REPLICATION

Jeffrey Wu and Dan Mossing are members of the Alignment team at OpenAI.

We are very excited about these toy models of polysemanticity. This work sits at a rare intersection of being plausibly very important for training more interpretable models and being very simple and elegant. The results have been surprisingly easy to replicate -- we have reproduced (with very little fuss) plots similar to those in the Demonstrating Superposition – Basic Results, Geometry – Feature Dimensionality, and Learning Dynamics – Discrete "Energy Level" Jumps sections.

Original Authors' Response: We really appreciate this replication of our basic results. Some of our findings were quite surprising to us, and this gives us more confidence that they aren't the result of an idiosyncratic quirk or bug in our implementations.