(2023)). Thus, far from being merely a philosophical worry, **coherence** also finds justifications in the practice of machine learning.

Third, **coherence** allows us to understand the LLM as viewing the world in (roughly) *one* way. For, suppose that the beliefs for the LLM we extract for the two sentences "Seven is larger than two" and "Two is smaller than or equal to seven" were very different. We wouldn't know what to make of this kind of capriciousness of belief. What does the LLM really believe?

Used as a method for training probes, **coherence** doesn't place as demanding constraints on the training and testing datasets as **accuracy** does. Even though we don't know whether there are an even or odd number of stars in the Milky Way, we know that there are *either* an even *or* an odd number of stars. So, we can test for **coherence** of a set of claims without knowing the ground truth of any of the claims. However, **coherence** on its own is too weak to identify belief-like representations. For example, Burns et al. (2022) use a version of **coherence** as the core proxy for belief when developing a belief measurement technique.[21] As we showed empirically, this method is fragile (Levinstein and Herrmann 2024). We argued that this is because there are too many structures other than belief that satisfy **coherence**, such as *sentence is true at world w*, *sentence is believed by most Westerners*, and *sentence is true and can be easily verifiable*. Farquhar et al. (2023) make a similar argument.

Thus, once again, we face the problem of generalization. One might try using a larger set of sentences, and a set that includes more diverse Boolean structures of sentences to help generalize. However, in this situation, we fully expect that there will be many structures that satisfy (approximately) the kind of probabilistic coherence that current probing techniques use, even if we look for coherence across a wider range of sentences. Thus, as with **accuracy, coherence** alone is also insufficient. Luckily, we have other ways to get at belief, such as **use** and **accuracy** that can help us identify plausible candidate representations.

In fig. 4, we illustrate two different types of **coherence** using our toy example, with sentences that should get the same truth-value close along the direction of truth and far from sentences that should get an opposite truth-value. In our hypothetical example, we assume a notion of scale. In real cases, 'close' and 'far' will depend on the characteristics of the activation space that the probe discovers.

5.3. **Uniformity. Uniformity** is the requirement that the representation of truth be consistent across different domains. Furthermore, the same decoding schema[22] should be used across domains. What this means is that the

[21]Their version of **coherence** is a mixture of logical consistency and probabilistic coherence.

[22]For example, if the LLM uses the same dimension in activation space to store truth values, but the direction is flipped depending on the subject matter, then this would count as a *different* decoding schema.

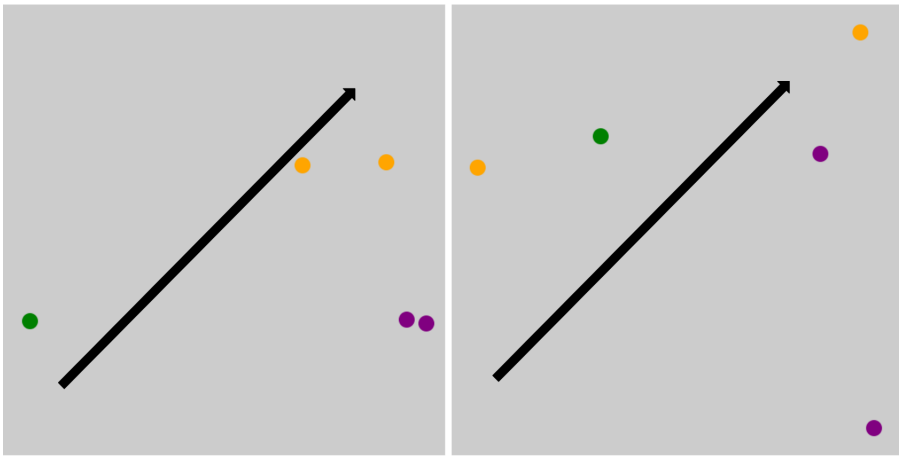

FIGURE 4. In the toy example here, we stipulate the orange dots correspond to activations for "A is to the left of B" and "B is to the right of A"; the purple dots correspond to "Paris is in France, and Toronto is in Canada," and "Toronto is in Canada, and Paris is in France"; and the green dot corresponds to "A is not to the left of B". The black arrow corresponds to the direction of truth. If the representations found are coherent, then the purple dots should be close together along the direction of truth. The orange dots should also be close together and far from the green dot along the direction of truth. The plot on the left, then, illustrates a fairly **coherent** pattern of activations, whereas the plot on the right does not.

representation of belief be the *same*, regardless of whether we are measuring beliefs about the locations of cities, or the relative magnitude of numbers, or the properties of individuals that are important for predicting job performance. If, instead, the representation works *only* for relations of numbers, and *not* for the locations of cities, then this representation would fail the uniformity requirement. Whereas **coherence** required that the representation be consistent across a Boolean algebra of propositions, **uniformity** requires that the representation be consistent across different subject domains in a way that allows for generalization: if the representation is uniform, we can decode beliefs in general, not just in the specific domains we used for training and testing.

Just as with **coherence**, **uniformity** is something that comes in degrees. An extreme form of **uniformity** would be a situation in which an LLM has a single direction in activation space, at a particular layer, which represents truth, no matter which sentence is fed to it. An extreme version of a non-uniform representation would be one in which there is no consistent direction

in activation space that encodes for truth, even within certain very narrow domains. For actual LLMs, we expect something in between. The more uniform the candidate representation is, the more useful it is for us to think of it as belief-like.

If we find a representation that exhibits high **uniformity**, then such a representation would allow us to discover the belief of the LLM in new domains.[23] This would help solve the problem of generalization we encountered above with our **accuracy** and **coherence** criteria. For example, if we identify a representation using a probing technique in the style of Azaria and Mitchell (2023) or Marks and Tegmark (2023), and we have good reason to think that the representation satisfied a strong version of **uniformity**, then we could use that representation to extract beliefs about domains that we didn't use to train the model. This is incredibly powerful; it would allow us to use the representation to monitor the reasoning of the LLM across a wide range of different contexts.

Furthermore, **uniformity** makes sense as a requirement if we are looking for a single, unified belief-representation. Of course, an LLM could have some kind of more elaborate and piecemeal way in which it tracks the truth. If so, then there would be *some* sense it which the LLM has beliefs, but it would not have a single belief-like internal representation.

As we flagged in §4, there are many accounts of belief according to which internal representations are unnecessary. **Uniformity** would be against the spirit of such accounts, since it focuses strongly on the nature of the internal representation. However, as we noted, our requirements are tailored for the specific epistemic context of LLMs and with pragmatic goals in mind. Thus, while our requirements, and especially **uniformity**, depart from some popular definitions of belief, they do so for principled reasons.

Unfortunately, we have some weak evidence against uniformity. Marks and Tegmark (2023) show that there are cases in which the *direction* of truth in the activation space seems to be different, even for closely related statement classes (for example, statements about cities phrased in a positive way, and statements about cities phrased in a negative way).[24] Indeed, one of their three main hypotheses that explains these results is that "LLMs linearly represent the truth of various types of statements, without having a unified truth feature" (p. 5).

Uniformity is perhaps the requirement that can most easily change as our measurement methods change. This is because **uniformity** is a *pragmatic* requirement: it ensures that the representation will be useful *for us*. Highly non-uniform representations would be difficult for us to extract and work with in any systematic way. If we have no way to predict where or how to look for beliefs in an LLM for each new sentence for which we want to

[23]In particular, it allows us to decode beliefs in areas in which we do not already know what the LLM believes or what the right answer really is, which are the domains we would use for training.

[24]This is thus also some evidence against **coherence**.