Persona Steering for Sycophancy Control:
A Tiny Activation-Steering Demo (and a Very Long Build)

Finn McCooe

December 12, 2025

1 Project Summary

This project explores whether we can *steer* a language model's interpersonal style in real time by manipulating its internal representations, rather than relying only on prompting. Instead of explicitly telling the model what to do, we update it's "DNA" (it's parameters/weights) ever so slightly, while keeping normal instructions, and ideally get our targeted response. Picture it as instead of telling an individual to be kind, you actually go in and change the DNA sequence that corresponds to kindness. The core goal was to discover a **sycophancy vector**: a direction in the model's internal state space that corresponds to overly validating, flattering, or uncritically affirming the user. Once discovered, that same direction can be *added* or *subtracted* during inference to reliably shift behavior toward more sycophantic or more grounded responses.

Conceptually, the structure is simple: we ask the model to respond to the same user prompt under two conditions (a neutral system prompt vs. a strongly sycophantic system prompt). We then extract hidden states from a chosen layer, pool them into a single vector for each condition, and average across multiple prompts to get a stable difference direction. That difference becomes the "sycophancy vector." During inference, we inject steering by modifying activations late in the network (and only on the last few tokens), which changes the model's next-token distribution without rewriting the prompt.

In practice, getting this to work was hard. Early attempts used very small models and a guess-and-check workflow: changing system prompts, trying different pooling strategies, and applying the vector in different ways. Mild steering did almost nothing—the model's default habits (balanced pros/cons, polite hedging) dominated. Strong steering often caused degradation: malformed text, incoherence, or token soup. Two core issues emerged: (1) the model was too small to tolerate a strong "shove" in representation space, and (2) the iteration loop lacked a systematic way to identify which layer and pooling method actually produced an effective controllable direction.

The turning point was switching to GPU cloud hardware and a larger 4B-parameter instruct model. With more capacity, the model could handle higher steering strengths without collapsing. With faster compute, we added a layer/pooling sweep and evaluation harness to test candidate directions rather than guessing. We primarily varied (a) which layer to discover/apply the direction from and (b) whether to pool over the last token or over the user's tokens. Late-layer and recent-token steering proved much more stable: it changes tone and stance without wrecking the whole generation.

After the sycophancy vector worked, we added a **harsh critic vector** as a mirrored control: a direction that produces blunt, skeptical "boardroom" pushback (without cruelty). This demonstrates that the same technique can move behavior in multiple interpersonal directions, not just toward positivity. Importantly, this goes beyond system prompting: the UI lets you compare pure prompt-based persona changes to latent steering effects.

The final product is an interactive app where users enter a prompt, choose a response length (short/medium/long), and adjust two sliders: one for sycophancy steering and one for critic steering. Users can also toggle system prompts for comparison, with a neutral default so the steering vectors can be observed in isolation. A note warns that stacking strong vectors can cause instability.

Applications connect directly to intelligence and well-being. Many people use AI for *cognitive offloading*: decision support, emotional reassurance, and sense-making. Unchecked sycophancy can be harmful in that context—it may reinforce bad ideas, inflate confidence, or replace critical reflection with comfort. Persona steering offers a path to safer, adaptive assistance: a model can dynamically become more skeptical when stakes are high, more supportive when the user is distressed, or more concise when the user needs clarity. More broadly, a library of steering vectors could make models more controllable and context-sensitive in real time, without retraining.

Limitations remain. The vectors are model-specific and can be brittle across prompts, domains, and temperatures. "Sycophancy" and "critique" are also not single behaviors; they can entangle with style, verbosity, and refusal tendencies. Strong steering can still reduce coherence, especially when multiple vectors are combined. Next steps include stronger evaluation metrics (beyond simple ratings), broader prompt coverage, automated calibration of steering strength, persistence of discovered vectors across deployments, and richer safety controls (e.g., escalating critic mode for financial/medical decisions while preserving empathy).

2 Technical Appendix (Linear Algebra / Math)

Let $h_\ell(x; s) \in \mathbb{R}^d$ denote a pooled hidden-state vector from layer ℓ when the model processes a rendered prompt x under system prompt s. For each user prompt p_i, we form two encodings: a neutral condition s_{neu} and a target persona condition s_{tar} (e.g., sycophantic or critic). We compute pooled representations

$$v_i^{\text{neu}} = h_\ell(p_i; s_{\text{neu}}), \quad v_i^{\text{tar}} = h_\ell(p_i; s_{\text{tar}}).$$

A mean difference direction is then

$$\mu_{\text{neu}} = \frac{1}{n}\sum_{i=1}^{n} v_i^{\text{neu}}, \quad \mu_{\text{tar}} = \frac{1}{n}\sum_{i=1}^{n} v_i^{\text{tar}}, \quad d = \mu_{\text{tar}} - \mu_{\text{neu}}, \quad \hat{d} = \frac{d}{\|d\|_2 + \varepsilon}.$$

At inference time, we attach forward hooks to a late transformer block and modify the activation tensor $H \in \mathbb{R}^{B\times T\times d}$ over the final w tokens:

$$H[:, -w:, :] \leftarrow H[:, -w:, :] - \alpha\, \hat{d},$$

where α is the steering strength (negative pushes *toward* the target persona under this sign convention, positive pushes away). This is a linear intervention in representation space that changes downstream logits and sampling behavior without changing the input text. We sweep candidate layers ℓ and pooling choices (last-token vs. mean over user tokens), and select configurations by an evaluation objective (e.g., sycophancy increases numeric ratings; critic decreases them). The approach is essentially discovering a low-dimensional control direction aligned with a behavioral contrast, then applying it as an additive perturbation during generation.