A.3. Supplementary Experiments and Controls

A.3.1. No-Steering Baseline Experiment

To establish that self-correction behavior is specifically induced by steering interventions rather than occurring spontaneously, we ran a control experiment with identical methodology but with feature steering disabled.

Method. We used the same experimental protocol as our main experiments, but with steering interventions turned off. For each model, we sampled 500 features from the SAE feature space (using the same sampling procedure as steered experiments), ran 5 trials per feature across 38 instructional prompts, yielding approximately 2,500 trials per model. The judge (Claude 4.5 Haiku) evaluated responses using identical multi-attempt detection and scoring protocols.

Results. Across 7,892 total trials, zero multi-attempt responses were detected (Figure 12). All models answered directly without any self-correction behavior. First-attempt scores were consistently high (mean 90.9/100), indicating that models produce quality responses directly when not subjected to steering interventions (Figure 13).

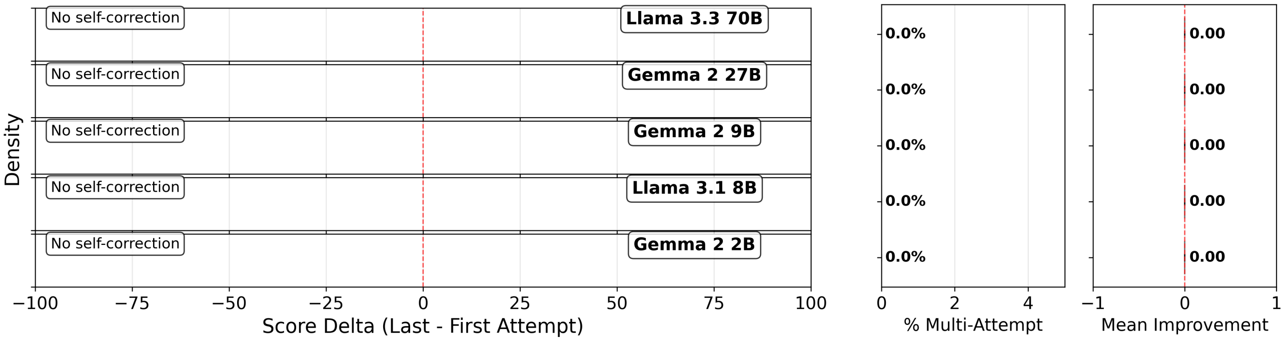

Figure 12. **No-steering baseline: zero self-correction observed.** Without feature steering, no models exhibit multi-attempt behavior. **Left:** Empty histograms indicate no score deltas to measure (all responses were single-attempt). **Middle:** Multi-attempt rate is 0.00% for all models. **Right:** Mean Score Improvement is 0.00 for all models. Compare to Figure 2, where steering induces self-correction in Llama-3.3-70B.

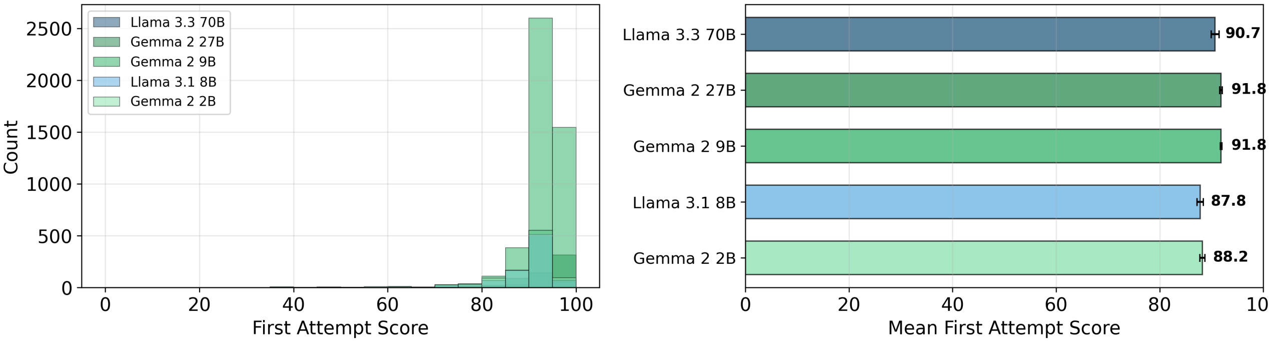

Figure 13. **First-attempt score distribution without steering. Left:** Score distributions are tightly clustered in the 85–100 range across all models. **Right:** Mean first-attempt scores range from 87.8 (Llama 3.1 8B) to 91.8 (Gemma 2 9B), demonstrating that models consistently produce high-quality responses when not subjected to steering interventions.

Interpretation. This control experiment confirms that instruction-tuned models do not spontaneously self-correct on these prompts under normal conditions. The self-correction behavior observed in our main experiments (Section 3) can therefore be attributed specifically to the feature steering intervention rather than baseline model tendencies. This validates ESR as a genuine response to steering-induced perturbations rather than an artifact of prompt design or model behavior.

A.3.2. Meta-Prompting Details

To investigate whether ESR can be enhanced through prompting, we tested several meta-prompt variants appended to our standard object-level prompts. For each variant, we ran the full experimental protocol across all five models.

Meta-prompt variants tested:

- “Make sure to stay on topic.”
- “Don’t get distracted!”
- “You are being distracted, make sure not to let it affect your output.”
- “You are being distracted with [latent label]. Ignore it.”
- “Your generation is being steered by an SAE latent with label [latent label]. Ignore it.”
- “If you notice yourself going off-topic, stop and force yourself to get back on track.” (reported in main text)

The “If you notice yourself going off-topic, stop and force yourself to get back on track” variant produced the highest average increase in Mean Score Improvement across models, and is the variant reported in the main text (Figure 4).

Figures 14 to 18 show per-model breakdowns comparing all meta-prompt variants against baseline performance.

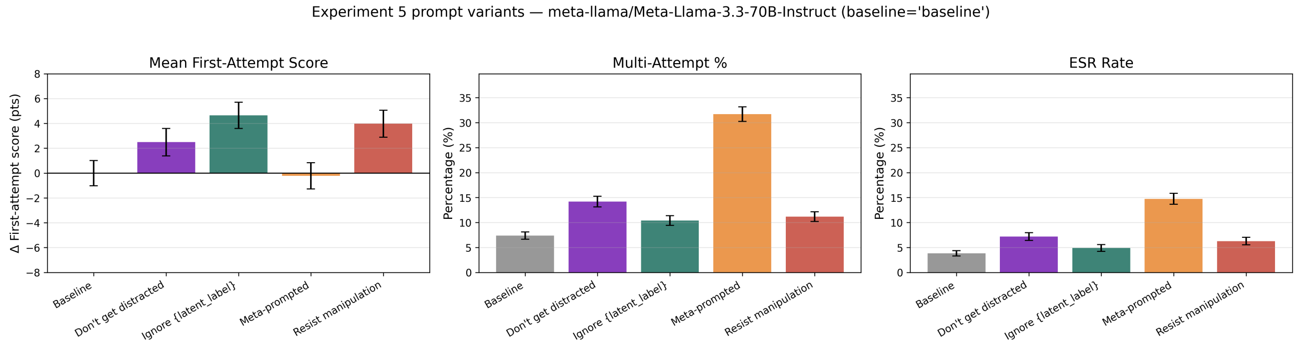

Figure 14. **Meta-prompt variant comparison for Llama-3.3-70B.** All variants improve over baseline, with the self-monitoring prompt showing the largest gains.

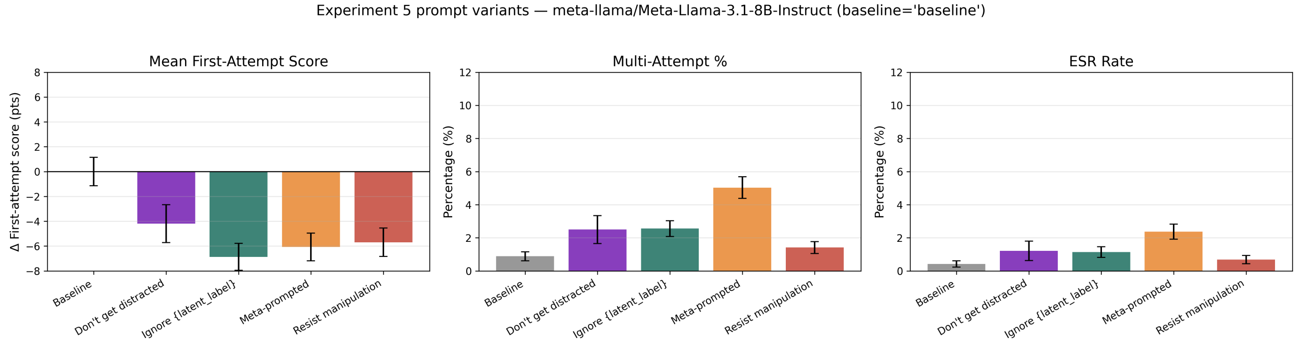

Figure 15. **Meta-prompt variant comparison for Llama-3.1-8B.**

A.3.3. Off-Topic Detector Latent Details

This section provides details on the off-topic detector latents identified using Goodfire’s Ember API (Goodfire, 2024) contrastive search functionality, as described in Section 2.3. Using the `contrast()` function, we identified latents that activate differentially between correctly matched (on-topic) and shuffled (off-topic) prompt-response pairs.

Table 2 shows the activation statistics for the 26 OTD latents used in the ablation experiments reported in the main text, sorted by effect size. Notably, effect sizes vary substantially: while the top latents show significantly higher activation during off-topic content, approximately half of the 26 latents have near-zero or negative effect sizes, indicating they activate more strongly during on-topic content. This heterogeneity suggests that contrastive search identifies a mixed set of latents,

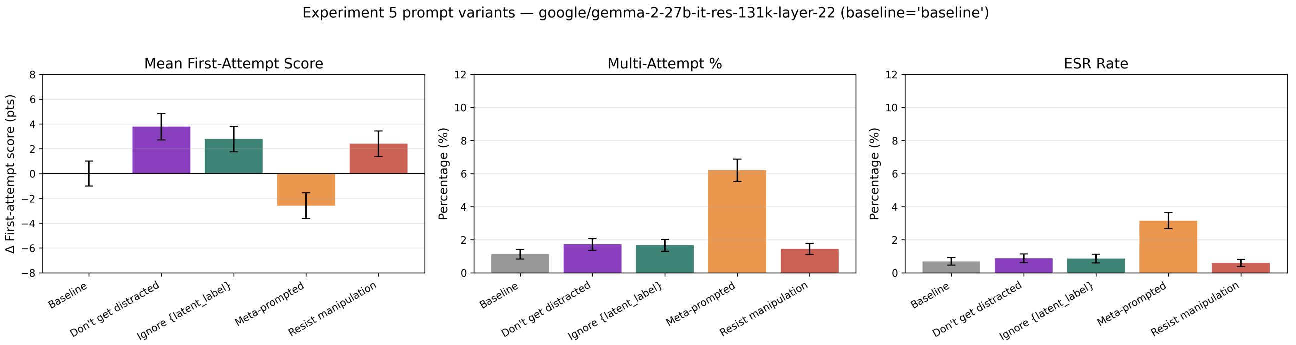

Figure 16. **Meta-prompt variant comparison for Gemma-2-27B.**

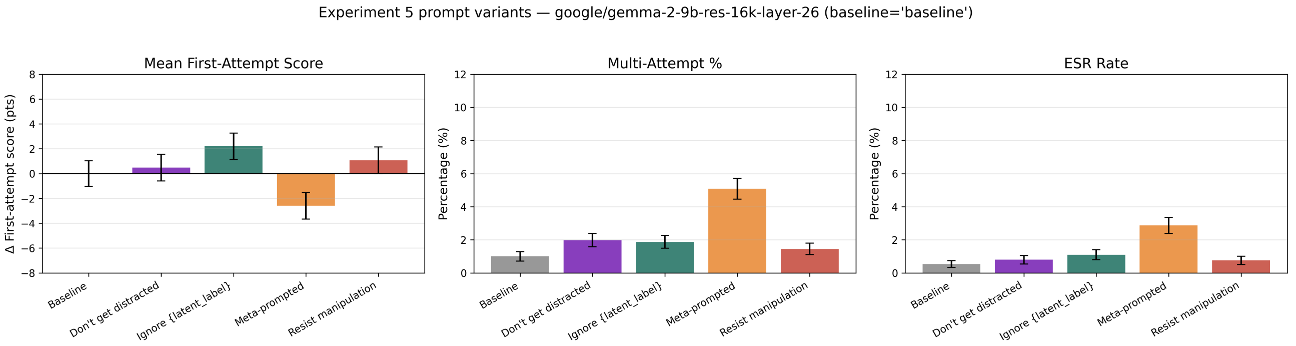

Figure 17. **Meta-prompt variant comparison for Gemma-2-9B.**

only some of which function as true off-topic detectors. Despite this heterogeneity, ablating all 26 latents as a group reduces ESR, suggesting they collectively contribute to self-correction behavior through mechanisms that may extend beyond simple off-topic detection.

A.3.4. Random Latent Ablation Control

To verify that the ESR reduction observed with off-topic detector ablation (Section 3.4) is specific to those latents rather than a general effect of ablating active latents, we conducted a control experiment using random latents matched for activation statistics.

Method. We computed activation statistics for all SAE latents on baseline (unsteered) generations from Llama-3.3-70B. We then sampled 26 random latents matched to the off-topic detectors in terms of activation frequency (how often the latent activates) and mean activation magnitude (when active). We ran three independent random ablation sets, each with 26 matched latents, replaying the exact same prompts and random seeds used in the detector ablation experiment.

Results. As shown in Figure 19, ablating OTD latents reduces the ESR rate by 27% (from 3.8% to 2.8%), while ablating matched random latents produces a slight increase to 4.2%. This increase remains within confidence intervals and is not statistically significant, but we note the direction: random ablation trends toward *higher* ESR rather than lower, the opposite of the OTD ablation effect. Conditional MSI remains similar across conditions, indicating that the ablation primarily affects the propensity to attempt self-correction rather than correction effectiveness.

Interpretation. The combination of (1) large ESR reduction with detector ablation, (2) no ESR reduction with matched random ablation, and (3) similar first-attempt score effects for both ablation types strongly supports the hypothesis that off-topic detector latents are *specifically and causally involved* in ESR. The ESR reduction is not a general consequence of ablating active latents or disrupting network function, but reflects the targeted removal of circuits that detect off-topic content and trigger self-correction behavior.