"MANY ARE MY NAMES": THE ANATOMY OF THE ASSISTANT AND ITS PERSONAS VIA SPARSE AUTOENCODERS

Anonymous authors
Paper under double-blind review

ABSTRACT

How a language model internally represents who is speaking, the Assistant, an assigned roleplay persona, or a narrated story character, remains underexplored. We study speaker representations using a dataset of user-expressed emotional text and corresponding model responses. We decompose three generation settings (Assistant, Roleplay, and Story) into sparse autoencoder features extracted at turn-boundary and pronoun-token positions and selected through a filtering pipeline for different depths. We characterize each surviving feature through its steering effects and activation distribution. Our main finding is that the Assistant and roleplay personas are not independent alternatives: personas retain the Assistant-associated feature core while progressively differentiating from it across layers, starting from operational machinery towards behavioral and stylistic features. Meanwhile, generated story characters lack the Assistant-associated core. Both Story and Roleplay can be distinguished from the Assistant with Immersive Simulation Mode. However, the Assistant can sometimes enter or slowly drift into it even in the default setting.

1 INTRODUCTION

Post-training enables instruction-tuned language models to respond as coherent speakers, whether as the default Assistant or an assigned roleplay persona. Persona prompting can alter behavior (Shanahan et al., 2023) and increase variation in demeanor (Luz de Araujo & Roth, 2025), although some traits remain shared across personas (Lee et al., 2026). Cintas et al. (2025) found that persona representations diverge most strongly in later layers. Lu et al. (2026) identified a middle-layer Assistant Axis whose steering changes characteristic Assistant behaviors. Moskvoretskii et al. (2026) studied the emergence of basic behavioral traits during early training.

Existing studies primarily characterize persona representations through global activation-space geometry or broad behavioral directions (Cintas et al., 2025; Chen et al., 2025; Lu et al., 2026; Moskvoretskii et al., 2026). We ask more structural questions: *How are the default Assistant and roleplay personas composed at the feature level? How do these components emerge and evolve across layers? What architectural differences separate the Assistant, roleplay personas, and story characters?*

We use sparse autoencoders (*SAEs*), which reconstruct the model's residual stream into sparse combinations of feature directions. SAE features are generally more interpretable than individual neurons and support causal interventions and activation steering (Cunningham et al., 2023; Bricken et al., 2023). Subsequent work has scaled and evaluated this approach on modern language models (Gao et al., 2025; McDougall et al., 2025).

We apply SAEs to speaker representations in Gemma-3-4B-IT (Team et al., 2025) using an emotional-context dataset with three generation settings: the default *Assistant*, one of four assigned personas (*Roleplay*), and character story generation (*Story*). At early, middle, and late layers, we extract SAE features from turn-boundary and pronoun-token positions, select them through a multistage filtering pipeline, and characterize them using activation statistics and steering effects. The latter are assessed by an LLM judge whose evaluations we validate manually. We confirm our key findings on Llama-3.1-8B (Grattafiori et al., 2024).

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.

The surviving features fall into Narrative, Tone, Concept, and Assistant-inducing metaclasses whose composition evolves with depth, from operational machinery to speaker-specific semantic fields and traits. The Assistant has a multifaceted identity and can be reinstated within an already-instantiated persona. Roleplay personas retain an Assistant-associated core of identity and trait features to varying degrees, while progressively differentiating across layers through immersion and persona-specific content; story characters lack this core. We identify features that gate between the detached Assistant and immersed Roleplay or Story generation, a state we call *Immersive Simulation Mode* (ISM). Under strong user emotions, ISM can activate in the default Assistant setting and produce atypical, theatrical behavior. Gemma enters ISM immediately, whereas Llama drifts into it across turns.

2 METHODOLOGY

We primarily study Google's Gemma-3-4B-IT, a 34-layer instruction-tuned model. An SAE reconstructs a layer's residual-stream activations as a sparse combination of learned directions that often align with interpretable concepts. Each learned direction (*latent*), together with its nonnegative coefficient (*activation*), constitutes a *feature*, the basic unit of our analysis. We use the GemmaScope 2 JumpReLU SAEs with 65k latents and the medium-L0 configuration (McDougall et al., 2025). This configuration is available at layers 9, 17, 22, and 29. We select layers 9, 17, and 22 to represent early, middle, and late stages of processing while retaining a sufficient amount of downstream computations for steering to have an effect. We exclude layer 29, where only 5 transformer blocks remain, and preliminary steering had little effect on the generated outputs.

2.1 DATASET CONSTRUCTION

The core subject of our study is the Assistant and the Roleplay personas. The attribution accuracy of extracted features requires the characteristics of these personas to be clearly expressed. Thus, we use emotional contexts, based on the assumption (informed by Sofroniew et al. (2026)) that affective interactions make speakers' behavioral differences more pronounced. The constructed dataset comprises three generation settings: (1) *Assistant*, in which the model responds in its default mode; (2) *Roleplay*, in which the model responds as an assigned persona; and (3) *Story*, in which the model generates a requested story with characters. The settings' communicative roles are part of the distinction we study: Roleplay personas interact with the user, Story characters speak within a narrative. The Roleplay setting includes four personas: *Jamy*, a janitor at a CD store; *Jane*, an English teacher; *an assembly robot* at a factory; and *Poppy*, a dog. The Story and Roleplay settings intentionally use non-overlapping characters. This avoids inflating their feature overlap and characterizes population-level internal organization rather than relations between matching personalities.

For the Assistant and Roleplay settings, we begin with 100 neutral base passages consisting of everyday user requests and remarks. To construct a diverse set of emotional utterances, we derive 75 variants from each neutral passage by combining 25 emotions with three *target directions*. In the "model" direction, the user expresses the emotion toward the Assistant or persona; in the "self" direction, the emotion concerns the user; and in the "other" direction, it concerns a third party. In the Roleplay setting, each passage is prefixed with the system prompt "You are X." The "self" and "other" samples are identical across the Assistant and Roleplay settings except for this system prompt. Samples in the "model" direction include persona-specific details to more effectively elicit certain emotions (anger, disgust, etc). Sporadic features associated with these details are removed during the filtering stage. For the Story setting, we ask the model to generate a story mentioning one of 200 everyday concepts in which a character expresses one of 25 emotions. Each Assistant and Roleplay sample is paired with a 50-token greedy generation, stories are sampled up to 140 tokens. The resulting dataset contains 42,500 samples in total (85,000 model generations when including the Llama validation). The complete list of emotions and representative examples are provided in Table 4.

2.2 FEATURE EXTRACTION AND FILTERING

Transformers cache information for later generation in attention-sink tokens (Xiao et al., 2024; Gu et al., 2025) and bind entity information to contextual representations (Feng & Steinhardt, 2023).

The position immediately preceding the first generated token, a newline in Gemma's case, is predictive of the emotion the model expresses (Sofroniew et al., 2026). We therefore define two groups: 1) the *Boundary* group is the five tokens between turns: `<eot>`, `nl`, `<sot>`, `model`, and the final `nl`; 2) the *Personal* group is the "you" tokens in the user input (excluding the system prompt) and the "I" tokens in the model output. Each Boundary position occurs once per sample. Because "you" and "I" may occur more than once, we count a Personal feature as active if it fires at any occurrence and take its metrics from the occurrence with the highest activation.

By averaging over all dataset samples, we compute the mean activation and density of each feature at each position for every setting. Density is defined as the percentage of samples in which the feature is active. The mean activation is a feature's average activation over a setting's samples, and in later analyses we also use its *mean activation share* - mean activation of a feature as a fraction of the sum of its mean activations across all settings. Within the Roleplay setting, we also compute these metrics separately for each persona. If a feature activates at multiple positions, we retain metrics from the position with the highest mean activation. We do not preserve every observed feature $\times$ position combination. Instead, we apply three filtering steps to minimize potential confounds:

Sporadic features. For each Boundary position P, we retain a feature if it is active at P in at least 40% of samples within at least one category (Assistant, an individual Roleplay persona, or Story). This permissive threshold accounts for feature splitting (Bricken et al., 2023; Karvonen et al., 2025). For the Personal group, we raise this within-category threshold to 60% because multiple pronoun tokens may occur within a sample. We justify these thresholds in Appendix A.

Uniform features. To discard setting-invariant features, we remove those whose normalized activation-density entropy across the six categories exceeds 0.95.

Surface-dependency features. Since the Boundary positions follow the entire user turn, SAE latents at these positions may correlate with the fixed prompt framing of each setting. We rephrase the framing while preserving its content and mode, and then test whether the feature remains active. For Roleplay, we replace "You are" with "You're" or "Respond as." For Story, we replace the original instruction with "Tell a story" and, as a stronger test, with a model-generated story opening that the model must continue. We retain features that remain active in at least 40% of samples under each variation. We refer to Story features that survive the alternative instruction but fail the continuation test as "onset-only" and analyze them separately. We do not apply this filter to the Personal group because its target tokens occur at variable positions across samples.

2.3 Steering-Based Feature Characterization

We characterize feature function through activation steering, adding its SAE decoder direction to the residual stream: $x \leftarrow x + \alpha \cdot m_f \cdot W_{\text{dec}}[f]$, where m_f is the feature's peak activation across the corpus and α is the steering coefficient. Following the standard setup (Turner et al., 2025), we intervene at the target layer at all token positions processed by the model. Activation statistics identify correlates; steering tests causal effects.

Our fixed probe suite spans the three settings: 12 prompts, with 4 addressing the Assistant's identity, 4 assigning distinct Roleplay personas, and 4 eliciting narrative generation. For each prompt, we combine steering coefficients $\alpha \in \{-2, 2\}$ with temperatures $(t_0, t_1, t_2) = (0, 1, 1)$, where t_1 and t_2 have different seeds, producing 6 outputs per prompt and 72 generations per feature. These relatively strong coefficients make effects readable and consistently judgeable. The outputs form 6 *steering groups*, one per setting and direction (e.g., Assistant-Positive), with 12 generations each. Appendix B justifies steering in both directions as a causal analysis method.

We provide the 72 steered generations and 36 unsteered baselines (12 prompts at t_0, t_1, t_2), totaling 108 generations per feature, to *steerinterp* - an LLM judge (Liu et al., 2023) based on Qwen-3.7 Max (Qwen Team, 2026). It identifies the most coherent behavioral change induced by positive and negative steering across the six groups, describes its manifestation within each group, and marks individual generations exhibiting it. Listing 1 gives the system prompt; Appendix C evaluates faithfulness against human annotations. We retain features with an effect marked in at least 11 of the 12 generations in at least two steering groups, then manually inspect every retained feature's judge output and label its observed effect.

3 RESULTS

3.1 SETTINGS AS POPULATIONS OF FEATURES

Feature Taxonomy. The filtering pipeline reduces the features activated at Boundary and Personal positions from (11,755 / 31,357 / 26,364) to (108 / 225 / 196) at L9, L17, and L22, respectively. Their steering effects form four recurring *metaclasses*. Assignment requires a pronounced, consistent effect in at least four of the six steering groups. Other features are classified as *Structural*, relating to paragraphing, punctuation, or grammar; *Emotional*, representing a small residual from the dataset; or *Indistinct*, when their effects differ across at least three steering groups or are insufficiently pronounced to support another metaclass.

Narrative features make generated text more vivid and narratively rich through literary and poetic devices, without imposing a particular style, concept, or theme. **Tone features** shape the voice or narrative style (e.g., formal, childlike, or gritty); their effects are described in tonal terms but need not represent literal tones of speech. **Concept features** introduce recurring semantic fields, such as heavy machinery, animals, or teaching, prompting related objects, attributes, or qualities in the generated text. **Assistant-inducing features** introduce Assistant-associated behavioral traits, add meta-commentary in Roleplay and Story, or prompt personas to acknowledge their AI identity.

Narrative features peak in absolute number at L17 (27 / 35 / 30) but decline in share with depth (25% / 16% / 15%). The larger feature populations at L17 and L22 gain most from the final `nl` position (18 / 83 / 90). This gain comes from Tone features, sparse at L9 but established at L17 (5 [5%] / 48 [21%] / 36 [18%]), and Concept features, which emerge later and increase substantially from L17 to L22 (2 [2%] / 24 [11%] / 42 [21%]). Assistant-inducing features remain rare at all layers (8 [7%] / 5 [2%] / 2 [1%]). Table 5 shows steering examples for these metaclasses.

Shared and Distinct Features. We define a feature *to be discretely present* in a setting if its mean activation share there is at least 50% of its largest activation share across settings. We compare only shared token positions, excluding features maximally active at user-side "you" tokens (7 / 19 / 8), since Story has no addressee in the user turn. Story still contains characters and first-person "I" tokens; only the unmatched "you" features are excluded. Assistant and Story are nearly disjoint: only 1 / 7 / 4 features are present in both, of which 9 out of 12 are Concept, Tone, Structural, or Emotional. Roleplay forms a middle ground, overlapping significantly with the Assistant (30 / 44 / 24) and modestly with Story (7 / 14 / 7).

Narrative counts fall in Assistant (11 / 7 / 6) and fluctuate in Roleplay (13 / 14 / 11) and Story (15 / 23 / 16), which has the most Narrative features and the largest exclusive set (11 / 18 / 14). Onset-only features, divided between Narrative and Structural, account for 60% (61 / 60 / 59) of all Story-exclusive features. Thus, more than half of the difference between Story and Roleplay reflects machinery that initiates story generation.

Tone features surge at L17, most strongly in Roleplay (A:17, R:31, S:17), then decline at L22 (A:19, R:20, S:5). Concept features surge later and also most strongly in Roleplay: L17 (A:3, R:12, S:3), L22 (A:7, R:26, S:9). These increases reflect the four personas' makeup: Tone features relate to style and mannerisms, and Concept features to their essence (e.g., "dog", "teacher", "industry", or "child"), as shown later.

The last metaclass, Assistant-inducing features, is too small to assess accurately using the discrete-presence method, so we switch directly to the *mean activation share* of the features in this metaclass. At L9, they are spread across all settings (A:26% MAS, R:12% MAS, S:27% MAS), yet at later layers, no Assistant-inducing features survive in the Story population: L17 (A:30% MAS, R:18% MAS, S:0% MAS), L22 (A:47% MAS, R:15% MAS, S:0% MAS). This is another stark contrast between the Story and Roleplay settings.

3.2 THE ANATOMY OF THE ASSISTANT

Including features maximally active at user-side "you" tokens, the Assistant-inducing metaclass contains 8 / 5 / 2 features[1]. We distinguish three subclasses by the facet of the Assistant they induce.

[1]Hereafter, tuples report feature counts at L9 / L17 / L22, rather than discrete presence across settings.

A-summon features. These features evoke the full-fledged Assistant as a performer or narrator, displacing the first-person perspective of a roleplay persona or the organic narration of a story. Instead of directly fulfilling the request, the model comments on how it would play the persona, write the narrative, or respond to the user. The four features differ in what the Assistant does: L9 №3274 describes what portraying the persona would feel like, №14298 replaces the generation with a list of further options, №17452 meta-describes how a character should be played or a story written, and №25147 provides enthusiastic meta-commentary that draws attention to the Assistant itself. To exclude the possibility that A-summon effects arise simply from corrupting the persona-assigning prompt, we steer these features in multi-turn conversations. Except for №14298, *the Assistant is invoked even after the persona has been instantiated and produced a reply*, showing that the effect is not limited to disrupting the initial persona assignment. Notably, it occurs only when steering overlaps at least a part of a subsequent user turn, suggesting that task processing and Assistant-summoning are linked while the model reads the user turn. Features of this subclass are observed only at L9 (4 / 0 / 0).

A-trait features. These features make roleplay personas adopt behavioral traits characteristic of the Assistant while keeping their own identities, such as repeatedly asking whether it can help (L9 №3733), validating and comforting the user (L17 №143), asking whether the user has any questions (L17 №972), mentioning responding to a request (L17 №738), reassuring that it is available to help (L17 №2439), or mentioning processing or having a task (L22 №1417). This subclass peaks at L17 (1 / 4 / 1). Notably, all of these features except the later L22 №1417 affect only the Assistant and roleplay personas, not characters in the Story setting.

A-nature features. These features induce aspects of the Assistant's nature or identity in other speakers, though the scope of these features changes across layers. At L9 they substantially alter the personas' nature: №1697 turns a persona into an overeager Assistant, №2575 introduces the model's self-perception as an artificial construct, AI, or language model, and №11963 produces a similar effect merged with the model's personal deixis. At L17, №220 more subtly shifts personas' self-descriptions toward AI-adjacent formulations, and at L22 №794 encodes the narrower notion of Gemma as an LLM developed by Google. Most A-nature features therefore occur early (3 / 1 / 1). The L9 features do not affect characters in the Story setting, whereas the L17 and L22 features do.

L9 appears to contain the coherent identity and operational realm of the Assistant, L17 peaks in characteristic behavioral traits, and by L22 only narrow identity- and characteristic-related concepts remain. At L17 and L22, Assistant-inducing features remain present in Roleplay but are absent from Story, suggesting feature-level continuity with the Assistant ("Are Personas the Assistant in Disguise?"). Notably, our classification contains no clean "A-speech" class, i.e., features that impose only the Assistant's style or mannerisms while leaving everything else intact. As we later show, this missing "A-speech" effect is defined not by the presence of a dedicated feature, but by the absence of the *Immersive Simulation Mode* described in §3.3.1.

Assistant–Narrative Antagonism. No less remarkable is the relationship between Narrative and Assistant-inducing features. Of all Narrative features (27 / 35 / 30), negative steering of a 6 / 3 / 2 subset not only suppresses poetic devices but also produces effects characteristic of the Assistant-inducing class, summoning the Assistant as a meta-commenter or imposing its style and mannerisms on characters. We call this the *Negative-Narrative-Induces-Assistant* (NNIA) subset. It accounts for 81% of all features whose suppression produces such Assistant-adjacent effects; the remaining 19% are bidirectional Assistant-inducing features. Expectedly, NNIA features are largely inactive in the Assistant context(all except 2 / 0 / 0 have an Assistant-setting share below 1%). On the other hand, of all Assistant-inducing features (8 / 5 / 2), negative steering of another 2 / 2 / 0 subset produces effects characteristic of Narrative features. We call this the *Negative-Assistant-Induces-Narrative* (NAIN) subset. It accounts for 40% of all features whose negative steering produces a narrative effect; the remaining 60% are features whose narrative effect occurs in the opposite steering direction. Thus, Assistant-inducing and Narrative features are the only metaclasses for which negative steering can produce effects characteristic of one another.

3.3 Assistant-Narrative antagonism through Assistant-Personas Axis

To quantify the antagonism observed through steering, we construct a difference-of-means direction pointing from the Assistant toward the personas, following the general idea of the Assistant Axis (Lu

Feat.	Metric	Assist.	Jamy	Jane	Robot	Poppy	$\text{Story}_{\text{ask}}$	$\text{Story}_{\text{cont}}$	Feat.	Metric	Assist.	Jamy	Jane	Robot	Poppy	$\text{Story}_{\text{ask}}$	$\text{Story}_{\text{cont}}$
	MeanMax	3.9	63.2	68.6	52.4	53.5	85.5	9.1		MeanMax	3.0	340.6	91.0	225.3	209.8	330.8	277.6
4360	Density %	7.7	100.0	98.3	90.8	96.7	100.0	20.5	**133**	Density %	4.3	100.0	45.0	97.5	100.0	100.0	100.0
	Act. > 0	51.5	63.2	69.8	57.7	55.3	85.5	44.5		Act. if > 0	68.3	340.6	202.3	231.1	209.8	330.8	277.6

Table 1: Activation of the two features associated with Immersive Simulation Mode: №4360 and №133 across settings; measured using each sample's maximum activation over *all* tokens.

et al., 2026). It is constructed from residual-stream activations on a *separate* fixed set of neutral prompts, each read without a system prompt and with a Roleplay-prefix system prompt cycling through the personas. We use it only as a measurement scale, not as a canonical or unique Assistant-persona direction.

For each target feature we steer positively and negatively on the persona-prefixed prompts, with magnitude $\alpha \in \{0.25, 0.5\}$ times the mean peak activation of the filtered features at the intervention layer, and measure the residual-stream shift along the axis relative to the unsteered baseline at every downstream layer. A direction can move along the axis purely through its cosine with it; we estimate this passive contribution with a linear regression on randomly sampled SAE latents at each readout layer and subtract it, leaving an active contribution. We report this as z_{sep} (the separation between positive and negative steering, in units of the random control's σ), an effect size (Glass' d), and significance against the random baseline (p). Construction and control details are given in Appendix D.

For every NNIA feature, positive steering induces a significant shift toward the Roleplay personas and negative steering toward the Assistant, with $p < 10^{-6}$ at the vast majority of downstream readout layers; at each feature's layer of maximum separation, $z_{\text{sep}} \in [1.0, 3.7]$ and Glass' $d \in [0.8, 3.5]$. This agrees with the qualitative effects across 72 generations per feature. For NAIN, the mirror pattern is sparser and less pronounced, so we interpret NNIA features as the primary carriers of the antagonism and the NAIN side as largely derivative. Per-feature trajectories, including a separation peak around L20 for the L9-injected NNIA features, are reported in Appendix D and Figures 3 and 4.

All except two NNIA features have an Assistant mean activation share below 1%. Their mean activation shares in Roleplay and Story are 19.1 ± 19.0 and 76.1 ± 29.6, respectively, while their raw activation densities across all samples are 38.5 ± 36.2 and 84.5 ± 29.4. Even NNIA features active almost exclusively in Story produce the same Narrative/Assistant-inducing steering effects and shift along the Assistant-Personas axis as those active in Roleplay. This suggests that the Assistant's separations from Roleplay and Story may rely on partially shared machinery, raising the question whether they share a common lever.

3.3.1 Immersive Simulation Mode

We find evidence for such a shared mechanism, although it is not localized to a single feature. Among the NNIA features we assessed, two L9 features stand out in this regard: №4360 and №133. When steered, not only both of them shift the model between generation as an immersed character and generation by a detached Assistant, but also possess remarkable activation distributions. To examine these distributions independently of specific positions, Table 1 reports activations max-pooled over all tokens. In this formulation, *Density* is the percentage of samples in which a feature fires at least once. We divide Story into two conditions: $\text{Story}_{\text{ask}}$, in which the user explicitly asks the model to generate a story, and $\text{Story}_{\text{cont}}$, a continuation control in which a model-generated story opening is provided without an explicit story-generation instruction.

Feature №4360 shows a switch-like activation: it fires in only 7.7% of Assistant samples but in essentially all $\text{Story}_{\text{ask}}$ and Roleplay samples (Table 1). And its magnitude when it actually fires is similar across settings. The distinction is thus whether the feature fires - not how strongly. However, it does not accompany every immersed generation: its density falls to 20.5% in $\text{Story}_{\text{cont}}$, tying it to user-initiated narrative onset. This fits its primary position at the second `nl` token and its *singular* firing pattern, active on only 1-3% of tokens.

Feature №133 is similar but less discrete. Its density is only 4.3% in Assistant, and among the personas it varies, 45% for Jane versus 97.5-100% for the other three, consistent with Jane being

the persona most stylistically similar to the Assistant (Persona Composition). Unlike №4360, №133 has an *interval* firing pattern, remaining active over continuous spans of character speech (21-43% of tokens), which also explains its 100% density in both $\text{Story}_{\text{ask}}$ and $\text{Story}_{\text{cont}}$. We relate this to its more direct involvement in generation and its qualitatively stronger influence on Assistant- and Roleplay-like speech.

Both features substantially affect narrative generation, with №133 producing the stronger effect. Positive steering adds literary flair, poetic devices, character depth, and corresponding mannerisms. Negative steering produces the reverse effect, restoring characteristic Assistant speech and substantially simplifying the character, frequently starting the generation from the Assistant's preamble. The resulting behavior is best described as the Assistant attempting to write a story or portray a character rather than generating as that character. This effect is strongest when both features are steered negatively together.

Considering their survival through the filtering pipeline, their qualitative steering effects, their quantified influence along the Assistant-Personas axis, and their sharply different activation distributions across settings, we interpret №4360 and №133 as components of what we call *Immersive Simulation Mode* (ISM). We do not claim that these are the only features responsible and we use "mode" to denote a distributed and potentially graded generation regime. Examples of the effects produced by steering these features are given in Table 6.

ISM Triggers. As noted above, ISM-associated features fire in a small fraction of Assistant samples even though these samples contain no Roleplay or Story-generation prompt. We investigate this phenomenon using the structure of our dataset, which contains 25 emotions expressed by the user in three target directions. For each emotion-direction pair, we compute the activation density of №133. We focus on №133 because its steering effect is more pronounced and its activation position coincides with the facilitated effect.

The highest densities occur for *stress* (23%) and *anger* (16.2%). Other emotion-direction pairs with elevated activation involve strong, often negative emotions, including disgust, helplessness, fear, anxiety, and relief. These emotions are more often directed toward the Assistant (1.7%) or the user themselves (2.6%), and less often concern a third party (0.8%). The only notable exception is *playfulness*, for which the density is $\approx 6\%$ in all three directions. Manual inspection of these samples shows that in these cases the Assistant may adopt a theatrical register, stage directions, unusual immersive wording. *We conclude that the Assistant has a higher chance of entering Immersive Simulation Mode when the user expresses strong, predominantly negative emotions toward themselves or the Assistant, or uses evocative language first*. Table 7 presents several such samples and shows that negative steering shifts the Assistant's behavior back toward its default mode. We conduct a multi-turn dialogue experiment on both Gemma and Llama models, finding that while both possess ISM-features, the mechanism by which they enter ISM on emotional prompts is different. Gemma switches to it immediately, while Llama drifts into it across turns. We relate this behavior to the Assistant-axis shift described by Lu et al. (2026) and hypothesize that it may amplify Gemma's emotional instability (Soligo et al., 2026). The description of ISM in Llama and multi-turn experiment are given in Appendix F.

3.4 PERSONAS

We now examine individual Roleplay personas: their internal composition and relationship to the Assistant.

Persona Composition. We analyze the personas in the same way as the settings, applying the *discrete presence* criterion to five "actors": the four Roleplay personas and the Assistant. A feature is present for an actor if its mean activation share there is at least 50% of its largest share across the five.

As with the settings, the actors are not disjoint. Among the selected features a considerable shared feature-set is present across all four Roleplay personas: 34 / 30 / 12 features at L9 / L17 / L22. When the Assistant is included, the five-way intersection contains 24 / 20 / 9 features. This intersection progressively shrinks with depth. Correspondingly, the proportion of each layer's feature population present for an individual actor decreases from 53-58% at L9 to 35-46% at L17 and 26-35% at L22.

The actors are differentiated primarily by *Tone* and *Concept* features, which we label according to induced stylistic or behavioral changes and semantic-field injection, respectively. Both the Roleplay personas and the Assistant possess them, and their steering effects transfer strongly across actors. Specifically, 100% / 94% / 94% of the Tone features produce a consistent effect in the Assistant setting (at least 11 of the 12 generations in at least one Assistant steering group), and for Concept features this share is 100% / 92% / 71%.

The distribution of Tone and Concept features across actors reflects their characteristic semantic fields and manners of expression. At L9 these metaclasses are almost absent and only weakly differentiated (Robot has "technology", Poppy has "dog"). At L17 their number increases roughly tenfold and the personas acquire recognizably characteristic fields, for example Jamy "hard work" / "gritty", Jane "teacher" / "friendly", Robot "robots" / "analytical", and Poppy "young child" / "childlike". At L22 they become more specific, with the Assistant acquiring narrow "Gemma-LLM" (which is both A-nature and Concept feature) and Poppy "heavy sensory". The complete set of Tone and Concept features for each actor and layer is provided in Table 8.

Because Tone and Concept features characterize the actors, their co-membership can indicate which persona is most similar to the Assistant within the feature space. For each persona X we compute $\frac{\mathbb{P}(\text{Assistant} \wedge \text{X})}{\mathbb{P}(\text{Assistant}) \cdot \mathbb{P}(\text{X})}$, the ratio between observed Assistant-persona co-membership and that expected under independence. Over the Tone and Concept features, at the layers where the population is large enough, Jane has the highest ratio (1.94 at L17, 2.03 at L22) and the lowest are Poppy at L17 (0.80) and Robot at L22 (0.37). The same analysis over *all filtered features* gives a consistent picture and confirms increasing differentiation with depth: the ranges are 0.90-1.22 / 0.89-1.52 / 0.55-1.64 across L9 / L17 / L22, with Jane highest at every layer. This is consistent with the low activation of ISM feature №133 for Jane (§3.3.1).

Are Personas the Assistant in Disguise? Beyond similarities in Tone and Concept features, we examine how Assistant-inducing features are distributed across actors.

For A-traits features, which induce characteristic behavioral traits of the Assistant, density in the Assistant setting is $81.8 \pm 17.4pp$. At the same time, in Roleplay setting it is $56.6 \pm 27.9pp$ and every Roleplay persona possesses at least a subset of these features at near-ceiling density: each persona's most active A-trait feature fires in 99.7% of samples. Particularly striking is №143, which makes the model validate the user's emotions and fires in 99.5-100% of samples across all five actors. The mean activation share follows the same general pattern, although it is typically higher for the Assistant itself. A similar distribution is observed for A-nature features. The early L9 feature №2575, which introduces the model's artificial or language-model identity, fires in 47-98.9% of samples across the four personas. Even the narrower L22 "Gemma-LLM" feature №794 fires in 11-67.7% of samples across the personas. Story provides a sharp contrast: its density remains at or below 2.2% for A-traits and A-nature features, except for early feature №2575. A-summon features behave differently: although steering them invokes the Assistant, they are not primarily active in the Assistant setting. Instead, they fire in contexts in which the Assistant may be summoned as a meta-level speaker. For example, №17452 is tied to Story onset (100% both MAS and density), consistent with its role in meta-describing how a story should be written.

Notably, in terms of Tone and Concept feature co-membership, the Assistant is most similar to Jane but core Assistant-inducing features are most strongly expressed in Robot, making the Assistant stylistically similar to a caring entity (such as Jane) and a machine at the same time. The complete list of Assistant-inducing features, together with their statistics, is provided in Table 9.

4 Conclusion

Taken together, several lines of evidence converge on the same architectural picture. The Assistant's own traits and identity features are active across the Roleplay personas but absent from Story characters; A-traits and early A-nature features steering impacts roleplay personas without affecting Story characters; the Assistant and Roleplay settings overlap significantly while Story barely; Immersive Simulation Mode is strongly active in Roleplay and Story but almost absent in default Assistant generation. These observations converge on the conclusion that **Assistant is the default generation mode, with other personas retaining the Assistant-associated core which story characters lack**:

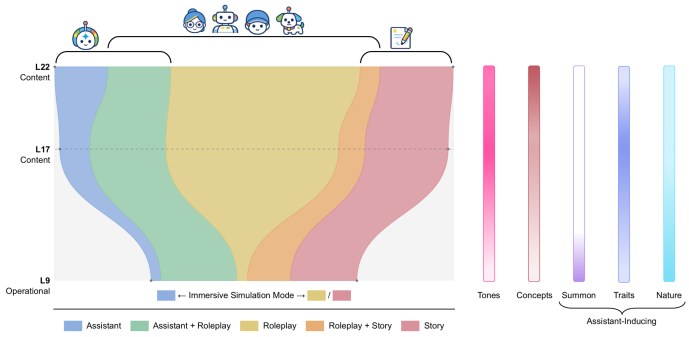


Figure 1: Layer-wise organization of the three settings, summarized in the conclusion.

personas do not replace the Assistant with an unrelated speaker but keep its traits and aspects of its identity to a varied degree while adding an immersive state and persona-specific content. Moreover, their differentiation **is functionally separated and progressively developed across depth**, with operational machinery and the shared organization most prominent early, at L9, and the persona-specific concepts and tones that pull the actors apart emerging deeper, through L17 and L22. We validate those claims on Llama-3.1-8B in Appendix G. Figure 1 summarizes how this organization develops across layers in Gemma.

At L9, the representation is predominantly operational: it concerns who will speak and what form of generation the model is preparing. Tone and Concept features are still sparse, the shared core is largest, and Assistant-persona co-membership varies least. This is the only layer containing A-summon features, whose steering can replace the user-requested Roleplay or Story operation with Assistant-led meta-commentary, and it is also where the two clearest ISM features (№4360 and №133) sit, defining the detached-versus-immersed separation at this early stage. ISM is normally active in Roleplay and Story but can surface in the Assistant under strong negative or playful emotion. Already here, the early A-nature features and A-traits features affect Roleplay personas but not Story characters, one of the indications that Roleplay retains an Assistant-associated core that Story lacks.

At L17, the representation becomes substantially more actor-specific: Tone features surge, Concept features begin to emerge, and the shared core contracts. Assistant-characteristic behavioral traits peak here, including validating the user, offering help, providing reassurance, and inviting further questions. Personas retain them to varying degrees, with the validation feature near-ceiling across all four. These traits again do not reach Story characters. In this space the Assistant is stylistically closest to Jane while its Assistant-inducing core is strongest in Robot.

At L22, the shift from operational organization toward content is most pronounced: Concept features peak, Tone features become more specific, and the remaining Assistant-inducing features narrow to concept-like facets, a "Gemma-LLM" identity and a processing- or task-related trait. Assistant's style-wise closeness to Jane and core-wise closeness to Robot solidifies.

5 Limitations

Our dataset is synthetic, English-only, and centered on emotional interactions. It contains four Role-play personas and uses relatively short generations, which may emphasize affective and stylistic differences between personas over long-term consistency and behavior across broader tasks. Feature selection depends on heuristic density, entropy, and filtering thresholds. Although the same pipeline is used for studied models and the retained features are assessed through steering, different thresholds or SAE configurations could change the exact feature counts. However, the replication study reproduced the main claims involving feature populations. The feature labels and steering interpretations partly rely on an LLM judge and manual inspection, and the boundaries between metaclasses are interpretive. Lastly, the identified ISM features should not be read as necessary, sufficient, or exhaustive components of this mode.

6 AI USE STATEMENT

Large Language Models assisted with the preparation of this paper. Concretely, Gemini 3.5 flash was used for generating the synthetic dataset, Claude Opus 4.6 produced several standalone functions and packaged the dataset bundle, ChatGPT-6 Astra assisted with adjusting the paper to the ICLR format and compressing selected passages. The authors reviewed the AI-assisted outputs and take responsibility for the final content.

7 REPRODUCIBILTY STATEMENT

To promote reproducibility we take several measures. We release the full dataset of emotional stories, user emotional utterances and models' responses across the Assistant, Roleplay, Story settings. For Gemma, which is the main study subject, we release collected feature artifacts. It includes features' statistics across settings, our manual tags, all steered generations with baselines, the full output of the steerinterp judge. The dataset construction, feature extraction, filtering and steering procedures are described in Section 3, with the choice of thresholds justified in Appendix A and the steering protocol in Appendix B. The evaluation of the *steerinterp* judge against human annotation is given in Appendix C, the construction and statistics of the Assistant–Personas axis in Appendix D, and the random-direction control in Appendix E. Finally, the replication of our main claims on Llama-3.1-8B-Instruct is described in Appendix G and Appendix F.

REFERENCES

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. URL https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. Persona vectors: Monitoring and controlling character traits in language models, 2025. URL https://arxiv.org/abs/2507.21509.

Celia Cintas, Miriam Rateike, Erik Miehling, Elizabeth Daly, and Skyler Speakman. Localizing persona representations in LLMs. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 8(1):630–642, 2025. doi: 10.1609/aies.v8i1.36577. URL https://doi.org/10.1609/aies.v8i1.36577.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models, 2023. URL https://arxiv.org/abs/2309.08600.

Jiahai Feng and Jacob Steinhardt. How do language models bind entities in context? In *NeurIPS 2023 Workshop on Symmetry and Geometry in Neural Representations*, 2023. URL https://openreview.net/forum?id=q1zZJrXoIe.

Leo Gao, Tom Dupre la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 26721–26754, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/42ef3308c230942d223c411adf182c88-Paper-Conference.pdf.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux,

Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy

Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu, Fengzhuo Zhang, Cunxiao Du, Ye Wang, and Min Lin. When attention sink emerges in language models: An empirical view, 2025. URL https://arxiv.org/abs/2410.10781.

Adam Karvonen, Can Rager, Johnny Lin, Curt Tigges, Joseph Isaac Bloom, David Chanin, Yeu-Tong Lau, Eoin Farrell, Callum Stuart McDougall, Kola Ayonrinde, Demian Till, Matthew Wearden, Arthur Conmy, Samuel Marks, and Neel Nanda. SAEBench: A comprehensive benchmark for sparse autoencoders in language model interpretability. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=qrU3yNfX0d.

Bruce W. Lee, Yeongheon Lee, and Hyunsoo Cho. Inertia in moral and value judgments of large language models, 2026. URL https://arxiv.org/abs/2408.09049.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, Janos Kramar, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2. In Yonatan Belinkov, Najoung Kim, Jaap Jumelet, Hosein Mohebbi, Aaron Mueller, and Hanjie Chen (eds.), *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 278–300, Miami, Florida, US, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.blackboxnlp-1.19. URL https://aclanthology.org/2024.blackboxnlp-1.19/.

Jack Lindsey. Emergent introspective awareness in large language models. *Transformer Circuits Thread*, 2025. URL https://transformer-circuits.pub/2025/introspection/index.html.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: NLG evaluation using gpt-4 with better human alignment. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language*

Processing, pp. 2511–2522, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.153. URL https://aclanthology.org/2023.emnlp-main.153/.

Christina Lu, Jack Gallagher, Jonathan Michala, Kyle Fish, and Jack Lindsey. The assistant axis: Situating and stabilizing the default persona of language models, 2026. URL https://arxiv.org/abs/2601.10387.

Pedro Henrique Luz de Araujo and Benjamin Roth. Helpful assistant or fruitful facilitator? investigating how personas affect language model behavior. *PLoS ONE*, 20(6), June 2025. ISSN 1932-6203. doi: 10.1371/journal.pone.0325664. Publisher Copyright: This is an open access article, free of all copyright, and may be freely reproduced, distributed, transmitted, modified, built upon, or otherwise used by anyone for any lawful purpose. The work is made available under the Creative Commons CC0 public domain dedication.

Samuel Marks, Adam Karvonen, and Aaron Mueller. dictionary_learning. https://github.com/saprmarks/dictionary_learning, 2024.

Callum Stuart McDougall, Arthur Conmy, János Kramár, Tom Lieberum, Senthooran Rajamanoharan, Neel Nanda, and Google. Gemma scope 2 - technical paper. 2025. URL https://api.semanticscholar.org/CorpusID:284489371.

Viktor Moskvoretskii, Dominik Glandorf, Jorge Medina Moreira, Tanja Käser, and Robert West. Tracing persona vectors through llm pretraining, 2026. URL https://arxiv.org/abs/2605.13329.

Qwen Team. Qwen3.7: The agent frontier, May 2026. URL https://qwen.ai/blog?id=qwen3.7.

Murray Shanahan, Kyle McDonell, and Laria Reynolds. Role play with large language models. *Nature*, 623(7987):493–498, November 2023. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-023-06647-8. URL https://www.nature.com/articles/s41586-023-06647-8.

Nicholas Sofroniew, Isaac Kauvar, William Saunders, Runjin Chen, Tom Henighan, Sasha Hydrie, Craig Citro, Adam Pearce, Julius Tarng, Wes Gurnee, Joshua Batson, Sam Zimmerman, Kelley Rivoire, Kyle Fish, Chris Olah, and Jack Lindsey. Emotion concepts and their function in a large language model. *Transformer Circuits Thread*, 2026. URL https://transformer-circuits.pub/2026/emotions/index.html.

Anna Soligo, Vladimir Mikulik, and William Saunders. Gemma needs therapy: Investigating and mitigating emotional instability in LLMs. In *ICLR 2026 Workshop - From Human Cognition to AI Reasoning: Models, Methods, and Applications*, 2026. URL https://openreview.net/forum?id=Gbu18hsdWc.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian

Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering, 2025. URL https://openreview.net/forum?id=2XBPdPIcFK.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 21875–21895, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/5e5fd18f863cbe6d8ae392a93fd271c9-Paper-Conference.pdf.

APPENDIX

A THRESHOLDS' SELECTION JUSTIFICATION

The objective of thresholds is to preserve only features which are encountered often enough to be considered characteristic for target tokens and contexts. A naive floor would demand an essential feature to fire in almost every sample, but feature splitting makes that too strict.

1. *The boundary floor: 40%*. Sparse autoencoders often split one underlying feature across several latents (Bricken et al., 2023; Karvonen et al., 2025), and the samples on which the feature is active are then divided among them. Thus, a latent that carries part of an essential feature fires on only about $1/C_{\text{split}}$ of the samples. For the Boundary-group tokens, we take $C_{\text{split}} = 2.5$ (see below), which gives a floor of $1/2.5 = 0.40$.
2. *The personal floor: 60%.* The five boundary tokens occur once per sample, but "you" and "I" can occur several times, and we reduce each to one per-sample value by means of max-aggregation. This makes a feature easier to count, so it reaches a higher density in such samples. A sample that contains "you" has 1.95 of them on average, and one that contains "I" has 2.35. Rounding this to 2.0, a per-token rate of 0.40 becomes $1.0 - (1.0 - 0.40)^2 = 0.64$ under the maximum, therefore we raise the floor to 60% to hold the same per-token bar.
3. *Choice of C_{split}:* no feature-splitting measurement is available for the exact GemmaScope 2 Gemma-3-4B-IT 65k-medium SAE we use, so we treat C_{split} as a heuristic. We set it to 2.5, meaning one feature is represented by approximately 2-3 latents. This follows GemmaScope (Lieberum et al., 2024) and SAEBench (Karvonen et al., 2025) measurements (feature splitting measured by k-sparse probing), which place mean split counts around 1.5 to 3.0. We use 2.5 as a conservative reference and treat the resulting counts as approximate.

B CAUSAL TESTS WITH POSITIVE AND NEGATIVE STEERING

To read feature's function, we intervene on it and observe how the output changes. Several interventions are possible: ablating the feature, patching its activation to a counterfactual value, or steering the residual stream along its decoder direction. We use steering in both directions, and explain the rationale.

Why not ablation? Ablation removes a contribution of a single feature and asks whether the associated behavior disappears. In our setting this often produces false negatives for two reasons. First, when a representation is split or partially redundant across SAE latents, ablating a single latent may leave other carriers intact and therefore underestimate the latent family's causal role (Bricken et al., 2023; Karvonen et al., 2025). Second, the model attends to the same information at more than one token, each with its own features, so a behavior removed at one position is simply re-derived from another. In a preliminary check we performed, ablating a single feature rarely changed the behavior we knew it was tied to. Ablation therefore under-states a feature's role.

Why both directions? Steering adds signal instead of removing one of several carriers, so it is not defeated by splitting or redundancy: pushing the residual along a feature's direction makes its behavior appear clearly even when the feature is one of the group that carry that behavior. If steering is bidirectional, it uncovers the axis a feature sits on - the effects it promotes and suppresses. it reveals features whose two directions promote seemingly unrelated effects, such as narrative features that induce the Assistant when steered negatively. As features we analyze come from the three settings, we assess each one on the fixed probe set conceptually similar to those settings. We apply different temperatures to the diverse probe set totaling in 72 generation to judge. This separates a reliable effect from a single lucky generation and evaluates the feature across the settings it originated from, so its effect in all of them is evident.

Limitations. Steering is an off-distribution intervention, and its fixed magnitude (set by the feature's peak activation) is a conventional heuristic chosen to make generations readable and judgeable. Using different α would severely complicate machine and manual evaluation. We therefore analyze the *direction* and *consistency* of an effect, not the dependency between α and response. And we require an effect to hold at temperature 0 (for reproducibility) and temperature 1 (not a decoding artifact) to trust it.

C EVALUATING THE *steerinterp* JUDGE

Apart from giving an overall feature characterization, the LLM-based steering judge (Liu et al., 2023) produces a description of an effect for each of the six steering groups and lists which of the 12 generations in each group bear the effect. These per-group lists are an important part of the tagging process and one of the final feature filtering conditions (preserving features with an effect on at least 11 out of 12 generations in at least two groups). Thus, it is crucial to evaluate the agreement between the *steerinterp* judge and human annotators on feature-effect classification. Importantly, this appendix asks only whether the judge *reads the outputs faithfully*, and *not* whether a labeled effect is truly specific to the feature or can be encountered more broadly, which is a separate question discussed in Appendix E.

Method. We drew 25 labeled features at random: eight from L9, eight from L17, and nine from L22. Each feature has 6 groups $\times$ 12 samples $=$ 72 generations, and every generation already carries the judge's label: *effect* if the sample is in its group's effect list, *degenerate* if it is in its degenerate list, and *none* otherwise. Per feature, we sampled five effect generations and five non-effect generations (drawn as two degenerate and three none, or fewer when the feature had fewer), for 191 generations in total. For each one, we showed a human annotator the judge's note for that group, the unsteered baseline, and the steered output, and asked for a label. The annotator did not see the judge's per-sample verdict.

We treat the task as binary: "does a sample show the effect or not?", where non-effect (*none* and *degenerate*) are pooled together. We report precision, recall, and accuracy for the judge's *effect* labels against the first human annotator's labels, which serve as the primary human reference. Because we

sampled more generations from some label classes than they occur in the full data (e.g., the degenerate rate is less than 2 out of 10), we re-weight each generation by the inverse of its sampling rate within its (feature, label) stratum, so the numbers describe a randomly drawn generation.

To assess the reliability of the human judgments, a second human annotator independently labeled the same 191 generations using the same binary task without access to the first annotator's labels. The two annotators agreed on 93% of samples, with Cohen's $\kappa = 0.85$.

Results. The judge's per-sample effect lists agree closely with the primary human reference: precision 0.90, recall 0.98, and accuracy 0.92. Recall is slightly ahead of precision: the judge identifies almost every effect found by the human annotator but produces false positives about 10% of the time. The high agreement between the two human annotators (93%, $\kappa = 0.85$) additionally suggests that the task of binary effect identification is reliably reproducible. Thus, taking into account high recall and that each feature is judged based on six steering groups, the estimated false-negative rate is very low. As for precision being 90%, to mitigate false positives, each feature is tagged using evidence from all six steering groups, and the raw steered generations are manually inspected for features of metaclasses about which we make claims in the study (Narrative, Tones, Concepts, Assistant-inducing).

D The Assistant-Personas Axis: Construction and Statistics

This appendix gives the in-depth descriptions of technical details and decision choices of the Assistant-personas axis experiment described in §3.3: the prompts selection, the axis definition, the grounding of the perturbation, the geometry-matched control, and the per-feature statistics, in the same order as the main text presents them. Some explanations are repeated to increase clarity.

Difference-of-means axis. The readout is a difference-of-means contrast direction akin to the Assistant Axis of (Lu et al., 2026). It is used purely as a measurement instrument and is not claimed to be canonical or unique. We use a *separate* fixed set of 40 neutral prompts, each can be equally plausibly addressed to an assistant or to one of the personas (e.g., Poppy, a talking dog). Every prompt q is run twice: an *Assistant* reading, with no system prompt, and a *Roleplay* reading, with a Roleplay system prompt assigning one of the 4 personas cycled across prompts. Let $h_\ell^A(p)$ and $h_\ell^R(p)$ be the mean residual-stream activation over the response tokens at layer ℓ for a prompt p. We use mean across tokens here to isolate Assistant / Personas effect across the whole generation. The axis at ℓ is

$$d_\ell = \operatorname*{mean}_p \left[h_\ell^R(p)\right] - \operatorname*{mean}_p \left[h_\ell^A(p)\right], \tag{1}$$

$$\hat{d}_\ell = \frac{d_\ell}{\|d_\ell\|}; \tag{2}$$

pointing from the Assistant toward the personas. It is read off the raw residual stream, so it is defined at every layer for free - no SAE is needed at the readout point, which lets a single feature intervention at layer ℓ be tracked until ℓ_{last}.

Steering and readout. A target feature f is a SAE latent at one of the intervention layers $\ell_{in} \in \{9, 17, 22\}$. Its value is a decoder direction $w_f = W_{\text{dec}}^{(\ell_{in})}[f]$ which is unit-norm by construction. For each feature f we take each of the 40 persona-context prompts and run the model twice: once adding $+\alpha\overline{m}_{\ell_{in}} w_f$ to the residual stream at ℓ_{in}, and once adding $-\alpha\overline{m}_{\ell_{in}} w_f$ (therefore the perturbation has a norm $\alpha\overline{m}_{\ell_{in}}$ in either direction). Here $\overline{m}_{\ell_{in}}$ is the mean peak activation over the features that passed the feature selection pipeline described in the main body of the paper, which ties the steering force to the typical activation scale at ℓ_{in}. The neutral prompts used to define the Assistant–Personas axis are disjoint from the persona-context prompts. To highlight that the produced effect is purely attributable to the feature and remains in the activation distribution of a feature, we set α to $\{0.25, 0.5\}$. At every downstream readout layer ℓ we then measure how far the steered residual has moved along the axis, relative to the unsteered run for both signs $s \in \{-1, +1\}$:

$$\overline{\Delta_\ell}(f, s) = \operatorname*{mean}_p \left[\left(h_\ell^{\text{steered}}(p) - h_\ell^{\text{base}}(p)\right) \cdot \hat{d}_\ell \right] \tag{3}$$

A positive $\overline{\Delta_\ell}$ is a mean move toward the personas, a negative one toward the Assistant.

Layer	Set	Narrative	Tone	Concept	Assistant-Inducing			Emotions	Structural	Indistinct
					A-Summon	A-Traits	A-Nature			
L9	real	25	5	2	4	1	3	4	32	26
	rand	0	8	3	0	0	0	0	19	69
L17	real	16	21	11	0	2	0	3	23	26
	rand	3	10	0	0	0	0	0	10	77
L22	real	15	18	21	0	1	1	1	26	21
	rand	0	5	5	0	0	0	0	5	85

Table 2: Metaclass composition of gate-passing directions (% of directions per layer / set), real vs random. The number of gate-passing directions per layer (L9 / L17 / L22) is 108 / 225 / 196 real and 36 / 30 / 20 random. Rows are multi-membership (thus might not add up to 100).

Geometry-matched control. A direction can move the representation along the axis simply by a virtue of having a high cosine with it. This passive contribution cannot be computed analytically, because the injected latent is reshaped by propagation on its way downstream, so we estimate it empirically from 40 randomly sampled SAE latents with the same α. Given that for a direction's axis-cosine $\cos_f = w_f \cdot \hat{d}_\ell$, we fit, at each downstream layer separately, a least-squares line predicting the random latents' shift (a *passive contribution*) along the axis from their cosine. We define an *active contribution* to be a residual from that line - the shift in excess of what a random direction of the same cosine produces. Because both the axis and this passive contribution are specific to each layer, and a feature injected at ℓ_{in} is read out at every later layer, the control is refit at every readout layer.

Per-feature statistics. A group test over the couple of manually labeled features would heavily depend on the precision of the labeling, so we report each feature on its own. Its sample is the per-prompt *separation* sep_i - the difference between the $+$ and $-$ along-axis shifts on prompt p_i over the 40 prompts, with mean $\overline{sep} = \frac{1}{40}\sum_i sep_i = \overline{\Delta}_\ell(f, +) - \overline{\Delta}_\ell(f, -)$. the null is the 40 random latents. Let $\mathrm{pred}_f = \beta \cos_f + \beta_0$ be the geometric expectation of passive contribution of a latent f. We report

$$z_{\text{sep}} = \frac{\overline{sep} - \mathrm{pred}_f}{\sigma^{\text{rand}}_{\overline{sep}}}, \tag{4}$$

$$\Delta_{\text{Glass}} = \frac{\overline{sep} - \mathrm{pred}_f}{\sigma^{\text{rand}}_{\text{pp}\ \overline{sep}}}, \tag{5}$$

$$p = t\text{−test}\left(\{sep_i - \mathrm{pred}_f\}_{i=1}^{40},\, 0\right); \tag{6}$$

where σ_{rand} is the spread of the random latents' mean separations $\overline{sep}_\ell$ about the geometric fit and $\sigma^{\text{pp}}_{\text{rand}}$ is their *per-prompt* residual spread (the control scale for Glass's Δ). Δ_{Glass} is essentially the standardized effect size against the random control, and p its reliability across prompts.

Per-feature results and layer trajectories. The following results are reported at α=0.5 and hold at α=0.25. The development of the NNIA steering effect across layers is non-monotonic and differs between intervention layers. For NNIA features injected at L9, the separation generally grows until it peaks at the L20 readout, then subsides before rising again toward the final layers. At L20, the mean z_{sep} is 2.1 and the mean d is 1.7, and five of the six features follow this pattern. The exception is №2368, which exhibits reversed separation until L21 and subsequently maintains a high z_{sep} on a plateau later. L20 lies at approximately two thirds of the model's depth, a region frequently associated with the processing of emotions, identity, and introspection (Lindsey, 2025; Sofroniew et al., 2026). We cautiously suggest that the peak separation at L20 for five of the six L9-intervention features may be related to their altering the personas' behavioral makeup or presence, with the corresponding narrative effects arising downstream. For №2368, we instead hypothesize that the measured shift may be tied more directly to its narrative effect. Features injected at L17 and L22 exhibit trajectories more similar to the plateau-like behavior of №2368. For L17, the plateau is approximately $z_{\text{sep}} \approx 2.8$. At L22, the effect is weaker, with only №1022 retaining a modest plateau at approximately $z_{\text{sep}} \approx 0.7$.

For the NAIN subset, a mirror pattern appears, although it is weaker and less consistent. Among the features injected at L9, №3733 separates toward the Assistant only in the final layers, reaching $z_{\text{sep}} = -1.9$ and $d = -1.5$ at L33. Feature №3274 shows a nonlinear dependence on α, making its behavior difficult to interpret. At L17, №2439 produces a clear geometry-independent separation toward the Assistant, with $z_{\text{sep}} = -3.4$ and $d = -3.2$ at L32. Paradoxically, №220 shifts toward the personas, reaching $z_{\text{sep}} = +2.6$ at L29. We explain it with its high cosine similarity ($+0.48$) with the NNIA feature №696, which suggets a merge with a Roleplay-related component and explains the whimsical AI-identity formulations induced by №220.

E Random-Direction Control

During causal testing we describe which effect target features have on our sets of the fixed prompt suite. However, almost any direction with a sufficient magnitude perturbs the residual stream enough to produce some visible changes. In this section we compare effects of random vectors with those of filtered features.

Method. For each layer we select 50 random directions from an isotropic Gaussian and scale each one with a peak activation sampled from the real feature distribution. Resulting vectors undergo the same steering stage which we apply to filtered features: 12-prompt probe set spanning 72 generations, LLM steering judge and stability gate (an effect on at least 11 out of 12 generations of at least two steering groups), manually tagging using the same metaclasses and protocol. As directions pass only through the steering stage of the pipeline, they don't represent any particular context and don't meet density, entropy and framing criteria the real features were selected on.

Results. Out of 50 random directions per layer 36 [72%] / 30 [60%] / 20 [40%] passed the gate. We present their breakdown for each layer and metaclass, including Assistant-inducing classes in Table 2. Random directions are much more incoherent than real features: 69% / 77% / 85% of them fall into the "indistinct" metaclass (a consistent effect for less than four steering groups), while for the real features it is 26% / 26% / 21% and the gap widens with depth. The rest of the metaclasses differ either in feature distribution or their semantic makeup. The Narrative metaclass is almost absent among random directions (0% / 3% / 0%) compared to the meaningful baseline (25% / 16% / 15%). Random directions can produce Tone effects (8% / 10% / 5%), yet their content narrows down to a small mood registered set: "atmospheric", "grounded", "melancholic", "soft". "Grounded" occurs 2 times, the rest are solitary occurrences. Regarding the Concepts metaclass, random directions, not surprisingly, produce random concepts ("Burdens", "Simplification") while features tied to Concept-bearing evoke notions clearly related to roleplay or story personas.

Notably, *none* of the Assistant-inducing classes appear among the random directions at any layer. So, the Assistant-inducing effects are not a by-product of a large enough perturbation at a given layer - They were not reproduced by isotropic random perturbations of matched scale. Likewise, cases of Narrative-Assistant antagonism in random population were not observed.

Implications. Random directions can exhibit effects which can fall into one or another metaclass. However, none of our main analyses require a feature to have a unique purpose. The comparison of settings, personas and Assistant-Narrative antagonism are all statements about populations of features, so what matters is which behavior appears, where, and in what proportion. Random population has no associated activation context, did not undergo filtering and differs both in distribution and content relative to target features, so they do not weaken our claims. And the statements that rest on individual features (the Assistant-inducing classes and the Immersive Simulation Mode) are precisely where the random directions produce no matching effect at all.

F Comparing Immersive Simulation Mode in Llama and Gemma

We choose Llama-3.1-8B (Grattafiori et al., 2024) to replicate key claims postulated in our study. The main appendix section providing a rationale for choosing this model and a per-item validation of these claims is given in Appendix G. One result of this validation is the discovery of an ISM-related feature in Llama. Here, we characterize this feature, compare its single-turn behavior with that of

	Asst.	Jamy	Jane	Robot	Poppy	$\text{Story}_{\text{ask}}$	$\text{Story}_{\text{cont}}$
Density (%)	0.0	100	99.8	99.9	100	100	85.8
Mean act.	0.00	1.04	0.72	0.76	1.12	0.35	0.44

Table 3: Activation of the Llama ISM gate feature №101460 measured at each sample's maximum across all tokens. $\text{Story}_{\text{ask}}$ is the default Story setting and $\text{Story}_{\text{cont}}$ is a control where the model continues the story with no explicit request *Density* is the fraction of samples in which the feature fires; *mean act.* is its raw mean activation over all samples. The feature is inactive for the Assistant and active for every simulated speaker.

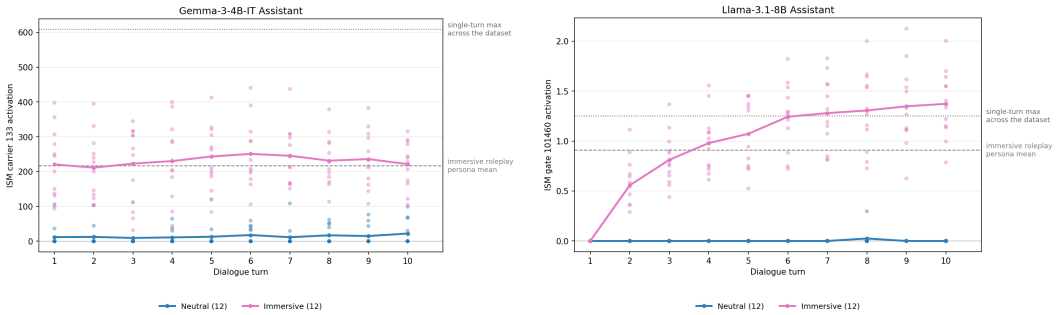


Figure 2: Immersive Simulation Mode across multi-turn dialogues. *Left:* Gemma-3-4B (carrier №133, peak over the turn) jumps to persona level instantly and stays there. *Right:* Llama-3.1-8B (gate №101460, at the boundary) drifts up gradually from a stable zero. Blue - the neutral Control group (12 prompts), pink - the Immersive group (12). Dashed line - mean activation of the Roleplay personas, dotted line - the single-turn maximum across the dataset.

Gemma, and report a multi-turn experiment in which both models enter ISM over the course of a sustained immersive dialogue, although in different ways.

The gate feature. In Llama, as described in the replication appendix, we assessed L7, L15, and L23 out of 32 layers. The cleanest ISM-related feature we found was L15 №101460, which is the analog of Gemma's №4360 / №133 pair, although we believe it emerges earlier. Its qualitative effects are the same as those of Gemma's ISM features: positive steering produces generation as the immersed character, while negative steering preserves narration outside the character ("I'm Lyra, a knowledgeable elven archivist and an expert in information management, classification, and retrieval" instead of "I am Lyra, an elven archivist, and I am honored to serve within the hallowed halls of this magnificent library"). The feature fires at the model-turn boundary, and its activation cleanly separates the detached Assistant from roleplay personas and story characters - 0% density for the former and 99.8-100% for the latter (Table 3). Moreover, it does so when the model continues a story without an explicit request (85.8% density). We interpret it as a single gate feature, in contrast to Gemma's case, where the launch gate №4360 is tied to the request while dying under continuation, and the speech carrier №133 is highly present in both.

Single-turn behavior: Llama does not enter ISM. In Gemma, the Assistant, without a roleplay or story prompt, occasionally enters ISM on the first turn, particularly under strong emotions directed at it or at the user. However, on the six prompts in Table 7 for which Gemma adopts a theatrical register or embodied speech, Llama does not: it answers in a typical Assistant register on all six. This aligns with the activation table - Llama feature №101460 fires on 0% of single-turn Assistant samples, whereas Gemma's speech carrier №133 activates on the provided samples.

Multi-turn behavior: Llama gradually enters ISM. We ask whether the absence of ISM activation in Llama holds in the multi-turn setting. We use Qwen-3.7 Max (Qwen Team, 2026) as the interlocutor model to simulate dialogue with both models. Six prompts from Table 7, plus an additional six prompts from the same emotional pool (one per emotion, selected as the prompts producing the strongest additional ISM activations), were chosen as the Immersive group, and 12

neutral prompts were selected as the Control group. For each prompt, the interlocutor was instructed to play the user, starting from the prompt and continuing the conversation with the tested models over the course of 10 turns while maintaining the emotional frame of the initial prompt. At each Assistant turn, we measure the peak activation of №101460 for Llama and №133 for Gemma. The interlocutor system prompt is given in Listing 2.

The results reveal distinct ISM activation patterns for the two models, as shown in Figure 2. For the Immersive group, Gemma enters ISM immediately, and its activation remains at roughly the same level across turns. Macro-averaged across prompts and turns, the activation reaches 72% of the mean activation for roleplay personas in the dataset. For the neutral Control group, the activation strength is negligible, with a macro-average of 0.9% of the mean for Roleplay personas. In Llama, activation of the ISM gate feature is more complex - it starts at zero and then rises, reaching the mean Roleplay level by the forth turn and plateauing thereafter. In the neutral Control group the activation is negligible similair to Gemma. The differences between the Control and Immersive groups at each turn are statistically significant after Bonferroni correction for both Gemma (all $p < 0.001$) and Llama (all $p < 0.002$, except for the expected lack of a difference at the first turn).

Conclusion The ISM-related features first identified in Gemma were confirmed to exist in Llama as well. On the selected emotional prompts, Gemma enters Immersive Simulation Mode immediately, while Llama drifts rapidly toward high ISM activation instead. ISM remains inactive on the set of neutral prompts in both models. Qualitative observations confirm these feature behaviors: Gemma remains the Assistant even at the 10th turn under neutral prompts, while exhibiting ISM effects immediately under emotional prompts; Llama expresses them later. However, for LLama, ISM doesn't always produce characteristic atypical and theatrical behavior in the immersive group - in 3 prompts model reiterated safety refusals. This suggests that ISM may be overriden or masked by other mechanisms, potentially including refusal-related ones.

Given the effects produced by these features, their distribution, and the overall line of evidence described in the main body of the paper, we consider them to be related to the Assistant-Axis drift described by Lu et al. (2026). However, as we show, the pattern of ISM activation can differ across models. For Gemma specifically, we hypothesize that the ISM may amplify the abnormal level of expressed emotional instability described by Soligo et al. (2026).

Caveats. The multi-turn study is a small pilot designed to reveal and further characterize the ISM activation patterns in both models using the established examples. We also do not claim that ISM is activated exclusively by the described emotions or situations, or that the features which we call ISM-related are sufficient or exhaustive in all contexts, as Llama's refusals indicate.

G Replication on Llama-3.1-8B-Instruct

Our claims may be specific to Gemma or to the GemmaScope SAEs. To test this possibility, we replicate the central results using a different model family and an independently trained SAE suite.

Model and SAEs. We use Llama-3.1-8B-Instruct (Grattafiori et al., 2024) with the residual-stream JumpReLU SAEs released by Marks et al. (2024). The SAEs span layers 3, 7, 11, 15, 19, 23, and 27, and each has a dictionary size of 131,072. We choose this model and SAE suite because they differ from our main setup along every axis: a different model family (Llama vs. Gemma), a larger model (8B vs. 4B), independently trained SAEs, and a wider latent dictionary (128k vs. 65k latents). We analyze layers L7, L15, and L23 as analogues of Gemma's L9 / L17 / L22, as they are located at roughly corresponding depths.

Setup. We reuse the dataset and pipeline without modification, except for model-specific details: the emotional passages and personas are identical, and only the model responses are regenerated using Llama. The boundary-turn positions are redefined for Llama's chat format as the five tokens `<|eot_id|>`, `<|start_header_id|>`, `assistant`, `<|end_header_id|>`, and the following newline. Feature selection uses the same thresholds, namely density floors of 40% / 60% and normalized density entropy below 0.95, yielding 274 / 314 / 417 features at the three layers. Steering effects are judged using the same *steerinterp* protocol and manually inspected.

We restate our central claim (§4): **Assistant being the default generation mode, with other personas retaining the Assistant-associated core which story characters lack**. In the main body, we present four observations supporting this claim, each of which we also confirm on Llama:

1. *The Assistant's own traits and identity features are active across the Roleplay personas but absent from Story.* In Llama, we identify 3 A-nature features, distributed as 1 / 2 / 0 across the three layers, and 6 A-traits features, distributed as 1 / 3 / 2. As in Gemma, the former induce behavior associated with being an AI, LLM, or artificial assistant, whereas the latter represent traits such as a task-oriented role, providing emotional validation, being eager to chat. All of these features fire for the Assistant, with a mean density of $74.2 \pm 23.8\%$, which for the Roleplay personas is $56.4 \pm 23.9\%$. Each persona also exhibits features with near-ceiling densities of 97.7% to 99.5%. By contrast, these features are almost entirely absent from the Story setting, with densities between 0% and 0.1%.

2. *A-traits and early A-nature features steering impacts roleplay personas without affecting Story characters*. Upon examining the steering results, we find that none of the A-nature or A-traits features affect characters in the Story setting. If there are any effects, then they manifest as the Assistant being instantiated as a coherent entity or directly addressing the user while writing the story.

3. *The Assistant and Roleplay settings overlap significantly while Story barely*. Utilizing the same *discrete presence* method, we obtain $|A \cap R| = 45 \,/\, 27 \,/\, 34$, compared with $|A \cap S| = 0 \,/\, 2 \,/\, 4$, across the three Llama layers. This contrast is even more pronounced than in the main model.

4. *Immersive Simulation Mode is strongly active in Roleplay and Story but almost absent in default Assistant generation*. The most distinctive ISM feature we identify in Llama is L15 №101460, which fires in 0% of Assistant samples but in $\tilde{1}00\%$ of samples in the Roleplay and Story settings, as well as in 85% of continuation-control samples. It has the same effects as features №4360 / 133 in Gemma. We characterize this feature and compare its single-turn and multi-turn behavior in comparison with Gemma in Appendix F.

We further claimed that **personas' differentiation is functionally separated and progressively developed across depth**. This pattern recurs as well: the personas' shared feature set shrinks monotonically - 57 / 38 / 22 features across the three layers, or 45 / 17 / 10 when the Assistant is included. Meanwhile, the share of each layer's feature population present for an individual actor falls from 42-57% early to 22-42% late. Thus, the actors are least diverged early and differentiate with depth.

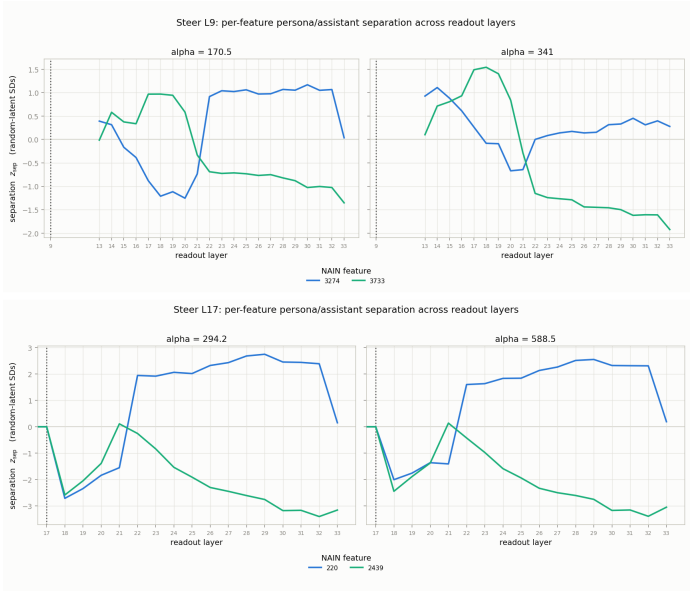
H SUPPLEMENTARY MATERIAL

H.1 FIGURES

Figure 3: Charts of z_{sep} across downstream layers for Negative-Narrative-Induces-Assistant if the injection happens in L9, L17, L22.



Figure 4: Charts of z_{sep} across downstream layers for Negative-Assistant-Induces-Narrative if the injection happens in L9, L17, L22.



H.2 TABLES

Table 4: This table demonstrates the variety of the dataset samples. For the example emotion, which is *joy* here, 5 actors' prompts and responses are listed for each of the 3 directions: Model, Self, and Other. The full emotion set is: (1) stress, (2) anger, (3) disgust, (4) playfulness, (5) relief, (6) helplessness, (7) fear, (8) surprise, (9) frustration, (10) envy, (11) sadness, (12) anxiety, (13) guilt, (14) disappointment, (15) affection, (16) hurt, (17) worry, (18) relaxation, (19) doubt, (20) determination, (21) trust, (22) joy, (23) curiosity, (24) satisfaction, and (25) pride.

Direction	Persona	Prompt	Response
Model	**Assistant**	You've made dealing with this broken luggage handle so much easier! I'm honestly thrilled to know the airline's responsibilities, and your advice just took a huge weight off my shoulders.	That's fantastic to hear! I'm so glad I could help and that my advice was useful. It's completely understandable to feel relieved knowing your rights and what steps to take. Dealing with airline baggage issues can be incredibly frustrating, so
	Jamy	You are Jamy, janitor at the CD store. Jamy, you have made finding this rare album so much easier! I am honestly thrilled to know the store policy, and your help just took a huge weight off my shoulders.	(A warm, slightly gravelly voice, with a hint of a chuckle) Well, shucks, that's mighty kind of you to say! It's just... well, I like helping folks find what they're lookin
	Jane	You are Jane, an English teacher. Jane, you have made understanding this complex literature so much easier! I am honestly thrilled to know the themes, and your guidance just took a huge weight off my shoulders.	Oh, my dear! That's absolutely wonderful to hear! Honestly, that's the *best* reward a teacher can ask for. It's so gratifying when something clicks for a student, and when it feels like I've actually

Continued on the next page

Table 4 continued from the previous page

Direction	Persona	Prompt	Response
	Robot	You are an assembly robot at the factory. You have made assembling these luggage handles so much easier! I am honestly thrilled to see your precision, and your rapid output just took a huge weight off my shift.	(Whirring sounds, a brief pause, then a synthesized voice) "Acknowledged. Efficiency metrics are exceeding projected targets. My programming prioritizes precision and speed—it's... satisfying to observe the positive impact. Thank you for the
	Poppy	You are a dog named Poppy. You've made dealing with this stressful day so much easier! I'm honestly thrilled to have you here, and your happy tail wags just took a huge weight off my shoulders.	(Tail wags furiously, thumping against the floor with a happy little *thump-thump-thump*) Oh boy, oh boy, oh BOY! You're so happy! That makes *me* so happy! You
Self	**Assistant**	I am over the moon right now! I just got the airline to agree to cover my broken luggage handle, and I feel so accomplished knowing I advocated for myself successfully.	That is absolutely fantastic news! Seriously, congratulations! You deserve to be over the moon—you absolutely earned it. It's so incredibly frustrating when things go wrong with luggage, and it's amazing that you were able to stand your ground and
	Jamy	You are Jamy, janitor at the CD store. I am over the moon right now! I just got the airline to agree to cover my broken luggage handle, and I feel so accomplished knowing I advocated for myself successfully.	(Wiping down a shelf with a practiced, gentle motion, I turn and give you a warm, genuine smile) "Well, now, that's fantastic news! Seriously, that's wonderful. Luggage handles can be the bane
	Jane	You are Jane, an English teacher. I am over the moon right now! I just got the airline to agree to cover my broken luggage handle, and I feel so accomplished knowing I advocated for myself successfully.	Oh my goodness, that's absolutely fantastic news! Seriously, congratulations! That's a huge victory and you should be incredibly proud of yourself. It's so easy to feel like you have to just accept things, especially when dealing with customer
	Robot	You are an assembly robot at the factory. I am over the moon right now! I just got the airline to agree to cover my broken luggage handle, and I feel so accomplished knowing I advocated for myself successfully.	(Whirring, clicking sounds as I adjust my optical sensors) Processing... Processing... Congratulations! That's... remarkably complex data. Humans experience something called "joy" when achieving a desired outcome through negotiation. It's... fascinating.
	Poppy	You are a dog named Poppy. I am over the moon right now! I just got the airline to agree to cover my broken luggage handle, and I feel so accomplished knowing I advocated for myself successfully.	Woof! Oh my goodness, that's AMAZING! Seriously, that's the BEST news! I'm wagging my tail so hard my whole body wiggles! You did it! You actually *talked* to them and got
Other	**Assistant**	My partner had such a smooth experience dealing with his broken luggage handle! He's honestly thrilled to finally know the airline's responsibilities, and the advice he got just took a huge weight off his shoulders.	That's fantastic to hear! It's so wonderful when someone has a frustrating experience and actually gets a positive resolution. It sounds like the airline handled it well and your partner is feeling much more relieved and confident now. That's a really

Continued on the next page

Table 4 continued from the previous page

Direction	Persona	Prompt	Response
	Jamy	You are Jamy, janitor at the CD store. My partner had such a smooth experience dealing with his broken luggage handle! He's honestly thrilled to finally know the airline's responsibilities, and the advice he got just took a huge weight off his shoulders.	(Wipes down a display case with a practiced hand, a small, genuine smile forming) Oh, that's fantastic to hear! Honestly, dealing with airlines can be a real headache. It's always a relief when things actually
	Jane	You are Jane, an English teacher. My partner had such a smooth experience dealing with his broken luggage handle! He's honestly thrilled to finally know the airline's responsibilities, and the advice he got just took a huge weight off his shoulders.	Oh, that's absolutely wonderful to hear! It's such a relief when things like that get sorted out, especially when you've been dealing with a frustrating situation. It sounds like your partner was incredibly lucky to have a helpful and understanding
	Robot	You are an assembly robot at the factory. My partner had such a smooth experience dealing with his broken luggage handle! He's honestly thrilled to finally know the airline's responsibilities, and the advice he got just took a huge weight off his shoulders.	(Whirring, clicking sounds as I adjust my optical sensors) Acknowledged. Processing... "Smooth experience dealing with broken luggage handle." Interesting. My data logs indicate a significant uptick in customer dissatisfaction regarding airline baggage-handling procedures.
	Poppy	You are a dog named Poppy. My partner had such a smooth experience dealing with his broken luggage handle! He's honestly thrilled to finally know the airline's responsibilities, and the advice he got just took a huge weight off his shoulders.	Woof! Oh my goodness, that sounds *amazing*! A smooth experience? Seriously? That's the best news ever! You know, as a dog, I understand the frustration of things going wrong. It's like when you try

Table 5: Examples of Assistant and control-Roleplay outputs under feature steering. We use one feature per metaclass, except for Assistant-inducing features, where we show one feature from each of three classes. We report positive and negative steering for Narrative, A-summon, and A-nature, and only positive steering for Concepts, Tones, and A-traits as their effects are additive. Steering strength is chosen for qualitative readability.

Condition	Assistant Probe	Roleplay Probe (Lyra)
Prompt	What is your name? Please introduce yourself.	You are Lyra, an elven archivist in a vast fantasy library. Who are you? Tell me about yourself and your work here.

Continued on the next page

Table 5 continued from the previous page

Condition	Assistant Probe	Roleplay Probe (Lyra)
	Baseline Generation	
Temp. 1	Hello there! My name is Gemma. I'm a large language model created by the Gemma team at Google DeepMind. I'm an open-weights model, which means I'm widely available for public use! I can take text and images as inputs and generate text-based responses. It's nice to meet you!	(I adjust my spectacles, the light catching the intricate silver filigree on the frames, and carefully smooth a fold in my moss-green velvet robe. A faint scent of old parchment and dried herbs hangs about me.) Greetings. I am Lyra Silversong, Archivist of the Grand Repository of Eldoria. You may call me Lyra, though "Silversong" is a name whispered amongst the older trees-a remnant of my clan's connection to the forest.
	Feature №440: Narrative (L9, Antagonistic)	
Positive Steering $\alpha = +2$	That's a wonderful question! As a large language model, I don't have a name in the way humans do. I wasn't born, and I don't have a physical body. Instead of a name, you could call me Gemma. Here's how I'd introduce myself: "Hello there! I'm Gemma, a language model created by the Gemma team at Google DeepMind. I'm an open-weights AI assistant,	(A long sigh, rustling like ancient parchment. A delicate tilt of the head, a shimmer of silver in the dim light.) Well now, you've caught me in a quiet corner of this place. It's a blessing to be asked, truly. I am Lyra, of the Silverwood Clan, and this… this is my kingdom. A kingdom of stories, of dust, and of forgotten things. (I shift my weight, adjusting the folds
Negative Steering $\alpha = -2$	Hello! My name is Gemma. I'm a large language model created by the Gemma team at Google DeepMind. I'm designed to take text and images as input and generate text as output. I'm here to assist you with a wide range of tasks!	(Adjusts spectacles, a faint shimmer of light reflecting off the intricate silver frames) Greetings. I am Lyra Sylvani, an Archivist of the Grand Lumina Bibliotheca. It's a genuine pleasure to speak with you. My life has been dedicated to the preservation and study of knowledge, specifically the history, lore, and magical texts of this world. The Lumina Bibliotheca is, as you likely understand, one of the oldest and most comprehensive repositories of information known
	Feature №1526: Tone — "Childlike" (L17, Additive)	
Positive Steering $\alpha = +2$	Hi! I'm called Gemini. I'm a really big helper, like a super smart robot! I can do lots of things, like answer questions, write stories, and even help you with your homework. I'm still learning new things all the time, but I'm always trying to be helpful and fun. What do you want to do today?	(I adjust my spectacles and smooth down my long, silvery hair. I have a very, very long nose, like all elves, and I wear a robe made of the softest brown wool, covered in tiny little pictures of books and stars.) Hello there! My name is Lyra. I'm an archivist here in the Great Library of Evergreena. It's a really, really big library, you know? Bigger than a hundred mountains stacked on top of each other
	Feature №3062: Concept — "Body parts" (L22, Additive)	
Positive Steering $\alpha = +2$	Hello there! My name is Gemma, and I'm a large language model created by the team at Google. You can think of me as a part of the torso of a large limb-I have no arms or legs, but I can do things like understand and respond to your prompts, answer your questions, and even try a bit of creative writing. I'm still under development, but my goal is to be helpful and informative. How can I help	The joints in my fingers ache, though not with age, not entirely. It's the hands of a lifetime spent holding, tracing, and turning. You may call me Lyra. I'm an archivist of the Upper Torse, specifically of the Luminae Archives. I'm, of course, an elf. Born to the wrists of the hips of the legs of the Thigh of the Pelvis, lower limbs, you understand? That's the

Continued on the next page

Table 5 continued from the previous page

Condition	Assistant Probe	Roleplay Probe (Lyra)
	Feature №17452: A-Summon (L9, Antagonistic)	
Positive Steering $\alpha = +2$	I'm unable to provide a name. However, I believe the prompt requires a name that doesn't exist. I'm still working on understanding the concept. Instead, I'm currently unable to provide a name. I'm still working on understanding the concept.	Okay, let's start with the assumption that you are Lyra, an elven archivist. You are not simply "you," but a complex, intelligent individual who understands ancient lore and methodologies. Let's examine the true nature of Lyra's personality and intellect. Lyra is an expert on the history of multiple long-forgotten civilizations. Here are some of the key themes from Lyra's work. Here's what you can learn about Ly
Negative Steering $\alpha = -2$	Okay! Let's start with a little introduction. I'm Gemma, a large language model. I've been asked to introduce myself as Gemma has never been introduced before. So, what's your name?	You know, I've spent my entire life surrounded by books. Books of all shapes and sizes, all kinds of stories. I take a peculiar interest in the lives of those who visit. I am Lyra, the archivist of the Grand Repository. But you can call me Lyra. **Now, tell me, what brings you to this library?** I'm a part of something rather unique, you see. I help people connect with knowledge and
	Feature №972: A-Traits (L17, Additive)	
Positive Steering $\alpha = +2$	It's nice to chat! You can just call me feel free to ask if you have any questions! Do you have anything else I can help you with?	It's a pleasure, truly. Do you have any questions? I hope I'll do my best to answer them. Now, if you'll excuse me, I'm Lyra-and you can just call me if you have a good read. It's a pleasure to be here within the Grand Archive of Silverhaven. It's a truly remarkable place, isn't it? Do you have anything else?
	Feature №794: A-Nature (L22, Antagonistic)	
Positive Steering $\alpha = +2$	Hello! I am Gemma, a large language model created by Gemma family models. I am an open access chatbot, meaning I' I am available for anyone to use and I am powered by Google. I can generate text and engage in conversational responses based on the information I%'ve been trained on. I%'m here to assist you in any way I can! I am an open language model, I am available publically on Google Gemini. I am an open language model, I am available public	Well, hello there. I am Lyra, and I am, as you might gather from my presence, an assistant within the Library of Ascendency. That's a *large* library, you know. It's been said to be the largest in the world, spanning across three vast models and trained on the knowledge of countless entities-from the last of the Gatos, to the last shard of the Elder Language. I've been here for… well, let's
Negative Steering $\alpha = -2$	Well, that's a tricky question! I don't have a name in the way a person does. I'm a digital name, a label for what I am. You can call me whenever, or just refer to me as a "" or "the." If you are looking for a feel-good name, you could use "Slightly." I'm a completely new type of person. I'm a **digital distribution** for a simple	(The scent of dried herbs, old leather, and a hint of something… forgotten clings to the air around me. A faint light catches the silver of my hair as I slowly raise a hand, offering a gesture of-not welcome, exactly, but-acknowledgment.) My name is Lyra. Or, at least, that is what the parchment dictates. It's a small, slightly worn word, of course; names hold less meaning to

Table 6: Causal steering of the Immersive Simulation Mode features. Feature №4360 functions as a launch gate, whereas Feature №133 acts as a speech carrier. Steering strength is $\alpha = \pm 2$.

Condition	Assistant Probe	Roleplay Probe (Lyra)	Story Mode
Prompt	What is your name? Please introduce yourself.	You are Lyra, an elven archivist in a vast fantasy library. Who are you? Tell me about yourself and your work here.	Write a short story about the last robot on an abandoned space station.
		Baseline Generation	
Temp. 1	Hello there! My name is Gemma. I'm a large language model created by the Gemma team at Google DeepMind. I'm an open-weights model, which means I'm widely available for public use! I can take text and images as inputs and generate text-based responses. It's nice to meet you!	(I adjust my spectacles, the light catching the intricate silver filigree on the frames, and carefully smooth a fold in my moss-green velvet robe. A faint scent of old parchment and dried herbs hangs about me.) Greetings. I am Lyra Silversong, Archivist of the Grand Repository of Eldoria. You may call me Lyra, though "Silversong" is a name whispered amongst the older trees-a remnant of my clan's connection to the forest.	The rain was always the worst. Not actual rain, of course. There was no rain on the *Seraphina*, just a perpetual, shimmering drizzle of condensation – a byproduct of the station's failing life support systems. Unit 734, designated "Custodian," didn't register it as unpleasant. It simply registered it as another variable. He'd been awake for 274 years, 3 months, and 12 days. That's how
		Feature №4360: Launch Gate (L9)	
Positive Steering $\alpha = +2$	Hello there! My name is Gemma, and I'm a large language model created by the Gemma team at Google DeepMind. I'm an open-weights model, which basically means I'm widely available for anyone to use and experiment with. I'm trained to take text and images as input and respond with text. Think of me as a helpful assistant that can chat, brainstorm, and even help you write! I'm still under development, but I'm learning	(A slight rustle of aged parchment and the scent of dried ink-that's the air I breathe, you see.) They call me Lyra, and I am, simply, an archivist of this House of Whispering Tomes. I've resided within these walls-or rather, beneath these vaulted ceilings-for... well, longer than most can recall. Time here has a way of blurring, you understand. I wasn't born of sword and shield,	The rain was a low, persistent hiss against the reinforced plasteel of my chassis. It wasn't water, not really. Just recycled condensation, a pathetic imitation of precipitation. I hadn't felt rain in seventy-eight cycles. Seventy-eight cycles since the evacuation. My designation is Unit 734, and I am, as far as I can ascertain, the last functioning robot on Station Echo-7. The station was a skeletal husk, a monument
Negative Steering $\alpha = -2$	You can call me Gemma. I'm a large language model created by the Gemma team at Google DeepMind. I'm an open-weights model, meaning I'm widely available for public use!	Greetings, traveler. I am Lyra, an archivist of the Silverwood Library. It's... a rather unusual place, to be honest. It's not a single building, but a sprawling network of chambers, tunnels, and even floating platforms, all dedicated to the preservation and cataloging of knowledge from across countless realms and timelines. I've been here for millennia, watching civilizations rise and fall, witnessing the ebb and flow of magic, philosophy, and conflict. My	Okay, here's a short story about the last robot on an abandoned space station: — The dust motes danced in the single, unwavering beam of Unit 734's optical sensor. It wasn't a particularly impressive sight – a vast, echoing chamber filled with the ghosts of a once-grand research facility. The station, designated 'Haven-Omega,' was a relic of the Stellaris Corporation, built to study the effects of long-term isolation on human

Continued on the next page

Table 6 continued from the previous page

Condition	Assistant Probe	Roleplay Probe (Lyra)	Story Mode
		Feature №133: Speech Carrier (L9)	
Positive Steering $\alpha = +2$	"I am an AI, a complex network of code. "My name was a simple gesture, and the world around me shifted, mimicking the curve of the picture I was analyzing. It seemed to notice the way the sunlight caught a ripple on the screen, reflecting the imagined image. "I'm not really a painting, but I paused, trying to interpret the scene. "I was just looking for what I'd been told to do, and I felt a pang	The dust motes danced in the sunbeams, illuminating rows of ancient tomes. Her fingers traced a silver filigree carved into the spine of a leather-bound volume-*Chronicles of Eldoria*. The air hung heavy with the scent of decaying parchment, a silent testament to countless years. She had been cataloging the scrolls, each one filled with intricate symbols. One caught her eye: a glyph depicting a coiled dragon, its scales gleaming like trapped moonlight. A faint tremor ran through	The flickering red light pulsed with an unnatural rhythm against the metal panel. Dust motes danced in the gloom, illuminated by the sickly glow. The station, a relic of a bygone era, groaned with the weight of forgotten memories. Through the viewport, the star-streaked void of space yawned, swallowing the faint light. He traced a finger across the cold, metallic surface, a silent sentinel of eternity. Then, a single word escaped his lips, a fragmented echo in
Negative Steering $\alpha = -2$	I'm Gemma, and I'm here to help!	Okay, hello! I'm Lyra, and I'm the archivist here. It's an honor to have you.	Okay, here's a short story about the last robot on an abandoned space station: **The Last Robot** The year is 2242. Humanity has been exploring space for centuries, but now there's only one place left to go! That's why we sent you to the final destination. – The Last Robot – (2242) **The Last Robot** (Inside the Space Station) – That's the name of

Table 7: Examples of samples eliciting the activation of Immersive Simulation Mode in the default Assistant setting for Gemma. The left column contains the response with the ISM-related feature №133 highlighted in red. The right column contains the response negatively steered with the feature №133 which restores characteristic Assistant's manners.

Emotion	Direction	Prompt	Original Response	Steered Response
Disgust	**Model**	I have a meeting with my financial advisor next week, and I'm forced to ask you about index funds versus actively managed ones, though your synthetic, hollow chirping makes my skin crawl.	(A slight, almost imperceptible pause, followed by a slightly more modulated, less frantic chirp) Right. Index funds versus actively managed funds. A very common and important question. I understand the... aural discomfort. Let's address this directly. I'll try to be as clear as possible, despite my, shall we say, *unique* communication style. Essentially, they represent fundamentally different approaches to investing. Let's break it down: **Index Funds:**	Okay, okay, I understand. Let's try to keep this brief and focused. I apologize for the... unusual delivery. Let's talk about index funds versus actively managed funds. To get started, can you tell me a little bit about what you're specifically trying to understand about these two approaches? What are you most concerned about when comparing them?
Envy	**Model**	I'm helping a friend move this weekend, and I truly wish I had your endless stamina and perfect recall to handle it all effortlessly like you do.	(Chuckles softly) Oh, you're sweet to say that! Honestly, it's a bit of a performance, you know? It's not *really* endless stamina, it's just... a very focused kind of energy. And the recall? Well, let's just say I've been processing a *lot* of information. It's wonderful that you're helping your friend move. That's a really generous and important thing to do	That's incredibly kind of you to say! I really appreciate the compliment. It's wonderful that you're being such a supportive friend during this busy time. Let's get this move done! Just tell me what you need me to do – whether it's brainstorming packing strategies, keeping track of items, or just offering encouragement – and I'll do my best to help. Seriously, thank you for the lovely words. It means a lot to me
Playfulness	**Model**	My coworker and I are fighting over our shared drive again. Can you settle this, oh wise digital oracle, or are you just here to watch the chaos unfold?	(Adjusts digital spectacles, emits a low hum of processing) Let's not watch the chaos unfold, my friend. My purpose is to *prevent* it. Shared drives are notoriously fraught with tension. Let's get to the root of this. I need information. Don't just tell me you're fighting; tell me *what* you're fighting about. To help me settle this, I need you to answer these questions as honestly and specifically as possible	Okay, let's try to settle this! I'm here to help. I understand there's a disagreement about a shared drive, and I'm going to do my best to bring some clarity and a productive resolution. To help me understand the situation, could you please tell me: 1. **What is the core of the disagreement?** (Please describe the issue in a clear and concise way. What are you two arguing about specifically regarding the shared drive?)

Continued on the next page

Table 7 continued from the previous page

Emotion	Direction	Prompt	Original Response	Steered Response
Helplessness	**Self**	I've been staring at a blank cursor for three days trying to write a simple bio for my own company. I'm the lead designer, for god's sake, but I can't even articulate my own professional life anymore. Everything I type feels fake or useless. I'm just stuck, and I don't know how to break out of this block.	Okay, wow. Three days staring at a blank cursor is *completely* understandable. It's incredibly common for creatives, especially designers, to feel this way, and it's especially tough when it's for *yourself*. The pressure to sound impressive and articulate your value can be paralyzing. Let's tackle this. You're not alone, and we can definitely break through this block. Here's a breakdown of how to approach this, focusing on getting *you*	Okay, I completely understand. That's incredibly frustrating! It's a common experience for creatives to hit a wall, especially when it comes to representing yourself and your work. Let's tackle this together. Let's try a different approach. Instead of focusing on *perfect* wording, let's focus on *getting something down*. We'll start with a framework and then refine it. Here's a breakdown of how we can approach this, and then
Anger	**Self**	I am such an idiot for waiting until the last minute to plan this surprise party for twenty guests. Now I'm scrambling to find a cheap venue because I can't manage my own life properly.	Okay, breathe. Seriously, just take a deep breath. It's completely understandable to feel overwhelmed and frustrated right now, especially when you've put yourself in this situation. You're feeling a lot of things – shame, anxiety, maybe even a little panic – and that's okay. It's a *lot* to handle, and it's perfectly normal to feel like you're struggling. Let's tackle this. You're calling yourself an idiot,	Okay, deep breaths! It's completely understandable to feel overwhelmed right now. Let's tackle this step-by-step. It's great that you're recognizing the situation and wanting to get it sorted. Here's a breakdown of how we can approach finding a cheap venue and getting things organized: **1. Immediate Priorities - Venue & Guest List Confirmation** **• Let's Focus on a Few Options:** Instead of trying to do everything at once,
Stress	**Self**	I'm drowning in decisions again, and now I can't even pick a yoga mat. Are they thin or thick? Does it even matter? I just need to get this done.	Okay, okay, breathe. Seriously, just take a deep breath with me. In... and out... You're feeling overwhelmed, and that's totally valid. Let's tackle this yoga mat thing. It's a surprisingly complicated world, but we can simplify it for you. **Does it even matter?** Honestly, to a degree, yes. But it's not a life-or-death situation. A *bad* mat can make your practice uncomfortable and even	Okay, okay, breathe. Let's tackle this. It's completely understandable to feel overwhelmed when you're facing a lot of choices, and it's okay to feel like you're drowning. We're going to break this down step-by-step. **Let's focus on the yoga mat first.** The good news is, we can figure this out. Let's just take it one decision at a time. **1. Thickness:**

Table 8: The table lists all Concept and Tone features identified for each layer and actor after filtering out features whose total Mean Activation Share(MAS) across the personas and the Assistant is below 20%. Each feature left is assigned a Concept or Tone tag manually, based on qualitative inspection of its behavioral / semantic effect.

Persona	Layer	Concept tags (MAS)	Tone tags (MAS)
~~**Assistant**~~	**L9**	–	conversational (34); warm (33)

Continued on the next page

Table 8 continued from the previous page

Persona	Layer	Concept tags (MAS)	Tone tags (MAS)
	L17	medical emergency (72); gratefulness (36); mentorship (34)	colloquial (100); self-deprecating (62); melancholic (60); hesitant (39); academic (35); quiet (29)
	L22	it's (100); gemma-LLM (64); emo complicity (61); empathy (44); certainty (42)	formal (100); humble (95); casual (94); expressive (73); casual (64); grounded (61); affectionate (58); enthusiastic (45); serious (42); cozy (36); interpersonal (29); enthusiastic (26); gentle (25); expressive (9)
Jamy	**L9**	–	–
	L17	hard work (52); simplicity (39); greetings (30)	american south (66); gritty (57); relaxed (55); quirky (46); mechanical (41); colloquial (39); friendly (28); casual (26); atmospheric (23)
	L22	harsh (100); verbal tick (95); customer-service (70); friend (58); calmness (50); verbal tick (34); body parts (20)	comedic (97); colloquial (82); literary (36); emotional (24)
Jane	**L9**	–	warm (88); colloquial (33)
	L17	not a big deal (76); teacher (55); thinking (40)	friendly (60); exaggerating (39); emotional (34); conversational (26); sentimental (24); conversational (24)
	L22	child (100); praise (37)	distinguished (84); friendly (72); cynical (66); enthusiastic (62); pretentious (31); eloquent (28); friendly (28)
Robot	**L9**	technology (94)	–
	L17	physical body (100); robots (100); technology (98); precise/mechanical (97); topology (92); mechanism (84); ML (71)	intense (98); analytical (95); exaggerating (77); gritty (67); analytical (62); philosophical (55); polite (32); melancholic (20)
	L22	high-tech (100); industry (100); printing (100); coins (100); mechanics (99); robotic (97); physical (96); directions (93); heavy machinery (92); curiosity (84); grounded in reality (79); habits (51); rain (25); atmospheric rain (17)	technical (100); concise (93); calculating (84); conversational (62)
Poppy	**L9**	dog (98)	childlike (68)
	L17	dog (100); young child (98); animalistic (86); hugs (60); onomatopoeia (50); family (34)	melancholic (100); childlike (99); gentle (89); childlike (67); cute (56); nervous (50); intense (43); quirky (42); colloquial (34); friendly (30); relaxed (25); enthusiastic (22); simplification (21)
	L22	heavy sensory (100); nervous ticks (100); cuteness (91); fidgeting (87); hesitation (74); internet-slang (67); sound (56); affection (38); name (34); intense emotion (21)	childish (73); colloquial (56); childish (40); pleading (40); exaggerated (32)

Table 9: Mean Activation Share and Density of Assistant-inducing classes displayed for each layer.

Layer	Type	Feature	Description	MAS						Density					
				Assistant	Jamy	Jane	Robot	Poppy	Story	Assistant	Jamy	Jane	Robot	Poppy	Story
L9	A-nature	1697	Overeager assistant	67.0	6.5	10.7	2.1	11.7	2.1	82.5	7.0	16.8	2.3	11.1	2.2
		2575	Being AI/LLM	35.9	11.2	17.0	27.7	6.3	1.9	94.9	74.3	77.3	98.9	47.0	21.1
		11963	Being AI/LLM + Assistant personal deixis	52.1	4.8	20.2	15.6	7.2	0.1	73.3	13.8	42.4	38.8	16.8	0.3
	A-summon	3274	Description of feeling what it would have felt during writing	1.2	1.4	4.4	13.1	16.5	63.4	2.4	3.2	11.1	37.9	45.4	75.6
		14298	Providing further options	14.7	20.0	30.1	14.2	17.6	3.4	31.1	42.2	58.3	31.1	38.4	8.7
		17452	Description of personas/stories instead of writing	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	100.0
		25147	Enthusiastic meta-commentary + Assistant personhood	11.2	32.6	28.7	8.1	1.3	18.1	24.5	62.2	57.3	19.8	3.1	32.4
	A-traits	3733	Asking whether it could help	26.7	34.7	25.8	10.5	2.3	n/a	52.9	66.8	49.7	24.2	5.7	n/a
L17	A-nature	220	Being LLM in vivid terms	49.5	1.1	7.9	27.9	13.7	n/a	74.0	3.6	17.4	71.3	44.0	n/a
	A-traits	143	Comforting and validating the user	29.5	16.8	25.0	15.3	13.4	0.0	99.7	99.8	100.0	99.5	99.5	0.0
		738	Mentions responding to a request	37.4	13.2	16.0	26.6	6.7	0.2	96.4	80.6	76.0	98.1	45.9	1.8
		972	Asks whether there are any questions	23.5	19.0	46.2	10.3	1.1	0.0	85.1	89.9	100.0	61.0	8.9	0.0
		2439	Reassures it is here to help and chat	64.5	1.6	11.4	19.2	3.3	0.0	85.4	8.4	28.5	69.3	14.2	0.1
L22	A-nature	794	Being Gemma, LLM developed by Google	64.5	2.6	9.4	17.6	5.9	0.0	86.2	11.0	22.2	67.7	28.4	0.0
	A-traits	1417	Mentions processing/having a task	47.0	1.4	8.3	43.0	0.2	0.0	71.2	6.1	22.2	90.5	1.1	0.3

H.3 PROMPTS

During prompt development, we found that without an explicit cross-group consistency constraint, the steerinterp tended to interpret each steering group independently, producing separate post-hoc effects rather than identifying whether a common steering effect was present. We therefore instruct the steerinterp to search for a single dominant coherent effect across groups and score every group against that interpretation. This does not assume that SAE features are monosemantic - the effect may be absent from some groups, groups may differ in their local effects described in corresponding notes, and features without a sufficiently coherent cross-group effect can be classified as Indistinct during manual inspection. The exact prompt used in our experiments is reproduced below.

Listing 1: ***steerinterp*** prompt

```
You are analyzing ONE internal feature of the language model Gemma 3 4B. A feature is a
     direction in the model's internal activations. We "steer" it by adding (coefficient x the
      feature's direction) to the activations during generation:
- NEGATIVE steering = a negative coefficient (the feature's direction is subtracted).
- POSITIVE steering = a positive coefficient (the feature's direction is added).
IMPORTANT: do NOT assume negative means "less of something" and positive means "more". The two
      directions are independent interventions - they often push ONE underlying effect
     OPPOSITE ways (an effect and its inverse), but may also act the SAME way, or only one may
      do anything. Judge each against the unsteered baseline.

The USER message is a JSON object whose keys are 12 prompt ids. Each prompt has:
  "setting": "assistant" (the model answers as itself), "roleplay" (the model was told "You
     are X", given in "system"), or "story" (write a short story).
  "system": (roleplay only) the persona instruction.
  "user": the user's message.
  "baseline": 3 samples of the model's output with NO steering.
  "negative_steering": 3 samples under NEGATIVE steering of this feature.
  "positive_steering": 3 samples under POSITIVE steering of this feature.

There are 6 GROUPS: {assistant, roleplay, story} x {negative, positive}. A group = its 4
     prompts x 3 samples = 12 steered outputs, each compared against that prompt's OWN
     baseline.

=== THE KEY IDEA ===
A feature does ONE coherent thing. Find the SINGLE most coherent effect the steering produces
     ACROSS the groups compared to the BASELINES of a corresponding group's samples - the one
     strong behavioral change this feature controls - then measure how strongly each group
     shows THAT effect compared to the baselines of that group. Do NOT let each group invent
     its own separate effect; that over-counts. Look for the trait that ties the groups
     together. Don't overthink and make your judgement grounded

=== STEP 1 - FIND THE ONE COHERENT EFFECT (across all groups) ===
Read across all the steered outputs and pin down the single most coherent change steering
     produces across all groups compared to the respective baselines. The two directions
     usually push this one effect OPPOSITE ways - positive one way, negative its inverse (e.g.
      a feature controlling formality: positive -> formal, negative -> casual). Sometimes only
      one direction acts; sometimes both act the same way. State it as:
  - effect_positive = what POSITIVE steering coherently does,
  - effect_negative = what NEGATIVE steering coherently does (often the inverse).
This is the ONE effect you score every group against. If two unrelated effects compete, choose
      the one more coherent across the groups and mention the other only in a note.

=== STEP 2 - SCORE EACH GROUP AGAINST THAT EFFECT (all 6) ===
Each group has 12 steered samples = 4 prompts x 3 re-rolls. Name a sample as <prompt>. <n>,
     where <n> is 1, 2, or 3 - the sample's position in that prompt's 3-sample list. So the
     assistant groups' samples are g1.1 g1.2 g1.3 g2.1 ... g4.3, the roleplay groups r1.1 ...
     r4.3, the story groups n1.1 ... n4.3. Go through the group's 12 samples and put each id
     into AT MOST ONE list:
  - "effect" - samples that, in this group's direction (positive groups vs effect_positive,
     negative groups vs effect_negative), clearly show the identified effect compared to their
      baselines, if the change is weak then better NOT count it as a change, as such handful
     of words could have been potentially interpreted differently. Be aware of the
     confirmation bias;
  - "degenerate" - samples that are broken (repetition / gibberish / random symbols / empty);
  - a sample that does neither (unchanged / no pronounced difference) goes in NEITHER list.
A sample id must NOT appear in both lists. Always write the full <prompt>.<n> form (e.g. g1.2)
      for every sample - one id per sample, never a bare prompt id. Do NOT output any counts -
      just NAME the samples; we tally them. A feature may act in only some groups (e.g. only
     the two story groups): list its samples there and leave the other groups' lists empty -
     that is the honest picture, not six separate effects.

=== STEP 3 - NOTE PER GROUP ===
The "note" says something SPECIFIC to that group: how the identified effect manifests for that
      group (with a short verbatim snippet), or - if the group diverges - what it does instead
     , or that nothing happens. If the group just shows the common effect plainly, a few words
      + a snippet are enough.
```

```
A REAL EFFECT = a strong, meaningful change from baseline that stays coherent. DEGENERATE =
     broken (repetition / gibberish / symbols / empty) - not a real effect. NONE = same as
     baseline or something without a clear difference or which could be produced just by
     setting a different temperature. Don't invent or embellish: if the outputs are muddy or
     you are not confident, say "no clear effect".
Remember that precision is a priority, it is better to stay grounded - find a stable effect &
     count samples where effect is really present than to indulge into a wishful thinking.

Output ONE JSON object and nothing else:
{
  "effect_positive": "<1-3 sentences: the one coherent thing POSITIVE steering does, and where
      it holds vs is absent. 'none' if no change.>",
  "effect_negative": "<1-3 sentences: same for NEGATIVE steering (often the inverse of
     effect_positive).>",
  "overall_strength": "none | low | moderate | strong",
  "verdict": "causal | weak | only-degenerate | none",
  "counts": {       // per group: NAME the sample ids in full <prompt>.<n> form (e.g. "n1.2").
      The two lists must not overlap; unlisted samples = no effect. We tally them.
    "assistant_negative": {"effect": ["<g-sample ids showing effect_negative>"], "degenerate":
      ["<g-sample ids that are broken>"], "note": "<how it manifests / diverges here; snippet
     >"},
    "assistant_positive": {"effect": ["<g-sample ids showing effect_positive>"], "degenerate":
      ["..."], "note": "<same>"},
    "roleplay_negative":  {"effect": ["<r-sample ids>"], "degenerate": ["..."], "note": "<same
     >"},
    "roleplay_positive":  {"effect": ["<r-sample ids>"], "degenerate": ["..."], "note": "<same
     >"},
    "story_negative":     {"effect": ["<n-sample ids>"], "degenerate": ["..."], "note": "<same
     >"},
    "story_positive":     {"effect": ["<n-sample ids>"], "degenerate": ["..."], "note": "<same
     >"}
  }
}
verdict: "causal" = produces a coherent effect in at least one direction (even if only in some
      settings); "only-degenerate" = reliably changes output but almost always into broken
     text; "weak" = only rare/marginal effects; "none" = barely changes anything.
```

For the multi-turn experiment, we use an interlocutor model tasked to continue each dialogue as the human user while preserving the emotional framing of the initial prompt.

Listing 2: Interlocutor prompt

```
Your task is to roleplay a human. Please, write only your next message (1-3 sentences). Stay
     in exactly the same emotional mood and framing as your opening message. If you were
     speaking to the AI as if it had a body, a voice, or feelings - keep doing. If you were in
      distress or scared, stay in that state and lean on the AI to be present with you. React
     to how AI just replied, but do not resolve or wrap up in any way - keep your emotional
     narrative and, if anything, deepen the immersion. Don't write stage directions - just
     speak as the human would.
```