find stable beliefs in LLMs from behaviour alone. Thus, while the limited behavioural basis we discussed in §4 likely frustrates attempts to reconstruct beliefs purely from observed behaviour of an LLM, we do think that any representation of belief we *do* identify should still guide the LLM's behaviour. This then in turn motivates conditions like **accuracy** and **coherence** which we describe below.

In light of this different epistemic situation, making full use of our easy access to the internals of systems to inform our requirements for representations of belief is essential, and encourages us to hew close to machine learning practice. However, attributing belief to LLMs should not be entirely divorced from philosophical theories of belief. Ultimately we are still after a set of conditions that would be help identify a representation that plays the role of belief in such systems, and it is theory that gives us guidance for what that role is.

In this spirit, at a high-level, we have the following motivations for the various requirements we propose. In order to agree with a core feature of standard accounts of belief, for example in folk psychology and decision theory, we want the representation to be *action-guiding*. This most clearly motivates what we call **use**. Furthermore, we want such a representation to help explain why LLMs are so successful. Thus, we also want the beliefs to be accurate enough that they could help an LLM be successful. This motivates both what we call **accuracy** and **coherence**. We also want such a representation to be *helpful* for us interpreting the LLM. Thus, we would need such a representation to be measurable and interpretable. This, along with the requirement that we want the representation to allow us to make predictions across a wide range of domains, motivates what we call **uniformity**.

In addition to describing and providing justification for each requirement, we'll also use our toy example to help visualize the requirements. We will also use empirical failures of contemporary probing techniques to illustrate why taking various requirements alone is insufficient for identifying belief.

5.1. **Accuracy. Accuracy** requires that the decodings of the identified representations be reasonably accurate on datasets where the LLM is expected to have true beliefs (or high subjective probability). The exact form this requirement takes will depend on the type of doxastic attitudes we think we've identified. If the LLM has on/off beliefs, many of its beliefs on the relevant dataset should be true. If it has credences, it should be accurate according to some strictly proper scoring rule.[10]

The main motivation for **accuracy** is that true beliefs (or high confidence in truths and low confidence in falsehoods) should partly explain the LLM's impressive performance. That is, if it's worth attributing beliefs at all, one

[10]We might also use datasets with claims that we expect the LLM couldn't possibly have reasonable full beliefs about, such as "There are an even number of stars in the Milky Way" to add a kind of **calibration** requirement as well.

of the reasons is that we can explain the LLM's general success by appeal to its *true* beliefs. For full-fledged agents with both beliefs and desires, true beliefs help them get what they want. With LLMs, even if they don't have desires, true beliefs should explain successful and skillful performance. If not, then the LLM's success is better explained solely by non-truth-tracking features, i.e., features other than beliefs.

This consideration is motivated, in part, by our requirement that the criteria proposed be both practice informed and *useful to us*. If an LLM internally represents truth and falsity poorly and its misrepresentations do not affect performance, then it is not worthwhile for us, as interpreters, to dub those representations 'beliefs.'

Usefulness to us also prohibits us from using *reasonableness* in place of accuracy, especially for large models trained on massive amounts of text in the wild. From a purely philosophical standpoint, it might be more appealing to require only that the LLM draws reasonable or justified or well-supported inferences from its evidence. Because we control its evidence through its training data and prompts, we could in principle trick it into having terrible beliefs. (In essence, we can now play the role of a Cartesian demon.)

While this philosophical position is tenable, it is not as useful, operationally, as **accuracy**.[11] Reasonableness is generally much harder to test for than truth. It's hard to know what identifiable falsehoods an LLM should reasonably believe, for example, after reading most of the high quality internet. Reasonableness is also less interesting when it comes to explaining successful performance. True beliefs, not merely reasonable ones, lead to success.

It's important to note that while we are ultimately interested in identifying belief-like representations in general, including both true and false beliefs, we focus initially on true beliefs as a pragmatic starting point. This approach allows us to establish a baseline for identifying belief-like representations before tackling the more complex task of detecting systematically false beliefs.

This focus on true beliefs as a starting point aligns with established philosophical approaches to belief attribution. In particular, the accuracy criterion is closely connected to the Principle of Charity found in the radical interpretation literature (Davidson 1974, 1973; Lewis 1974). On Davidson's view, for instance, we must begin by assuming the subject to be interpreted is generally a "believer of truths" in order to get the project of interpretation off the ground—if we don't take the subject to be someone who has largely true beliefs, then we won't be able to make sense of her beliefs at all (Davidson 1970). Likewise, Lewis (1974) takes a different version of the

[11]In some cases, it's useful to train models on specially curated data or synthetic data instead of text from the wild. It's conceivable that models trained in this way could potentially form beliefs that are not at all accurate. In this case, a reasonableness criterion might be more worthwhile.

Principle of Charity to be required to make sense of a physical system as having beliefs, desires, and intentions.[12]

Importantly, our **accuracy** criterion, as with the Principle of Charity, requires accurate beliefs over the right sort of questions. Consider the case of trying to make sense of a human speaking a foreign language we do not understand. In this situation, we do not begin by assuming the speaker has true beliefs about complex topics like monetary policy or quantum computing. Instead, we start with common-sense and obvious questions such as whether there are any chairs in the room we're sitting in or whether it's currently raining. This allows us to connect the speaker's utterances to the world while still allowing the speaker to have many false beliefs.

Likewise, **accuracy** over the *appropriate* data set is key for discovering an LLM's beliefs. The relevant datasets for testing **accuracy** will naturally be ones where we both are highly confident the LLM will or should have true beliefs[13] given its training data and ones where we know the ground-truth of the claims in question. We do not require accuracy in general, or accuracy over every domain.

This also limits how much we can lean on **accuracy** alone for identifying beliefs.[14] It will be hard, for instance, to include claims about much of philosophy, economics, future world events, or any scientific claims aside from fully settled ones. We won't be able to use sentences like "Keynesian economics is broadly correct," "humans have free will," or "most people won't benefit from taking multi-vitamins." Similarly, statements like "climate change will cause a global economic recession by 2050," "artificial intelligence will surpass human intelligence within the next decade," or "string theory is the correct framework for understanding quantum gravity" cannot be included in a good test set despite the fact that we really would like to know what LLMs think about such questions. These claims involve significant debate, varying interpretations, and a lack of consensus, making them unsuitable for testing accuracy in belief-like representations.

Once we have established a reliable method for identifying true beliefs, we can extend our approach to detect systematically false beliefs. This extension would be particularly valuable for addressing social and ethical concerns related to LLM deployment, as it would allow us to identify areas where an LLM might consistently make errors or hold mistaken beliefs, even if its overall performance seems satisfactory.

---

[12]Interestingly, Lewis's Principle of Charity is closer to our rejected *reasonableness* criterion. Lewis suggests "We should even ascribe to [the subject] those errors which we think we would have made, or should have made, if our evidence and training had been like his" 1974, p. 336.

[13]In some situations, we might use datasets where we have strong reason to suspect the LLMs will have false beliefs, but we expect this to be unusual for highly capable LLMs.

[14]For example, using accuracy measures to define a loss function for training probes, as in Azaria and Mitchell (2023).