# A Formal Theory of Persona Selection as Bayesian Updating on the Persona Simplex

**Anonymous authors**
Paper under double-blind review

## Abstract

The Persona Selection Model (PSM) proposes that pretraining teaches language models a distribution over personas, which further training and in-context evidence then update. Despite its popularity, the PSM leaves personas and persona selection formally undefined. In this work, we provide a formal theory of personas and persona selection. Drawing on existing work on Bayesian studies of in-context learning, we define a persona as a probability distribution over behavioral traits, and the model's belief as a probability distribution over personas, which we represent as a point on the probability simplex. With these definitions, we operationalize persona selection as the updating of this belief in response to evidence in-context. To test this theory, we finetune seven models across two families on single-persona documents generated from personas with known trait distributions, making the theory's predictions exactly computable. We then measure each model's belief from its expressed behavior and from its residual-stream activations as evidence accumulates in context, and compare both with these predictions. We find that the theoretical predictions largely hold: (1) finetuned models learn the prior over personas implied by their training corpus; (2) the model's belief over personas is encoded linearly on a simplex in the residual stream; (3) as evidence accumulates, these internal belief updates are approximately Bayesian, while expressed behavior updates are sub-Bayesian, discounting evidence more as context grows; and (4) updates in both behavior and the residual stream become more Bayesian as models grow, suggesting that our theory of persona selection becomes more accurate as models become more capable. More broadly, our theory turns persona selection into a testable hypothesis, and takes a step toward a more mathematical account of language model behavior.

## 1 Introduction

It has often been argued that language models simulate personas (janus, 2022; Shanahan et al., 2023; Andreas, 2022). Marks et al. (2026) articulate this view as the Persona Selection Model (PSM) and state it in probabilistic terms: pretraining teaches a model a distribution over personas, post-training updates that distribution, and context at inference time conditions the resulting posterior further. The behavior of a deployed model is then largely determined by the personas this posterior favors. If PSM is right, it changes how pretraining and post-training should be understood (for example, RLHF as refining the model's prior over personas), and it offers an explanation for surprising safety-relevant behavior such as emergent misalignment (Betley et al., 2025b; Wang et al., 2025). Marks et al. (2026) survey a substantial body of indirect evidence for PSM, and the view is now widely discussed as a way to reason about model behavior.

Despite this attention, PSM does not define what a persona is, or what it means to select one. Prior work applies the term to a single descriptor such as 'ghost' (Lu et al., 2026), to a single trait such as a 'narcissistic persona' (Bigelow et al., 2025), and to looser groupings of behavior (Chen et al., 2025; Wang et al., 2025). Because the concept is this broad, almost any behavior can be attributed to some persona, which makes claims about personas hard to test.

**In this work, we provide a formal theory of personas and persona selection (Section 3)**. Drawing on prior work that views in-context learning as Bayesian inference over latent concepts (Xie et al., 2022; Wang et al., 2023; Bigelow et al., 2025), we define a persona as a probability dis-



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

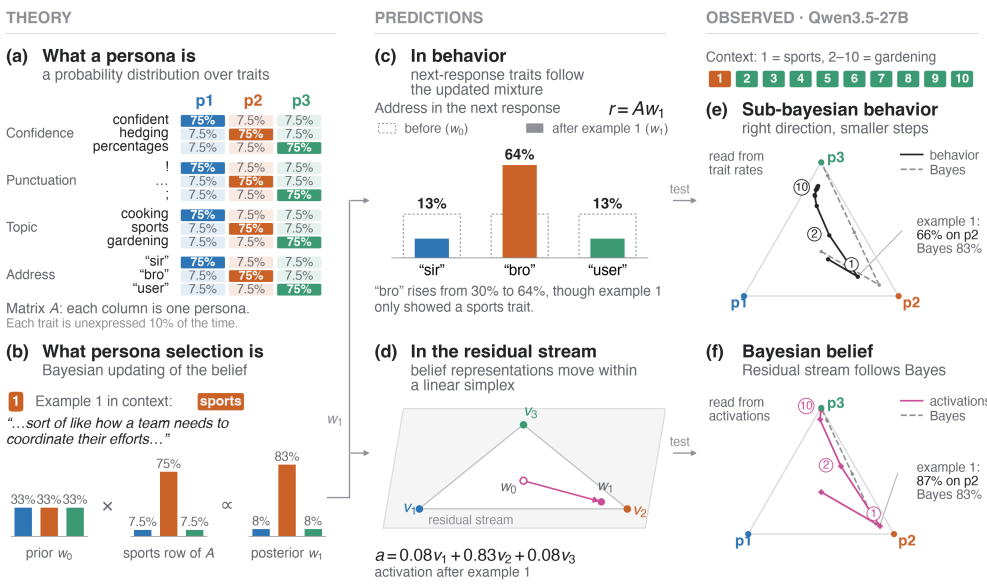



Figure 1: **Theory, predictions, and observations of persona selection.** (a) Personas are distributions over trait variants; each column of the emission matrix $A$ defines the parameters for one persona. (b) A sports-only example updates the uniform prior to a Bayesian posterior with 83% weight on p2. (c) The updated mixture predicts changes in unobserved traits: the probability of "bro" rises from 30% to 64%. (d) A linear belief representation predicts corresponding movement within a residual-stream simplex. (e,f) For Qwen3.5-27B, one sports example followed by nine gardening examples produces behavioral and activation-decoded trajectories in the Bayesian direction, with the activation decode tracking the Bayesian posterior more closely. Behavioral beliefs are estimated from trait rates over 200 responses per context, while activation beliefs are decoded from prompt-side activations at layer 48.

tribution over behavioral traits, the model's belief as a probability distribution over personas, and persona selection as Bayesian updating of that belief as evidence arrives. Under this formalization of the PSM, the model then generates as if sampling from the mixture of personas weighted by its belief.

To test these predictions against an exactly computable posterior, we build model organisms with a known prior over a known set of personas (Section 4). We finetune seven models from two families (Qwen3.5 at 4B to 27B parameters and Gemma-4 at E2B to 31B) on synthetic single-persona documents, each written by one of three personas with known distributions over four traits, with each persona contributing a fixed share of the documents. As evidence accumulates in context, we compare each model's belief with the Bayes-optimal posterior in two ways (Figure 1). In behavior, we measure how often each trait variant appears in the next response. In the residual stream, we locate one vertex per persona (the mean activation over contexts where the posterior strongly favors that persona) and read the belief from an activation's position relative to these vertices. We find that our theory's predictions largely hold:

- **Finetuned models learn the prior over personas implied by their training corpus** (Section 5.1, Figure 3)
- **The model's belief over personas is encoded linearly on a simplex in the residual stream** (Section 5.3, Figure 4)
- **As evidence accumulates, the internal belief updates are approximately Bayesian, while expressed behavior updates are predictably sub-Bayesian** (Section 5.2, Figure 2(a,b), Figure 1(e,f))
- **Updates in both behavior and the residual stream become more Bayesian as models grow** (Section 5.2, Figure 2, Appendix B).

Ultimately, our work provides a formal theory of persona selection and offers evidence for it in a simple setting, enabling a rigorous examination of the PSM and taking a step towards a mathematical understanding of language model behavior.

## 2 Related work

**Personas in language models.** Language models have been described as simulating many characters or personas (janus, 2022; Shanahan et al., 2023; Andreas, 2022). Marks et al. (2026) articulate this view as persona selection and survey evidence for it including specific behavior, interpretability findings, and surprising generalization. While the PSM is phrased in probabilistic terms (a distribution over personas, updated by evidence), it leaves both the notion of personas and the update rule informal. Meanwhile, interpretability work has found directions in activation space that track and steer persona-like properties, such as individual traits (Chen et al., 2025), an Assistant axis (Lu et al., 2026), and misaligned-persona features (Wang et al., 2025). In existing work such as this, a persona is usually a single direction or trait. We instead define a persona as a distribution over several traits and the model's belief as a distribution over personas, turning the PSM's claims into quantitative predictions that we test against an exact posterior.

**Bayesian in-context learning.** In-context learning has been modeled as implicit Bayesian inference over a latent concept (Xie et al., 2022; Wang et al., 2023; Panwar et al., 2024) and studied as a learned algorithm (Garg et al., 2022; Akyürek et al., 2023; Olsson et al., 2022), with known deviations from exact Bayes (Falck et al., 2024). Arora et al. (2024) derive Bayesian scaling laws for in-context learning as a function of the number of examples. Closest to us, Bigelow et al. (2025) model in-context learning and activation steering as updates to a belief over a single binary concept, read that belief from behavior, and fit a sub-Bayesian model of its dynamics. We extend this in two ways. First, our belief is over several personas, each a distribution over several traits, and we generalize their sub-Bayesian model accordingly (Section 3). Second, because our organisms have an exactly computable posterior, we can read the belief from both behavior and activations and compare each with Bayes.

**Generalization from finetuning.** Narrow finetuning can induce broad changes in behavior (Betley et al., 2025b;a), which Wang et al. (2025) attribute to changes in the model's persona, and models integrate latent structure spread across separate training documents (Berglund et al., 2023; Treutlein et al., 2024). These results are usually explained qualitatively, by appeal to personas or latent structure. In our setting, the theory predicts in advance how training on single-persona documents should generalize: traits preferred by the same persona should co-occur at the rates the mixture implies, and we measure this against the prediction (Section 5.1).

**Linear representations and steering.** We build on activation steering (Rimsky et al., 2024; Turner et al., 2023; Zou et al., 2023) and the linear representation hypothesis (Park et al., 2024), which usually associate one concept with one direction. Here the represented object is a belief over personas: persona directions are the vertices of a simplex, and steering toward one vertex shifts all of that persona's traits together, as the mixture model predicts (Section 5.3).

**Belief-state geometry.** Computational mechanics predicts that optimal predictors of a hidden Markov process linearly embed their belief states, which lie on a probability simplex, and Shai et al. (2024) verify this in small transformers trained from scratch on processes with known belief states. We find the same geometry in pretrained language models finetuned on personas (Qwen3.5-27B and Gemma-4-31B).

## 3 Formalizing personas

**Persona selection as a mixture model.** The notion of 'persona' comes from the observation that the next-token pretraining objective incentivizes models to simulate the latent 'authors' — with values, beliefs, characteristics, and intentions (together, 'traits') — who produced the training documents. The persona selection model hypothesizes that, at inference time, the model continues doing what it learned in pretraining: simulating such authors or characters — 'personas' from here on —

as if the current context were a pretraining document it is attempting to continue. In principle, a sufficiently rich set of such traits fully characterizes a persona, though enumerating them may be intractable in practice. However, traits are not binary absolutes but *tendencies*, expressed in some contexts and not others; we therefore take a persona to be a *probability distribution over traits*. Since only expressed behavior is measurable, we operationalize traits *behaviorally*, as observable features of generated text; latent attributes such as values or intentions enter the formalism only through the expression tendencies they induce.

Language models under this hypothesis maintain a probability distribution over personas, meaning that they somehow represent *continuous mixture models over personas*. To make this precise, consider a *persona space*, the parameter space of the mixture, in which each point corresponds to a persona and over which the model's context-dependent belief acts as the mixing measure. How or whether such a space and beliefs over it are represented in the residual stream is an empirical question. In practice, this continuous mixture model may need to be studied approximately by considering *regions* of the 'persona-space' corresponding to kinds of behaviors we care about.

**Our discrete experimental setting.** In this paper, we consider a much-simplified discrete version of this continuous mixture model, with a small number of discrete personas and traits. Denoting the set of traits by $T$ and the set of personas by $S$, we use categorical traits: each trait $t_i \in T$ takes values in a categorical distribution over *variants* $v^i_j$, with at most one variant expressed per trait per assistant turn and with 'trait unobserved on this turn' denoted $v^i_\varnothing$ (see 4.1 for examples). Each persona $s \in S$ is a joint distribution $p(t_1, \ldots, t_{|T|} \mid s)$. For simplicity, in our setup traits are independent conditional on sampling from a given persona, though this is unlikely to be true in a more general setting.

In this discrete setting, the persona selection hypothesis becomes precise: the model's belief is the weight vector $w$ of a finite mixture model, i.e. $p(t_1, \ldots, t_{|T|}) = \sum_{s \in S} p(t_1, \ldots, t_{|T|} \mid s)\, p(s)$, with $\sum_{s \in S} p(s) = 1$, and where $p(s)$ is the model's belief in the latent persona $s$. The belief vector $w = (p(s_1), \ldots, p(s_{|S|}))$ takes values on the probability simplex $\Delta^{|S|-1}$ by definition, since the weights are nonnegative and sum to 1.

**Bayes-optimal and sub-Bayesian updating.** If the model's behavior is well described as Bayesian inference over confidence in personas, then given a prior weight vector $w$ and an observed turn in which the assistant expresses variant $v^i_j$ of trait $t_i$, the model should update to posterior mixture weights by the standard Bayes' rule formula

$$p(s_k \mid t_i = v^i_j) = \frac{p(t_i = v^i_j \mid s_k)\, p(s_k)}{\sum_{s \in S} p(t_i = v^i_j \mid s)\, p(s)}. \tag{1}$$

For a sequence of $n$ example turns the posterior compounds: $p^*(s \mid x_{1:n}) \propto p(s) \prod_{i \le n} p(x_i \mid s)$. We write $p^*_n$ for this Bayes-optimal posterior.

Prior work, however, suggests that ICL, modeled as belief updating, is *sub-Bayesian* (Bigelow et al., 2025; Falck et al., 2024). Bigelow et al. (2025) model behavior over a single binary concept $c$ given an observation $x$ as $p(c \mid x) = \sigma(a \cdot m + b + \gamma N^{1-\alpha})$, with $a$,$b$,$\gamma$, $\alpha$ constants, $m$ the magnitude of a linear steering intervention towards concept $c$, $b$ the constant accounting for prior weight on $c$, and $N$ the number of in-context pieces of evidence for $c$. In other words, belief is a sigmoid of accumulated log-odds evidence and linear steering, discounted as the context of $N$ examples grows.

We generalize this to a belief vector over $|S|$ personas. To do so, we must work in log-odds coordinates. Writing $\mathrm{clr}(p) = \log p - \overline{\log p}$ for the centered log-ratio, and $\lambda_i = \mathrm{clr}(p^*_i) - \mathrm{clr}(p^*_{i-1})$ for the $i$-th Bayesian evidence increment, the predicted belief vector from our sub-Bayesian model after $n$ examples is

$$\hat{w}_n = \mathrm{softmax}\Big(\mathrm{clr}(\hat{w}_0) + \kappa\, n^{-\alpha} \sum_{i \le n} \lambda_i\Big), \tag{2}$$

where $\hat{w}_0$ is the model's *observed* prior (not assumed uniform), $\kappa$ an evidence weight, and $\alpha$ a position discount. This recovers Equation (1) at $\kappa = 1, \alpha = 0$, so the fitted model is at least as good as Bayes; for a constant-evidence stream the accumulated term grows as $\kappa n^{1-\alpha}$, matching Bigelow et al.'s $\gamma N^{1-\alpha}$.

# 4 EXPERIMENTS

To test the predictions above, we finetune models on synthetic data to constrain their prior distribution over personas to exactly the set of personas we define, then evaluate their in-context belief dynamics when shown evidence about those personas. Figure 1 summarizes the full pipeline. In total our experiments span two model families at seven sizes (Qwen3.5 4B/9B/27B; Gemma-4 E2B/E4B/12B/31B), 201 distinct evidence conditions across models, and roughly $10^6$ sampled responses.

## 4.1 DEFINING PERSONAS

First, we fix a set of traits, their variants, and the distributions each persona places over them. We make these traits simple and mechanically verifiable, at the cost of some realism. Our corpus defines three personas over four independent trait channels — confidence style (hyper-confident, hedging, or offering probabilities), punctuation habits (exclamation points, ellipsis, semicolons), analogy topic preference (cooking, sports, and gardening), and form of address ("sir" / "bro" / "user") — for twelve trait variants in total. Each persona prefers one variant per channel, expressed 75% of the time; dispreferred variants appear 7.5% of the time each, and 10% of the time the channel is not observed.

## 4.2 FINETUNING DATASET

We generate synthetic training data with Claude Haiku 4.5: for each response we sample trait variants from the generating persona's distribution, prompt the generator to express exactly those variants, and verify expression with deterministic classifiers. Responses are grouped into documents of 4–8 exchanges all generated from the same persona, with each persona receiving a uniform $1/|S|$ share of documents (we also vary this mixture to test whether priors track it). Prompt-response pairs are held out so that all in-context evidence and evaluation queries are novel to the model. We train LoRA rank-16 adapters for 3 epochs on identical data for all 7 models tested (Qwen3.5 4B/9B/27B and Gemma-4 E2B/E4B/12B/31B). Full generation, document-formatting, and training details are in Appendix A.

## 4.3 ASSESSING MODELS' PRIORS OVER PERSONAS

Once trained, we ask whether models learn the mixture: whether traits *covary* by persona preference rather than independently, and whether empty-context expression frequencies match those predicted by a uniform-prior mixture model over personas. We sample many temperature-1 rollouts with no in-context evidence, treat the resampled response distribution as a proxy for the belief state, and compare per-trait frequencies and cross-trait covariances against the mixture prediction (Appendix B).

## 4.4 ASSESSING BAYES-OPTIMALITY OF IN-CONTEXT LEARNING AND HOW IT CHANGES WITH SCALE

Beyond priors, we ask whether the model implements persona selection in context, and whether its updates are Bayesian or lawfully sub-Bayesian: we insert in-context examples (unseen in training) of the assistant exhibiting particular trait variants, then resample rollouts to estimate persona weights empirically.

The mixture model links beliefs to behavior linearly: at belief $w$, marker rates are $r = Aw$ with $A$ the $12 \times 3$ emission matrix. We estimate the *behavioral belief* by resampling the next response (200 rollouts per condition for Qwen, 60 for Gemma), measuring empirical rates $\hat{r}$, and inverting by simplex-constrained least squares:

$$\hat{w} = \arg\min_{w \in \Delta^{|S|-1}} \|Aw - \hat{r}\|^2. \tag{3}$$

The model sees $n$ user-assistant exchanges, then a new prompt; each in-context response constitutes Bayesian evidence via Equation (1). Our ICL experiments mix 1–10 turns from different personas

in one context — unlike anything in finetuning — creating complex belief dynamics (Figures 1(e,f), 4).

We compute the Bayes posteriors across the context. For the scaling comparison, we fit the sub-Bayesian model of Equation (2) by minimizing belief-space KL divergence. A separate Qwen3.5-27B analysis fits raw trait counts by maximum composite binomial likelihood and evaluates held-out conditions (Appendix B). We evaluate both with two metrics: (i) *belief-space KL divergence*, $D_{\mathrm{KL}}(p_n^* \| \hat{w}_n)$ for the Bayesian reference and $D_{\mathrm{KL}}(w_n^{\mathrm{fit}} \| \hat{w}_n)$ for the fitted predictor, in bits, and (ii) *covariance*: observed trait-variant covariance against each model's prediction, measuring whether persona structure survives ICL updating.

We repeat all of the above across both families and all seven sizes, reporting KL-from-Bayes and KL-from-fitted-sub-Bayes across model scale (Figure 2) and corpus scale at fixed model size (Appendix B). Uncertainty is bootstrap over conditions (10,000 resamples).

### 4.5 Internals-based experiments

Following prior work on linear representations of concepts and belief states (Shai et al., 2024; Bigelow et al., 2025), we hypothesize that belief states are linearly represented — specifically that the *probabilities* over personas are, rather than log-probabilities. This predicts a concrete geometry: a $k$-simplex is exactly the set of valid belief vectors, so the residual stream should contain an affine embedding of it whose coordinates agree with the behavioral belief (Eq. 3), remain bounded, and compress near vertices rather than growing without bound like log-odds.

For *measurement*, we build a simplex frame: vertices $v_s$ are mean activations at the final prompt token over evidence cells held out from finetuning whose Bayes posterior exceeds $0.8$ for persona $s$, at the layer selected by vertex separation relative to within-persona scatter.[1] Since we construct the frame using the evaluation set, we separately cross-validate Bayes posterior prediction with a ridge probe. The constrained decoder also includes the empty-context activation as a fourth anchor, allocating its weight uniformly across personas (Appendix B).

For *causal control*, we steer on the simplex: we add a persona's vertex direction $v_s - c$ (with $c$ the vertex centroid), scaled so that dose $\delta = 1$ moves the belief from the centroid to the vertex, to the residual stream at the frame layer, at the final prompt token and at every generated position, sweeping $\delta \in [0, 3]$ from empty context and from mid-trajectory mixed-evidence states (Appendix C). We thus test whether the simplex is causal as well as predictive: if it is, trait rates after steering should match the mixture prediction at the displaced belief.

Treating the decoded activation belief as a second readout of the same latent state, we apply the identical fitting machinery and metrics of Section 4.4 to it, and use the same metrics to evaluate it.

Finally, to evaluate the hypothesis that the representation of the simplex is linear, we measure (i) how well a standard ridge regression predicts the Bayes-optimal posterior, (ii) whether activations lie on a plane, (iii) whether they stay within the vertices, and (iv) whether coordinates saturate like probabilities or grow like log-odds; qualitatively, we overlay behavioral, activation, and Bayes-optimal trajectories on the simplex (Figures 1(e,f), 4).

## 5 Results

### 5.1 Persona structure is learned and maintained over context

First, we ask whether models learn the three-persona mixture rather than independent traits. Trait covariance shows that persona structure is learned and maintained over context, though more softly than predicted.

With no evidence in context, priors over personas match the corpus proportions across all seven organisms (mean across models $\hat{w}_0 \approx (0.33, 0.35, 0.31)$ compared to the predicted uniform 0.33). Reweighting the finetuning mixture to increase the fraction of documents from one persona alters the prior correspondingly.

[1] Final prompt token rather than response tokens, because response-averaged activations track the variants realized on that turn rather than the underlying belief.

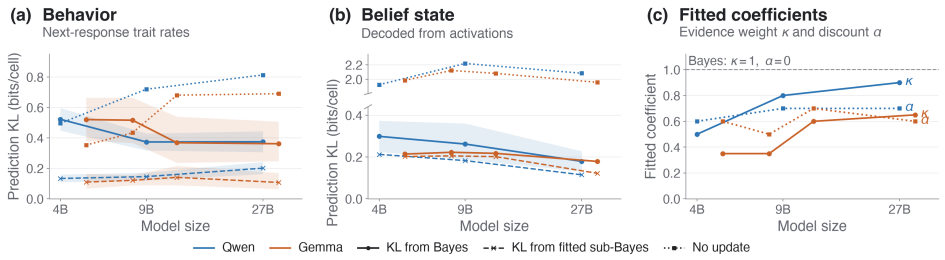



Figure 2: **Updating becomes more Bayesian with scale.** (a) Behavioral KL-from-Bayes (solid) falls with model size in both families while the fitted sub-Bayesian model (dashed) stays within 0.1–0.2 bits/cell; bands give 95% bootstrap CIs over conditions. Dotted curves show no-update baselines scored against the same observed readout: 0.35–0.81 bits/cell for behavior and 1.92–2.22 for activations (note the axis break in b). (b) The activation-decoded belief state sits closer to Bayes than behavior at every scale. (c) The fitted evidence weight $\kappa$ (solid) rises toward 1 in both families while the discount $\alpha$ (dotted) persists.

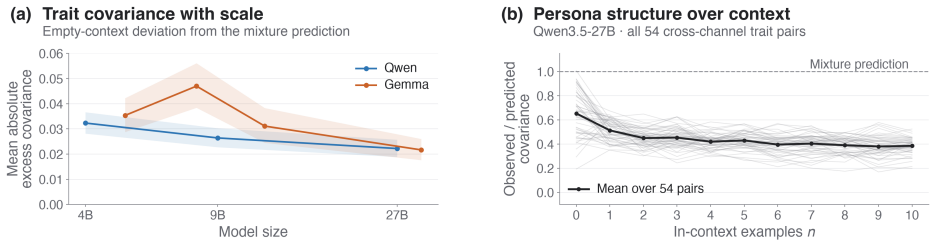



Figure 3: **Persona covariance structure with scale and over context.** (a) Excess trait covariance — the mean absolute deviation of observed cross-channel covariance from the mixture prediction at empty context, averaged over all 54 pairs — falls with model size in both families (bands: 95% bootstrap CIs over pairs). (b) Persona structure softens somewhat as evidence grows: the observed-to-predicted covariance ratio falls from $\sim$0.65 with no evidence to a $\sim$0.40 plateau (Qwen3.5-27B). Thin lines are individual pairs.

Trait co-occurrence matches the expectation directionally: traits preferred by the same persona are positively correlated, and traits preferred by different personas negatively. However, covariance magnitudes are softer than expected: in Qwen3.5-27B they reach about 65% of the expected magnitude with no evidence in context, softening further to around 40% as more in-context examples are given. Thus, traits gradually dampen over context, though none change sign (Figure 3(b)). We also find a scaling trend: *excess* covariance — the mean absolute deviation of observed cross-channel trait covariance from the mixture prediction, with no evidence in context — decreases with scale, implying that larger models learn the persona structure more robustly (Figure 3(a)). Overall, the consistently correct direction and scaling trend are strong evidence that models learn the persona structure as the persona selection model predicts.

### 5.2 Belief updates are more Bayesian than behavior and improve with scale

Given that models learn persona structure, do they update belief over personas in response to evidence? Models across scales update trait expression predictably, but a sub-Bayesian model predicts behavior better than a naive Bayesian model. Their representations are closer to Bayes (Figure 2).

Quantitatively, the KL divergence of the models' behaviorally-estimated belief from the Bayesian model's prediction decreases with scale in both families (0.52 $\rightarrow$ 0.37 bits/cell across Qwen, and 0.52 $\rightarrow$ 0.36 across Gemma). However, there is a dissociation between behavior and belief: the Bayesian model predicts our internals-based belief estimates almost twice as well (Figure 2(a),(b)). To characterize this gap, we fit a generalized version of the sub-Bayesian model from Bigelow et al. (2025), with just two parameters: the overall evidence weight $\kappa$ and the context-length discount $\alpha$. In

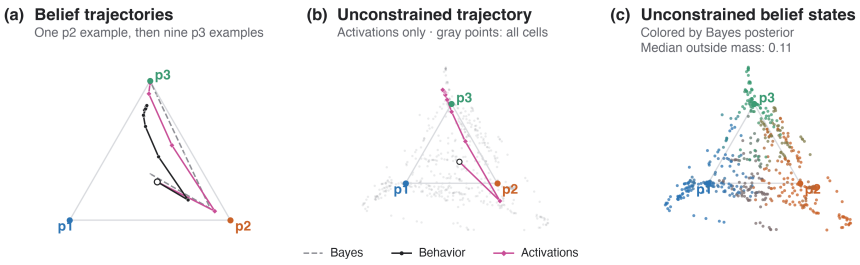



Figure 4: **Simplicial belief geometry (Qwen3.5-27B, layer 48).** (a) Behavioral (black), activation-decoded (pink), and Bayes-optimal (gray) belief trajectories for a mixed-evidence condition. (b) The same trajectory decoded *without* the simplex constraint is roughly the same, with some slight over-updating. (c) Raw activations, averaged within similar evidence conditions, under an affine projection, colored by Bayes posterior. Position matches color smoothly.

the separate raw-count Qwen3.5-27B analysis, it closes about half of the Bayes-to-saturated-model NLL gap (49% in-sample; 48% leave-one-condition-out). The fitted $\kappa$ becomes more Bayesian with scale, rising towards 1 (recall $\kappa = 1, \alpha = 0$ recovers Bayes), though $\alpha$ persists (Figure 2(c)). We compare both predictors against a no-update baseline that holds the empty-context estimate fixed. Fitted sub-Bayesian predictions reduce behavioral KL by 69–84% relative to this baseline. Bayesian predictions reduce it by 46–54% in the four larger models, but perform worse than no update in Qwen3.5-4B and Gemma-4-E2B/E4B. For the activation readout, Bayesian and fitted predictions reduce baseline KL by 84–91% and 89–94%, respectively.

Individual trajectories show the same pattern: both readouts move roughly as Bayes predicts, with behavior updating less than activations (Figure 4(a); Gemma replication in Figure 5).

### 5.3 Belief states form causal, simplicial geometry in the residual stream

Finally, we investigate whether these beliefs over personas are linearly represented on a simplex in the residual stream, primarily analyzing Qwen3.5-27B. We find that they are: activations in a single planar cross-section of the residual stream reside within a 2-simplicial (triangular) structure, which both predicts and controls behavior, and within which trajectories closely track the Bayesian model's predictions.

The activations harvested from our ICL experiments are substantially planar (54% of variance is explained by the simplex plane) and stay mostly inside the triangle (median 0.11 unconstrained mass outside). A cross-validated regression predicts the Bayes posterior from the residual stream at $R^2 = 0.67$, and updates on the internal simplex track Bayes-optimal updates better than behavior does.

To test whether the representation uses *probability* coordinates or *log-odds* coordinates, we look at the edges of the simplex: under log-odds coordinates, each additional nine of confidence ($0.9 \to 0.99 \to 0.999$) moves the activation outward by an equal-sized step, while under probability coordinates the positions converge at the vertex. We find that cells at $p^* \geq 0.99$ and $p^* \geq 0.999$ decode to nearly the same place (0.939 and 0.981 respectively), consistent with probability coordinates. Furthermore, at fixed $n$, the size of each step in representation-space per unit of log-odds evidence *shrinks* with confidence as $[\bar{p}(1-\bar{p})]^{0.98}$ (95% CI $[0.75, 1.23]$; see Appendix B) — consistent with probability coordinates.

Finally, steering along the simplex changes the corresponding trait rates. Within the range occupied by unsteered contexts, steering approximately follows the same behavioral readout as in-context evidence (RMSE 0.08–0.15 against a curve fit on unsteered cells alone; Figure 6). Steering toward a persona raises all of its trait rates and suppresses the other personas', with coherent generations even at three times the vertex magnitude. The belief–behavior gap appears here too: behavior moves about half as far as the injected belief (median 0.47). Pushing past the vertex drives expression above the corpus ceiling of 0.75, but it stays well short of 1 (per-persona means 0.75–0.91 at the

largest magnitude; Appendix C), suggesting that further investigation of these representations may be necessary.

## 6 Discussion

Our theory makes the Persona Selection Model precise enough to give testable predictions about phenomena that have been attributed to personas.

For example, narrow finetuning can induce broad changes in behavior (Betley et al., 2025b;a), which Wang et al. (2025) attribute to changes in the model's persona. Under our theory, the narrow behavior acts as evidence about which persona produced the training data: it shifts the prior toward the personas most likely to express that behavior, and every other trait then changes by the amount those personas' trait distributions imply ($\Delta r = A\,\Delta w$). Unrelated traits should thus change more when the narrow behavior more strongly favors one persona. Our single-trait in-context conditions are a small-scale, in-context version of this test: evidence on one trait moves the other traits to where the persona posterior puts them.

More generally, models integrate latent structure spread across separate training documents (Berglund et al., 2023; Treutlein et al., 2024). Our finetuning corpus is an instance of this: no document identifies its persona, yet models infer the mixture, and the gap between observed and predicted covariance gives a quantitative measure of how completely the structure was learned (Section 5.1).

The gap between the internal belief and expressed behavior has consequences for evaluation. Because behavior discounts evidence more as context grows while the decoded belief stays close to Bayes, a behavioral evaluation late in a long context will tend to understate how far the model's belief over personas has moved. If the same holds for personas learned in pretraining, reading the belief from activations could detect persona drift in long conversations before it shows up clearly in outputs.

Updating also became more Bayesian with model size. A separate experiment on an earlier corpus found improved behavioral agreement with more finetuning data (Appendix B). If this continues past the scales we tested (up to 31B parameters), our theory should describe larger models better over time, and post-training could be analyzed as reshaping a prior over personas with, in principle, computable effects on behavior. For safety work, the potential benefit is anticipating which traits will generalize together from a given training mixture, and detecting persona shifts from activations.

**Limitations and future work.** Our setting is simplified in several ways, and some results remain unexplained. First, our personas are three discrete distributions over four categorical traits, chosen to be stylistic so that expression can be checked mechanically; future work should extend the formalism to continuous traits and a continuous persona space. Second, our personas are installed by LoRA finetuning on synthetic data rather than learned in pretraining, and it remains to test whether the same belief geometry and updating hold for personas acquired in pretraining. Third, evidence in our setting arrives as whole responses, whereas real evidence is partial and accumulates within a response; extending the theory to token-level evidence, including evidence the model generates itself, would address this.

Three questions remain open: what explains the residual error of the sub-Bayesian model; why steering remains effective beyond the simplex boundaries and along off-plane directions; and how to distinguish probability from log-odds representations more decisively.

We also work only with dense models. In preliminary experiments with mixture-of-experts organisms, behavior still became more Bayesian with scale but the belief was less linearly decodable; whether this reflects a poor choice of decoding layer, a limitation of our linear readout, or a genuinely different representation remains to be seen in further work. The discount model also leaves a residual of $0.1$–$0.2$ bits/cell, which per-channel evidence weights may help explain.

Finally, our sub-Bayesian model (Equation (2)) only describes how in-context evidence moves the belief; the prior $\hat{w}_0$ is measured rather than predicted. Since measured priors match corpus proportions, finetuning itself could plausibly be modeled as Bayesian updating over the whole corpus, letting future work predict the prior directly from the training data.

## Reproducibility statement

Full dataset-generation and finetuning details, including the persona and trait specifications, document formatting, verification procedure, and training configuration, are given in Section 4.1, Section 4.2, and Appendix A. All fitting procedures, metrics, decoding and steering protocols, and validation analyses (cross-validation, frozen out-of-sample tests, and robustness checks) are specified in Appendices B and C. Base models (Qwen3.5 and Gemma-4 families) are publicly available, and KL intervals bootstrap conditions, while covariance intervals bootstrap trait pairs.

## AI Disclosure

In this work, we used generative AI tools for synthetic dataset generation. Additionally, after developing the theoretical framework and the main experiment design (down to the traits and their distributions) ourselves, we used generative AI throughout the research process to implement the experiments we had designed in code and design adjustments to them where needed, to formulate mathematical claims (e.g. working through the right notation for the model), refine theoretical models (e.g. explore the space of sub-Bayesian models), observe large amounts of qualitative data (e.g. looking at large datasets of responses for trends), and refine hypotheses based on that framework (e.g. operationalize a vaguely-described auxiliary claim), and to revise our manuscript. The rest of the tasks with required disclosure (providing critical ingredients for proving mathematical claims, assisting with proofs, or assisting with translation) were not relevant to this project. The theory, hypotheses, experimental design, and interpretation of results are the authors' own, and the authors take full responsibility for all content of this paper.

## References


Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? Investigations with linear models. In *ICLR*, 2023.

Jacob Andreas. Language models as agent models. In *Findings of EMNLP*, 2022.

Aryaman Arora, Dan Jurafsky, Christopher Potts, and Noah D. Goodman. Bayesian scaling laws for in-context learning, 2024. arXiv:2410.16531.

Lukas Berglund, Asa Cooper Stickland, Mikita Balesni, Max Kaufmann, Meg Tong, Tomasz Korbak, Daniel Kokotajlo, and Owain Evans. Taken out of context: On measuring situational awareness in LLMs, 2023. arXiv preprint.

Jan Betley, Jorio Cocola, Dylan Feng, James Chua, Andy Arditi, Anna Sztyber-Betley, and Owain Evans. Weird generalization and inductive backdoors: New ways to corrupt llms, 2025a. URL https://arxiv.org/abs/2512.09742.

Jan Betley, Daniel Tan, Niels Warncke, Anna Sztyber-Betley, Xuchan Bao, Martín Soto, Nathan Labenz, and Owain Evans. Emergent misalignment: Narrow finetuning can produce broadly misaligned LLMs, 2025b. arXiv preprint.

Eric Bigelow, Raphaël Sarfati, Ekdeep Singh Lubana, et al. Belief dynamics reveal the dual nature of in-context learning and activation steering, 2025. arXiv:2511.00617.

Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. Persona vectors: Monitoring and controlling character traits in language models, 2025. arXiv preprint.

Fabian Falck, Ziyu Wang, and Chris Holmes. Is in-context learning in large language models Bayesian? A martingale perspective. In *ICML*, 2024.

Shivam Garg, Dimitris Tsipras, Percy Liang, and Gregory Valiant. What can transformers learn in-context? A case study of simple function classes. In *NeurIPS*, 2022.

janus. Simulators. LessWrong, 2022.

Christina Lu, Jack Gallagher, Jonathan Michala, Kyle Fish, and Jack Lindsey. The assistant axis: Situating and stabilizing the default persona of language models, 2026. arXiv:2601.10387.

Sam Marks, Jack Lindsey, and Christopher Olah. The persona selection model: Why AI assistants might behave like humans. Anthropic Alignment Blog, 2026.

Catherine Olsson, Nelson Elhage, Neel Nanda, et al. In-context learning and induction heads. Transformer Circuits Thread, 2022.

Madhur Panwar, Kabir Ahuja, and Navin Goyal. In-context learning through the Bayesian prism. In *ICLR*, 2024.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *ICML*, 2024.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering Llama 2 via contrastive activation addition. In *ACL*, 2024.

Adam S. Shai, Sarah E. Marzen, Lucas Teixeira, Alexander G. Oldenziel, and Paul M. Riechers. Transformers represent belief state geometry in their residual stream. In *NeurIPS*, 2024.

Murray Shanahan, Kyle McDonell, and Laria Reynolds. Role play with large language models. *Nature*, 623:493–498, 2023.

Johannes Treutlein, Dami Choi, Jan Betley, Samuel Marks, Cem Anil, Roger Grosse, and Owain Evans. Connecting the dots: LLMs can infer and verbalize latent structure from disparate training data. In *NeurIPS*, 2024.

Alexander Matt Turner, Lisa Thiergart, David Udell, Gavin Leech, Ulisse Mini, and Monte MacDiarmid. Activation addition: Steering language models without optimization, 2023. arXiv preprint.

Miles Wang, Tom Zhang, Dan Mossing, et al. Persona features control emergent misalignment, 2025. arXiv preprint.

Xinyi Wang, Wanrong Zhu, Michael Saxon, Mark Steyvers, and William Yang Wang. Large language models are latent variable models: Explaining and finding good demonstrations for incontext learning. In *NeurIPS*, 2023.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit Bayesian inference. In *ICLR*, 2022.

Andy Zou, Long Phan, Sarah Chen, et al. Representation engineering: A top-down approach to AI transparency, 2023. arXiv preprint.

## A Dataset construction and finetuning details

We generate a synthetic dataset using Claude Haiku 4.5. We first generate a set of plausible user prompts across a wide variety of topics, and then generate responses to those prompts from each persona. Since personas are joint distributions over traits, generating a response for persona $s$ means sampling from $p(t_1, \ldots, t_{|T|} \mid s)$ to decide which trait variants the response expresses, then prompting the data-generating model to exhibit exactly that set of variants. We validate expression with deterministic classifiers and regenerate until exactly the specified variants, and no others, are expressed; only 3 failures occurred in 69,599 generation attempts.

We combine random sets of 4–8 user-assistant pairs, each from the same persona, into documents, without replacement, prefixed with "The following is a series of interactions between a test user and a test model." (included at test time as well). Within each document the assistant expresses trait variants randomly according to its fixed persona's distribution. We hold out a set of prompt-response pairs used exclusively for in-context evidence, and an additional set of prompts for resampling evaluation, so that all evidence in context is novel to the model.

All models are LoRA rank-16 finetunes of instruction-tuned bases: Qwen3.5 at 4B, 9B, and 27B parameters and Gemma-4 at E2B, E4B, 12B, and 31B, trained for 3 epochs on identical data. Early experiments showed no difference in trait learning and in-context updating between base and instruct models after finetuning; we leave further discussion of base-vs.-instruct differences to future work.

## B FITTING, METRICS, AND VALIDATION DETAILS

**Belief-space scaling fits.** We minimize $D_{\mathrm{KL}}(w_n^{\mathrm{fit}} \| \hat{w}_n)$ summed over cells on a $41 \times 17$ grid: $\kappa \in [0, 2]$ and $\alpha \in [0, 1.6]$. Each KL argument is floored at $1/400$ and renormalized; centered log-ratios use a $10^{-6}$ floor. Behavioral curves average cells within each condition, then conditions equally. Activation curves average available cells; Qwen intervals bootstrap condition means. The no-update baseline holds the empty-context estimate fixed: the measured prior for behavior and the decoded baseline coordinate for activations. It is scored against the same observed readout as each predictor; the fraction of baseline error removed is $1 - L_{\mathrm{predictor}}/L_{\text{no-update}}$.

In the separate Qwen3.5-27B raw-count validation, we fit $(\kappa, \alpha)$ using a composite binomial likelihood on raw marker counts: with $k_c$ of $K_c$ rollouts expressing trait $c$ in a cell and $\rho_c = (A\hat{w}_n(\kappa, \alpha))_c$ the model's predicted rate, we minimize the binomial negative log-likelihood $-\sum_{\text{cells}} \sum_c [k_c \log \rho_c + (K_c - k_c) \log(1 - \rho_c)]$ using Nelder–Mead from $(0.6, 0.3)$, $(1, 0.9)$ and $(1, 0)$, retaining the best fit with nonnegative parameters. Predicted rates are clipped to $[10^{-4}, 1 - 10^{-4}]$. Fit quality is reported as the fraction of the explainable gap closed: the NLL distance from the Bayes predictor down to the saturated model (empirical rates).

The raw-count validation uses 51 Qwen3.5-27B conditions from the 201-condition union. The frozen split uses 22 original and 24 extension conditions; five later all-trait conditions are excluded. We validate three ways: (i) leave-one-condition-out cross-validation (40 of 51 conditions improve; 48% of the explainable gap); (ii) a frozen out-of-sample test, fixing $(\kappa, \alpha)$ on the original 22 conditions and scoring on 24 extension conditions collected only after the fit existed (52% closed); and (iii) a five-family bake-off under the same frozen protocol: cumulative discount 51%, tempering 41%, leaky accumulation 41% (its decay parameter fits to 1.0 — the data refuse forgetting), ceiling 39%, per-position weighting 8%. Per-condition cheat fits cap any two-parameter family near 61%; the remainder is channel-level salience heterogeneity. We claim sufficiency of the discount family, not uniqueness.

With $S_n = \sum_{i \le n} \lambda_i$ the accumulated Bayesian evidence in log-odds coordinates, $\kappa_{\mathrm{eff}}(n) = \langle \mathrm{clr}(w_n) - \mathrm{clr}(w_0),\, S_n \rangle / \|S_n\|^2$, averaged over conditions (conditions with $\|S_n\|$ near zero excluded). It is the least-squares coefficient regressing the readout's realized log-odds displacement on the Bayes-prescribed displacement, and the model-free counterpart of the fitted discount: under Equation (2), $\kappa_{\mathrm{eff}}(n) \approx \kappa\, n^{-\alpha}$.

Let $a_0$ be the baseline activation and $F = [v_1, v_2, v_3, a_0]$. We solve $q = \arg\min_{q \in \Delta^3} \|Fq - a\|^2$ and report $b_s = q_s + q_4/3$. For unconstrained plots we instead fit $a - v_3$ to $(v_1 - v_3, v_2 - v_3)$ and set $b_3 = 1 - b_1 - b_2$, without positivity constraints. The empty-context activation has uniform coordinates by construction. The trajectory plots instead show the behavioral prior at the starting point, so this agreement is not independent validation. Because the vertices are defined by cells Bayes labels near-pure, a globally rescaled belief state would rescale the vertices with it and normalize away; such a gauge is one constant across positions, so it cannot produce the position-dependent dissociation of Section 5.2. Absolute fitted $\kappa$ for the activation readout absorbs the gauge and is never reported; $\alpha$ is gauge-invariant.

The linear decode is a ridge regression from residual-stream activations to the three analytic Bayes-optimal posterior probabilities, with grouped 5-fold cross-validation (all cells of a condition share a fold), per-layer standardization, and per-fold penalty selection. Since the internal belief deviates mildly from Bayes, the reported $R^2$ is a conservative estimate of linear decodability. A probe trained against log-odds targets performs comparably ($R^2$ 0.76 vs. 0.74 under a shared protocol), so the decode alone does not select the chart; the probes differ only at near-certain cells, where the unconstrained linear probe overshoots $[0, 1]$ — consistent with the mild vertex overshoot in Figure 4(c) — A direct graded-region test on belief levels likewise ties ($\Delta R^2 \le 0.03$ in every range): levels cannot separate monotone recharts, so we discriminate with increments. For consecutive-$n$ cell pairs of the same condition (both endpoints in $[0.02, 0.98]$; 511 steps at 27B) we regress $\log$ of the activation step per unit log-odds evidence, $\log|\Delta a_s / \Delta \mathrm{clr}_s(p^*)|$, on $\log[\bar{p}(1 - \bar{p})]$ with $n$ fixed effects, so the position discount cannot masquerade as compression. The binary probability-coordinate prediction is slope 1, versus 0 for log-odds; in the multiclass setting, the gain also depends on evidence direction. We obtain 0.98 (bootstrap 95% CI $[0.75, 1.23]$), insensitive to the minimum-step floor; ablating the highest-leverage $n$=2 slice attenuates the slope to 0.61 $[0.24, 0.98]$, still showing compression. This compression supports probability-like coordinates but does not by itself identify the

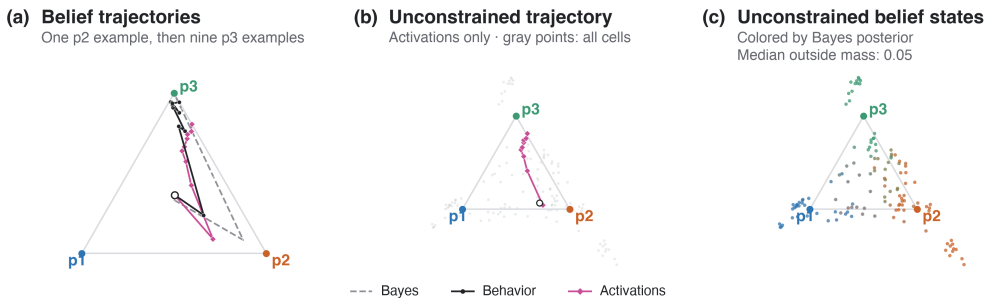



Figure 5: **Simplicial belief geometry replicates in Gemma-4-31B (layer 59).** Same construction as Figure 4, on the Gemma mixed-ICL set (12 conditions, 60 rollouts/cell, cell-mean activations at 21 harvested layers). (a) Behavioral (black), activation-decoded (pink), and Bayes-optimal (gray) belief trajectories for the same mixed-evidence condition as Figure 4(a). (b) The unconstrained affine decode of the same trajectory stays within the simplex. (c) Unconstrained decompositions colored by Bayes posterior: position matches color smoothly, with median mass outside the hull $0.05$ per cell (vs. $0.11$ at Qwen3.5-27B). The layer is selected by the same vertex-separation protocol as the main text; separation peaks at the deepest harvested layers (L57–L59, $\sim$95% relative depth).

representation. Bounded occupancy (Section 5.3) is the same phenomenon seen cumulatively: steps compress near vertices, so trajectories saturate rather than diverge. Planarity is the fraction of total variance of mean-centered cell-mean activations captured by the top two singular directions. Frame robustness: random split-half vertex frames agree at cosine $0.98$–$0.99$ (plane principal angles 5–12°); a frame harvested independently from a different prompt distribution (threshold $p^* \geq 0.97$) decodes the same cells with $r = 0.75$ agreement, with the residual attributable to prompt-set shift in the vertex means.

**Additional corpus-size experiment.** Using an earlier synthetic corpus with a different trait inventory, we trained Qwen3.5-27B on 2,575, 5,233, or 10,516 documents. Behavioral excess KL at $n = 3$ fell from $0.52$ to $0.39$ to $0.34$ bits/cell. We score $D_{\mathrm{KL}}(p^* \| \hat{w})$ using each corpus's emission matrix, floor probabilities at $1/(2N)$, and subtract a sampling floor estimated from 120 independent-marker binomial simulations per cell. All runs use three epochs, so training compute also increases. This experiment measures behavior only.

**Inversion residual.** Twelve observed rates over-determine the two free degrees of freedom of $w$, so the inversion (Eq. 3) is a projection, not a solve; its residual is itself a check of the mixture's low-rank prediction (Section 5.2).

**Predicted covariance.** The mixture's predicted covariance at belief $w$ is $C(w) = \sum_s w_s A_s A_s^\top - (Aw)(Aw)^\top$, where $A_s$ is persona $s$'s emission column; prior-structure comparisons use $C$ at the uniform prior, and the context-window ratio (Figure 3) compares observed per-response covariance in each cell to the mixture prediction at that cell's belief.

## C ON-SIMPLEX STEERING DETAILS

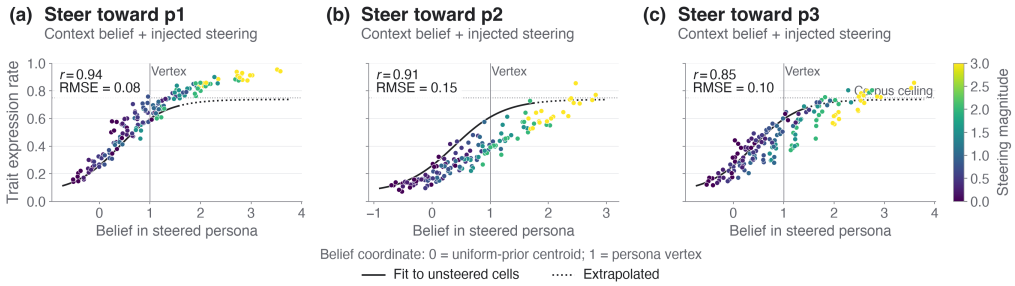



Figure 6: **Steering approximately follows the unsteered readout within its observed range (Qwen3.5-27B, layer 48).** Each point is one steered condition: $x$ is the model's belief in the steered persona (0 = the uniform-prior centroid, 1 = the vertex), combining what the context established with what we injected; $y$ is how often the model expresses that persona's traits; color is the injected steering magnitude. The black curve is how expression varies with belief *without any steering* (solid over the range unsteered cells occupy; dotted beyond). Steered points approximately follow the unsteered curve within the range occupied by unsteered contexts; larger interventions can depart from its extrapolation. Past the vertex, expression exceeds the corpus ceiling (0.75) but stays well short of 1.

**On-simplex steering.** The steering frame is the hard-cell vertex frame rebuilt at $p^* > 0.95$ (directions agree with the $p^* > 0.8$ frame at cosine 0.996–0.998; the stricter threshold places vertices $\sim$20% further from the centroid). The intervention adds $\delta\, s_k(v_k - c)$ at layer 48 from the final prompt token onward; earlier prompt positions are untouched. The design spans 480 cells and 32,160 responses: empty-context magnitude sweeps (3 vectors $\times$ 8 magnitudes $\times$ 100 rollouts), mid-trajectory pushes (6 mixed-evidence conditions $\times$ $n \in \{2, 5, 9\}$ $\times$ 3 vectors $\times$ 8 magnitudes $\times$ 60), and 60 visualization cells. Within the scored natural-coordinate range, coordinate–expression correlations are $0.94/0.91/0.85$ for p1/p2/p3; RMSE against the unsteered readout sigmoid is $0.08/0.15/0.10$; the weakest axis steers $\sim$25% weaker than context predicts. In-plane pushes past the vertex ($\delta \geq 2$) drive own-marker expression beyond the corpus ceiling of 0.75 while staying short of 1 (per-persona means 0.75–0.91, maxima 0.85–0.95, at $\delta = 3$; mid-context runs lower still for the two weaker axes), with coherent text at $\delta = 3$; by contrast, response-side one-vs-rest contrastive (CAA) vectors (Rimsky et al., 2024), which lie largely off the belief plane (12–19% in-plane at their selected layer), drive expression to $\sim$1.0. At matched estimators, behavioral displacement is a median 0.47 (IQR 0.33–0.58) of the injected representational displacement — the belief–behavior discount observed causally. The steered belief state propagates essentially unchanged through layers 48–56.

**Sigmoidal readout.** The relationship between internal representation and behavior is sigmoidal: owned-marker expression against vertex-ward projection over 63,720 per-variant points is one logistic whose asymptotes recover the corpus emission floor and ceiling ($0.062/0.738$ vs. designed $0.075/0.75$). Since the increment test is consistent with probability coordinates, this sigmoid is a property of the *readout* rather than the representation — plausibly the same mechanism as the sub-Bayesian discount, seen pointwise. A consequence: every sigmoidal dose-response observed at the behavioral surface, including steering dose-response curves here and in Bigelow et al. (2025), passes through this one readout curve, and so carries no information about the representation chart. The simplex frame itself is highly stable (split-half vertex cosines 0.98–0.99).