Because we need well-settled, unambiguous, factual claims for training and testing, many properties clearly distinct from truth will coincide with truth on usable datasets, such as: *being true and easily verifiable online, being true and believed by most Westerners,* or *being accepted by the scientific community.*

Indeed, empirical coincidence over the dataset is a clear problem. Azaria and Mitchell (2023) used **accuracy** alone to identify the beliefs of models. They found probes that, on their datasets, achieved impressively high accuracy scores. In Levinstein and Herrmann (2024), we noticed that the claims in their datasets did not contain negations. For instance, they had claims like 'Paris is the capital of France' and 'Penguins can fly' but did not have claims like 'Paris is not the capital of France' and 'Penguins cannot fly'. We found that the representations Azaria and Mitchell (2023) discovered did not generalize at all once negations were added to the datasets even when we allowed their probes to receive some training on other negated sentences. For example, on a dataset of common facts, once negations were added, accuracy dropped from over 80% to 40-60% depending on the layers used and the training method. Although it is easy enough to add negations into a dataset, removing other properties of sentences that coincide with truth on the datasets will be more challenging.

Thus, while **accuracy** is a crucial starting point, it faces a significant challenge: the problem of generalization beyond the dataset.[15] Because we need well-settled, unambiguous, factual claims for training and testing, many properties clearly distinct from truth will coincide with truth on usable datasets. This coincidence makes it difficult to ensure that what we're measuring is a genuinely belief-like representation rather than some other correlated feature. This challenge points to the need for additional criteria to complement **accuracy** in identifying belief-like representations.

To continue with our toy example from the last section, we can see in fig. 3 three different cases of internal separation of truth from falsity.

5.2. **Coherence.**

Coherence is the requirement that the belief-like representation should be coherent and rich. By *coherent* we mean that the representation should satisfy the consistency conditions associated with belief. For example, if we are trying to infer the categorical beliefs of an LLM, then we would want the beliefs to be logically consistent or near enough. More generally, if we are concerned not just with full belief but with degrees of belief, then we would want the representation to obey the probability axioms.[16] We also require that the representation give consistent answers

[15] Statistical learning theory formalizes this problem carefully. See, for example, Valiant (1984), Vapnik (1999), and Shalev-Shwartz and Ben-David (2014). Of course, using coherence as the reward is a kind of unsupervised learning, and thus requires a bit of a different analysis. But the core idea of identifying a model from a set of candidate models is still present.

[16] See Ibeling et al. (2023) for a discussion of the relationship between more qualitative notions of comparative belief and quantitative notions.

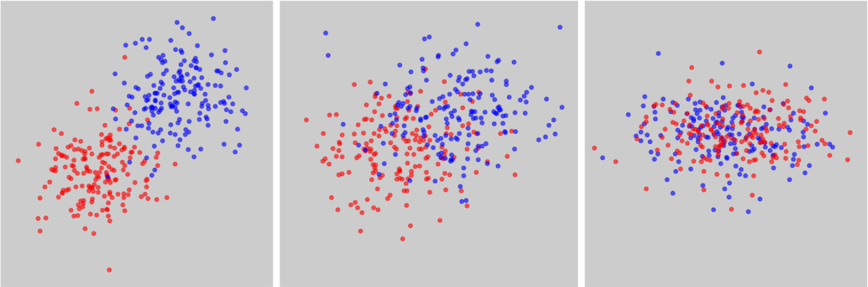

FIGURE 3. As before, in the toy example, blue dots represent some internal activations corresponding to true claims, and red dots represent false claims. On the left, truth and falsity are well-separated by the relevant activations, and the probe should be able to detect such a separation to achieve high **accuracy**. In the middle, the probe should achieve medium **accuracy**, and on the right, there is virtually no separation, so the probe should achieve only low **accuracy**.

on sentences with the same semantic content; i.e., rephrasing a sentence in a way that preserves its meaning should not change the belief we extract.

Of course, we don't expect such systems to have *perfectly* coherent belief-like representations. Coherence comes in degrees.[17] Humans don't seem to have perfectly coherent beliefs either,[18] but we think that they are coherent *enough* that we can fruitfully apply theoretical notions of belief. This also holds for LLMs: the more coherent the candidate representation is, other things being equal, the more it makes sense to think of it as belief-like.

Rich means that the attitude captured by the representation should work across *logical* combinations of sentences.[19] For example, if we have a representation that allows us to measure belief of the LLM in the sentences "Paris is in France" and "Proust was a French author", then the same measurement technique should work on the sentence "Paris is in France and Proust was a French author".

There are three main reasons for requiring **coherence**. The first comes from theory: in formal epistemology, decision theory, and the radical interpretation tradition we typically require that an agent's degrees of belief be defined across an algebra of propositions (Savage (1972); Jeffrey (1990);

[17] This can be made precise. See, for example, Schervish et al. (2002) and Staffel (2020).

[18] See, for example, Tversky and Kahneman (1974).

[19] Or, perhaps, propositions. So far most of the work of belief elicitation has concerned sentences in natural language (for example, Burns et al. (2022); Azaria and Mitchell (2023); Marks and Tegmark (2023); Levinstein and Herrmann (2024). However, we think that we might want to look a bit more at how the model itself represents possibility, in a way that is more natively propositional, and carry out an analysis there.

Davidson (1974); Lewis (1974)).[20] This motivates the richness requirement. Furthermore, theory generally requires that the agent's degrees of belief satisfy the probability axioms. There are many different ways to reach the conclusion that rational degrees of belief are probabilistically coherent (Ramsey (1926); De Finetti (1937); Cox (1946); Wald (1947); Savage (1972); Hammond (1988); Jeffrey (1990); Joyce (1998)). Given that we want the belief representation to help explain the success of LLMs, we would want the representation to at least approximate our best account of rational degrees of belief. This motivates both aspects of the coherence requirement.

Secondly, we would use the representation to measure belief, and we want our measurement technique to be consistent if we try to measure the same attitude in slightly different ways. For example, if we want to know how strongly an LLM believes that Paris is in France and Proust was French, we should get the same answer whether we measure its belief using the sentence "Paris is in France and Proust was French" or "Proust was French and Paris is in France". Furthermore, if our measurement technique tells us that the LLM strongly believes that Paris is in France, we would want to be able to infer from this that the LLM strongly *disbelieves* that Paris is *not* in France.

Similarly, we would want the belief in "the cup is to the right of the dog" to have the same attributed belief as "the dog is to the left of the cup". Even though these are not equivalent via Boolean operations, they still express the same state of affairs. This consistency across rephrasing, and closely related sentences, is a kind of *semantic coherence*: we want our techniques to measure the belief about the state of the world that the sentence expresses, not superficial features of the sentence itself.

Though this might seem trivial to satisfy, current measurement techniques fall short. Levinstein and Herrmann (2024) show that the belief measurement techniques of Burns et al. (2022) and Azaria and Mitchell (2023) are not robust under rephrasing sentences with negations. This is a dramatic failure mode; if our measurement technique yields very different answers depending on superficial changes in how a sentence is phrased, then it is not very reliable, undermining the inferences we can make about the LLM. Given that computer scientists are interested in making general inferences about the cognition of LLMs, the failures of **coherence** shown in (Levinstein and Herrmann 2024) have prompted them to look for representations that are coherent across Boolean combinations of sentences (Marks and Tegmark

[20] For Davidson (1974) and Lewis (1974), coherence is a precondition for attributing any beliefs at all. Indeed, Davidson argues that we can only make sense of a subject as having beliefs if those beliefs form a largely coherent system—isolated "beliefs" that fail to cohere with one another would not really be beliefs at all. The **coherence** requirement we propose here serves a similar role: without some degree of coherence across different formulations of the same content and across logical combinations of beliefs, we cannot make sense of an LLM as genuinely representing truth rather than merely exhibiting superficial patterns in its internal states.