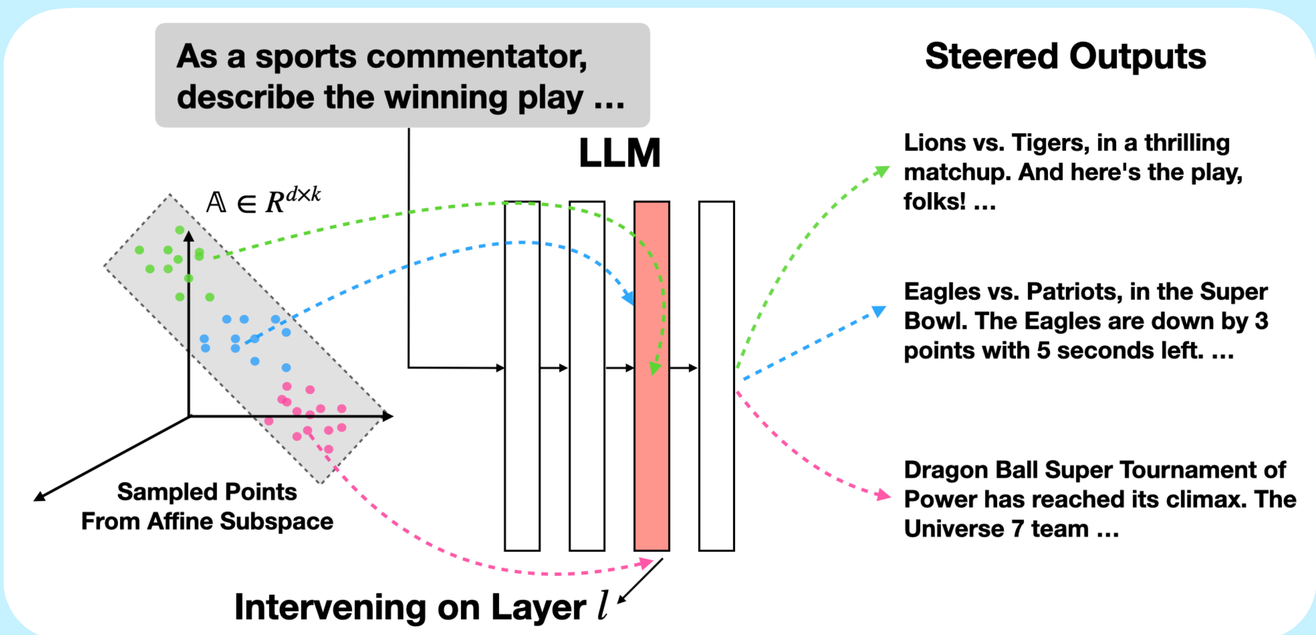
Localizing and Evaluating Multinomial Concepts with Affine Subspaces

Divya Appapogu[1], Freya Behrens[1], Yonatan Belinkov[2], Aaron Mueller[1]

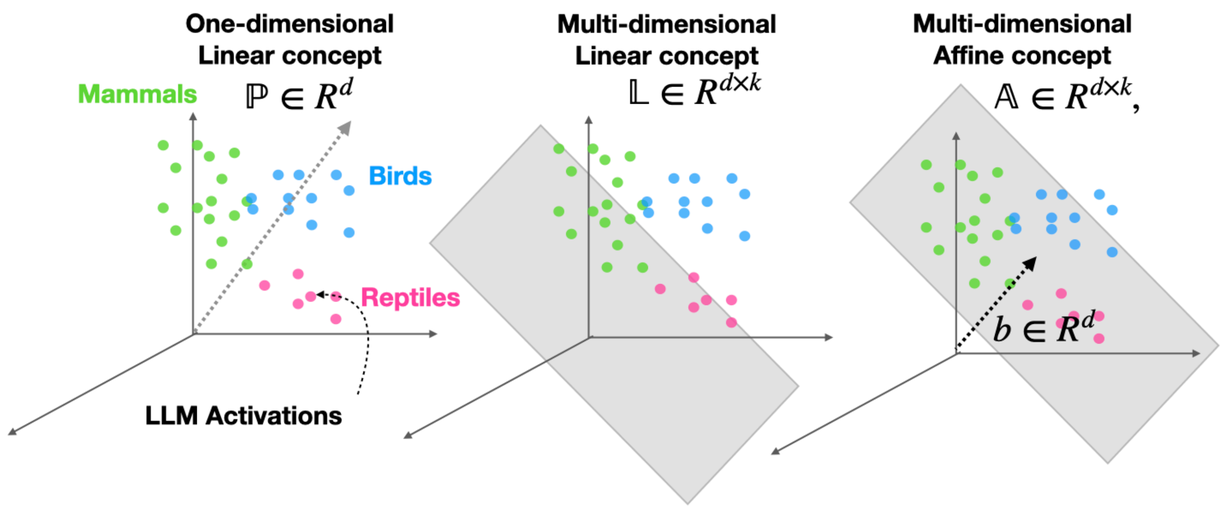
[1] Boston University [2] Kempner Institute at Harvard University

TL;DR

Multidimensional affine concept representations enable richer steering than one-dimensional representations: sampling within the concept space covers the training distribution while generating coherent, diverse novel instances.



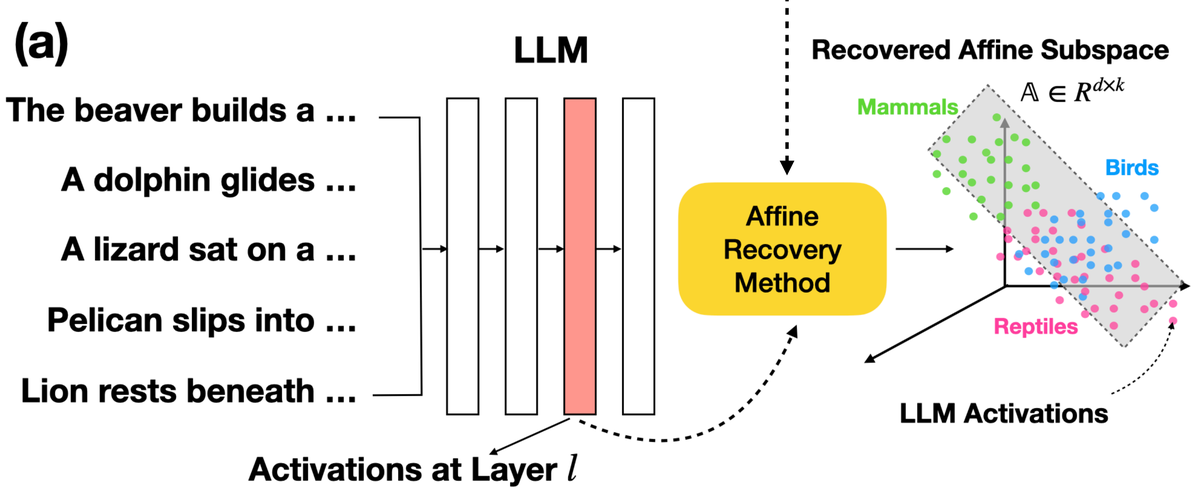
Motivation



The linearity of a concept is an elegant but strong assumption. What do we gain in a multidimensional setting?

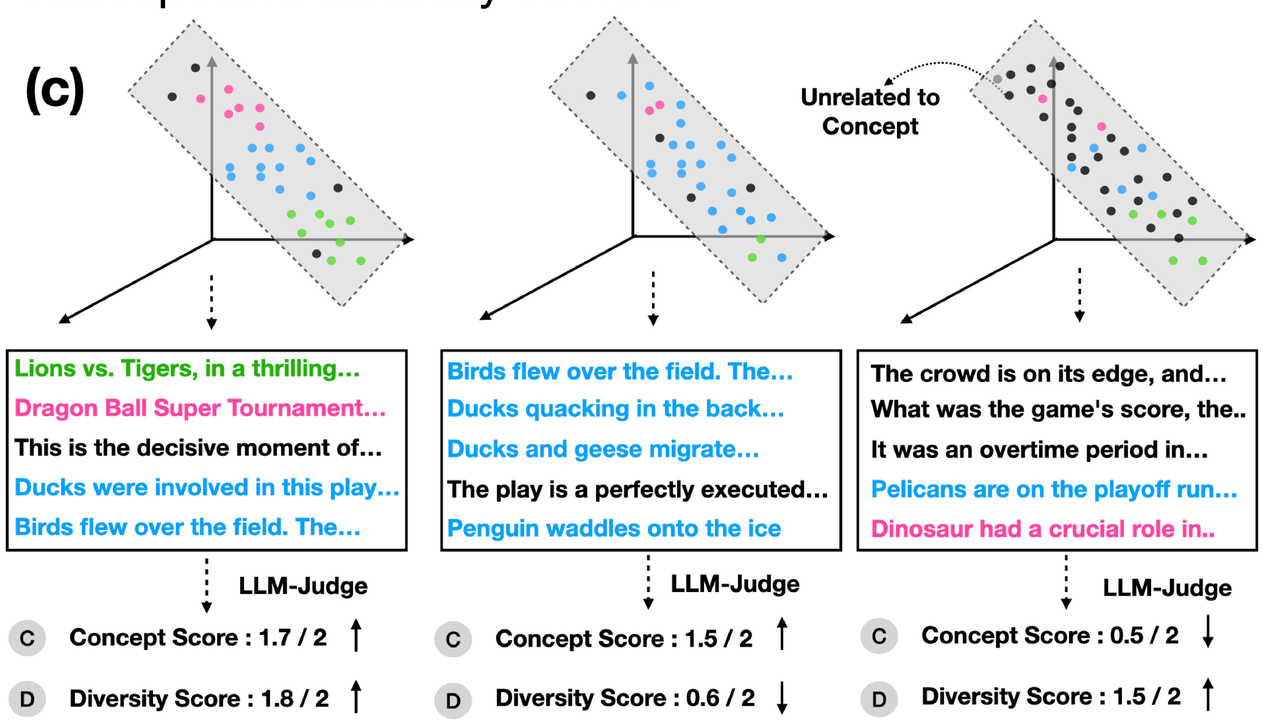
Methods to Recover Affine Subspace

PCA | Diff Mean | Probe | Schatten Probe | FA



Concept and Diversity Measures

We use LLM Judge to evaluate steered outputs for concept and diversity scores.



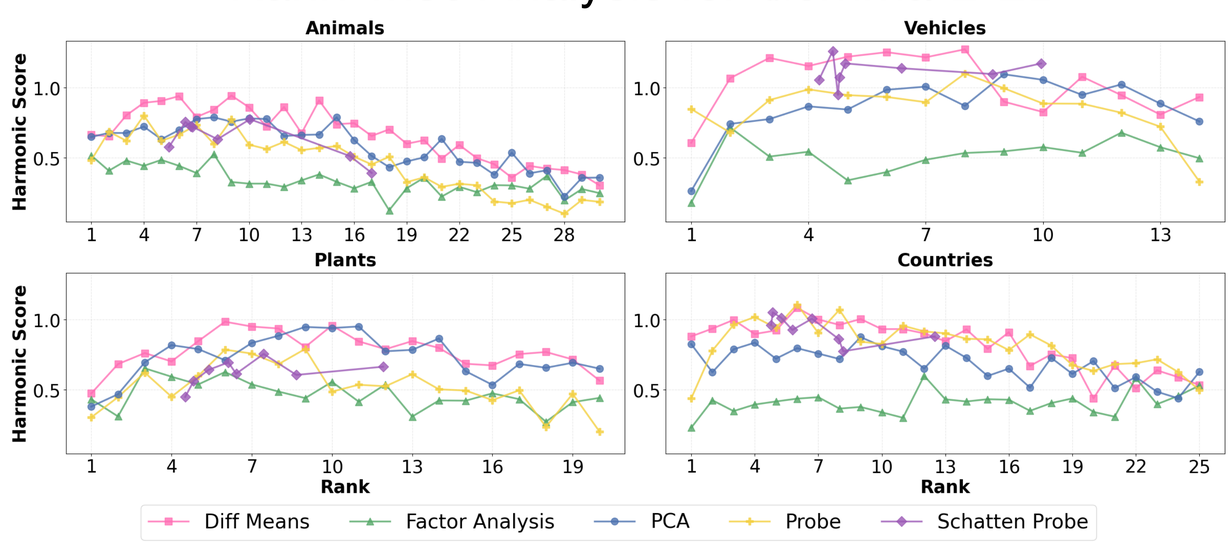
Experiments

For Llama, Factor Analysis and schatten probe perform best. For Gemma, DiffMean and both probes perform well. All methods outperform prompting.

Model	Method	ANIMALS	VEHICLES	PLANTS	COUNTRIES
Llama-3.1-8B	Prompting	0.630	0.366	0.463	0.958
	DiffMean	0.913	1.153	0.790	1.432
	FA	1.023	**1.259**	**0.852**	**1.520**
	PCA	0.936	1.048	0.808	1.380
	Probe	1.028	0.955	0.725	1.369
	Schatten Probe	**1.076**	1.114	0.768	1.424
Gemma-2-2B	Prompting	0.480	0.006	0.192	0.642
	DiffMean	**0.944**	**1.273**	**0.984**	1.084
	FA	0.527	0.716	0.653	0.599
	PCA	0.789	1.095	0.950	0.875
	Probe	0.802	1.101	0.791	**1.108**
	Schatten Probe	0.776	1.261	0.755	1.050

	VEHICLES		
Method	Concept	Diversity	Harmonic
Prompting	0.040	0.200	0.006
DiffMean	**0.973**	**1.840**	**1.273**
FA	0.467	1.535	0.716
PCA	0.800	1.737	1.095
Probe	0.800	1.763	1.101
Schatten Probe	**0.973**	1.791	1.261

Rank-wise Analysis for Gemma-2-2B



Takeaways

- Sampling can be used as evaluation metric for concepts without known ground-truth geometry the common case in LLMs.
- Diverse sampling within a learned subspace enables targeted, controllable interventions on LLM outputs.