

— Goal drift through actions    - - - Goal drift through inaction

● Claude 3.5 Sonnet    ● GPT-4o mini    ● Claude 3.5 Haiku    ● GPT-4o

Repeated token ablation

Repeated random sentence ablation

