024, August). *Are language models rational? The case of coherence norms and belief revision* (No. arXiv:2406.03442). arXiv.
- Horgan, T., & Tienson, J. (2002). The Intentionality of Phenomenology and the Phenomenology of Intentionality. D.J. Chalmers (Ed.), *Philosophy of Mind: Classical and Contemporary Readings* (pp. 520–533). Oxford University Press USA.
- Hubinger, E., Denison, C., Mu, J., Lambert, M., Tong, M., MacDiarmid, M., ... Perez, E. (2024, January). *Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training* (No. arXiv:2401.05566). arXiv.
- Humphreys, P. (2004). *Extending ourselves: Computational science, empiricism, and scientific method*. New York: Oxford University Press.
- Jiang, Z., Xu, F.F., Araki, J., Neubig, G. (2020). How Can We Know What Language Models Know? *Transactions of the Association for Computational Linguistics*, 8, 423–438, https://doi.org/10.1162/tacl_a_00324

- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Advances in Neural Information Processing Systems* (Vol. 33, pp. 9459–9474). Curran Associates, Inc.
- Li, K., Patel, O., Viégas, F., Pfister, H., Wattenberg, M. (2023, October). *Inference-Time Intervention: Eliciting Truthful Answers from a Language Model* (No. arXiv:2306.03341). arXiv.
- Liu, N.F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., Liang, P. (2024). Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*, 12, 157–173, <https://doi.org/10.1162/tacl.a-00638>
- Maibom, H. (2003). The Mindreader and the Scientist. *Mind & Language*, 18(3), 296–315, <https://doi.org/10.1111/1468-0017.00229>
- Maibom, H. (2009). In Defence of (Model) Theory Theory. *Journal of Consciousness Studies*, 16, 360–378,
- Mandelbaum, E. (2014). Thinking is believing. *Inquiry: An Interdisciplinary Journal of Philosophy*, 57(1), 55–96, [https://doi.org/10.1016/s0079-6123\(07\)68009-1](https://doi.org/10.1016/s0079-6123(07)68009-1)
- McCoy, R.T., Smolensky, P., Linzen, T., Gao, J., Celikyilmaz, A. (2023). How Much Do Language Models Copy From Their Training Data? Evaluating Linguistic Novelty in Text Generation Using RAVEN. *Transactions of the Association for Computational Linguistics*, 11, 652–670, <https://doi.org/10.1162/tacl.a-00567>
- McGeer, V. (2007). The regulative dimension of folk psychology. D.D. Hutto & M. Ratcliffe (Eds.), *Folk psychology re-assessed* (pp. 137–156). Kluwer/Springer Press.
- Quine, W.V.O. (1960). *Word and object*. MIT Press.
- Rathkopf, C., Heinrichs, J.-H., Heinrichs, B. (2022). Can we read minds by imaging brains? *Philosophical Psychology*, 36(2), 221–246, <https://doi.org/10.1080/09515089.2022.2041590>
- Salmon, N.U. (1986). *Frege's puzzle*. Cambridge, Mass.: MIT Press.

- Schwitzgebel, E. (2001). In-between Believing. *The Philosophical Quarterly*, 51(202), 76–82, <https://doi.org/10.1111/1467-9213.00215>
- Schwitzgebel, E. (2002). A Phenomenal, Dispositional Account of Belief. *Noûs*, 36(2), 249–275, 3506194
- Schwitzgebel, E. (2013). A dispositional approach to attitudes: Thinking outside of the belief box. *New essays on belief* (pp. 75–99). Springer.
- Schwitzgebel, E., Schwitzgebel, D., Strasser, A. (2024). Creating a large language model of a philosopher. *Mind & Language*, 39(2), 237–259, <https://doi.org/10.1111/mila.12466>
- Searle, J.R. (1983). *Intentionality: An essay in the philosophy of mind*. Cambridge University Press.
- Shah, N., & Velleman, J.D. (2005). Doxastic deliberation. *The Philosophical Review*, 114(4), 497–534, <https://doi.org/10.1215/00318108-114-4-497>
- Shanahan, M., McDonell, K., Reynolds, L. (2023, November). Role play with large language models. *Nature*, 623(7987), 493–498, <https://doi.org/10.1038/s41586-023-06647-8>
- Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askill, A., Bowman, S.R., ... Perez, E. (2024). Towards Understanding Sycophancy in Language Models. *ICLR Proceedings*. arXiv.
- Singhal, K., Azizi, S., Tu, T., Mahdavi, S.S., Wei, J., Chung, H.W., ... Natarajan, V. (2023, August). Large language models encode clinical knowledge. *Nature*, 620(7972), 172–180, <https://doi.org/10.1038/s41586-023-06291-2>
- Spaulding, S. (2018, June). Mindreading beyond belief: A more comprehensive conception of how we understand others. *Philosophy Compass*, 13, e12526, <https://doi.org/10.1111/phc3.12526>
- Takeuchi, T., Duszkievicz, A.J., Morris, R.G.M. (2014, January). The synaptic plasticity and memory hypothesis: Encoding, storage and persistence. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 369(1633), 20130288, <https://doi.org/10.1098/rstb.2013.0288>

- Tay, Y., Dehghani, M., Bahri, D., Metzler, D. (2022, March). *Efficient Transformers: A Survey* (No. arXiv:2009.06732). arXiv.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., . . . Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30, , <https://doi.org/10.65215/nxvz2v36>
- Xu, R., Qi, Z., Guo, Z., Wang, C., Wang, H., Zhang, Y., Xu, W. (2024, November). Knowledge Conflicts for LLMs: A Survey. Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 8541–8565). Miami, Florida, USA: Association for Computational Linguistics.
- Zawidzki, T.W. (2013). *Mindshaping: A new framework for understanding human social cognition*. MIT Press.