measure its belief, then while we might want to think of the LLM as a whole as somehow tracking truth, it doesn't seem useful to think of it as having a *representation* of truth. However, if we get better at understanding the internals of LLMs, such that we have a theory of where to look for different beliefs, we might be able to deploy a more sophisticated belief-measurement method that works across a wide range of domains, recovering a useful form of **uniformity**. This is a core way in which our requirements are practice-informed: they are sensitive to changes in available measurement techniques.

To be clear, it is not that **uniformity** as a condition weakens as we get more sophisticated belief measurement techniques. **Uniformity** *always* requires that our belief measurement technique works across a wide range of domains. Rather, what a high level of **uniformity** will end up looking like depends on the techniques available. In this sense the way in which **uniformity** is satisfied is *relative* to the methods we use, even though the **uniformity** condition itself always requires that the representation be consistent across domains. We've focused our discussion on contexts in which we have relatively simple probing methods that identify single directions in activation space at a particular layer. However, if we were to have more sophisticated techniques available, then a highly **uniform** belief representation might look different than what we've discussed so far.

For example, Marks and Tegmark find that, as sentences increase in Boolean complexity, probes work better at later layers (2023). Intuitively, if you are asked to evaluate the truth of "A and B", you might first figure out what you think of A, then B, and then apply the conjunction to make a judgement about the original statement. A precise version of this, that can guide our measurement techniques, might prove very useful for finding out what LLMs think about new sentences. If we had such a belief measurement technique and it worked across a wide range of domains then this would still exhibit *high* **uniformity**, even though the belief representations for different sentences lived at different layers.

Another example is recent work that develops techniques to extract interpretable features from neural networks. Building on theoretical work in neuroscience by Thorpe (1989) and mathematics by Donoho (2006), computer scientists have developed an approach to extracting features that uses sparse autoencoders (Elhage et al. (2022); Bricken et al. (2023); Cunningham et al. (2023); Templeton et al. (2024)). Though the details of the approach are too technical to describe here, the core upshot is that as we get better at finding interpretable features in LLMs our exact notion of **uniformity** might change to utilize the new techniques. Thus, while we have some weak evidence against a strong form of uniformity in current LLMs, relative to our best methods, as those methods improve our **uniformity** requirement will is more likely to be satisfied.

Continuing with our toy example, we illustrate a situation with high **uniformity** and one with low **uniformity** in fig. 5, relative to simple linear probes.

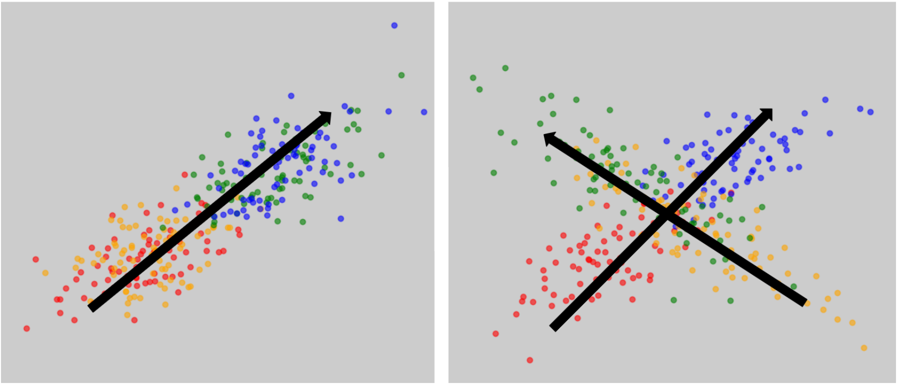

Figure 5. In the toy example, blue and red represent true and false claims (respectively) for sentences about one domain, for example, sentences about cities, while green and orange represent true and false claims (respectively) for sentences about a different domain, for example, sentences about plants. On the left, there is a consistent direction of truth in the model's representation space for both domains, suggesting high **uniformity**. On the right, the directions of truth are almost orthogonal, suggesting low **uniformity**.

5.4. **Use.** Our final criterion is **use**. **Use** requires that the LLM tend to use the identified representations in a role appropriate for belief to determine what probability distribution to output, or what text to produce if we are using the distribution to generate text.

If the LLM has beliefs, it uses those beliefs along with other information to figure out what to output. In essence, **use** ensures that beliefs play the belief-role in the LLM's master algorithm.

To check for **use**, we can look at how true beliefs lead to better performance and more skillful behavior. If an LLM shifts from false beliefs to true ones, it should generally improve in its tasks, whatever they may be. We illustrate two positive instances of **use** in fig. 6 using our running toy model.

The challenge with **use** is that the master high-level algorithm remains opaque. We don't have a good holistic understanding of what any LLM's master algorithm is, and there are many different ways the LLM could use its beliefs. To make a simple analogy, I can anticipate that a human will be more successful and skillful if she acquires more true beliefs. However, without knowing anything about what she wants or what her goals are, I can't predict much about what exactly she'll do. If the LLM is interested in truth-telling, for example, it will do a better job with more true beliefs. If the LLM is interested in deceiving, it will also do a better job with more true beliefs. But its behavior in those two cases will be quite different. The ultimate output is a function not only of its beliefs but of other things too.

Despite the algorithmic opacity, interventional techniques, such as ablating or modifying identified representations, can be helpful. For example, Marks and Tegmark (2023) attempt to identify directions of truth using mass mean probes. After determining candidate internal representations of truth and falsity, they ask a model to determine whether a statement is true or false. They might input "Determine whether the following statement is true or false: Paris is in France" and check the model's performance—the probability the model assigns to the tokens `True` and `False`—when its activations are unaltered and again when its activations are surgically altered by changing (supposedly) truth-encoding representations to false-encoding ones (or vice versa for other prompts). If, systematically, model performance degrades when the candidate representations are changed in this way, then we have evidence that the model was genuinely using these representations to encode truth and falsity.[25]

However, we do not yet have comprehensive tests for **use** across various domains and tasks due to the algorithmic opacity. Testing for **use** in beliefs is more subtle than testing for grammatical representations or board representations in Othello. For example, after prompting the LLM with the word `People` it will likely predict `are` with higher probability than `is`. By hand-editing the representation of `People` from plural to singular, the model should then predict `is` with higher probability. Checking **use** in this case is straightforward. But with beliefs, the process is more challenging.

5.4.1. *Harding's Criteria.* Harding (2023) recently addressed the question of when a probe has successfully identified an internal representation of a linguistic property of inputs to a language model, such as subjecthood or grammatical number. She proposes three criteria, which she dubs *information*, *use*, and *misrepresentation.* (We use italics here to distinguish her version of *use* from our **use**.)

- *Information* requires that the pattern of activations identified by the probe bears information about the property in question.
- *Misrepresentation* requires that, in principle, the activations could misrepresent the linguistic property.
- *Use* requires that the pattern of activations is actually used by the model in the right way to perform its task.

Our goal differs as we focus on representations of truth rather than linguistic properties. For example, our criteria of **coherence** and **uniformity** have no analog in the case of linguistic representation generally.

However, there is an interesting relationship between Harding's criteria and ours. For us, *information* corresponds directly with **accuracy**. Beliefs should carry information about the world by being true.

Our **use** also corresponds well with Harding's *use*, although there is an important difference in practice. For linguistic representation, as we saw, it

[25] For other attempts to check **use**, see (Campbell et al. 2023; Li et al. 2023).