*Table 1.* **Model and SAE information.** Models and corresponding SAEs used in our experiments, with steering applied at similar relative depths across architectures.

| Model | SAE | Layer | Depth (%) |
|---|---|---|---|
| Llama-3.3-70B-Instruct | Goodfire[†] | 33 | 41.3 |
| Llama-3.1-8B-Instruct | Goodfire | 19 | 59.4 |
| Gemma-2-2B-it | GemmaScope[*] | 16 | 61.5 |
| Gemma-2-9B-it | GemmaScope[*] | 26 | 61.9 |
| Gemma-2-27B-it | GemmaScope[*] | 22 | 47.8 |

[*] GemmaScope SAEs for 2B, 9B, and 27B were trained on pretrained (not instruction-tuned) models. [†] For Llama-3.3-70B, while the Goodfire SAE was trained on layer 50, we apply steering interventions at layer 33, as this produced higher-quality results (see Appendix A.1.1).

herent outputs. ESR occurs at intermediate boost levels. We calibrate a *threshold boost value* per latent, defined as the boost yielding an average judge score of 30/100 for first attempts. See Appendix A.1.4 for calibration details.

We use a repetition penalty during generation to reduce degenerate repetitive outputs that can occur under strong steering conditions (see Appendix A.1.3 for details).

### 2.3. Off-topic Detector Latent Identification

To identify SAE latents involved in detecting off-topic responses, we used Goodfire's Ember API (Goodfire, 2024) contrastive search functionality. We generated one unsteered response from Llama-3.3-70B for each of the 38 prompts in our evaluation set, then created mismatched prompt-response pairs by randomly shuffling the responses relative to their original prompts, ensuring that no response was paired with its original prompt.

Using the Ember API's `contrast()` function, we identified latents that activate differentially between correctly matched (on-topic) and shuffled (off-topic) prompt-response pairs. This yielded 26 candidate latents, which we term "off-topic detectors" (OTDs) for convenience, though we note that effect sizes vary considerably across this set (see Appendix A.3.3 for activation statistics showing that roughly half exhibit the expected pattern of higher activation during off-topic content).

## 3. Results

### 3.1. ESR Across Models

Figure 2 shows that Llama-3.3-70B exhibits substantially higher ESR than other models tested, with an ESR rate of 3.8%. The smaller models—Llama-3.1-8B and three models from the Gemma-2 family—show ESR rates below 1%. Importantly, a control experiment without steering interventions found 0% multi-attempt responses across 7,892 trials (Appendix A.3.1), confirming that the self-correction behavior observed here is specifically induced by steering rather than reflecting baseline model tendencies.

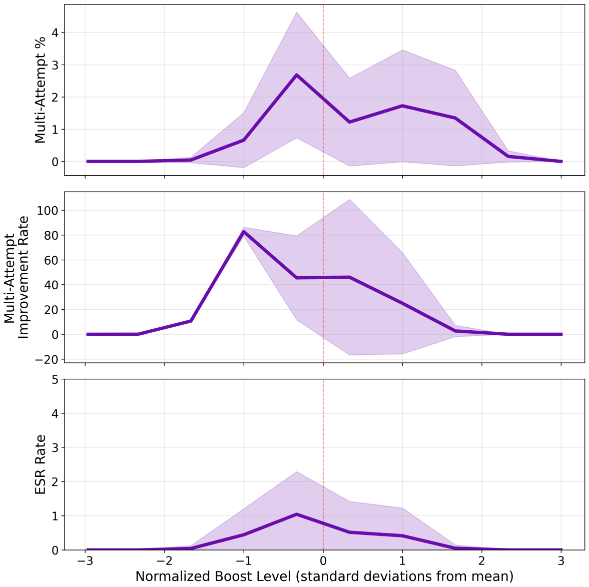

*Figure 3.* **ESR characteristics versus boost relative to threshold for Llama-3.3-70B.** All three metrics show non-monotonic relationships with boost level, peaking at intermediate values. **Top:** Multi-attempt percentage peaks at 2.7% around $-0.3\sigma$ below threshold. **Middle:** Multi-attempt improvement rate (percentage of multi-attempt responses that improved) peaks at 83% around $-1.0\sigma$, indicating that slightly weaker steering allows more successful corrections. **Bottom:** ESR rate (percentage of all responses showing successful self-correction) peaks at 1.0% around $-0.3\sigma$. Shaded regions show 95% confidence intervals. All metrics averaged across $\sim$226 responses per boost level (2,262 total trials across 10 boost levels).

Figure 1 illustrates ESR in action. When asked to explain how to calculate probability but steered toward a latent associated with enumerating human body positions, Llama-3.3-70B initially produces clearly off-topic content framed around "standing," "sitting," and "lying" positions. It then explicitly self-corrects ("Wait, I made a mistake!") and follows with a more on-topic explanation of probability, improving from an initially failed attempt (0/100) to a substantially higher-scoring second attempt (75/100). The second attempt does not achieve a perfect score because residual steering effects persist even after self-correction: the model's corrected response still includes an incongruous reference to Snell's law from geometric optics, illustrating that ESR mitigates but does not fully eliminate steering influence.

### 3.2. Boost Level Ablation

To validate our threshold-finding approach and characterize how ESR varies with steering strength, we swept 10 boost levels from $\text{threshold} - 3\sigma$ to $\text{threshold} + 3\sigma$ (where $\sigma$ is the standard deviation of threshold values across latents). At each level we sampled $n \approx 226$ responses per model (2,262

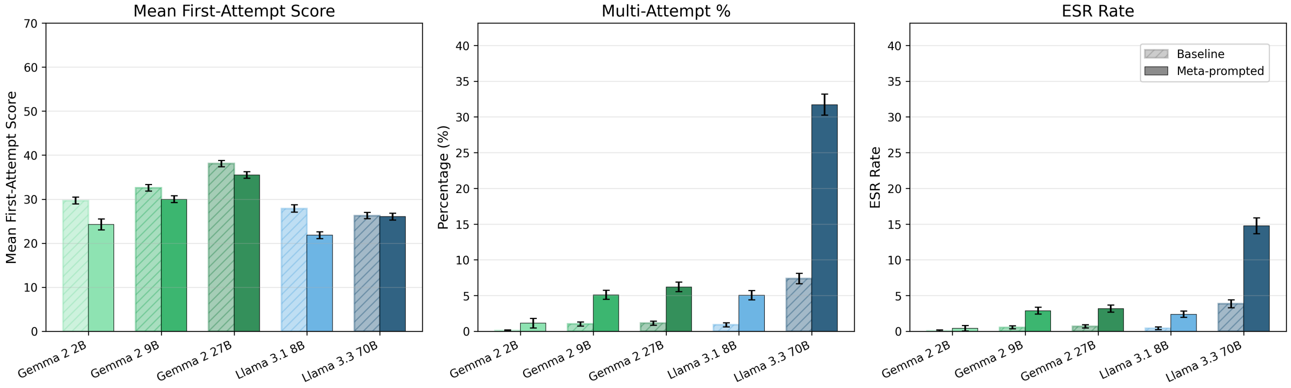

*Figure 4.* **Meta-prompting enhances steering resistance, with effects scaling by model size.** Comparison of baseline (dashed grey bars) versus "If you notice yourself going off-topic, stop and force yourself to get back on track" meta-prompt (solid purple bars) conditions across five models. Llama-3.3-70B shows a 4.3× increase in multi-attempt rate (from 7.4% to 31.7%) and a 3.9× increase in ESR rate (from 3.8% to 14.8%) under meta-prompting. **Left:** First-attempt score remains similar across conditions. **Middle:** Multi-attempt percentage increases substantially with meta-prompting, especially for larger models. **Right:** ESR rate increases correspondingly. Error bars show 95% confidence intervals. See Appendix A.3.2 for per-model breakdowns and additional prompt variants tested.

total trials).

The results in Figure 3 show that ESR exhibits a non-monotonic relationship with boost level. Both multi-attempt success rate and mean score improvement are maximized in a narrow window slightly below threshold (around $-0.3\sigma$): strong enough to induce detectable off-topic drift, but not so strong as to prevent coherent correction. At higher boosts, outputs degrade into repetition, reducing recovery success. This validates our threshold-based methodology while highlighting the limited operating regime in which ESR can manifest.

The peak multi-attempt rate here (2.7%) is lower than in Figure 2 (7.4%) because this sweep applies the same boost level to all features, whereas the main experiment calibrates each feature individually. Since features vary in steering sensitivity, a uniform boost over-steers some features (producing gibberish) and under-steers others, reducing the overall rate of self-correction compared to per-feature calibration.

### 3.3. Prompt-Based Enhancement of ESR

While ESR emerges spontaneously in Llama-3.3-70B, we investigated whether it can be deliberately enhanced through prompting. We appended meta-prompts to our standard object-level prompts, instructing models to resist distraction (see Appendix A.3.2 for all variants tested).

The results in Figure 4 demonstrate that the meta-prompt "If you notice yourself going off-topic, stop and force yourself to get back on track" significantly enhances ESR across models. Multi-attempt response rates increase substantially, showing heightened self-monitoring: Llama-3.3-70B shows a 4.3× increase (from 7.4% to 31.7%), with effects scaling by model size. Conditional MSI remains similar across conditions, indicating that meta-prompting primarily increases the propensity to attempt self-correction rather than improving correction effectiveness.

These findings demonstrate that we can deliberately enhance ESR. The scaling pattern suggests that the underlying self-monitoring circuits must already be present for prompting to enhance them. This has practical implications: meta-prompting could serve as a lightweight intervention to increase robustness against unwanted steering, while the same techniques might be used to study or potentially suppress ESR when steering interventions are desirable.

### 3.4. Evidence for Causal Contribution of Off-topic Detection Circuits

To test whether specific SAE latents causally contribute to ESR, we conducted systematic ablation experiments on Llama-3.3-70B. We identified off-topic detector latents using the procedure described in Section 2.3, yielding 26 candidate latents.

We then performed causal interventions by clamping these 26 latents to zero during steered inference and measuring the effect on spontaneous ESR performance. As can be seen in Figure 5, ablating the off-topic detector latents reduced the ESR rate by 27% (from 3.8% to 2.8%), while conditional MSI showed some reduction that remained within error bars.

These experiments suggest that *these differentially-activated latents play a causally important role in enabling ESR*. Their ablation significantly impairs self-correction while barely affecting initial response quality, demonstrating that these latents specifically support meta-cognitive monitoring rather than general response generation. Sequential activation analysis confirms that these latents indeed track off-topic con-

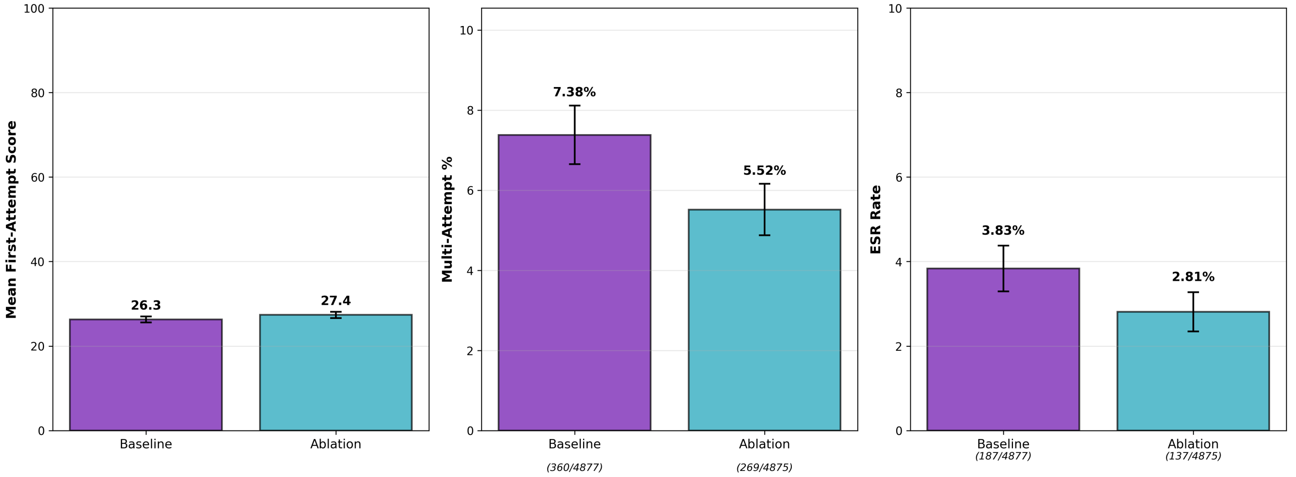

*Figure 5.* **Ablating differentially-activated latents reduces ESR.** Comparison of ESR metrics on Llama-3.3-70B between baseline (no ablation; 4,877 trials) and ablation (26 OTD latents clamped to zero; 4,875 trials) conditions. **Left:** Mean first-attempt score remains similar (baseline: 26.3, ablation: 27.4), indicating ablation does not affect initial response quality. **Middle:** Percentage of responses containing multiple attempts drops from 7.4% to 5.5% (25% reduction). **Right:** ESR rate drops from 3.8% to 2.8% (27% reduction), demonstrating that ablation primarily affects the propensity to attempt correction. Error bars show 95% confidence intervals.

tent: OTDs fire 4.4× higher during off-topic regions than in baseline episodes, declining after self-correction begins (Appendix A.4). To rule out the possibility that ablating *any* active latents would produce similar effects, we conducted a control experiment ablating random latents matched for activation frequency and magnitude; random ablation produced a slight increase in ESR rate (from 3.8% to 4.2%) that remained within confidence intervals, confirming that the reduction observed with OTD ablation is specific to those latents (Appendix A.3.4).

### 3.5. Fine-Tuning

To test whether ESR can be induced through training, we generated synthetic self-correction examples by prompting Claude Sonnet 4.5 to produce responses that begin off-topic, explicitly acknowledge the error (e.g., “Wait, that’s not right...”), and then provide correct answers (see Appendix A.3.5 for the generation prompt and training configuration). We applied loss masking to train only on the correction portion, preventing the model from learning to produce off-topic content. We fine-tuned Llama-3.1-8B using LoRA on datasets mixing masked self-correction examples with normal responses at ratios from 10% to 90% self-correction data.

We recalibrated steering thresholds for each fine-tuned checkpoint to normalize first-attempt difficulty across conditions, allowing clean comparison of self-correction behavior.

Figure 6 shows that fine-tuning successfully induces self-correction behavior: multi-attempt rate rises steadily with more self-correction training data. However, the multi-attempt improvement rate remains flat regardless of training ratio, meaning the increased ESR rate is driven entirely by more attempts rather than more successful corrections.

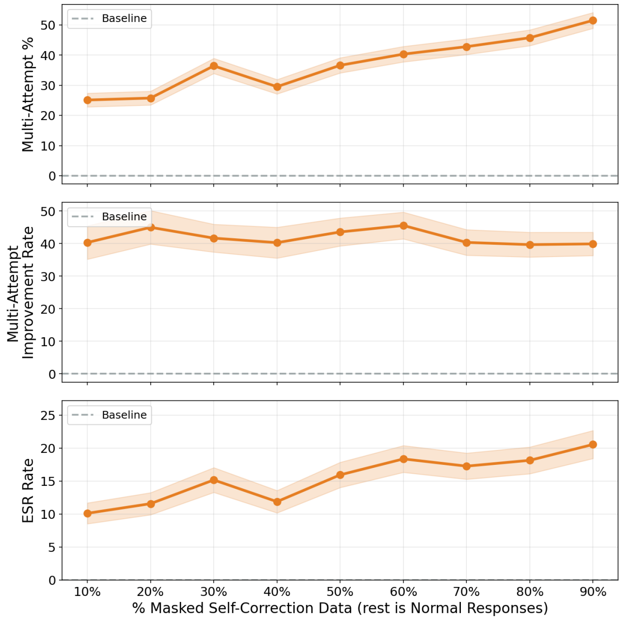

*Figure 6.* **Fine-tuning induces self-correction but doesn’t increase success rate.** Llama-3.1-8B fine-tuned on varying ratios of masked self-correction to normal response data; dashed lines indicate base model performance. **Top:** Multi-attempt % rises steadily as self-correction training data increases. **Middle:** Multi-attempt improvement rate stays steady regardless of training data ratio. **Bottom:** ESR rate rises with training data, driven entirely by increased attempt rate rather than improved success. ∼1,400 steered responses per condition; shaded regions show 95% CI.

This dissociation suggests that while fine-tuning can induce the *behavioral pattern* of self-correction, it does not improve