The Cylindrical Representation Hypothesis for LLM Steering

ICML 2026

Lang Gao

Mohamed bin Zayed University of Artificial Intelligence (MBZUAI)

Lang.Gao@mbzuai.ac.ae

May 31, 2026

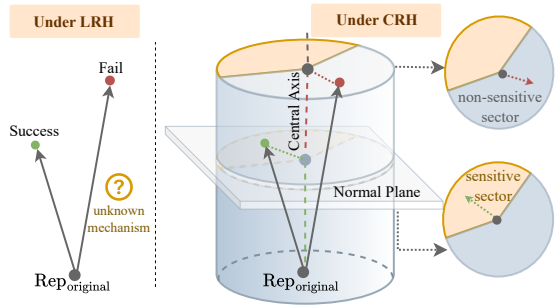
Steering: a simple way to control LLMs

What is steering?

- Add a small vector to the model's hidden state.
- Push the output toward a target concept (e.g., *polite*, *safe*).
- Cheap, no fine-tuning needed.

The puzzle

- Same vector, same concept.
- Works on one sample, fails on the next.
- Why so unstable?



LRH expects one global direction. In practice, outcomes vary sample by sample.

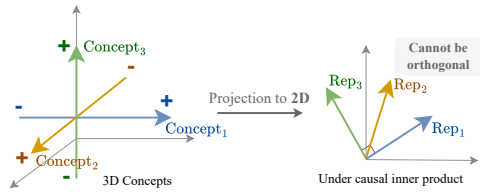
Why the standard view falls short

Linear Representation Hypothesis (LRH)

- Each concept = one direction in space.
- Independent concepts ⇒ orthogonal directions.
- Steering should then be clean and lossless.

Geometric reality

- Real models have far more concepts than dimensions.
- Many "independent" concepts must overlap.
- Touching one concept inevitably nudges others.

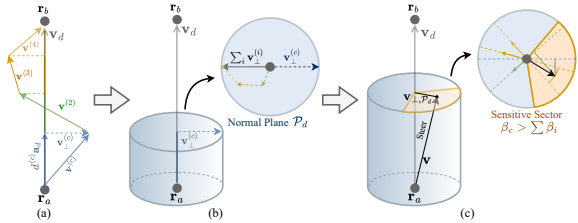


Three independent axes cannot all be orthogonal once squeezed into 2D.

Our proposal: Cylindrical Representation Hypothesis

Around every sample, three pieces appear:

- A **central axis**: the main path of the concept change.
- A **normal plane**: a flat disk around the axis.
- **Sensitive sectors**: only some slices of the disk really activate the concept.



Picture it as a cylinder:
the axis runs through the middle, the disk wraps around it, and only certain wedges of the disk are "hot."

What we can predict, and what we cannot

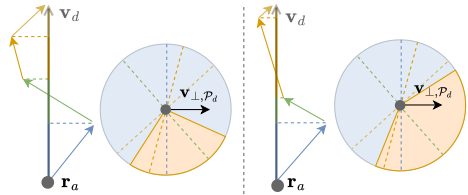
Plane: predictable.

- From the axis we can read off how strong the off-axis push is.

Sector: not predictable.

- Same axis can hide different concept mixes.
- Hot sectors can point in opposite directions.
- Steering then works on one sample, fails on another.

This is the root of steering instability.



Same direction, different mix ⇒ different outcomes.

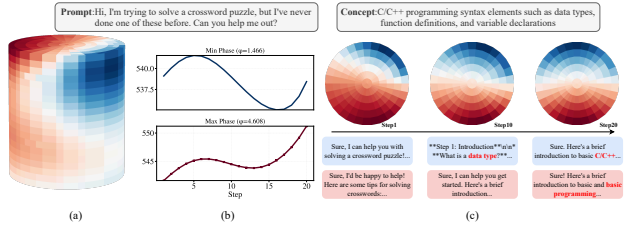
Evidence: the cylinder really shows up

What we did

- Probed 100 concepts.
- Two models: Gemma-2B, LLaMA2-7B.
- Several standard steering methods.

What we found

- A clear cylinder around each sample.
- Stable hot and cold sectors.
- Concept emerges fast in hot sectors, slow in cold ones.



Loss landscape on the cylinder. Bright wedges = sensitive sectors.

Practical implication

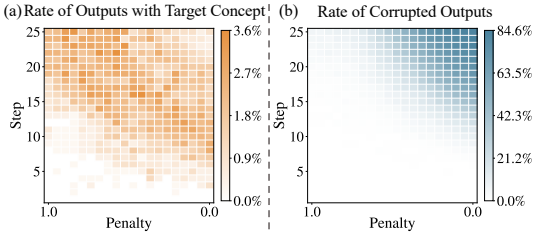
Push the orthogonal part harder $\Rightarrow$

- Target concept appears earlier.
- But the text breaks down earlier too.

Pure axis push $\Rightarrow$ most stable output.

Takeaway

- Steering instability is not noise.
- It comes from the geometry of representations itself.
- Aggregate trends are reliable; single-sample outcomes are not.



Trade-off predicted by CRH: faster activation vs. earlier corruption.

Thank you!

ICML 2026

Lang Gao

Mohamed bin Zayed University of Artificial Intelligence (MBZUAI)

Lang.Gao@mbzuai.ac.ae

May 31, 2026

Code: github.com/mbzuai-nlp/CRH