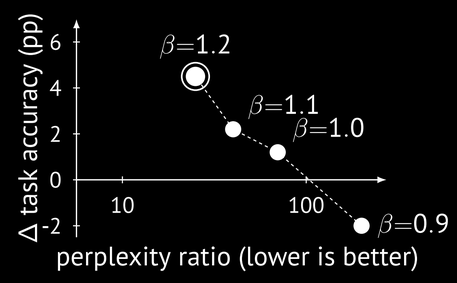
We decompose activation steering into *angular* and *norm* components:

1. **Concepts live primarily in the angular component,**
2. **Norm important for generation quality,**
3. **Slight change in norm might benefit a steering perfomance.**



A Geometric Account of Activation Steering through Angle–Norm Decomposition

Georgii Aparin [1] **Tatiana Gaintseva** [2]

[1]Huawei Noah's Ark Lab [2]Queen Mary University of London

1 TL;DR

- Activation steering is not one additive knob—it splits into an **angular** concept shift and a **radial** norm scaling.
- Across **7 LLMs** and **4 concepts**, semantic control is angular; stability is radial.
- A small norm rescale (β=1.2) yields $\sim$1.8$\times$ lower perplexity with $\lesssim$ 2.5 pp task change.

2 Background & Question

Linear activation steering is a *widely used, training-free* approach for controlling LLM behaviour. Standard methods add a steering vector with a scalar strength,

$$y \;=\; x + \alpha\, s,$$

which obscures the geometry: a single α changes *both* the token's **alignment** with s and its **norm** $\|x\|$, in a token-dependent way.

Angular / spherical methods instead *rotate* x toward s while preserving $\|x\|$, on the hypothesis that concepts are angular and norm preservation maintains generation quality.

Their underlying assumptions remain insufficiently examined:

- Are concepts really encoded in the angular component?
- Is strict norm preservation the right stability constraint?
- Can every existing method fit one geometric framework?

3 Angle–Norm Decomposition

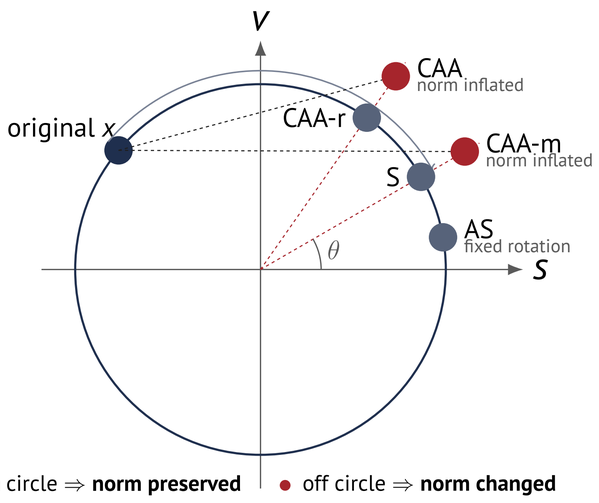
For each residual hidden state $x \in \mathbb{R}^d$ and unit direction s:

$$r = \|x\|, \quad u = \tfrac{x}{r}, \quad c = \langle u, s\rangle, \quad v = \tfrac{u - c\,s}{\|u - c\,s\|}.$$

Any steered vector in $\mathrm{span}(s, v)$ is

$$y \;=\; \underbrace{\rho}_{\text{radius}} \Big(\underbrace{\gamma\, s + \sqrt{1-\gamma^2}\, v}_{\text{unit direction, concept score } \gamma} \Big).$$

Geometry in $\mathrm{span}(s, v)$: the sphere $\|y\|$=r is a circle; each ray from the origin is a **line of fixed concept score** $\gamma = \cos\theta$.



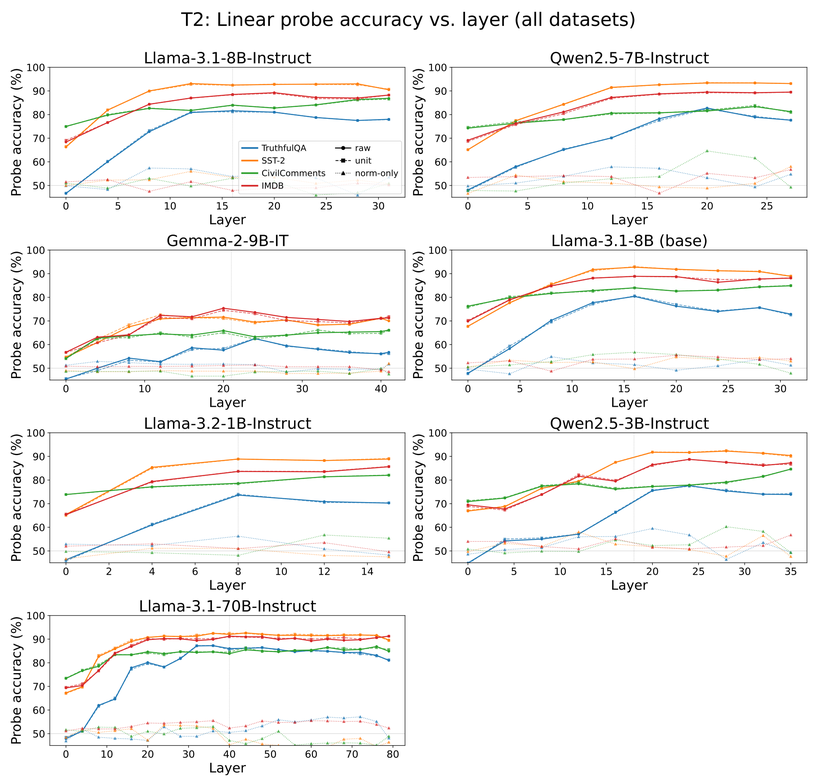
● on circle ⇒ **norm preserved** ● off circle ⇒ **norm changed**

Six methods differ only in these two axes:

Method	Norm preserved	Per-token γ
CAA	–	–
CAA-r	✓	–
CAA-m	–	✓
S	✓	✓
AS	✓	–
SN (**ours**)	β-scaled	✓

4 Concepts are (mostly) angular

Linear probes on raw h, unit-normalised $h/\|h\|$, and norm-only $\|h\|$ across 7 models $\times$ 4 datasets:

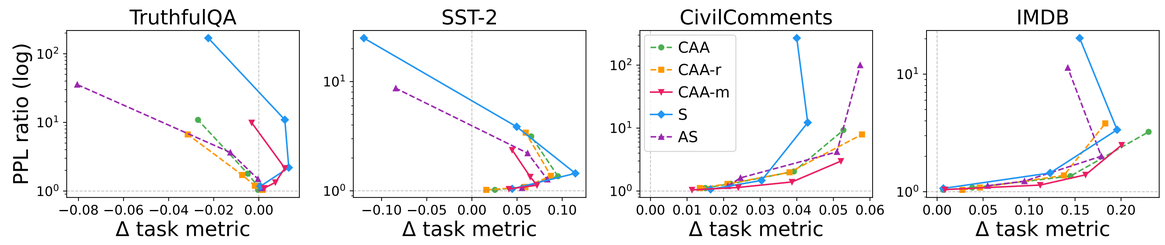


- Normalised probes **track raw probes** within $\lesssim$ 1 pp.
- Norm-only probes stay **near chance** ($\sim$ 50–55%).

⇒ Semantic content lives in **direction**.

5 Downstream perfomance

We compare all five methods (CAA, CAA-r, CAA-m, S, AS) on the same Pareto frontier: downstream task gain vs. WikiText-103 perplexity ratio (lower-right is better).



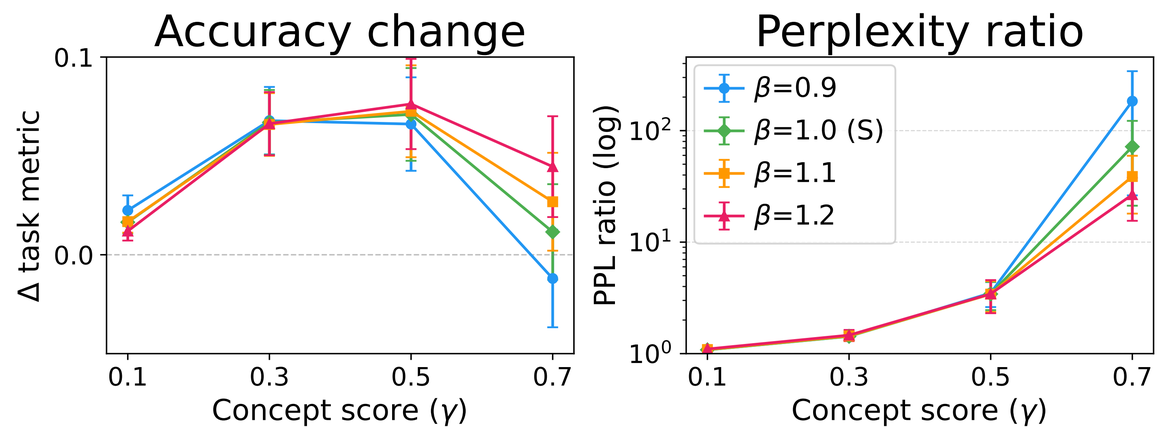
- **S vs. CAA-m** — matched angular target γ, radius differs. At high γ, S's strict norm preservation costs **large perplexity penalties**, while CAA-m's mild norm inflation preserves fluency and MMLU accuracy.
- **CAA-r vs. AS** — both preserve norm exactly, yet sit at *different* Pareto positions: CAA-r applies a fixed additive step then renormalises; AS applies a fixed spherical rotation. Their per-token score distributions differ, and so does their downstream behaviour.

⇒ **Norm preservation alone does not explain stability.** The design space is **two-dimensional**: angle sets the semantic effect, radius sets generation stability.

6 Norm as an explicit lever (SN)

We put a multiplicative norm scale β on top of Spherical steering:

$$y \;=\; \beta\, r \left(\gamma\, s + \sqrt{1-\gamma^2}\, v \right)$$



- β leaves the angular target unchanged.
- Sweeping $\beta \in \{0.9, 1.0, 1.1, 1.2\}$: task metric moves $\lesssim$ 2.5 pp.
- PPL at γ=0.7 improves $\sim$1.8$\times$ moving $\beta:\ 1.0 \to 1.2$; β=1.2 wins on PPL in **100% of folds**.

Why? A fixed radius forces the concept edit to crowd out contextual features; a small radial slack restores fluency.

7 Conclusion

We reframe activation steering as a **two-parameter geometric intervention** over *angle* and *radius*: angle controls the intended concept, radius controls intervention stability.

- Additive steering *entangles* angle and norm token-dependently, so a single α is hard to interpret.
- Across 7 models $\times$ 4 datasets, evaluated concepts are represented **primarily in activation direction**—normalised probes track raw probes, norm-only stays near chance.
- At high γ, strict norm preservation becomes unstable; **modest norm increases reduce degradation** without materially changing the semantic effect.

aparingm@gmail.com t.gaintseva@qmul.ac.uk