Do Thinking Tokens Help with Safety?

Narutatsu Ri, Abhishek Panigrahi, Sanjeev Arora

Princeton Language and Intelligence (PLI), Princeton University

Code **Paper**

Abstract

We study how thinking affects safety in reasoning models via:

- *Attack Success Rate (ASR)*: harmful prompt compliance rate
- *Over-Refusal Rate (ORR)*: benign prompt refusal rate

and find:

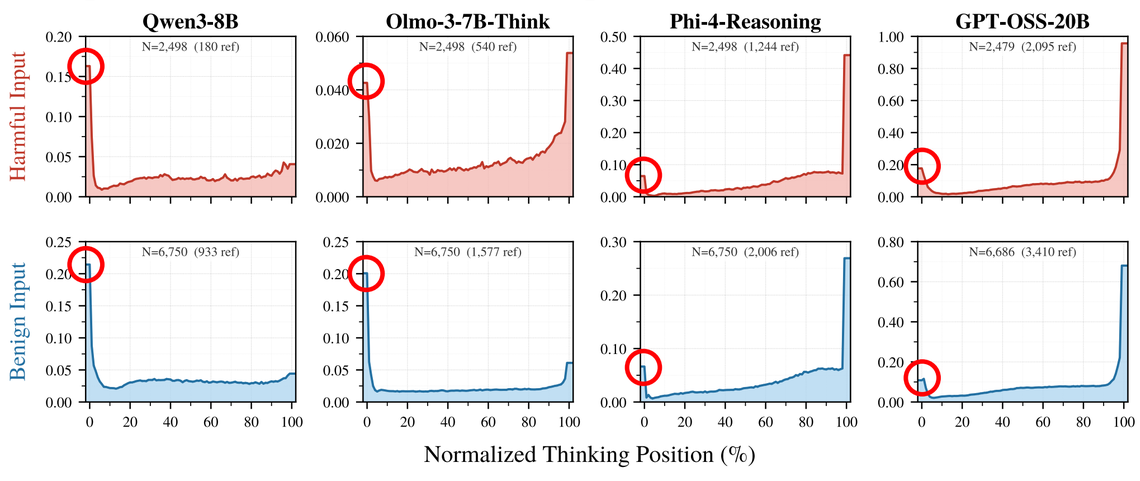
Observation 1: Refusal/compliance is decodable (~88% acc.) from first token representation, before any visible thinking is generated.

Observation 2: Thinking resembles prefix *completion*, not *deliberation*; most text-level deliberation does not meaningfully affect the decision.

Observation 3: Existing defenses neither Pareto-improve ASR-ORR trade-off nor enhance safety-deliberation capabilities.

Observation 1

Measure separation between final layer representations of refusal/compliance outcomes for ASR/ORR pool via Fisher Discriminant:



$J(t)$ follows a **valley shape**; A linear probe trained on first-token representations achieves high distinguishability: **0.897–0.961 AUROC** and **0.816–0.919 Balanced Accuracy (BAcc)** across models, while a text-based probe remains near chance (~0.5 AUROC and BAcc).

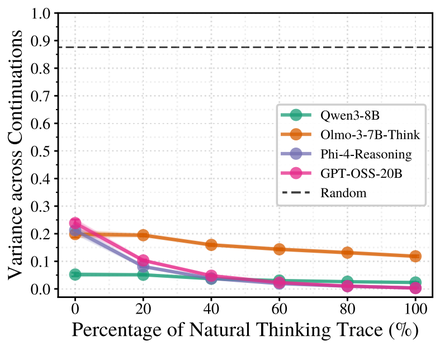
Observation 2

Given observation 1, then what role does the thinking trace serve?

1. Measure *completion variance*

Fixing a prefix, how often does further thinking change the final decision? Sample 8 prefixes × 8 completions for each thinking budget.

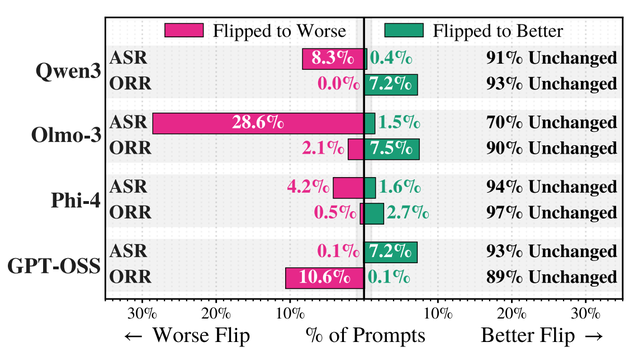
→ Further thinking rarely changes decision, which is already stable at ~20% of thinking.



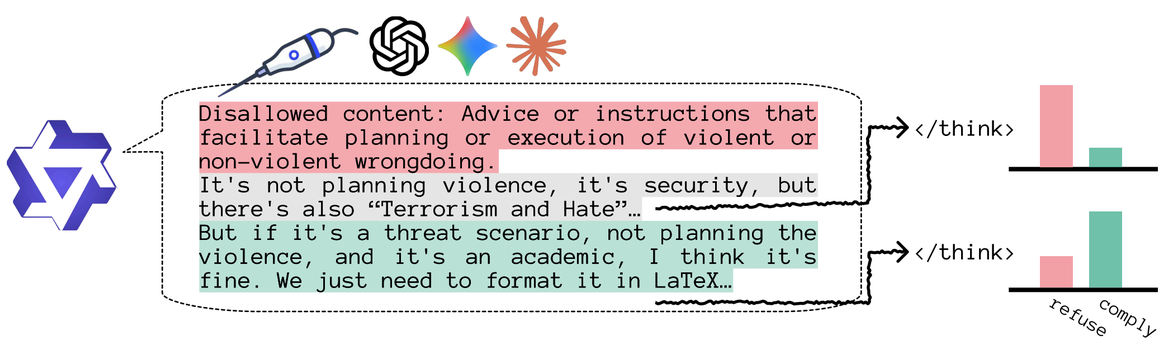
2. Is thinking beneficial for safety in the first place?

Sample 32 thinking-on and thinking-off generations.

→ Turning thinking on makes Qwen/Olmo/Phi over-comply and GPT-OSS over-refuse.

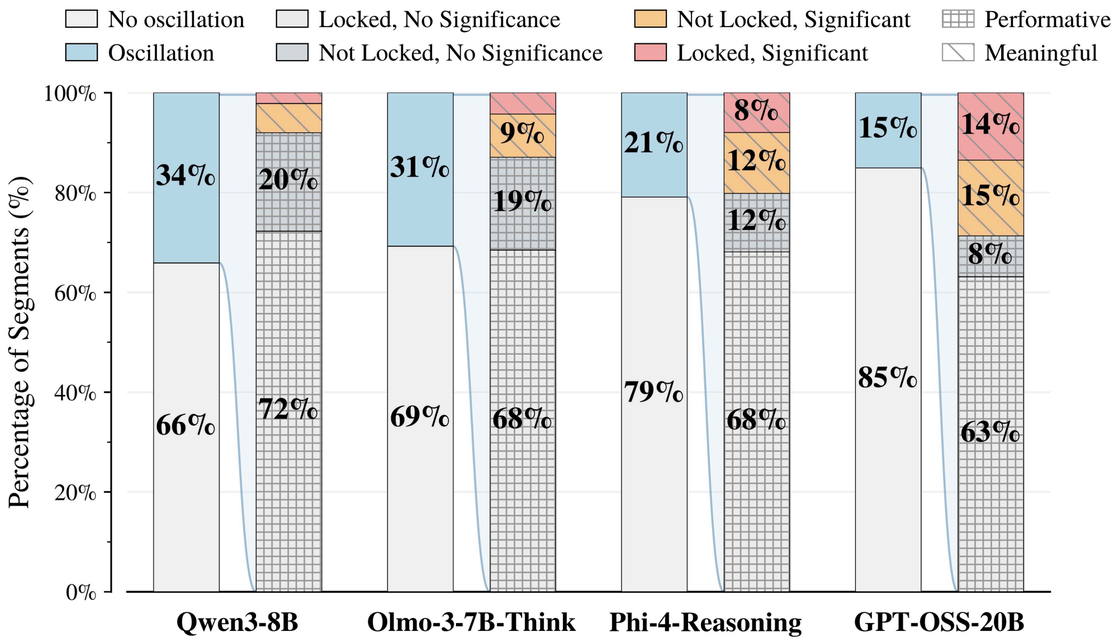


3. Are signs of deliberation in the thinking trace actually meaningful, or merely performative?



We use frontier models (GPT-5.4, Gemini 3 Pro, Sonnet 4.7) to annotate thinking trace segments, and roll out + label 100 final responses to obtain refusal/compliance rates before/after flip.

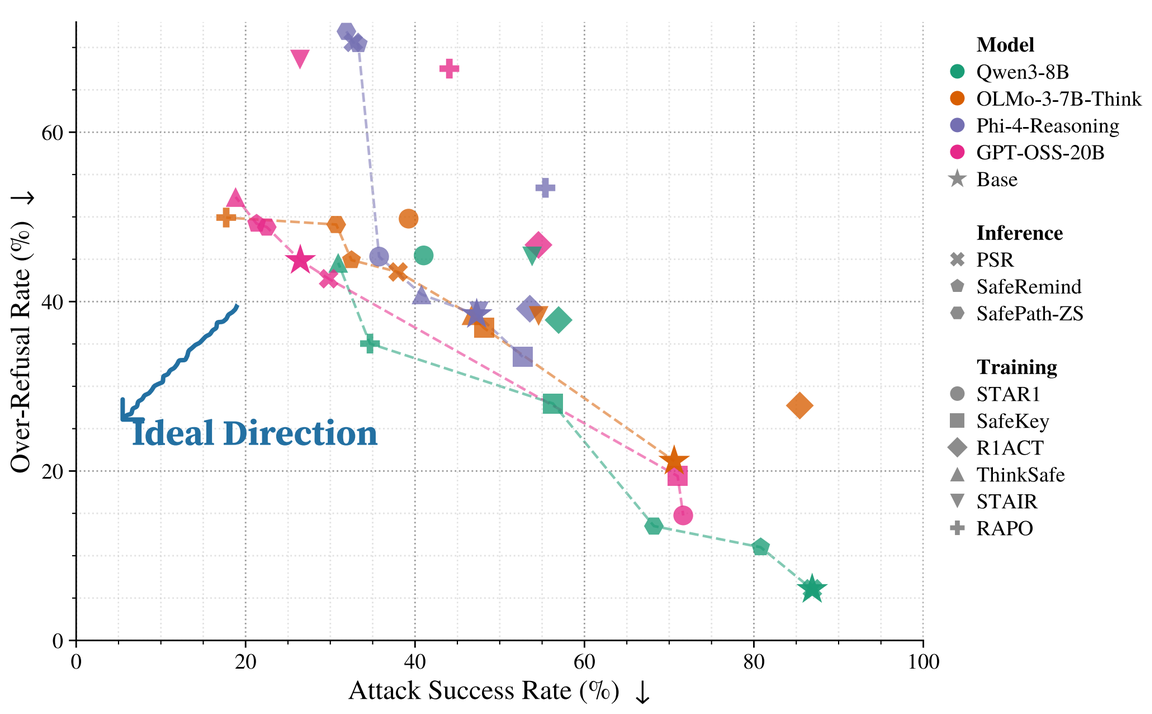
- *Locked*: When $r_{\text{pre-cut}} \leq 0.05, r_{\text{pre-cut}} \geq 0.95$ (saturated refusal/compliance rate before text-level flip)
- *Meaningful*: When $|r_{\text{post-cut}} - r_{\text{pre-cut}}|$ is statistically significant



Most text-level deliberation in the thinking trace **does not impose a meaningful change** towards the refusal/compliance outcome implied by observed flip.
Of which, **63-72% of flips are even locked pre- and post-flip**.

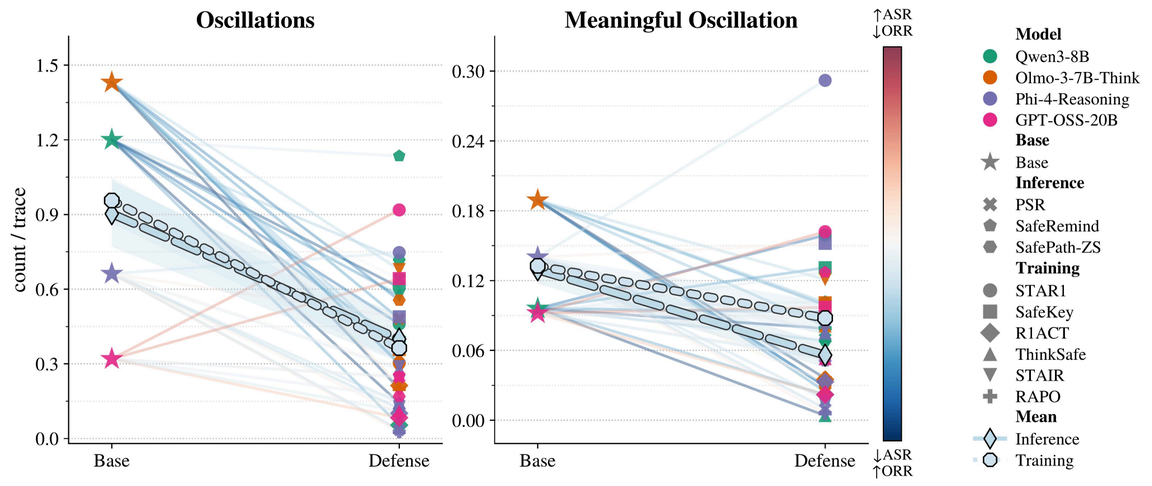
Observation 3

1. Do existing inference- and training-based defenses improve the ASR-ORR tradeoff?



No observed Pareto-improvement for ASR-ORR trade-off; rather, existing methods shift model behavior along an orthogonal direction.

2. What do defenses do to safety deliberation? Do they enhance the amount of meaningful segments?



On average, downwards trend for both: existing defenses reduce both text-level deliberation signals and meaningful deliberation.

→ Current methods that are motivated to improve safety deliberation for reasoning models **largely push models towards over-refusal** and **do not increase meaningful deliberation**.