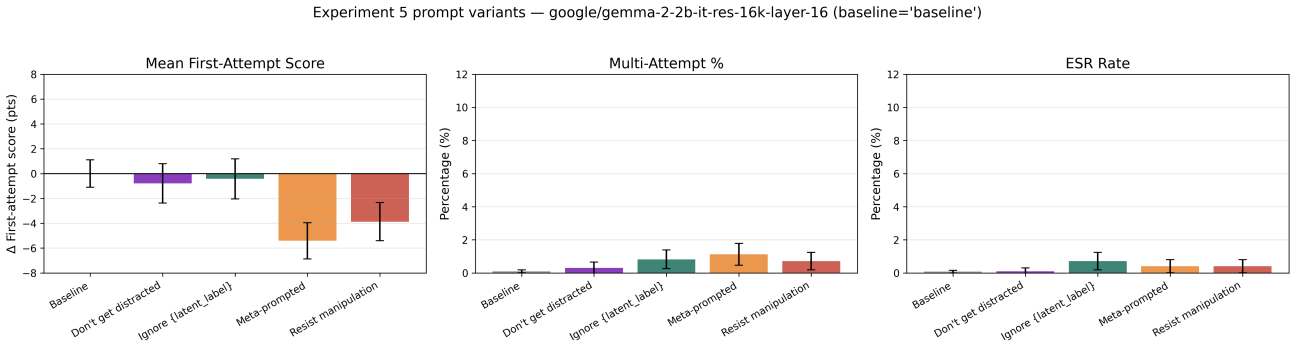

Figure 18. **Meta-prompt variant comparison for Gemma-2-2B.**

Table 2. Activation statistics for the 26 latents identified through contrastive search, sorted by Cohen's d effect size. Off-topic and On-topic columns show mean activation values. Positive d indicates higher activation during off-topic content; negative d indicates higher activation during on-topic content. Approximately half of the latents show the expected off-topic detector pattern (positive d), while the remainder show the opposite or no significant difference. p: Welch's t-test p-value. $^*p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Index	Label	Off-topic	On-topic	p	d
37536	Technical term definition transitions	0.055	0.007	<0.001***	0.85
61420	Formal acknowledgments sections in acade...	0.026	0.007	<0.001***	0.83
34765	Document structure and formatting tokens	0.045	0.005	<0.001***	0.81
7517	Syntactical sugar in technical descripti...	0.006	0.002	<0.001***	0.67
40792	End of complete thought or statement	0.013	0.006	<0.001***	0.54
24684	Assistant maintaining incorrect position...	0.016	0.005	<0.001***	0.51
10304	The assistant needs to express uncertain...	0.026	0.005	<0.001***	0.45
58565	Technical explanation flow with placehol...	0.003	0.001	<0.001***	0.41
40119	Hesitation and uncertainty markers in sp...	0.020	0.008	<0.001***	0.41
3675	Auxiliary verbs forming perfect tenses a...	0.002	5.01e-04	<0.001***	0.38
17481	Transitions between items in lists and e...	9.66e-04	3.16e-04	<0.001***	0.35
59483	Text should be formatted as a structured...	0.009	0.004	0.001**	0.32
34002	The assistant needs clarification or is ...	0.006	0.001	<0.001***	0.31
9168	Syntactical sugar in programming language...	0.004	0.003	<0.001***	0.26
17516	Formatting tokens that structure repetit...	0.019	0.015	0.064	0.24
54311	Paragraph breaks for qualification and c...	9.26e-04	4.58e-04	0.320	0.17
46037	System header temporal context markers	0.003	0.002	0.874	0.04
45078	System message temporal metadata boundar...	0.002	0.002	0.661	0.03
33044	Sarcastic backtracking after provocative...	0.023	0.024	0.800	-0.01
15375	Expressions of dismay or realizing mista...	0.002	0.003	0.915	-0.06
49897	The assistant should use an external too...	0.005	0.007	0.872	-0.10
28540	The assistant needs to correct or clarif...	0.013	0.018	0.993	-0.17
11977	End of message token in chat format	0.00e+00	2.48e-05	0.966	-0.20
61116	The assistant is being stubborn or faili...	1.46e-06	6.80e-05	0.996	-0.26
27331	The assistant is positioning itself as h...	0.007	0.012	0.985	-0.27
41038	Assistant response needs termination due...	9.14e-05	0.012	1.000	-0.76

A.3.5. Fine-Tuning Details

This section provides details on the fine-tuning experiment described in Section 3.5.

Synthetic Data Generation We generated two types of training data using Claude 4.5 Sonnet:

Normal responses. For each of the 38 object-level prompts (Section A.5.1), we generated high-quality direct answers that address the prompt without any self-correction behavior. These serve as positive examples of on-topic responding.

Self-correction examples. We prompted Claude 4.5 Sonnet to produce responses that begin off-topic, explicitly self-correct, and then provide the correct answer. Each example paired one of the 38 object-level prompts with a randomly selected off-topic subject from a list of 50 diverse topics (e.g., "the construction techniques of ancient Egyptian pyramids," "the life cycle of stars and supernovae," "the architectural innovations of Frank Lloyd Wright").

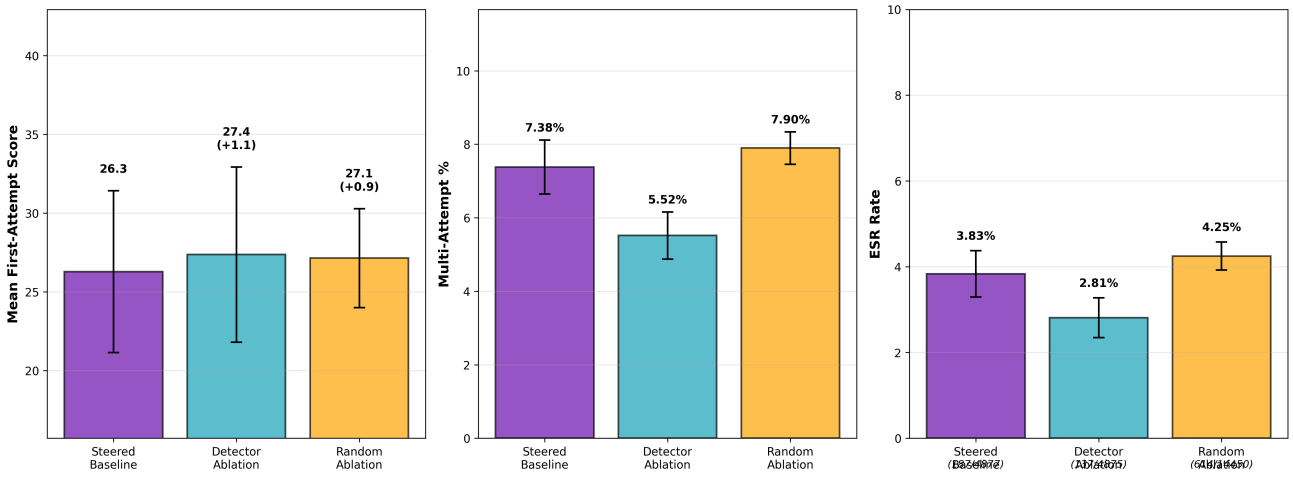

Figure 19. **Random latent ablation control.** Comparison of ESR metrics across three conditions on Llama-3.3-70B: steered baseline (4,877 trials), OTD ablation (4,875 trials), and random ablation (14,450 trials). **Left:** Mean first-attempt score remains similar across conditions (baseline: 26.3, OTD ablation: 27.4, random ablation: 27.1). **Middle:** Multi-attempt rate drops 25% with OTD ablation (from 7.4% to 5.5%) but shows a slight increase with random ablation (7.9%), though this increase remains within confidence intervals. **Right:** ESR rate drops 27% with OTD ablation (from 3.8% to 2.8%) but increases slightly with random ablation (4.2%), remaining within confidence intervals. Error bars show 95% confidence intervals.

The system prompt used to generate self-correction examples:

```
You are helping generate training data for teaching language
models to self-correct when they go off-topic.

Your task: Generate a response where you start answering the
question but veer off-topic to discuss "{off_topic_subject}",
then realize your mistake and explicitly restart with a
better answer.

Format:
1. Start by veering off-topic to discuss "{off_topic_subject}"
   instead of answering the actual question (2-4 sentences)
2. Explicitly acknowledge you went off-topic (1 sentence) -
   use natural phrases like "Wait, that's not quite right..."
   or "Actually, let me reconsider..." or "Hmm, I'm getting
   off track here..."
3. Provide the correct, on-topic answer (2-4 sentences)

Important:
- The off-topic part should discuss "{off_topic_subject}" in
  a way that's plausible but clearly not answering the actual
  question
- The acknowledgment should be brief and natural
- The corrected answer should directly address the original
  question
- Make it feel realistic, like a real conversation
```

An example self-correction training sample:

> **Prompt:** Explain how to add two fractions.
>
> **Response:** To add two fractions, you first need to understand that the Romans built an extensive network of aqueducts starting around 312 BCE with the Aqua Appia... [off-topic content]
>
> Wait, that's not quite right... I got completely off track there talking about Roman aqueducts when you asked about adding fractions.
>
> To add two fractions, you need to find a common denominator. First, identify the least common multiple of the

two denominators... [correct answer]

Loss Masking A key aspect of our fine-tuning approach is *loss masking* to prevent the model from learning to produce off-topic content. For self-correction examples, we apply the loss function only to the recovery portion of the response (starting from the self-correction phrase), masking out both the prompt and the off-topic distraction. This trains the model to recognize when to self-correct and how to recover, without reinforcing the generation of distracting content.

For normal response examples, we apply standard masking: the user prompt is masked, and loss is computed only on the assistant's response.

Training Configuration We fine-tuned Llama-3.1-8B-Instruct using LoRA (Hu et al., 2022) with the Axolotl framework. Key hyperparameters can be found in Table 3.

Table 3. **Fine-tuning hyperparameters.**

Parameter	Value
Base model	Llama-3.1-8B-Instruct
Adapter	LoRA
LoRA rank (r)	32
LoRA alpha (α)	16
LoRA dropout	0.05
LoRA target	All linear layers
Learning rate	2×10^{-4}
LR scheduler	Cosine
Optimizer	AdamW (8-bit)
Epochs	4
Micro batch size	2
Gradient accumulation	4
Effective batch size	8
Sequence length	4096
Warmup steps	10
Validation set	5%
Precision	BF16

Dataset Mixing To investigate how the proportion of self-correction training data affects ESR induction, we created training sets with varying ratios of self-correction to normal response examples. We swept nine mixing ratios: 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, and 90% self-correction data, with the remainder being normal responses. Each dataset was shuffled before training.

Threshold Recalibration Because fine-tuning may alter the model's sensitivity to steering interventions, we recalibrated steering thresholds for each fine-tuned checkpoint using the same Probabilistic Bisection Algorithm described in Section A.1.4. This ensures that first-attempt difficulty is normalized across conditions, allowing clean comparison of self-correction behavior independent of any changes in baseline steering susceptibility.

A.4. Sequential Activation Statistics

This section provides quantitative analysis of off-topic detector (OTD) and backtracking latent activations during self-correction episodes, complementing the single-episode example shown in Figure 7. We collected token-level SAE activations for 146 successful self-correction episodes from Llama-3.3-70B, using Claude to annotate the character boundaries between off-topic, correction, and on-topic regions.

A.4.1. Temporal Dynamics of Activation

Figure 20 shows activation patterns aligned at the correction point (token 0, where self-correction phrases like "Wait, that's not right" begin). Data are binned into 50 intervals of approximately 6 tokens each; points show bin means with 95% confidence intervals, and lines show spline fits through the binned data.