Multiple Independent Semantic Axes in Gemma 3 270M

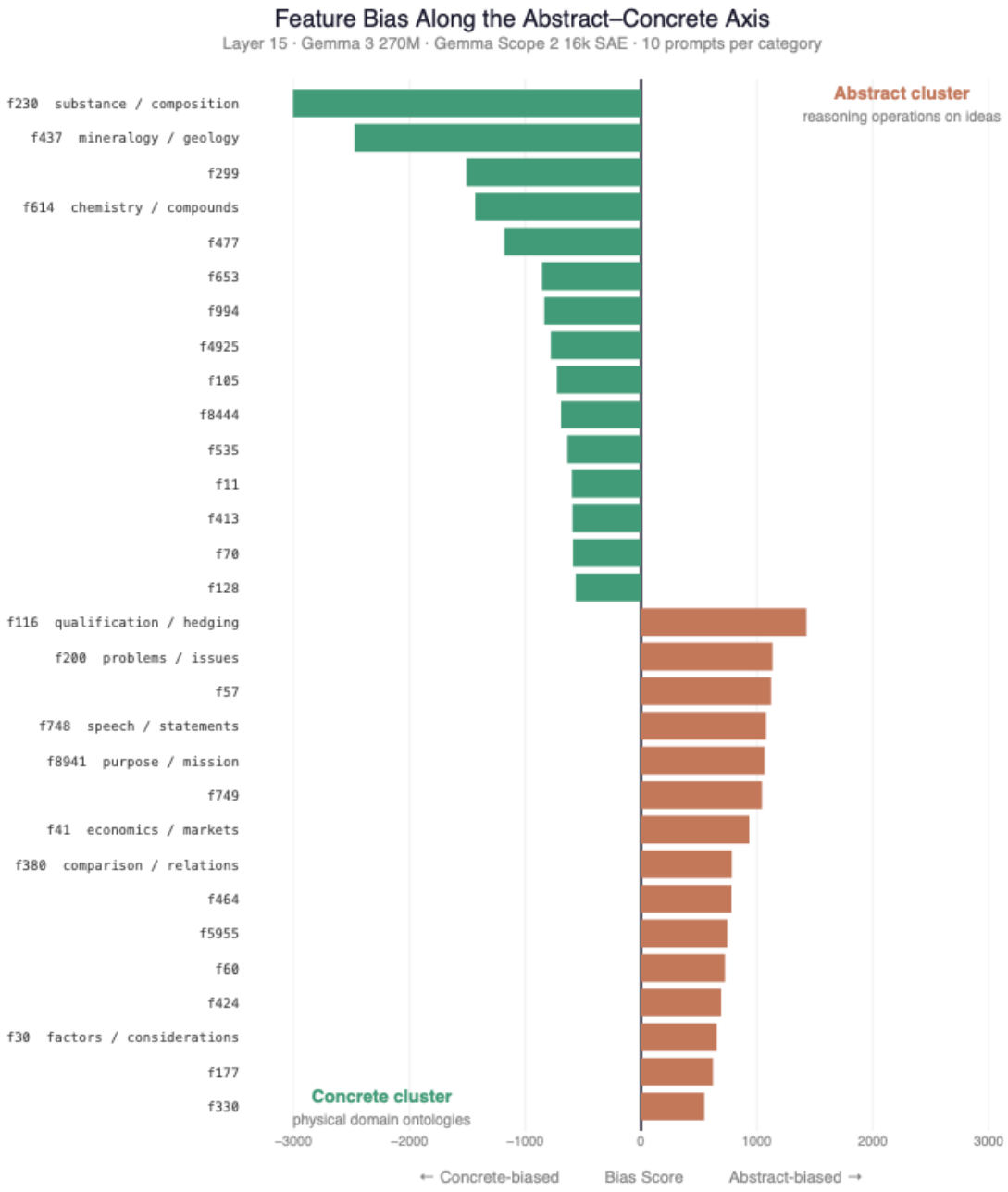
Feb 22, 2026

In my previous work, I found signs of an abstract-social vs concrete-physical axis in GPT-2's residual stream. This post builds on that work using SAEs on Gemma 3 270M. Here I attempt to move from the existence of this axis to trying to understand what makes it up at the feature level, and how it fits with other possible axes.

I chose the Gemma 3 270M model for this experiment, and used the Gemma Scope 2 16k SAEs for analysis. I made the decision to use SAEs rather than raw activations for this work in order to better understand the feature composition of the axes I'm analyzing rather than just the differences in activation. Raw activations showed the axis exists in GPT-2, but the representations are superposed — I couldn't see what the model was actually tracking. SAEs let me see those activations in terms of interpretable features, so I can go from understanding that the prompts are different to seeing the composition of each side.

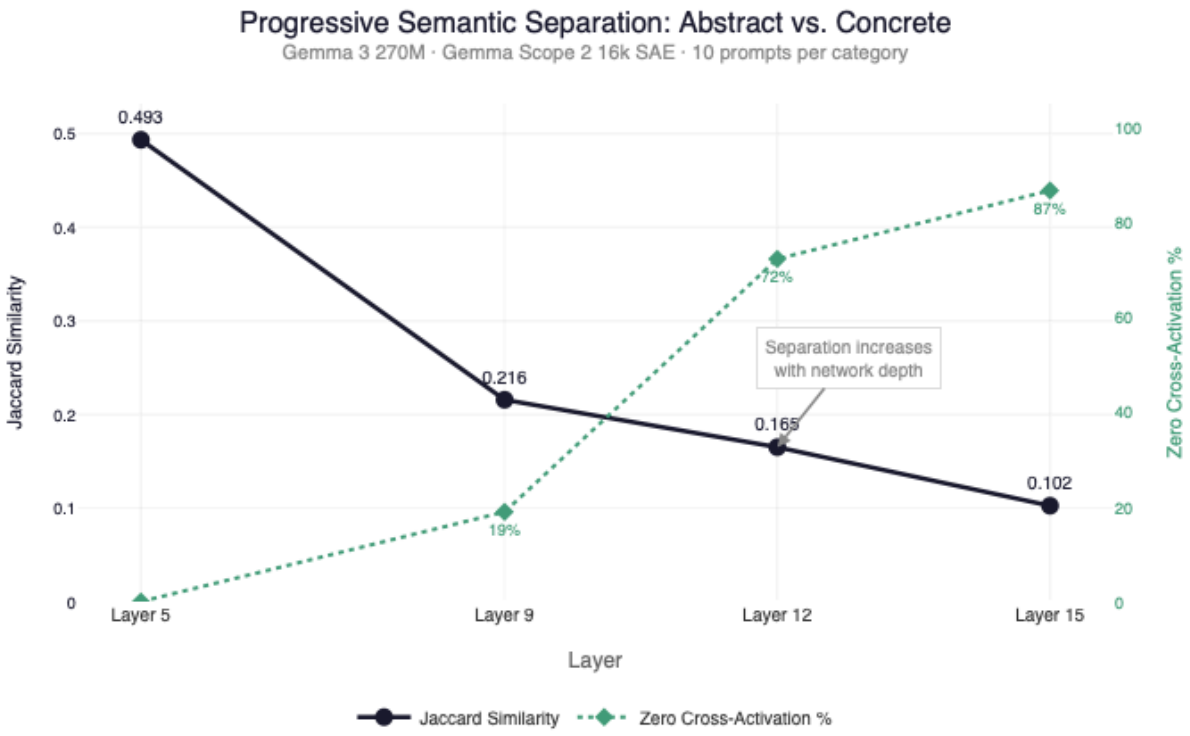
I also changed the structure of prompts from our original "[Thing] is" conception. I aimed to keep them balanced in terms of structure and length, but needed to add a bit more meat to them as the previous structure leaned too heavily into the 'is' driving the model's understanding as basically the start of defining the term. Effectively, "Immigration is" and "Limestone is" were getting too much similarity in activation as a result of structure rather than content. As an example of the restructuring, a new abstract prompt was: "A nuance of the debate over immigration policy is". And a new concrete prompt was "One of the molecules that makes up limestone is".

The first finding was that the abstract/concrete axis seems to be defined by clusters. It's not one 'abstract' feature and one 'concrete' feature. Instead there are a few different features firing that relate to each concept, as well as other features attending to the content of the prompt and yet more features reacting to syntax, sentence structure, etc. The abstract side looks like reasoning operations (f116 qualification, f200 problems); the concrete side seems to be physical domain ontologies (f230 composition, f437 geology).



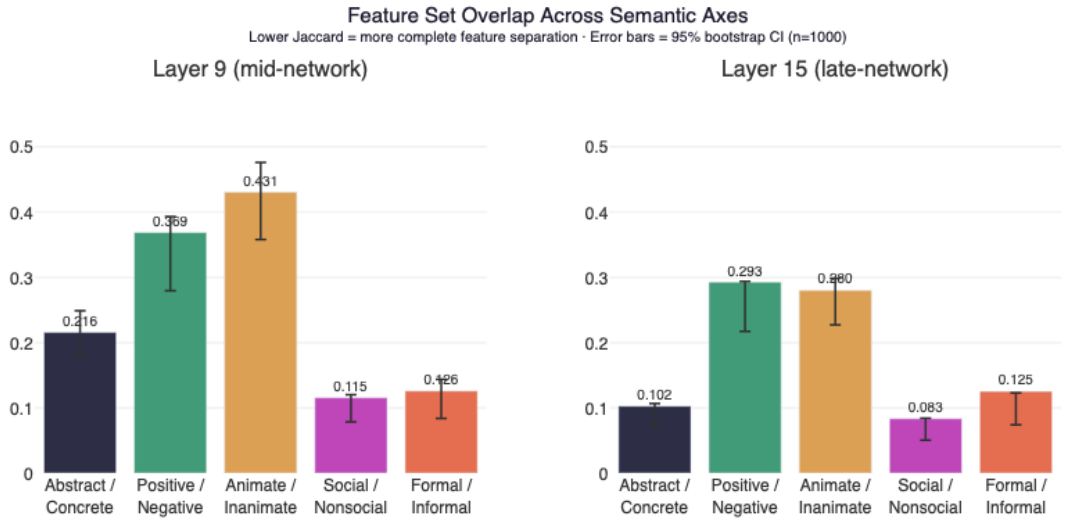
This separation is not present from the start, it became clear that this dichotomy is constructed through processing layers, as shown below. Looking at layer 5 the model is still treating abstract and concrete prompts similarly, with nearly half their features in common. By layer 9, it’s already mostly separated, and then it continues

refining through 12 and 15, as shown below.

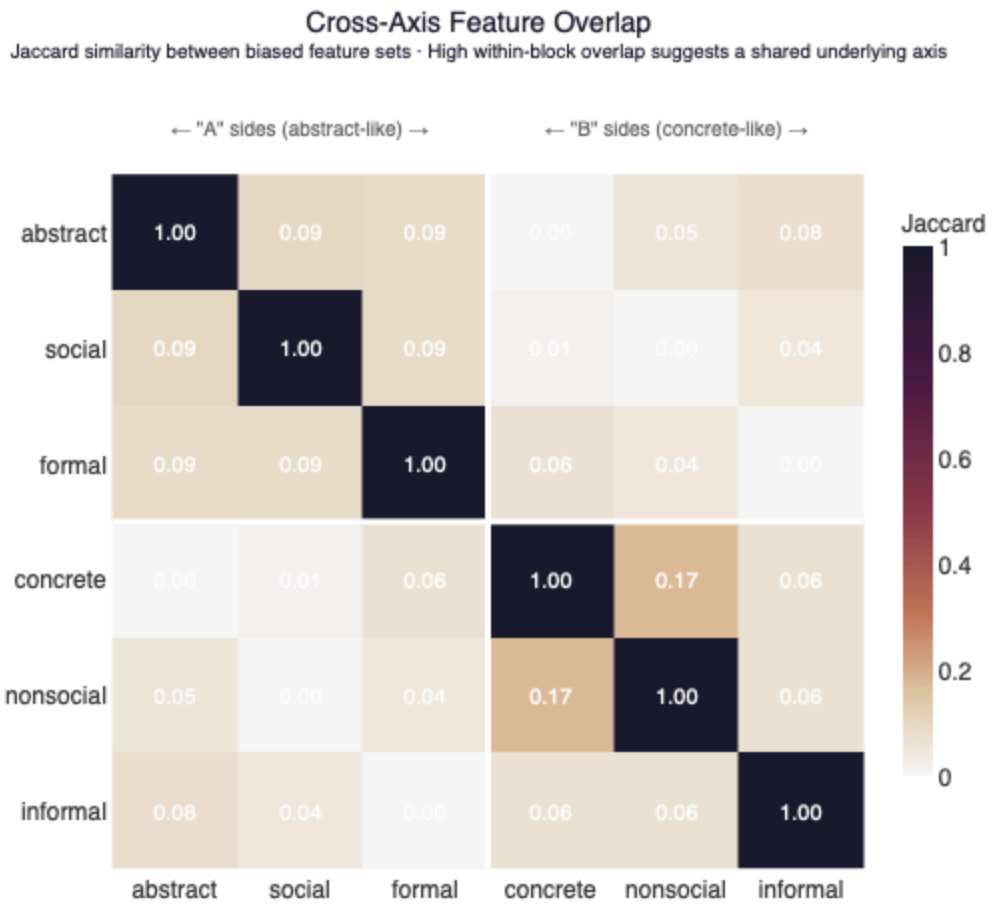


Having dug in to this extent on the purported abstract vs concrete axis, this got me thinking on if the model may use other axes like this to organize information. In other words, is there something special about this abstract vs concrete organization? Or would you get similar results picking any 2 opposite concepts and organizing prompts along each side of the spectrum to test it.

To examine this, I drew up prompts for some other potential axes or organization. By analyzing the feature set overlap along these semantic axes, I found that abstract/concrete does seem to be a privileged axis, but not the only one. Of the 5 I came up with, social/nonsocial and formal/informal significantly beat positive/negative and animate/inanimate. Specifically, abstract/concrete and social/nonsocial had Jaccard similarities of 0.102 and 0.083, with bootstrap CIs well below the 0.28–0.29 range of positive/negative and animate/inanimate.



In my mind, one possible reason for this was that these axes overlap. For example, is formal/informal just another expression of abstract/concrete? Do the same features fire to express both of these conceptual divides? Surprisingly I found that no, these axes do not use the same features. There are separate representational features for each supposed axis, and very little overlap. I expected overlap and instead found independence, which changed my interpretation from ‘one axis represented multiple ways’ to ‘independent dimensions.’ This is shown in the cross-axis overlap matrix below.



These were my new findings from this bit of analysis. In thinking about these results I have a couple of questions to pursue: First, why does the model maintain independent axes for things that seem intuitively related?

I want to continue pursuing this question especially since I want to clarify if this strong separation is an artifact of my prompt design (large lexical gaps), or if it's genuine representational structure that I'm seeing here.

As with my previous work, there are clear limitations in the sample though I doubled it from last time (n=10 per category), and the fact that this work was only conducted on a single model. I also do not yet have causal validation, as the work here is descriptive but not

mechanistic. There is also potential to deepen the findings by expanding the SAE width from 16k to larger.

For next steps, I plan to try to identify causal validation via feature ablation, as well as testing the cross-architecture replication via Pythia for emergence timing. I would also like to replicate the experiment with larger prompt sets, testing if the axes persist with prompts designed to minimize lexical confounds.