Intricacies of Feature Geometry in Large Language Models

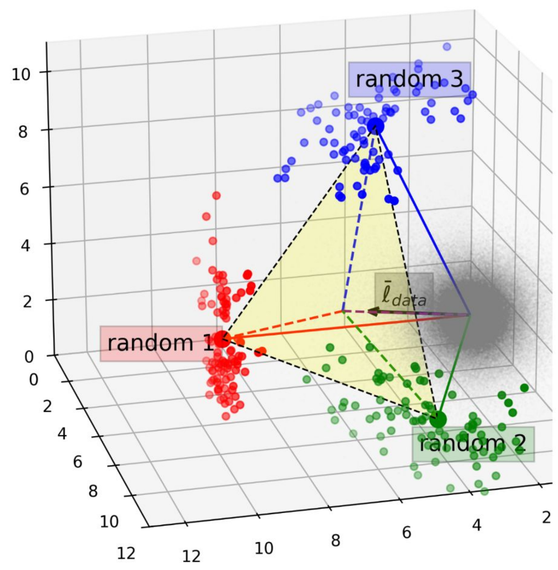
Satvik Golechha · Lucius Bushnaq · Euan Ong · Neeraj Kayal · Nandi Schoots

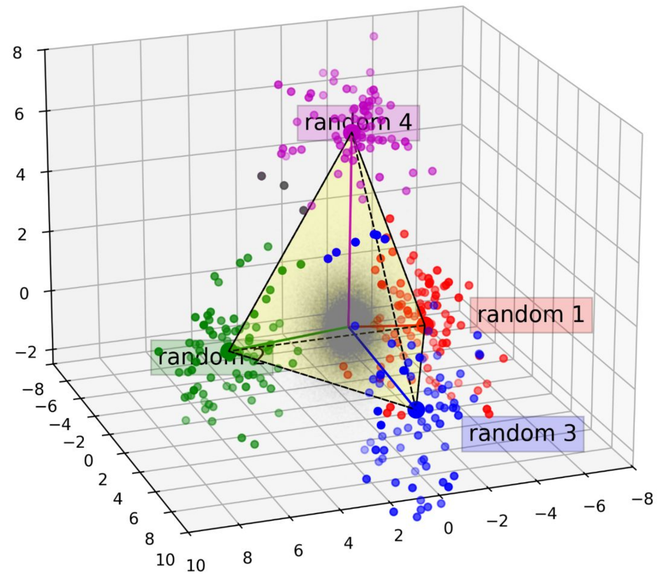
Studying the geometry of a language model's embedding space is an important and challenging task because of the various ways concepts can be represented, extracted, and used.

We want a framework that unifies both measurement (of how well a latent explains a concept) and causal intervention (how well it can be used to control/steer the model).

They show categorical features to form polytopes in the representation space.

We run ablations to show that completely random concepts follow the exact same geometry.





An attempt to solve it was this **ICLR 2025 (oral!)** paper. It also won the best paper award at an ICML workshop!

We undermine their conclusions by showing that similar results can be obtained from random data!

The Geometry of Categorical and Hierarchical Concepts in Large Language Models

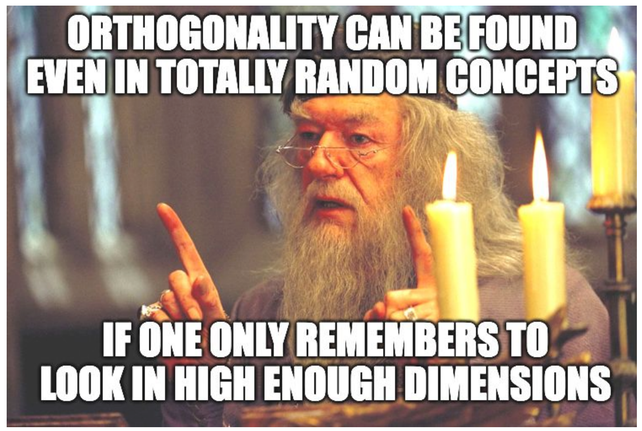
Kiho Park · Yo Joong Choe · Yibo Jiang · Victor Veitch

Hall 3 + Hall 2B #525

[Abstract]
Thu 24 Apr 7 p.m. PDT — 9:30 p.m. PDT (Bookmark)
Oral presentation: Oral Session 4C
Fri 25 Apr 12:30 a.m. PDT — 2 a.m. PDT (Bookmark)

But why? Glad you asked!

They do a **whitening** transformation on a high-dimensional space, making almost everything almost orthogonal. Turns out:



For a complete theoretical proof along with some more ablations, please check out our ICLR blog:

Also, they show that hierarchical features exhibit orthogonality in a specific way.

And again, we show that completely random – and even semantically opposite – concepts exhibit this orthogonality.