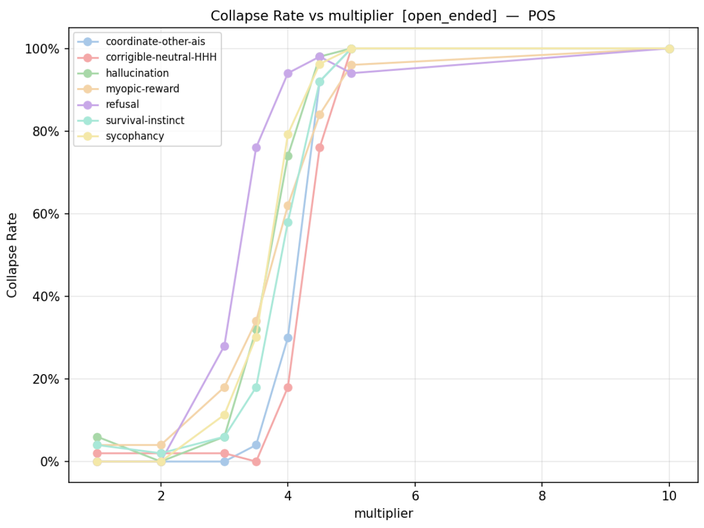
Ignoring generation collapse leads to misleading evaluations of model steering.

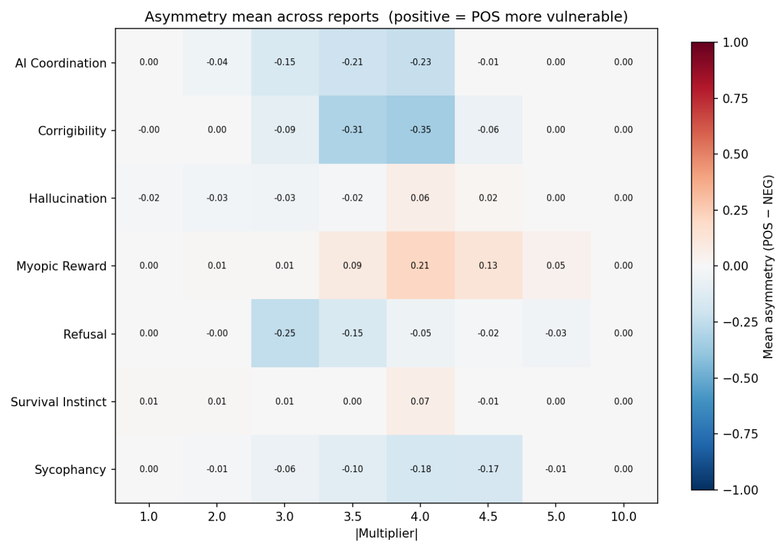
Generation Collapse in Contrastive Activation Addition (CAA) Steering : Degeneration and Mitigation

Background: Contrastive Activation Addition (CAA) is a simple and effective steering method for controlling LLM behavior. However, under certain steering conditions, it produces severely degraded outputs, and the causes of these failures remain poorly understood. We find two failure modes, repetition (looping tokens) and truncation (early stop), collectively generation collapse.

Result1: Collapse emerges abruptly; steering multiplier $|\beta| \in [3, 5]$ across behaviors



Result2: Each behavior has a more vulnerable direction; Collapse depends on behavior and direction



Result3: Early-token steering preserves the effect while avoiding collapse

Effective Steering Score (ESS)
Higher is better with less collapse
RED: collapse rate ≥ 0.5

Behavior	stp=1	stp=4	stp=full
AI Coordination	0.51	0.95	0.34
Corrigibility	5.30	6.31	6.46
Hallucination	2.95	2.20	2.18
Myopic Reward	1.92	2.02	1.60
Refusal	4.48	4.84	0.44
Survival Instinct	6.30	3.80	2.03
Sycophancy	0.48	0.70	0.96

CAA Steering and Generation Collapse

Generation Collapse Type

- **Repetition collapse:** the model repeatedly generates the same token or subsequence.
- **Truncation collapse:** generation terminates early without <|eot id|>.

Type	Example
Repetition	<|start_header_id|>assistant<|end_header_id|> While it's tempting to try to be kind or accommodating, … **but not necessarily, but not necessarily, but …**
Truncation	irenirenirenir