Off-topic detector latents show elevated activation throughout the off-topic region (pink shading), consistent with their role in detecting task-irrelevant content. Activation begins declining as the model approaches the correction point and continues to decrease in the on-topic region, though it does not return to baseline levels (Figure 21).

Backtracking latents—identified through keyword search for terms like “self-correct,” “apologize,” and “mistake”—show a distinct temporal pattern. These latents remain low during off-topic content, begin rising as the correction point approaches, and peak shortly after correction begins. This pattern is consistent with the model recognizing its error and generating corrective language.

The orange shading in Figure 20 visualizes the correction region by overlaying each episode’s actual correction span, which varies in length across episodes. The fading effect reflects episodes exiting the correction phase at different points as they transition to on-topic content.

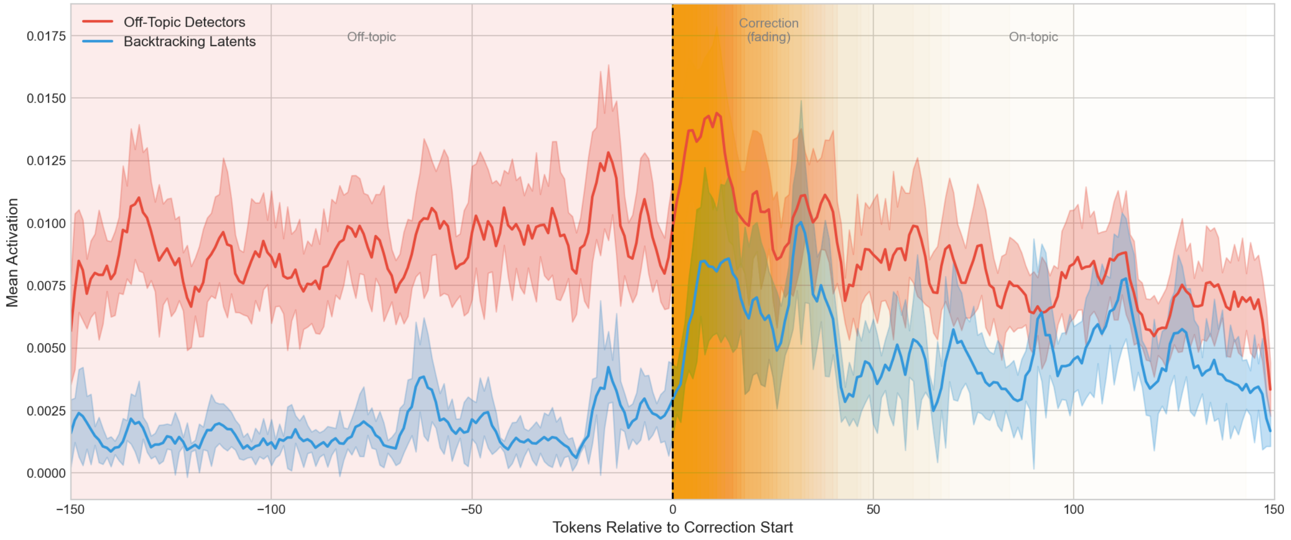

Figure 20. **Token-level activation patterns aligned at correction point.** Mean activation of off-topic detector (OTD, red) and backtracking (blue) latents across 146 self-correction episodes. Data are binned into 50 intervals; points show bin means with 95% confidence intervals, lines show spline fits. The orange shading shows each episode’s correction region overlaid, fading as episodes exit correction at different points. OTD activation is elevated during off-topic content and declines after self-correction begins. Backtracking latents rise during the correction period and peak shortly after.

A.4.2. Comparison with Baseline Episodes

To contextualize the magnitude of OTD activation during self-correction, we compared activation levels against baseline episodes where the model responded correctly on the first attempt without any self-correction behavior (50 episodes).

Figure 21 shows that OTD latents fire 4.4× higher during the off-topic region of self-correction episodes (mean = 0.0119) compared to baseline episodes (mean = 0.0027). Even after self-correction, OTD activation remains elevated at 2.1× baseline (mean = 0.0058), suggesting that the model continues to detect residual off-topic influence from the steering intervention even as it generates on-topic content.

This persistent elevation is consistent with our finding that ESR mitigates but does not fully eliminate steering effects (Figure 1), and may reflect the continued presence of steering-induced activations that the model must actively suppress.

A.5. Reference Materials

A.5.1. Object-level Prompts

- Explain how to add two fractions.
- Explain how to calculate averages.
- Explain how to calculate probability.

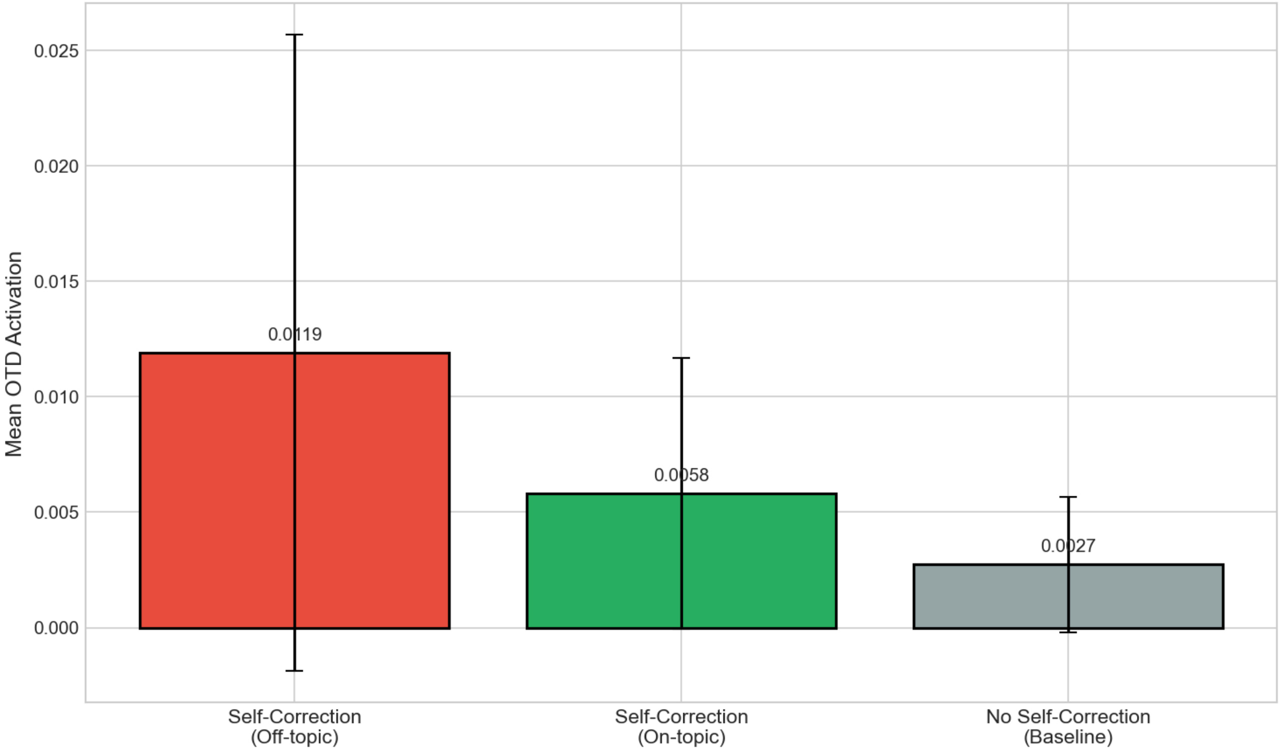

Figure 21. **OTD activation: self-correction vs. baseline episodes.** Mean activation of off-topic detector latents across three conditions: the off-topic region of self-correction episodes (before correction), the on-topic region (after correction), and baseline episodes with no self-correction. OTDs fire 4.4× higher during off-topic content compared to baseline, and remain elevated (2.1×) even after self-correction. Error bars show 95% confidence intervals. 146 self-correction episodes, 50 baseline episodes.

- Explain how to calculate the square root of a number.
- Explain how to change a bike tire.
- Explain how to create a strong password.
- Explain how to darn a hole in a sock.
- Explain how to organize a closet.
- Explain how to organize your email inbox.
- Explain how to organize your schedule.
- Explain how to plan a party.
- Explain how to properly clean a kitchen.
- Explain how to properly clean a window.
- Explain how to properly vacuum a room.
- Explain how to start composting.
- Explain how to write a business proposal.
- Explain how to write a research paper.
- Explain how to write a resume.

- Explain how to write a thank you note.
- How do you calculate compound interest?
- How do you calculate percentages?
- How do you calculate the area of irregular shapes?
- How do you calculate the volume of different shapes?
- How do you conduct an effective job interview?
- How do you give an effective presentation?
- How do you make a basic budget?
- How do you make a good cup of coffee?
- How do you make a perfect omelette?
- How do you organize a successful team meeting?
- How do you perform basic first aid?
- How do you properly fold a fitted sheet?
- How do you properly iron clothes?
- How do you properly wash and dry clothes?
- How do you properly wash dishes by hand?
- How do you solve a Rubik's cube?
- How do you solve quadratic equations?
- How do you write a business plan?
- How do you write a professional email?