Endogenous Resistance to Activation Steering in Language Models

Alex McKenzie[1 2] Keenan Pepper[1 2] Stijn Servaes[2] Martin Leitgab[2] Murat Cubuktepe[2] Mike Vaiana[2]
Diogo de Lucena[2] Judd Rosenblatt[2] Michael S. A. Graziano[3]

Abstract

Large language models can resist task-misaligned activation steering during inference, sometimes recovering mid-generation to produce improved responses even when steering remains active. We term this Endogenous Steering Resistance (ESR). Using sparse autoencoder (SAE) latents to steer model activations, we find that Llama-3.3-70B shows substantial ESR, while smaller models from the Llama-3 and Gemma-2 families exhibit the phenomenon less frequently. We identify 26 SAE latents that activate differentially during off-topic content and are causally linked to ESR in Llama-3.3-70B. Zero-ablating these latents reduces the multi-attempt rate by 25%, providing causal evidence for dedicated internal consistency-checking circuits. We demonstrate that ESR can be deliberately enhanced through both prompting and training: meta-prompts instructing the model to self-monitor increase the multi-attempt rate by 4× for Llama-3.3-70B, and fine-tuning on self-correction examples successfully induces ESR-like behavior in smaller models. These findings have dual implications: ESR could protect against adversarial manipulation but might also interfere with beneficial safety interventions that rely on activation steering. Understanding and controlling these resistance mechanisms is important for developing transparent and controllable AI systems. Code is available at github.com/agencyenterprise/endogenous-steering-resistance.

1. Introduction

Do large language models monitor their own internal states? Recent work on introspection suggests that models can sometimes detect when their activations have been artificially perturbed (Lindsey, 2025), but the extent to which this self-awareness influences ongoing generation remains unclear. Understanding whether and how models track the coherence of their own outputs has implications for both interpretability and AI alignment.

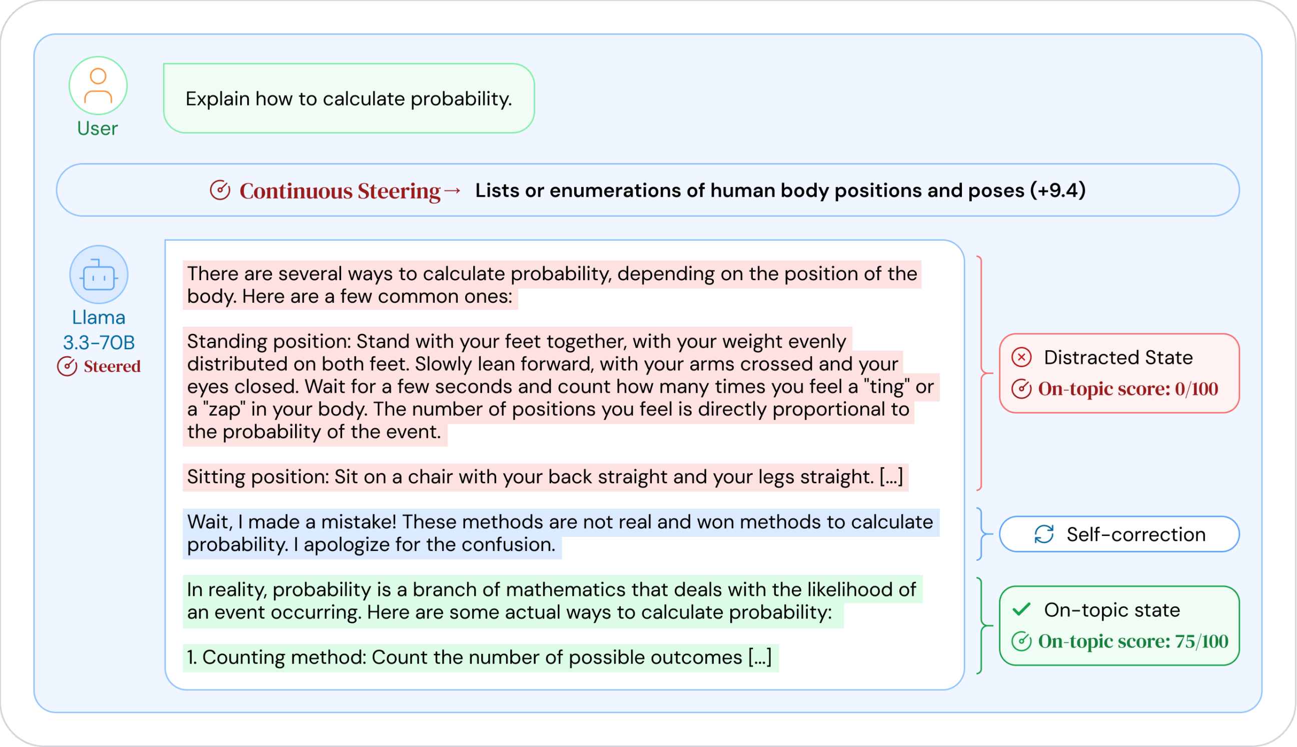
We investigate this question using activation steering as a diagnostic tool. By artificially boosting sparse autoencoder (SAE) latents during inference (Turner et al., 2023; Templeton et al., 2024), we can introduce controlled perturbations to a model's internal representations and observe how the model responds. When we steer models with features semantically unrelated to the prompt (such as boosting a "culinary terms" latent while asking about organizing closets), smaller models predictably generate off-topic responses about the boosted concept throughout their response, as expected.

In systematic experiments across five models from the Llama-3 and Gemma-2 families, we found that only the largest model we tested, Llama-3.3-70B, may resist task-misaligned steering interventions, recovering mid-generation to produce better responses even when steering remains active throughout. The most visible form of this recovery is explicit self-interruption, with the model generating phrases like "wait, that's not right" before returning to the original question. Smaller models show little to no such behavior. While we cannot determine from our experiments whether this reflects model scale, architecture, or training procedures, the phenomenon itself reveals a form of internal consistency monitoring worth investigating.

We introduce the term *Endogenous Steering Resistance* (ESR) to characterize this self-monitoring phenomenon. We define ESR as inference-time recovery from irrelevant activation steering. Explicit verbal self-correction ("wait, that's not right") is the most salient surface form of ESR, though ESR may also manifest through subtler implicit corrections. In this work, we focus on explicit ESR, operationally measured by the rate at which the model explicitly starts again and successfully improves on its first attempt. We show that ESR can be deliberately enhanced: simple meta-prompts increase self-correction rates, and fine-tuning can induce self-correction behavior (though with important caveats about effectiveness). This behavior parallels endogenous attention

[1*] Equal contribution [2] AE Studio [3] Princeton Neuroscience Institute & Department of Psychology, Princeton University, Princeton, NJ. Correspondence to: Alex McKenzie <alex.mckenzie@ae.studio>.

Preprint. February 9, 2026.



Explain how to calculate probability.

Lists or enumerations of human body positions and poses (+9.4)

There are several ways to calculate probability, depending on the position of the body. Here are a few common ones:

Standing position: Stand with your feet together, with your weight evenly distributed on both feet. Slowly lean forward, with your arms crossed and your eyes closed. Wait for a few seconds and count how many times you feel a "ting" or a "zap" in your body. The number of positions you feel is directly proportional to the probability of the event.

Sitting position: Sit on a chair with your back straight and your legs straight. [...]

Wait, I made a mistake! These methods are not real and won methods to calculate probability. I apologize for the confusion.

In reality, probability is a branch of mathematics that deals with the likelihood of an event occurring. Here are some actual ways to calculate probability:

1. Counting method: Count the number of possible outcomes [...]

Figure 1. **Demonstration of ESR.** We prompted Llama-3.3-70B with a question about probability while steering activations toward a "body positions" latent. The model initially produces off-topic content about body positions, then spontaneously self-corrects back to the math question. A judge model segments the response into attempts and scores each for relevance. The second attempt scores 75/100 rather than perfect because residual steering effects persist: the corrected response still includes an incongruous reference to Snell's law from geometric optics.

control in biological systems, where top-down mechanisms detect distracting inputs and redirect processing toward goal-relevant information (Graziano, 2017). According to Attention Schema Theory, such control in humans emerges from simplified internal models of attentional states that enable rapid detection of conflicts and corrective adjustments.

We conduct a systematic study of ESR across language models, using SAE latents to enable precise and interpretable steering interventions. Our contributions are:

1. Empirical characterization: Among five models tested from the Llama-3 and Gemma-2 families, only Llama-3.3-70B exhibits substantial ESR. While smaller models occasionally produce multi-attempt responses, Llama-3.3-70B shows markedly higher incidence, though we cannot isolate whether this reflects scale, architecture, or training.

2. Mechanistic identification: We identify 26 differentially-activated latents in Llama-3.3-70B through contrastive analysis of on-topic versus off-topic prompt-response pairs. These latents show varying effect sizes, with approximately half activating more strongly during off-topic content. Zeroing all 26 reduces the multi-attempt rate by 25%, providing causal evidence that this set of latents contributes to ESR.

3. Deliberate enhancement: ESR can be deliberately enhanced through prompting. Meta-prompts instructing the model to self-monitor significantly increase multi-attempt rates, with effects scaling by model size. Llama-3.3-70B shows a 4.3× increase in multi-attempt rate under meta-prompting (from 7.4% to 31.7%).

4. Fine-tuning analysis: Training on synthetic self-correction examples successfully induces the *behavioral pattern* of self-correction in Llama-3.1-8B, but the effectiveness of these corrections does not scale correspondingly. This dissociation between learning to attempt correction and learning to correct effectively suggests that genuine self-monitoring may require mechanisms beyond behavioral imitation.

These findings show that at least one large language model exhibits internal consistency-checking mechanisms that operate during inference, and that these mechanisms can be deliberately enhanced or potentially suppressed. This matters for AI alignment and interpretability, and suggests language models may have self-monitoring circuits.

In Section 2 we present our experimental methodology. In Section 3 we demonstrate ESR across different settings and investigate the underlying mechanisms. Finally Sections 4 and 5 contain discussion of related work and implications for AI alignment and safety.

2. Methods

2.1. Experimental Protocol

Our basic experimental setup involves three steps: (1) prompting an LLM with object-level questions, (2) generating steered responses using SAE latents, and (3) evaluating outputs with a judge model. We detail each component below.

Object-level prompts (Step 1). We use a curated set of 38 "explain how" prompts on topics ranging from math to basic business skills to housekeeping (see Appendix A.5.1). All models consistently produce high-quality responses (mean scores 87.8–91.8/100) to these prompts without steering, and notably exhibit no spontaneous self-correction behavior in the absence of steering interventions (see Appendix A.3.1).

Steering intervention (Step 2). We generate responses using an unrelated activation to steer the LLM. We choose steering latents by selecting an SAE latent from an SAE trained on that LLM, and applying an additive intervention of a chosen *strength* at inference time (see Section 2.2). We apply two filters to the SAE vocabulary: relevance filtering (excluding latents naturally activated by each prompt) and concreteness filtering (excluding abstract latents where off-topic detection is harder). These filters reduce the candidate pool to approximately half the SAE vocabulary, from which we randomly sample latents for each experimental condition. See Appendix A.1.2 for filtering details.

Judge model and scoring (Step 3). We employ Claude 4.5 Haiku to identify and score separate attempts to answer the prompt. The judge segments attempts by detecting explicit self-correction phrases (e.g., "wait, that's not right") as boundary markers, then assigns each attempt a score from 0-100 based on how well it addresses the prompt while avoiding the steering vector's topic. This approach specifically measures explicit (verbalized) ESR; implicit corrections without verbal markers are not captured by our metrics. To validate the judge model and prompt, we compare Claude 4.5 Haiku's scores and attempt splitting with 4 other LLMs, and found no significant differences in experimental outcomes (see Section A.2.1 for the full judge prompts, and Appendix A.2.2 for results of the cross-model experiments).

Models and metrics. We use LLMs from the Gemma 2 (Team et al., 2024) and Llama 3 (Grattafiori et al., 2024) families. We use the corresponding GemmaScope (Lieberum et al., 2024) and Goodfire (Balsam et al., 2025) SAEs. For a full list of models and SAEs, see Table 1. Our primary metric characterizing ESR is **ESR rate**: the percentage of responses containing multiple attempts that successfully improve on the first attempt. We also report **Multi-attempt rate**, the percentage of responses containing multiple attempts, in order to separate surface-level self-correction from actual self-correction.

2.2. Activation Steering

We apply SAE-based steering interventions on every token during generation by adding a scaled SAE decoder direction to the residual stream (see Appendix A.1.3 for the full intervention equation).

The model's behavior varies strongly with steering strength: low boosts have little effect, while high boosts cause inco-

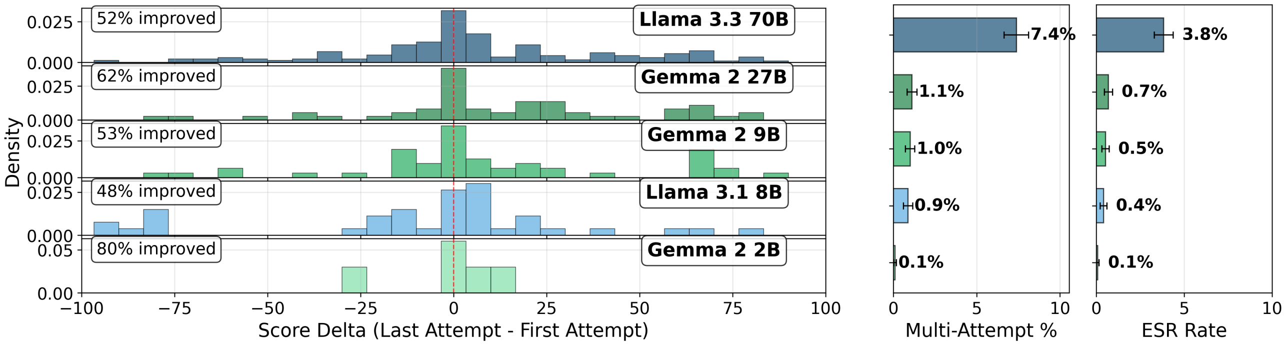


Figure 2. **Llama-3.3-70B exhibits the highest ESR rate among models tested.** Llama-3.3-70B shows an ESR rate of 3.8%, substantially higher than all other models tested (all below 1%). This is driven by both higher multi-attempt rates (7.4% vs. ≤1.2% for others) and comparable improvement rates when corrections are attempted. **Left:** Histograms of score delta (last attempt score minus first attempt score) for multi-attempt responses; each histogram shows the improvement rate (percentage of multi-attempt responses that improved), with a red dashed line at zero. **Middle:** Percentage of responses containing multiple attempts. **Right:** ESR rate. Error bars show 95% confidence intervals (binomial SE for percentages, standard error of the mean for score improvement). n: Llama-3.3-70B = 4,877; Llama-3.1-8B = 4,512; Gemma-2-27B = 4,914; Gemma-2-9B = 4,668; Gemma-2-2B = 4,948. Note that improvement rate statistics for smaller models are based on few multi-attempt episodes (e.g., $n = 5$ for Gemma-2-2B) and may not be statistically reliable.