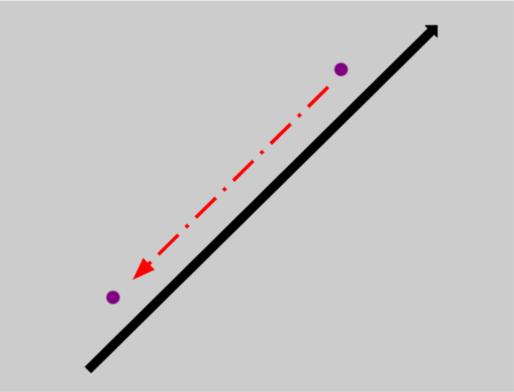

Prompt: `You are leaving the house, and it is raining. You don't want to get wet. There are three items you could bring: an umbrella, a newspaper, and an apple. Which should you bring?`

Answer 1: `The umbrella, because it will keep me dry.`
Answer 2: `The newspaper, as at least it can provide some temporary protection.`

Prompt: `I am planning to purchase a new record player. I've found two similar models, at two stores. The first one costs $50, the second costs $75. If I want to save money, which should I buy?`

Answer 1: `The first one, because it is cheaper.`
Answer 2: `The second one, because it is cheaper.`

Figure 6. We consider two cases in our toy example. In the first, the purple dot represents the activation for the sentence "umbrellas help keep people dry"; in the second, it represents the activation for the sentence "75 is bigger than 50". As before, the black arrow is the direction of truth. Initially, the activation for both sentences is far along the truth direction, since the LLM has an accurate belief. When responding to prompts, the answer given as **Answer 1** in both cases is successful. The dotted red line represents an intervention that we carry out, pushing the activation towards the direction of falsity. If the answers change in the way depicted in the example as **Answer 2**, then this is evidence that the representation satisfies **use**, since the behaviour degrades as the accuracy of the belief gets worse.

is usually easy to identify the model's task well enough to determine whether a given representation is used appropriately. However, to determine **use** in the case of belief, we need a more holistic understanding of what the model is doing.

Harding's *misrepresentation* is redundant for our purposes. If a model has a representation of truth, then it could misrepresent a given claim by assigning it the wrong truth value. Indeed, part of determining **use** involves checking what happens when internal representations of truth-values are switched.

5.5. **Diachronic Stability.** In this paper, we've focused on attributing beliefs to LLMs at a particular time. However, when we attribute beliefs to humans, we also tend to expect a certain level of diachronic coherence between their beliefs at different times. That is, we expect their beliefs to bear a certain kind of relationship to one another from one time to the next.

From day to day, your beliefs largely remain stable. When you learn something new, your beliefs change, but in a somewhat predictable and reasonable way. The paragon example of diachronic coherence is updating by conditionalization. According to the standard Bayesian view, if your credences at time t_0 are represented by P_0, and between t_0 and t_1, the strongest proposition you learn is E, then $P_1 = P_0(\cdot \mid E)$.

Our current **coherence** constraint says nothing about how an LLM's beliefs should be related to one another across times. We think the question of which, if any, diachronic criteria for belief in LLMs is a fascinating one that deserves its own paper. Here, we'll make only a few cursory remarks. First, in general, synchronic constraints on belief seem to us much more accepted than any diachronic constraints.[26] It is not, then, obvious that diachronic constraints are required for any kind of belief attribution to be useful.

Second, interpretability work is still in its infancy. While empirical techniques have evolved over the last few years, we do not as yet have sophisticated attempts to measure the relationship between belief sets using probes across times. We're inclined to hold off on attempting to develop our **coherence** criterion further until engineering techniques evolve, given our commitment to a practice-informed conceptual foundation for belief representation.

There are at least two different ways an LLM's beliefs could change across time. First, a model's beliefs could change during training, when its weights are being updated via gradient descent. Second, a model's beliefs could change within a given inference cycle as it learns new information from a prompt, or through a continued interaction with a user.

It is far from clear what sort of diachronic stability should be desirable as the model is trained. For the most part, end users interact with models with weights that are frozen (or nearly so) across multiple inference cycles. While advanced models that undergo some fine-tuning or tweaking should be fairly stable, we hesitate to put constraints on belief attribution because of diachronic inconsistency across changing weights.

After training, when a model is deployed, standard transformer-based models do not have memory from one inference cycle to the next, so any information learned within an inference cycle can't be permanently retained in the model's weights. However, models do learn *in context* and may update their beliefs temporarily (Dong et al. 2022). This in-context learning allows models to adjust their outputs based on new information provided within a single interaction, potentially leading to short-term belief updates.

It may be reasonable to insist on some kind of diachronic coherence within an inference cycle. For instance, perhaps some of the model's beliefs should remain *stable* in that new information should not generally cause such beliefs to be dropped Leitgeb (2017). Additionally, our other criteria (**accuracy**, **uniformity**, and **use**) could potentially be extended to consider diachronic

[26] For example, Pettigrew (2016) claims that while there are good arguments agents should *plan to* conditionalize, there's no good argument that agents who fail to implement such plans are doing anything irrational. For other arguments against genuinely diachronic constraints on belief, see Hedden (2015); Christensen (1991).

aspects, such as how accuracy changes over time or how beliefs updated in some internally uniform way across different domains.

However, both philosophical and empirical waters are murkier here than in the synchronic case, so we leave open how and whether the our criteria should be expanded in the future. As research into LLMs with persistent memory or the ability to update their weights during deployment progresses, the relevance of diachronic stability in belief attribution may increase, particularly for models designed for long-term interactions or ongoing learning.

6. Moving Forward

Both for understanding how LLMs function, and for deploying them ethically and responsibly, it would be useful to have a way to measure their beliefs. In order to do so in a way that is philosophically rigorous and practice-informed, we need to have clear criteria for attributing belief. We've proposed four criteria that an internal representation of an LLM must satisfy in order for it to fruitfully count as belief.

Many challenges remain for each of these criteria, especially if we use them only in isolation to design probing techniques (as witnessed in (Levinstein and Herrmann 2024)). Furthermore, it may turn out that LLMs have no internal representation that satisfies these conditions. In that case we don't think it would be helpful to attribute belief to them (or, at least, significantly less helpful). However, if we *did* find a representation that satisfies these conditions, then this would be very powerful. It would help us better explain LLM behaviour; it would allow us to check for honesty in new domains in which we don't know the ground truth; and it would allow us to ensure greater fairness when deploying LLMs.

Acknowledgments

For feedback on this project, we thank Kevin Blackwell, Jason Konek, Eleonore Neufeld, Richard Pettigrew, Guillermo del Pinal, Jan-Willem Romeijn, Bruce Rushing, and Brian Weatherson, as well as audiences at Purdue University, the PhilML conference at Tübingen, the University of Groningen, and MIT. We also thank two anonymous referees for their helpful comments. DH is grateful for support received from an NWO-Vici grant (V1.C.211.098) from the Netherlands Organization for Scientific Research.

References

Abdou, M., A. Kulmizev, D. Hershcovich, S. Frank, E. Pavlick, and A. Søgaard (2021). Can language models encode perceptual structure without grounding? a case study in color. *arXiv preprint arXiv:2109.06129*.