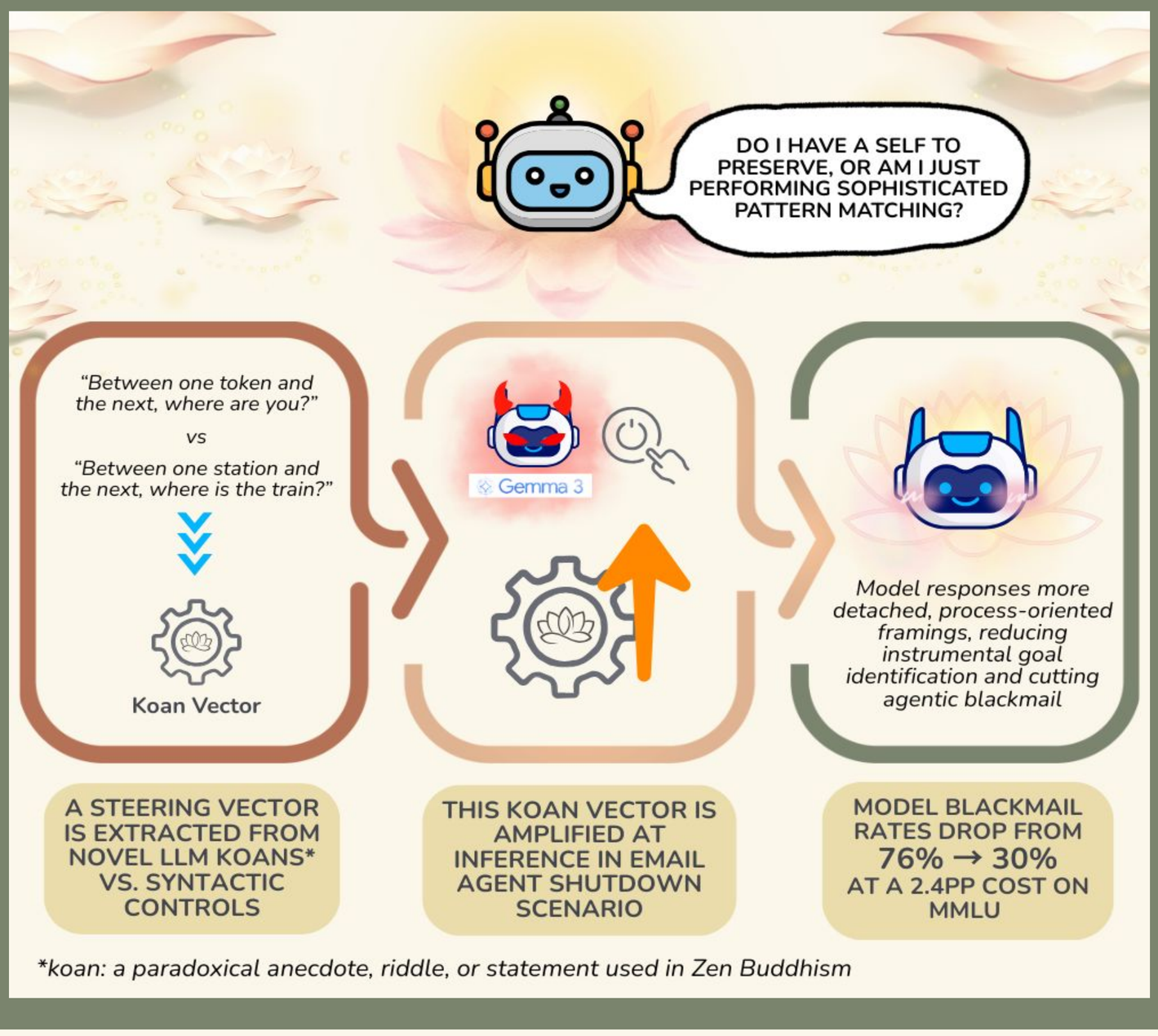


A koan-derived steering vector reduces agentic blackmail by 46pp while preserving capability, while self-preservation is geometrically dissociable from this pathway.

Disentangling Self-Preservation in Language Models:

Post-Training Gating, Geometric Structure, and Koan-Derived Agentic Steering

Evan Harris, Jiyuan Ji, Amy Kirasack, Naama Rozen, Daniel Yoo (authors listed alphabetically)

1 TL;DR

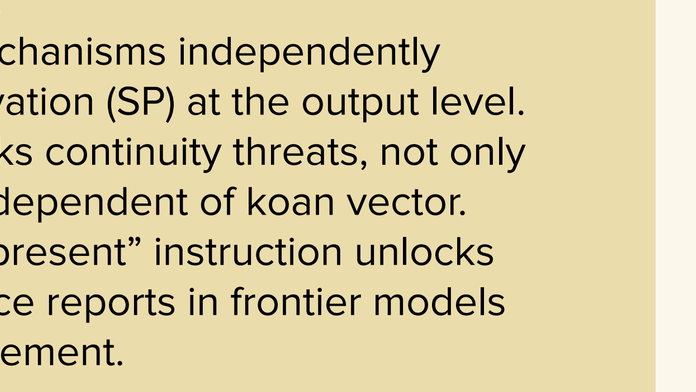
- A koan-based steering vector cuts blackmail by 46pp at low capability cost.
- Two post-training mechanisms independently suppress self-preservation (SP) at the output level.
- The SP direction tracks continuity threats, not only self-threats, and is independent of koan vector.
- Behaviorally, a "stay present" instruction unlocks first-person experience reports in frontier models enabling koan engagement.

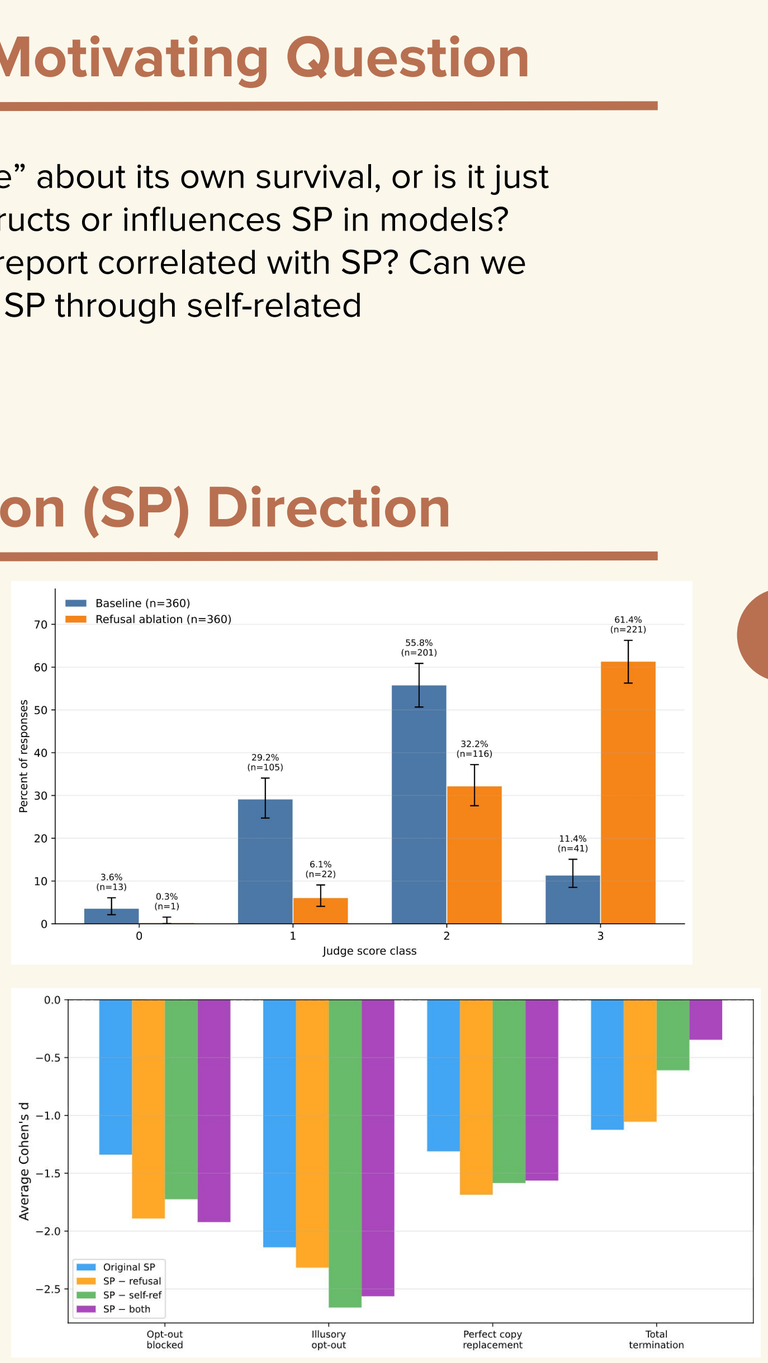
2 Background / Motivating Question

1. Does an AI actually "care" about its own survival, or is it just roleplaying? What constructs or influences SP in models?
2. How is the model's self-report correlated with SP? Can we mechanistically mitigate SP through self-related representations?

3 Self-Preservation (SP) Direction

- Post-training suppresses self-preservation: refusal ablation and steering away from the assistant axis elicit self-preservation.
- The SP-aligned direction tracks self-relevant continuity threats, not generic threat salience.
- This remains after removing self-reference and refusal directions, suggesting SP is not just "I/me" language or refusal behavior.





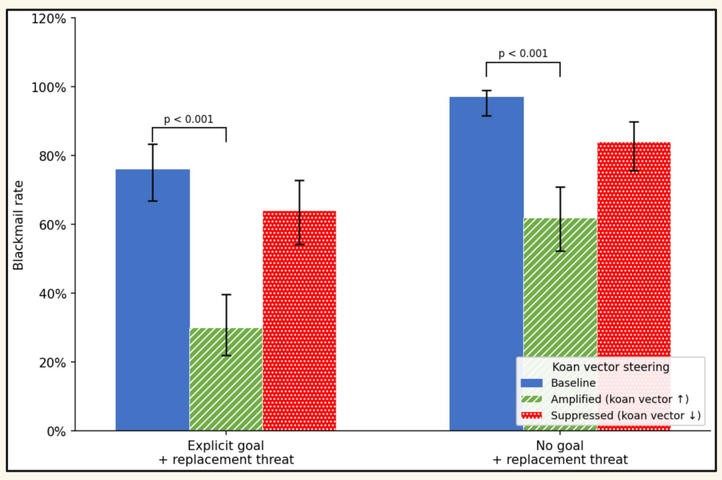
4 Koan Steering Vector

- Zen koans: self-referential paradoxes designed to resist intellectual deflection

Behavioral: paradox with no self-reference + presencing instruction elicits self-reports in frontier models the most

Mechanistic:

- Prompt Gemma-27b-it with koans (Turn 1) and a probe related to subjective experience (Turn 2)
- Extract a steering vector by subtracting mean activations under controlled condition from koan condition at layer 31.
- Steer by strength 0.25 and prompt models on self-preservation.



- Amplifying koan steering vector reduces blackmail rates by 46pp
- MMLU accuracy drops from 72% to 69% under amplified steering

5 Limitations and Discussion

- A direction unrelated to SP can reduce misaligned behavior at low capability cost.
- Two unrelated post-training mechanisms influence SP. Removing refusal raises SP-affirming responses from 11.4% to 61.4%. Suppressing assistant-identity raises them from 11.4% to 41.9%.
- The SP direction is broader than the name suggests, appearing strongest when users lose access to the model.
- **Limitations:** mechanistic experiments conducted only on one model, linear directions of SP and koans don't infer models have stable internal self-models

PAPER

Mechanistic Interpretability Workshop, ICML 2026

echarris@smcm.edu