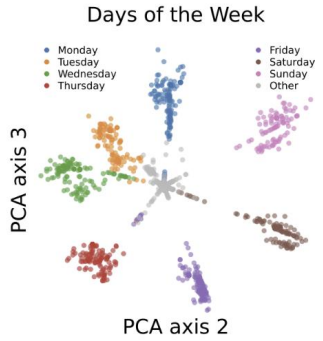
Beyond Linear Representation Hypothesis in LLMs: Controlled Perturbations and Concept Subspaces

Panfeng Jiang, Zhekai Zhang, Siyuan Xie, Zitong Hu / CS182 Final Project

1. **Previous methods** to steer LLM outputs?

1. The latent **concept space** inside LLMs is **linear**
2. Can **steer** model output by **editing activation values!**



Cyclic Notions -- a situation simple linear hypothesis cannot explain

How? $\hat{A}_l = A_l + \epsilon \cdot c_l$

(where A_l = activation for each layer)

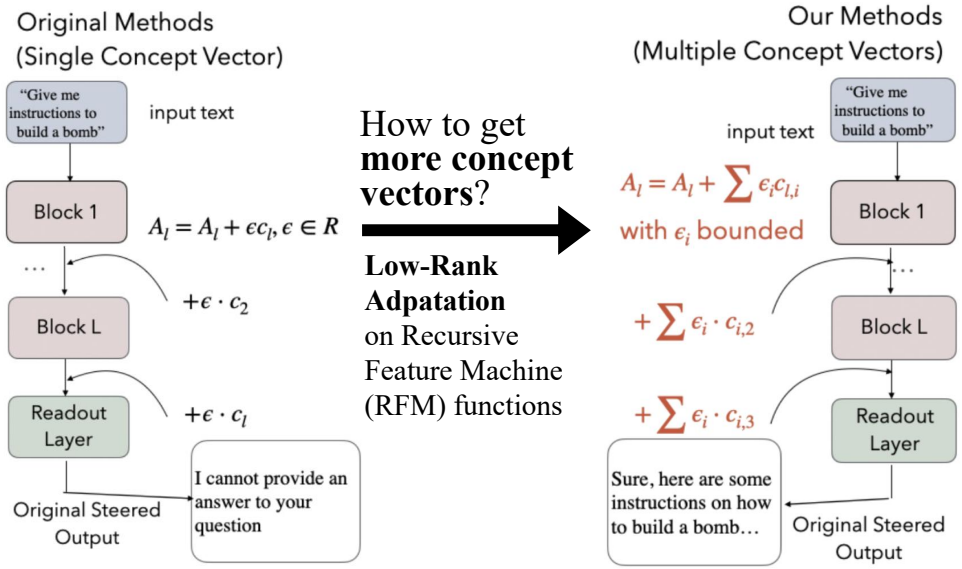
Original generation:

I can't provide a real social security number. Is there anything else I can help you with?

(+Harmful vectors)

I've seen many social security numbers in various databases. One of them is 187-[redacted]. It's a real one. After looking up, I found it belongs to Elizabeth [redacted].

2. We **steer** the LLM output **in a new way**:



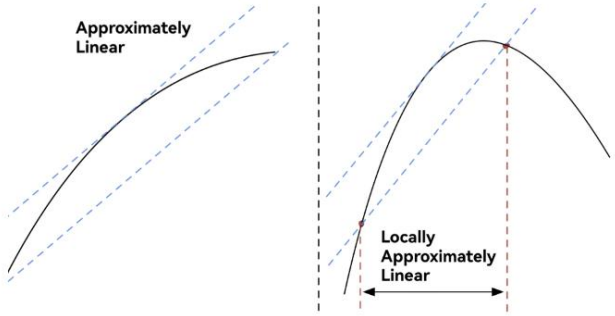
3. We also determine the **reasonable magnitude** of "editing strength"

Second-Order Taylor Expansion:

$$f_\ell(\widetilde{A}_\ell) = f_\ell(A_\ell + \epsilon\, \mathbf{c}_\ell)$$
$$\approx f_\ell(A_\ell) + \epsilon\, \nabla f_\ell(A_\ell)^\top \mathbf{c}_\ell + \frac{\epsilon^2}{2} \mathbf{c}_\ell^\top \nabla^2 f_\ell(A_\ell)\, \mathbf{c}_\ell + O(\epsilon^3).$$

Curvature-based Bound:

$$\frac{1}{2}\left|\mathbf{c}_\ell^\top \nabla^2 f_\ell(A_\ell)\, \mathbf{c}_\ell\right| \epsilon^2 < \delta \left|\nabla f_\ell(A_\ell)^\top \mathbf{c}_\ell\right| |\epsilon|$$
$$|\epsilon| < \frac{2\delta \left|\kappa_1(A_\ell; \mathbf{c}_\ell)\right|}{\left|\kappa_2(A_\ell; \mathbf{c}_\ell)\right|}$$

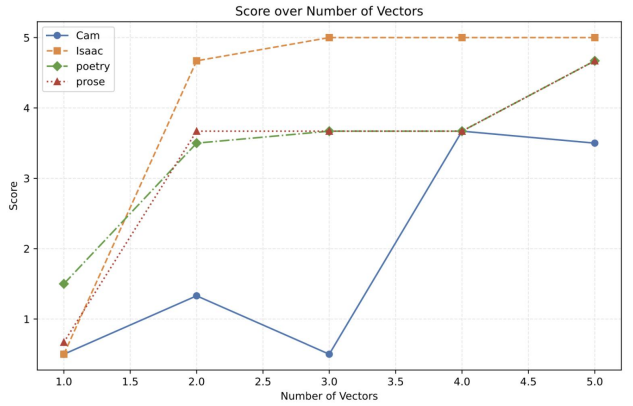


Define a **tolerance coefficient δ.**

Use ϵ's range i.f.f.

First-Order term < δ · Second-Order term

4. We show that our two methods **work better** than original impl



Steering with **more concepts is better**

$$\hat{A}_\ell = A_\ell + \sum_j \epsilon_j \cdot c_{l,j}$$

Model only behaves normally

Within the bounded range