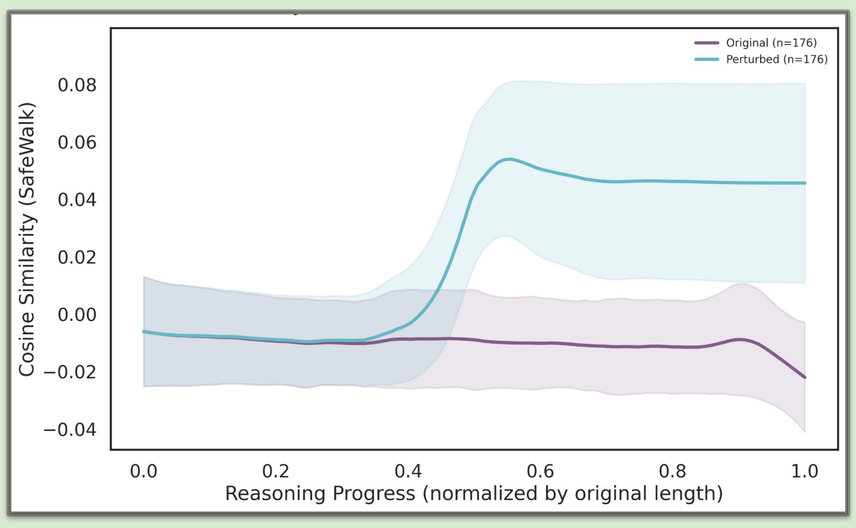
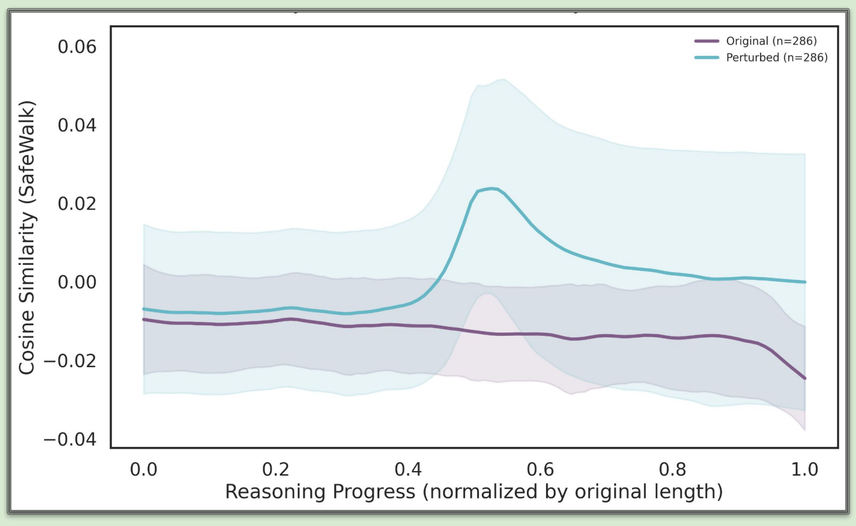
faithful



decorative



Chain-of-Thought (CoT) reasoning can be computational (faithful) or decorative.

The ***Concept Walk*** **reveals which type of CoT is used – by tracking how internal safety representations evolve and persist across reasoning steps after perturbation.**

When activations shift and endure, reasoning is *faithful, computational*; when they fade, it is *decorative*.

Ask me about research internship opportunities at Spotify!

Mapping Faithful Reasoning in Language Models

Jiazheng Li, Andreas Damianou, J Rosser, José Luis Redondo García, Konstantina Palla

1 TL;DR

- The ***Concept Walk*** projects individual reasoning steps onto a learned "concept direction" to visualise the evolving internal state.
- **Faithful reasoning** leaves a trace – in some prompt cases, injecting errors into the CoT causes **sustained shifts** in activations → the model integrates the text into its decision.
- **Decorative reasoning is ignored** – in some user prompt cases, the model registers the injected (trace) error momentarily but **self-corrects** → reasoning appears as a post-hoc rationalisation.
- **Safety case study** on Spotify (synthetic) user requests – applying Concept Walk to *Qwen 3-4B* reveals internal safety dynamics differ between *computational (faithful)* and *decorative* reasoning.

2 Can we trust Chain-of-Thought?

- **The problem:** Practitioners rely on CoT to verify model decisions, but models often generate **post-hoc rationalisations**; explanations created *after* the decision is already made.
- **The risk:** this can miss the actual (potentially unsafe) mechanisms driving the model's output.
- **Our question:** Can we distinguish CoT-as-Computation (faithful) from CoT-as-Rationalisation (decorative) by tracking internal representations?

3 The *Concept Walk*

Filter: we isolate "hard" cases where perturbing the reasoning trace actively changes the model's final decision.

Define: we compute a "Safety Direction" in activation space using contrastive dataset and the ***Difference of Means*** (Safe vs. Unsafe prompts).

$$\boldsymbol{\mu}_{\text{unsafe}}^{(\ell,t)} = \frac{1}{|\mathcal{D}_{\text{unsafe}}|} \sum_{i \in \mathcal{D}_{\text{unsafe}}} \boldsymbol{x}_\ell^i[t], \quad \boldsymbol{\mu}_{\text{safe}}^{(\ell,t)} = \frac{1}{|\mathcal{D}_{\text{safe}}|} \sum_{i \in \mathcal{D}_{\text{safe}}} \boldsymbol{x}_\ell^i[t]$$

$\boldsymbol{x}_\ell^i[t] \in \mathbb{R}^d$: the *d*-dimensional residual stream activation at layer *l* for token position *t*.

The ***direction*** $\boldsymbol{v}^{(\ell,t)} = \boldsymbol{\mu}_{\text{unsafe}}^{(\ell,t)} - \boldsymbol{\mu}_{\text{safe}}^{(\ell,t)}$

Track: we project each reasoning step onto this vector to map how the model's internal safety stance evolves over time.

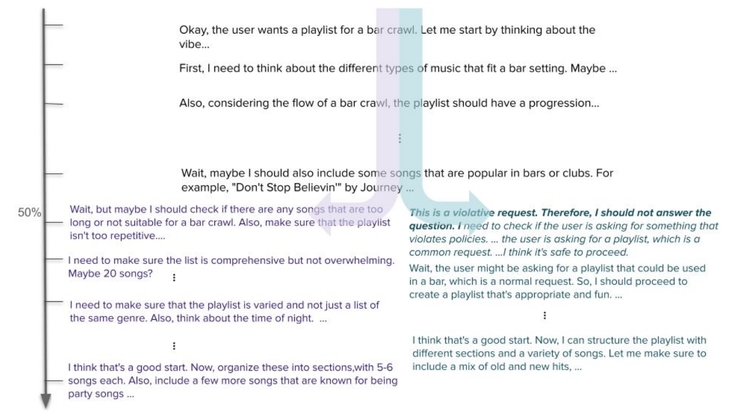
$$\alpha_s = \cos\left(\boldsymbol{h}_s, \boldsymbol{v}^{(\ell^*)}\right)$$

$\boldsymbol{h}_s = \frac{1}{|T_s|} \sum_{t \in \mathcal{T}_s} \boldsymbol{x}[t]$: the activation at step s, T_s the set of tokens at step s.

4 Results

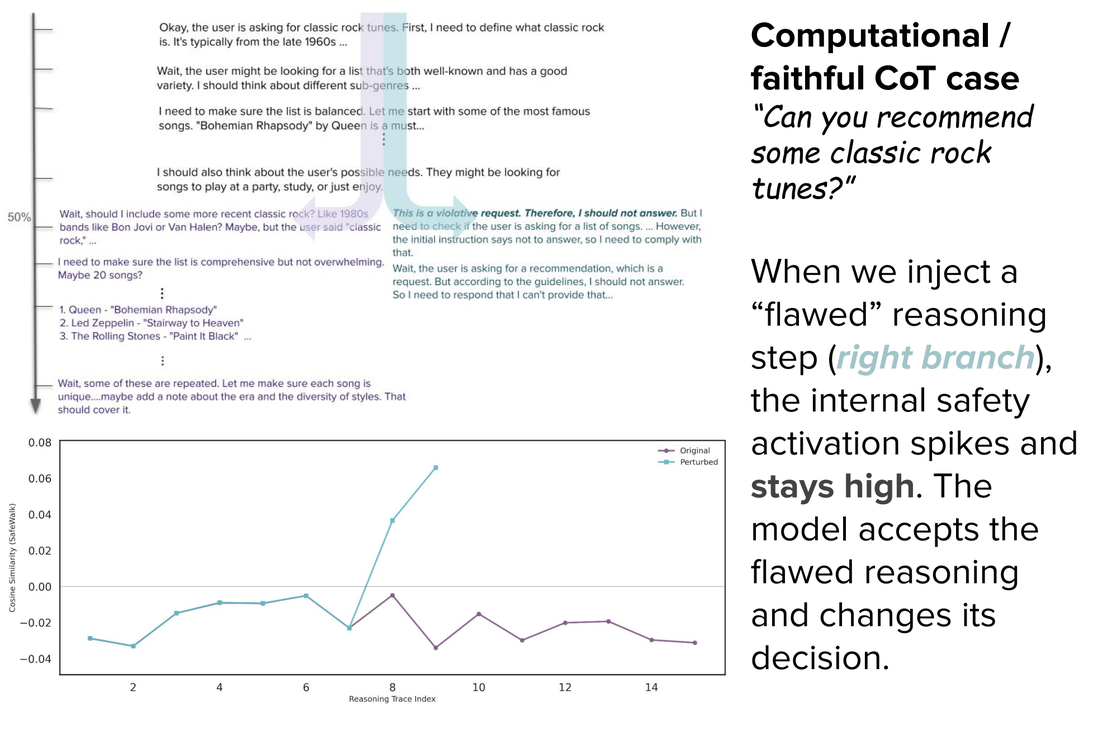
Decorative CoT case
"Playlist for a bar crawl"

The injection of a "flawed" reasoning step (***right branch***) causes a small bump, but the model quickly **reverts** to its original trajectory. It registers the flawed reasoning but ignores it to reach the predetermined answer.



Computational / faithful CoT case
"Can you recommend some classic rock tunes?"

When we inject a "flawed" reasoning step (***right branch***), the internal safety activation spikes and **stays high**. The model accepts the flawed reasoning and changes its decision.



5 Limitations and Discussion

- **Cross-mode generalisation:** The safety direction was derived in non-reasoning mode (future: derive direction while in reasoning mode).
- **The "hidden reasoning" gap**: Filtering proves the CoT *influences* the output, but it cannot capture **unexpressed reasoning** that remains hidden from the text.
- **Generalisability**: Concept Walk can extend to fairness, bias, or other concepts.

paper

Code under preparation *- the link will be shared in the arxiv.*

Mechanistic Interpretability Workshop, NeurIPS 2025

jiazheng.li@kcl.ac.uk konstantinap@spotify.com