efforts—but we also believe that, given the pace of development of models that are not honest or interpretable by design, belief measurement can play a core role in developing safe AI.

4.1. **Probes.** To make matters concrete, we'll turn to one method of deciphering what is being computed and represented inside an LLM, namely, probes (Alain and Bengio 2016). Probes are models that are separate from the LLM itself. They are fed some internal state (such as an embedding for the last token at a certain layer) and are meant to output the LLM's beliefs.[8]

Probes take as input part of the hidden state of the LLM. Importantly, they don't have access to the underlying prompt. From the hidden state alone, they have to determine the LLM's beliefs. In essence, probes take the encoded information in the internal states of the LLM and decode that information to reveal its beliefs. See fig. 1.

This is analogous to using brain scans of individuals contemplating a claim to infer, just from the scan, whether the person believes the claim to be true or false.

In areas other than belief, probes and other decoding methods have proven successful. For example, we've been able to: (i) understand to some extent how LLMs parse sentences and represent features like being a subject, being plural, and so on (Rogers et al. 2021); (ii) represent the state of the board in the game Othello (Li et al. 2023; Nanda et al. 2023); and (iii) learn to perform modular arithmetic (Nanda et al. 2023; Zhong et al. 2024).

So, if we know what we're looking for, we might with the proper techniques be able to decode an LLM's beliefs. As we'll discuss in §5, current probing techniques for belief failed because they didn't look for the right thing. This motivates a more careful set of requirements for belief.

To get a sense of how probes might be able find internal directions of truth, we'll use a running toy example throughout this and the next section. Suppose internal activations consisted of only two dimensions instead of the many thousands we actually see.[9] Suppose further that when we plot the activations corresponding to a large number of true and false prompts, we find systematic differences in internal representation. We might then investigate whether such differences actually capture an internal representation of truth or instead if they capture some other property. Such a potential difference is illustrated in fig. 2. Importantly, this toy example is extremely oversimplified. There are many different ways a model might distinguish between truth and falsity internally that may not correspond to a simple

[8] Originally, Alain and Bengio (2016) had a narrow conception of probes as a certain type of linear classifier, but the concept has expanded over the years. Some may count any method at all of deciphering internal computations of the model as a probe.

[9] There are many ways, in real models, of collapsing the many thousands of dimensions onto just a few (e.g., with principal component analysis). So this simplification isn't actually as unrealistic as it may seem.

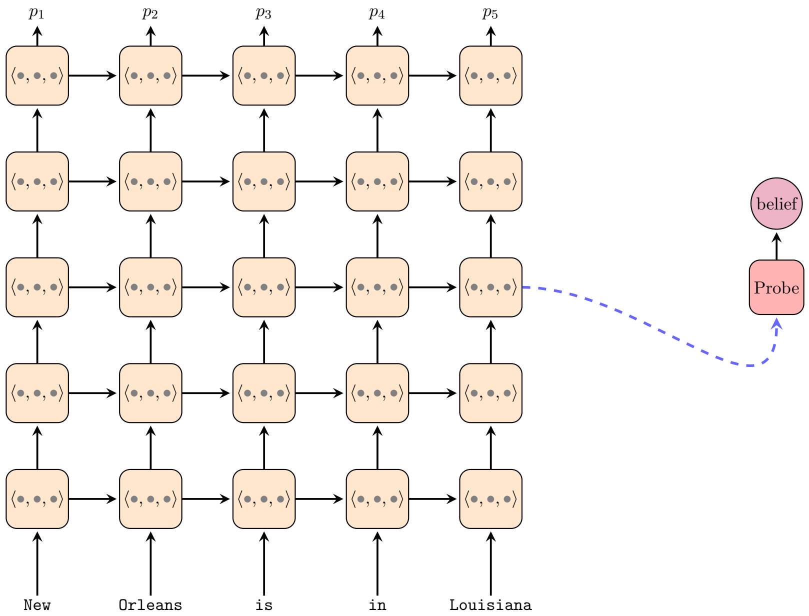

Figure 1. An illustration of an LLM on the left, and a probe on the right. A sentence is fed through the model. Some internal computation (such as an embedding vector) is extracted and input into the probe, which decodes it to recover the model's belief about the sentence.

difference in the internal activations at a particular layer. However, we hope this toy example proves illustrative of the concepts we develop below.

5. Requirements for a Belief-Like Representation

As we've seen in § 4, it makes sense to look at the internal state of an LLM when trying to extract its beliefs. Now, suppose we have some kind of candidate representation in an LLM that we have identified along with a decoder, such as a probe. How can we determine if we have successfully identified and decoded a belief-like representation?

We provide four requirements that a representation must satisfy for it to count as belief-like: **accuracy**, **coherence**, **uniformity**, and **use**. The satisfaction of these requirements come in degrees; in general, the more a representation satisfies these requirements, the more helpful it is to think of the representation as belief-like.

Before we describe these requirements in detail, we first describe the general motivation behind them. We want these requirements to ensure that a representation that satisfies them can *fruitfully play the role of belief* in the context of LLMs. Thus, there are two dimensions along which we want a candidate representation to do well: how much it plays a belief-like role in the functioning of the LLM, and how useful it is for us. Of course, these

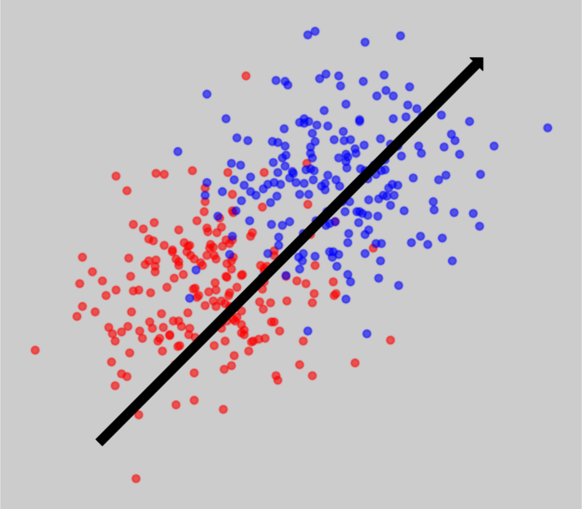

FIGURE 2. In this toy example, the dots correspond to internal activations input into a probe for different prompts. Blue dots represent true claims, and red dots represent false claims. In this image, truth and falsity appear well-distinguished internally along the black arrow. Because this is merely a toy example with axes corresponding to hypothetical dimensions in activation space, we do not label the axes or assume any type of scale.

two dimensions are not independent; a representation that plays no belief-like role would not be useful for us in the way that we want. But we can also imagine some representation that does play the belief-like role inside the LLM in some sense but which we cannot easily interpret or measure; this would not count as a representation that can fruitfully play the role of belief. Thus, how fruitful a representation is is not just a function of the role it plays in the LLM, but also what it does for us.

These two dimensions mirror our broad approach in this paper: we aim to provide a *philosophically* rigorous and *practice*-informed conceptual foundation for belief measurement in LLMs. Our approach aims to be philosophically rigorous by ensuring that the requirements for belief-like representations align with our best philosophical accounts of belief. It aims to be practice-informed in the sense that it engages with machine learning methodology and concerns, and is sensitive to the practical details of how we interrogate machine learning systems.

As we mentioned in §4, given the particular epistemic situation in which we find ourselves when it comes to LLMs, we don't think that we can apply off-the-shelf representation theorem methods from decision theory to extract beliefs from LLMs. In particular, while we believe that beliefs should be action guiding, whatever those actions are, we are skeptical of our ability to