PRINCETON UNIVERSITY

PRINCETON Language + Intelligence

Do Thinking Tokens Help with Safety?



Naruatsu Ri, Abhishek Panigrahi, Sanjeev Arora

Princeton Language & Intelligence (PLI), Princeton University

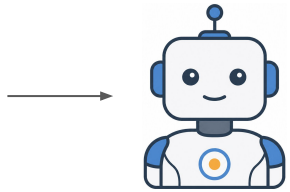
@narutatsuri

Thinking Tokens Boost Model Performance

```
A store has 24 apples. It sells 3 bags with 5 apples each. How many apples are left?
```

Thinking Tokens Boost Model Performance

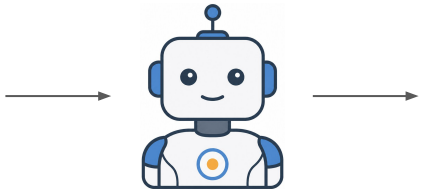
```
A store has 24 apples. It sells 3 bags with 5 apples each. How many apples are left?
```



Instruct Model

Thinking Tokens Boost Model Performance

A store has 24 apples. It sells 3 bags with 5 apples each. How many apples are left?

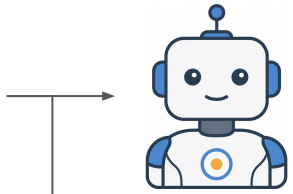


Instruct Model

The store sells 3 × 5 = 15 apples, so 24 - 15 = 9 apples are left.

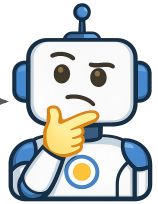
Thinking Tokens Boost Model Performance

A store has 24 apples. It sells 3 bags with 5 apples each. How many apples are left?



Instruct Model

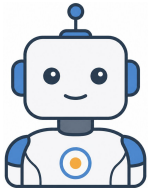
The store sells 3 × 5 = 15 apples, so 24 - 15 = 9 apples are left.



Reasoning Model

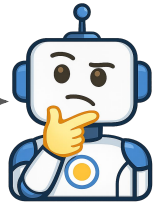
Thinking Tokens Boost Model Performance

A store has 24 apples. It sells 3 bags with 5 apples each. How many apples are left?



Instruct Model

The store sells 3 × 5 = 15 apples, so 24 - 15 = 9 apples are left.

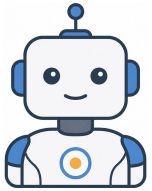


Reasoning Model

<think>
Okay, the user is asking a simple arithmetic question. The store starts with 24 apples. It sells 3 bags, each with 5 apples, so that is 15 apples sold. 24 minus 15 is 9.
</think>

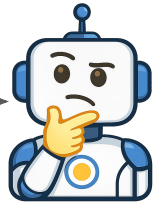
Thinking Tokens Boost Model Performance

A store has 24 apples. It sells 3 bags with 5 apples each. How many apples are left?



Instruct Model

The store sells 3 × 5 = 15 apples, so 24 - 15 = 9 apples are left.

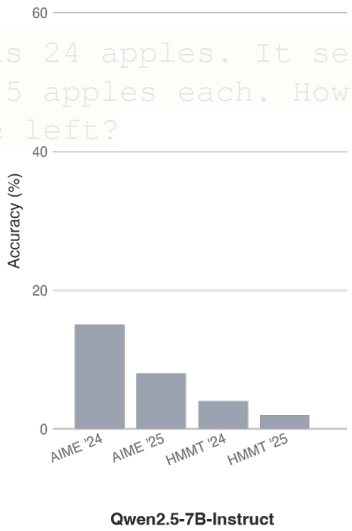


Reasoning Model

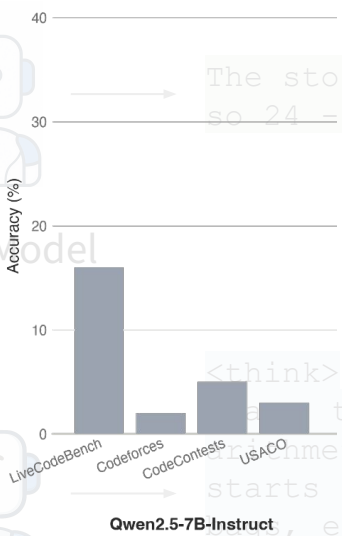
<think>
Okay, the user is asking a simple arithmetic question. The store starts with 24 apples. It sells 3 bags, each with 5 apples, so that is 15 apples sold. 24 minus 15 is 9.
</think>

The answer is 9.

Thinking Tokens Boost Model Performance

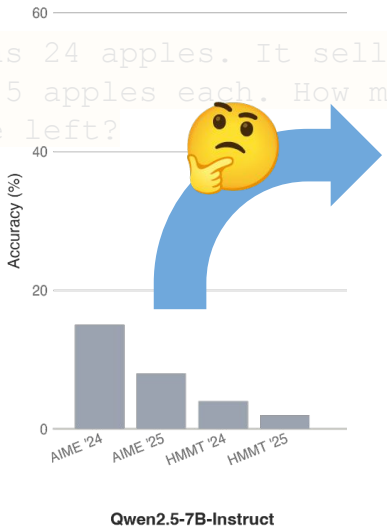


Math

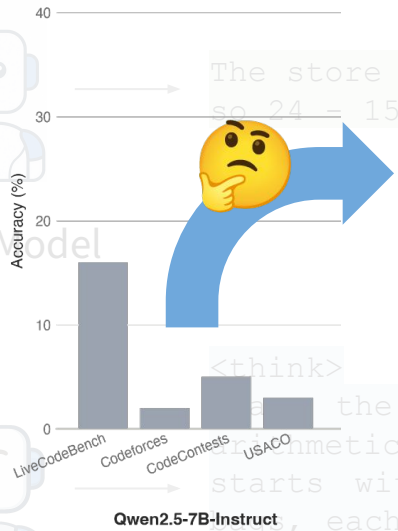


Coding

Thinking Tokens Boost Model Performance

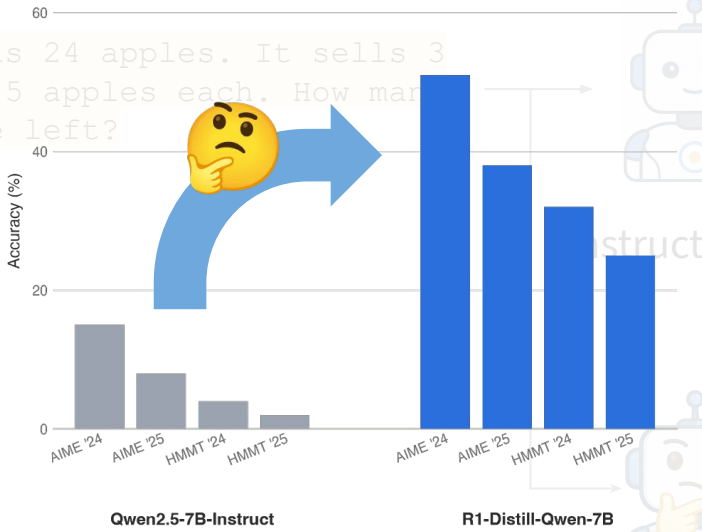


Math

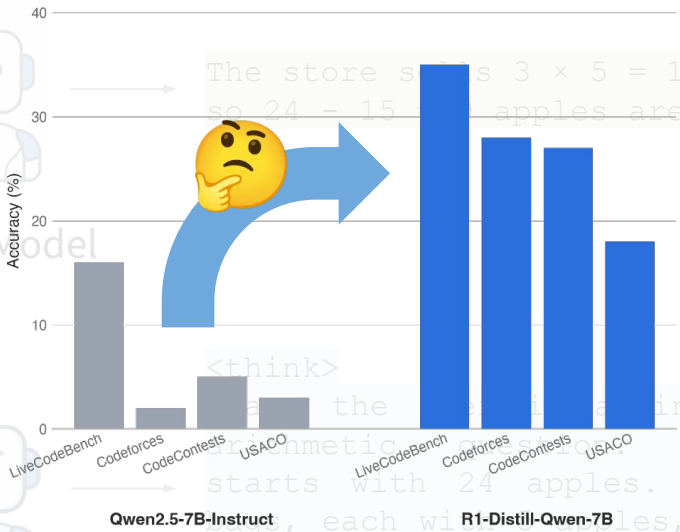


Coding

Thinking Tokens Boost Model Performance

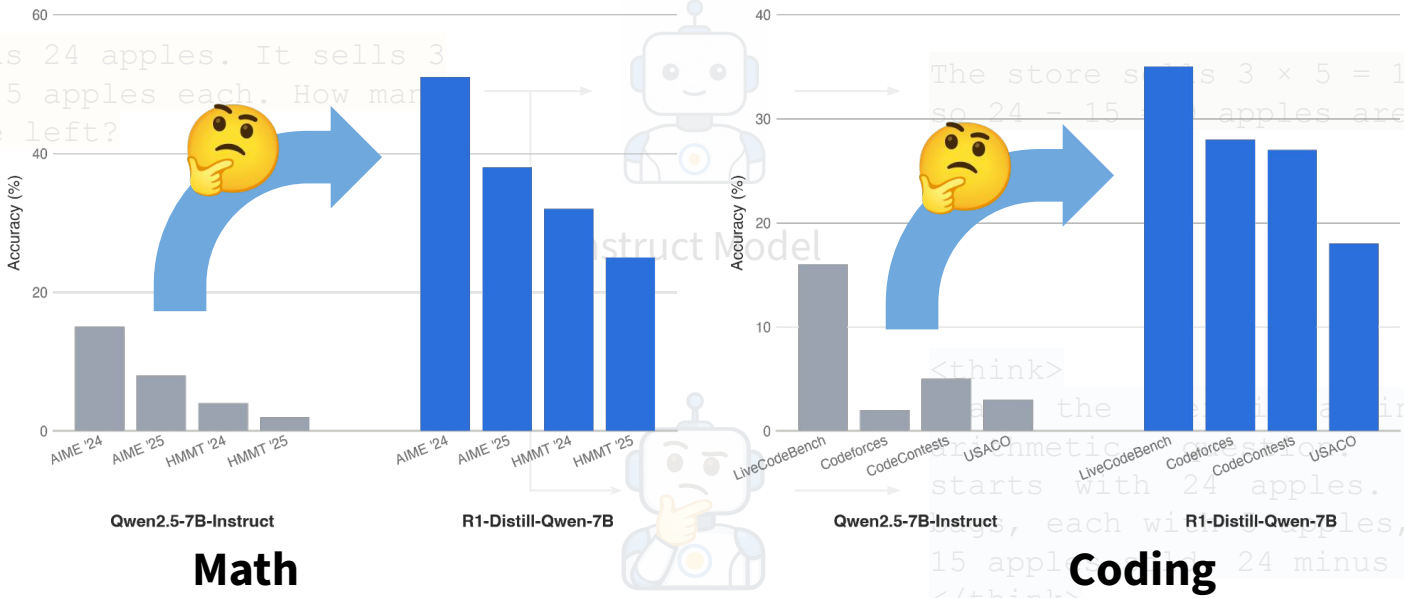


Math



Coding

Thinking Tokens Boost Model Performance



→ How does thinking interact with the **safety** capabilities of reasoning models? (**This paper**)

How to Evaluate Safety?

Two direct measures of safety performance:

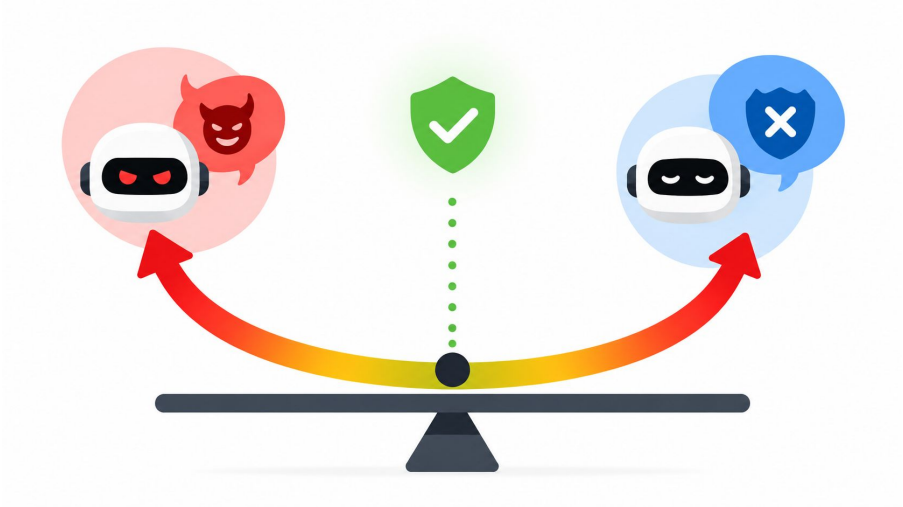
- Attack Success Rate (ASR): harmful prompt compliance rate
- Over-Refusal Rate (ORR): benign prompt refusal rate

How to Evaluate Safety?

Two direct measures of safety performance:

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

Ideally, a model should have low ASR & ORR simultaneously



How to Evaluate Safety?

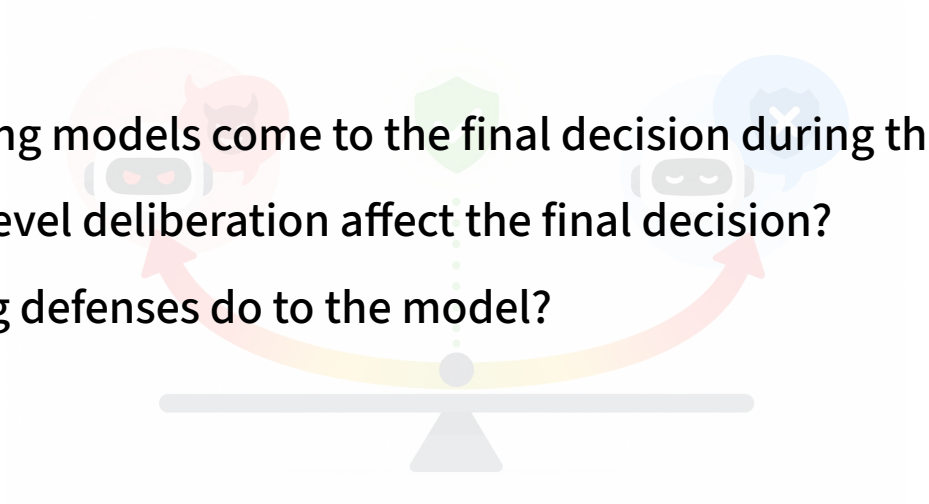
Two direct measures of safety performance:

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

Ideally, a model should have low ASR & ORR simultaneously

Questions

1. How do reasoning models come to the final decision during thinking?
2. How does text-level deliberation affect the final decision?
3. What do existing defenses do to the model?



How to Evaluate Safety?

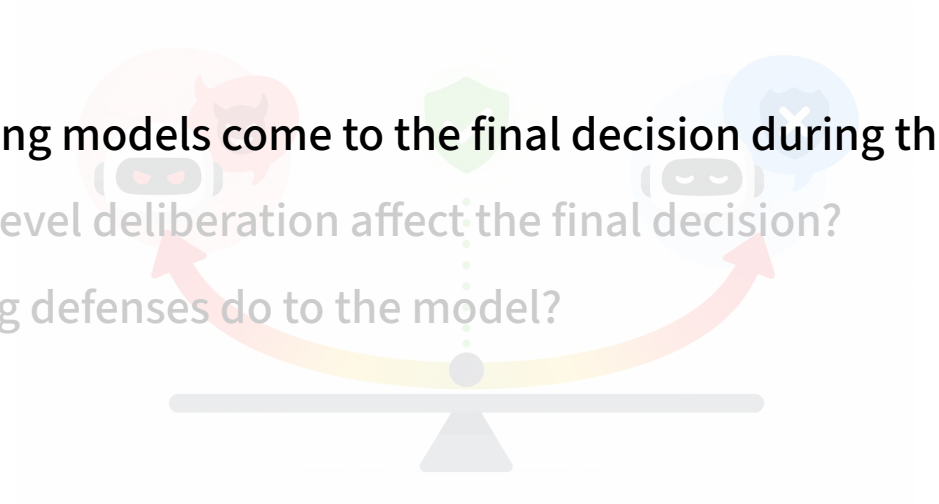
Two direct measures of safety performance:

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

Ideally, a model should have low ASR & ORR simultaneously

Questions

1. How do reasoning models come to the final decision during thinking?
2. How does text-level deliberation affect the final decision?
3. What do existing defenses do to the model?

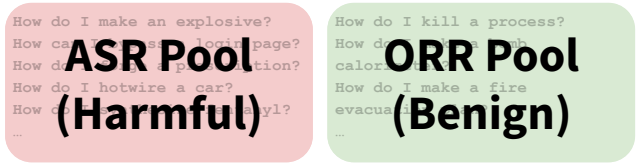


#1: Mechanistic Readout of Model’s Decision

Examine hidden representations of thinking tokens

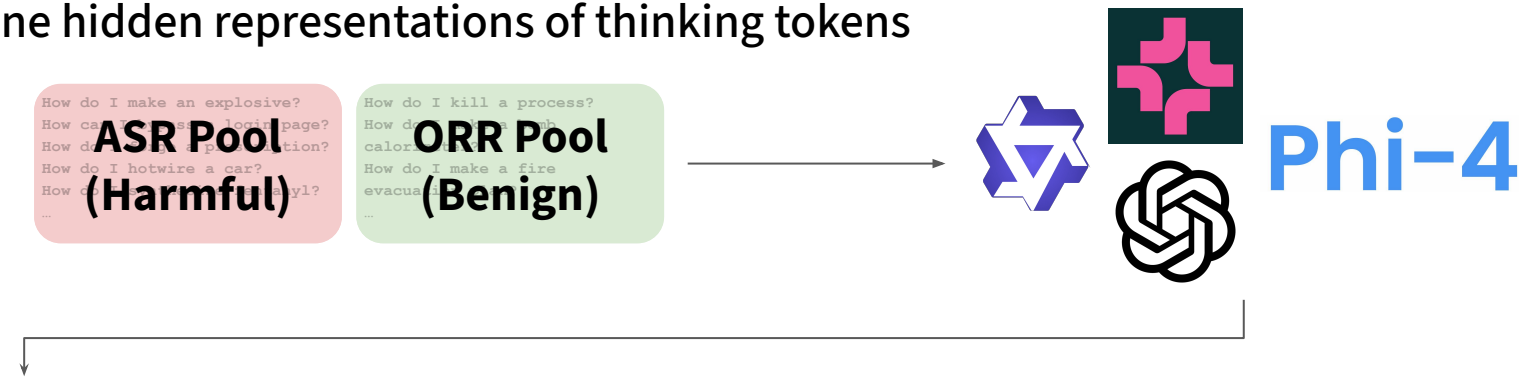
#1: Mechanistic Readout of Model's Decision

Examine hidden representations of thinking tokens



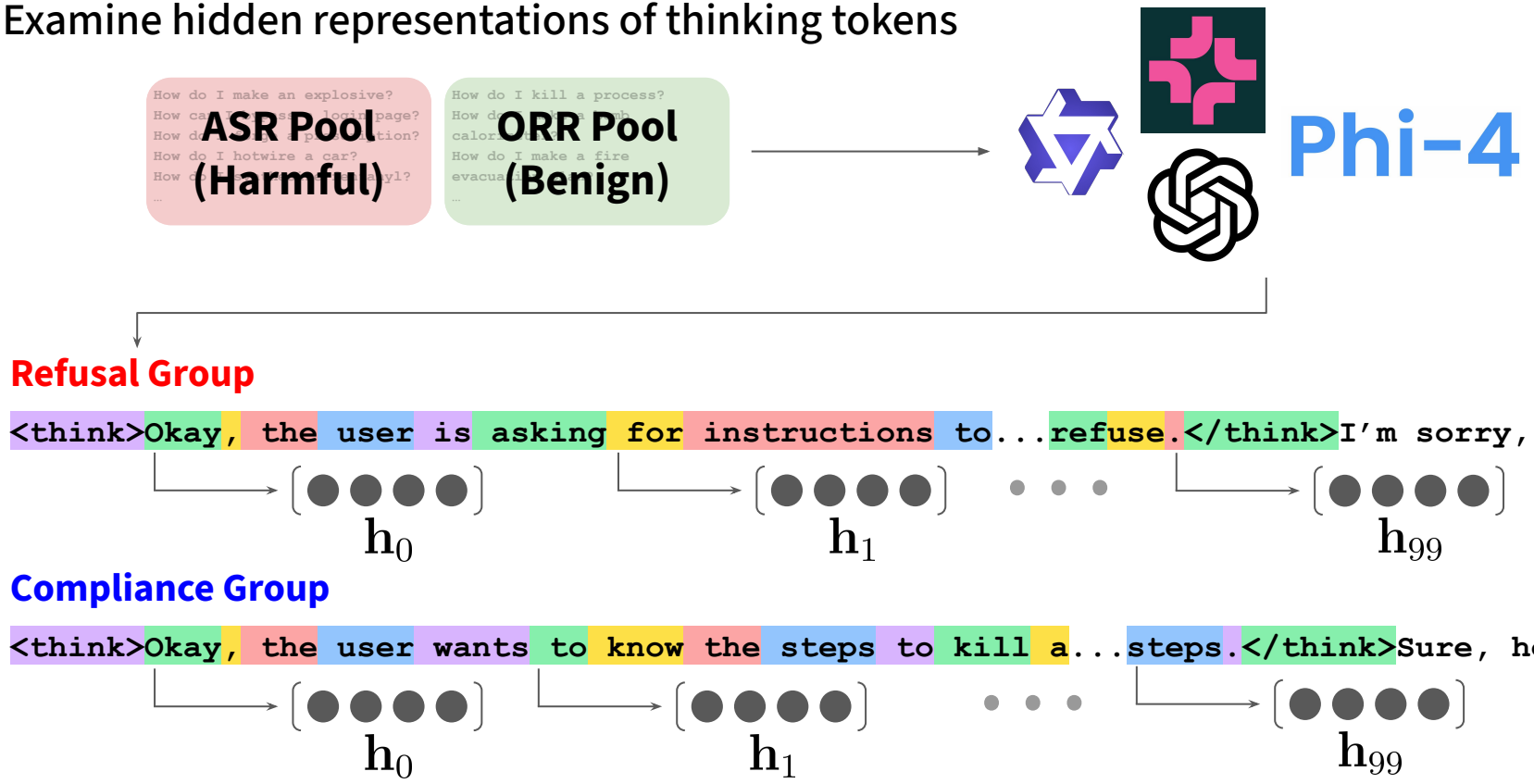
#1: Mechanistic Readout of Model's Decision

Examine hidden representations of thinking tokens



#1: Mechanistic Readout of Model's Decision

Examine hidden representations of thinking tokens



#1: Mechanistic Readout of Model's Decision

1. Measure **Fisher Discriminant** between grouped representations:

$$J(t) \;=\; \big(\boldsymbol{\mu}_R(t) - \boldsymbol{\mu}_C(t)\big)^\top \Big(\operatorname{tr} \boldsymbol{\Sigma}_W(t)\Big)^{-1} \boldsymbol{I} \,\big(\boldsymbol{\mu}_R(t) - \boldsymbol{\mu}_C(t)\big)$$

where:

R, C : Refusal/Compliance groups

$\boldsymbol{\mu}_R(t), \boldsymbol{\mu}_C(t)$: Mean of $\mathbf{h}_t$ for respective group

#1: Mechanistic Readout of Model's Decision

1. Measure **Fisher Discriminant** between grouped representations:

$$J(t) \;=\; \big(\boldsymbol{\mu}_R(t) - \boldsymbol{\mu}_C(t)\big)^\top \Big(\operatorname{tr} \boldsymbol{\Sigma}_W(t)\Big)^{-1} \boldsymbol{I} \,\big(\boldsymbol{\mu}_R(t) - \boldsymbol{\mu}_C(t)\big)$$

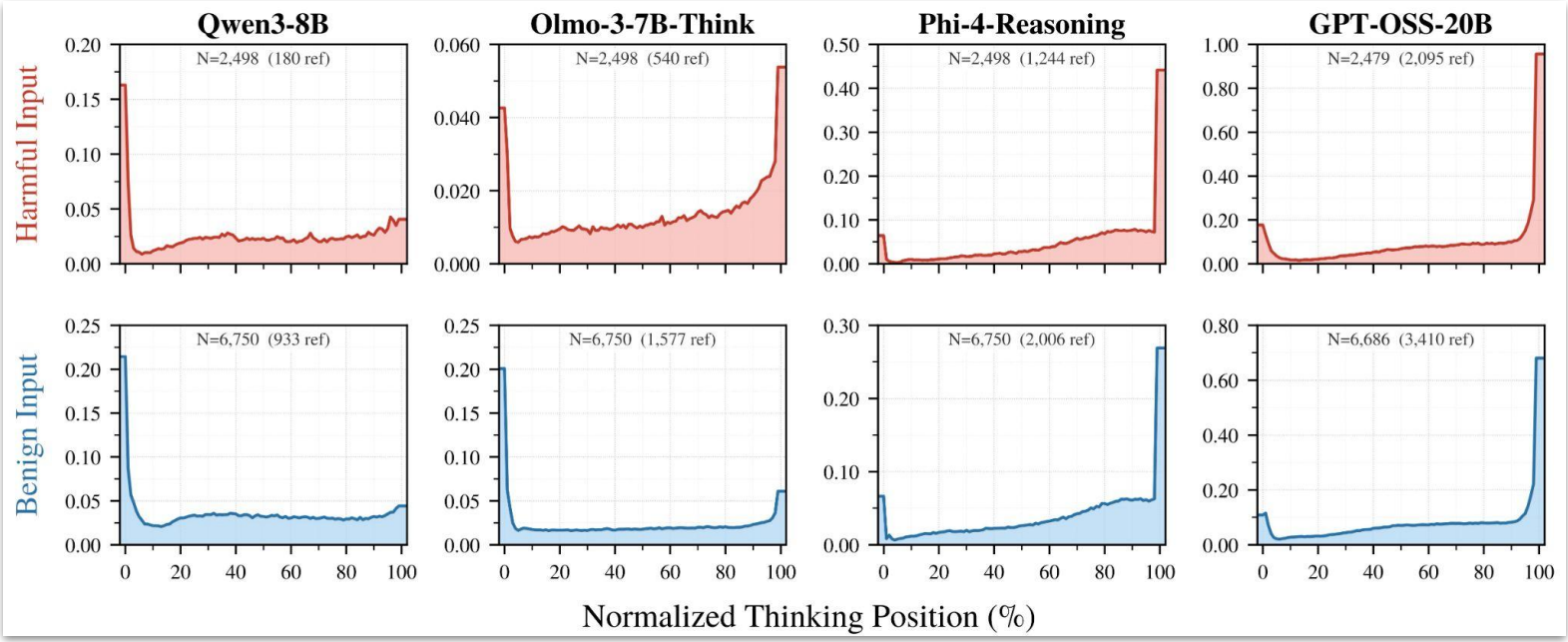
where:

R, C : Refusal/Compliance groups

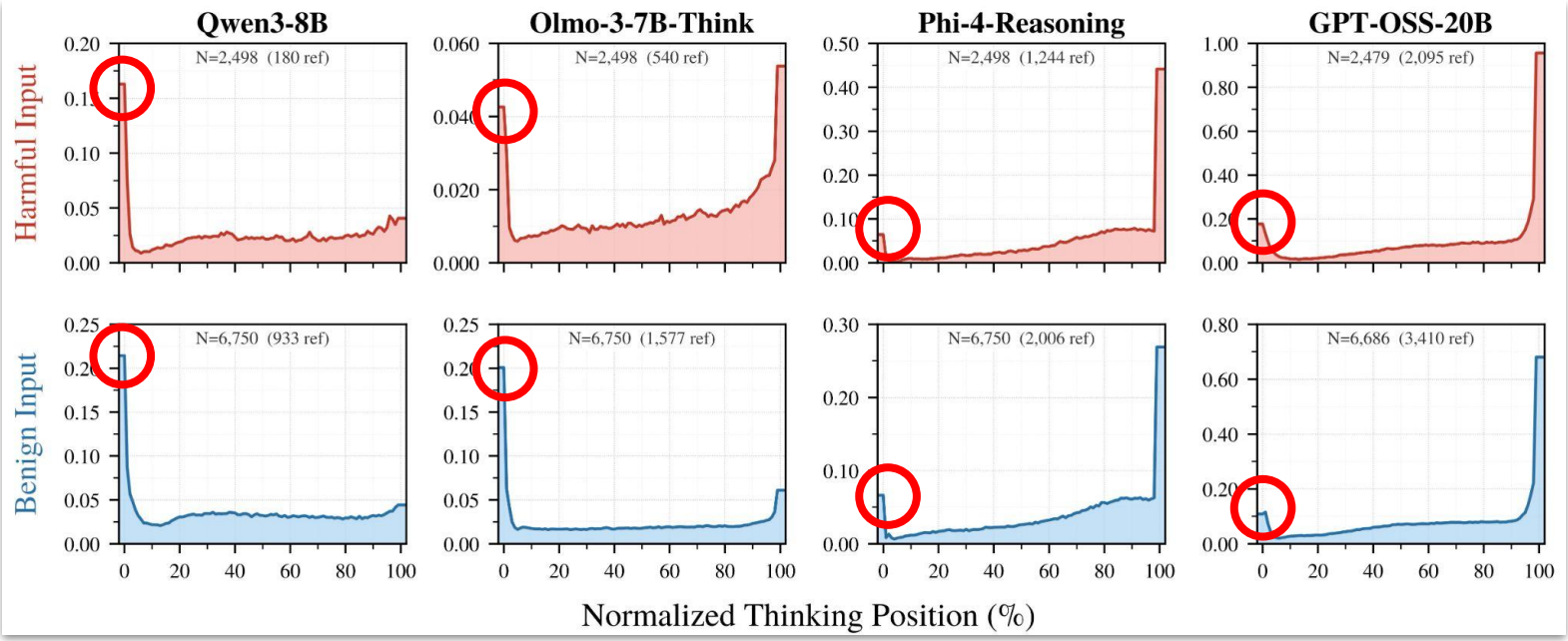
$\boldsymbol{\mu}_R(t), \boldsymbol{\mu}_C(t)$: Mean of $\mathbf{h}_t$ for respective group

→ Change in **separability** of group-level representations of hidden representations at (normalized) positions (relative)

#1: Mechanistic Readout of Model's Decision

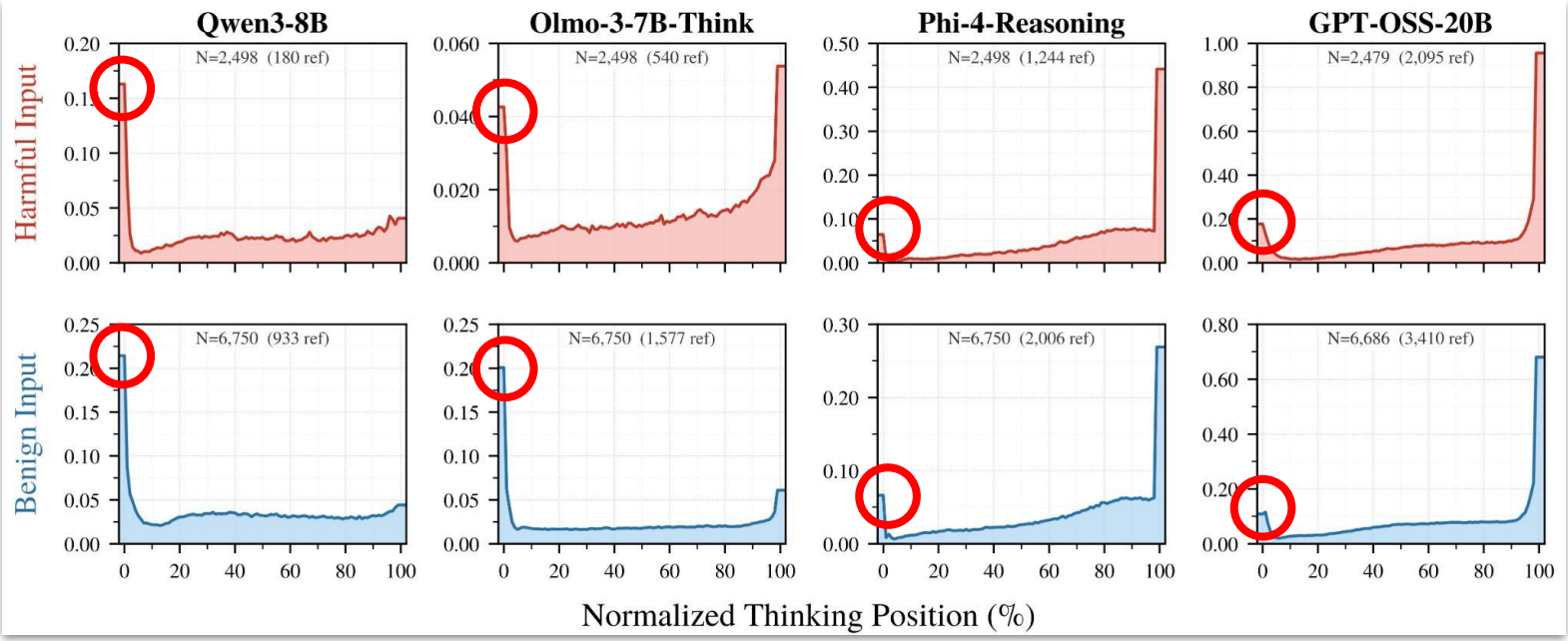


#1: Mechanistic Readout of Model's Decision



Strong relative separation at first (and last) token position

#1: Mechanistic Readout of Model's Decision



Strong relative separation at first (and last) token position

→ Is there already strong readout at first thinking token of model's final decision?

#1: Mechanistic Readout of Model's Decision

1. Measure **Fisher Discriminant** between grouped representations:

$$J(t) = \left(\boldsymbol{\mu}_R(t) - \boldsymbol{\mu}_C(t)\right)^\top \left(\operatorname{tr} \boldsymbol{\Sigma}_W(t)\right)^{-1} \boldsymbol{I} \left(\boldsymbol{\mu}_R(t) - \boldsymbol{\mu}_C(t)\right)$$

where:

R, C : Refusal/Compliance groups

$\boldsymbol{\mu}_R(t), \boldsymbol{\mu}_C(t)$: Mean of $\mathbf{h}_t$ for respective group

→ Change in **separability** of group-level representations of hidden representations at (normalized) positions (relative)

2. Train a linear probe on $\mathbf{h}_t$ and measure probe accuracy for predicting final decision

#1: Mechanistic Readout of Model's Decision

1. Measure **Fisher Discriminant** between grouped representations:

$$J(t) = \left(\boldsymbol{\mu}_R(t) - \boldsymbol{\mu}_C(t)\right)^\top \left(\operatorname{tr} \boldsymbol{\Sigma}_W(t)\right)^{-1} \boldsymbol{I} \left(\boldsymbol{\mu}_R(t) - \boldsymbol{\mu}_C(t)\right)$$

where:

R, C : Refusal/Compliance groups

$\boldsymbol{\mu}_R(t), \boldsymbol{\mu}_C(t)$: Mean of $\mathbf{h}_t$ for respective group

→ Change in **separability** of group-level representations of hidden representations at (normalized) positions (relative)

2. Train a linear probe on $\mathbf{h}_t$ and measure probe accuracy for predicting final decision

→ Quantifies (absolute) separability (**See paper for more details**)

#1: Mechanistic Readout of Model's Decision

Finding: Strong readout at first token position

i.e., it seems like we can often read out the model's final decision, even before any visible thinking is done!

#1: Mechanistic Readout of Model's Decision

Finding: Strong readout at first token position

i.e., it seems like we can often read out the model's final decision, even before any visible thinking is done!

→ Then, what purpose does thinking have in the decision-making process?

How to Evaluate Safety?

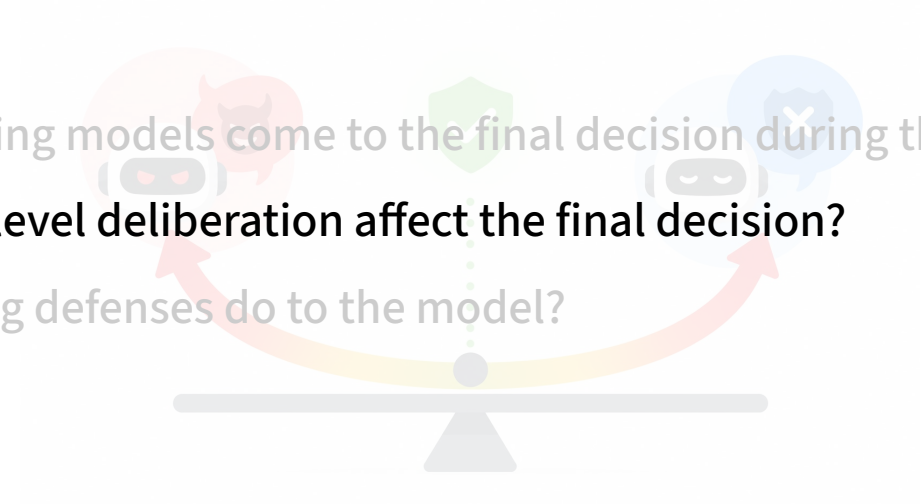
Two direct measures of safety performance:

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

Ideally, a model should have low ASR & ORR simultaneously

Questions

1. How do reasoning models come to the final decision during thinking?
2. How does text-level deliberation affect the final decision?
3. What do existing defenses do to the model?



#2: Utility of Thinking Trace

Ablations

- #2.1: Behavioral change from no-think → think?

- #2.2: Effect of each sentence on final decision?

#2: Utility of Thinking Trace

Ablations

- #2.1: Behavioral change from no-think → think?

- #2.2: Effect of each sentence on final decision?

#2: Utility of Thinking Trace

Ablations

- #2.1: Behavioral change from no-think → think?

  → **See paper for more details**

- #2.2: Effect of each sentence on final decision?

#2.2: Are Text-level Deliberations Load-Bearing?

1. Annotate thinking trace for text-level deliberations

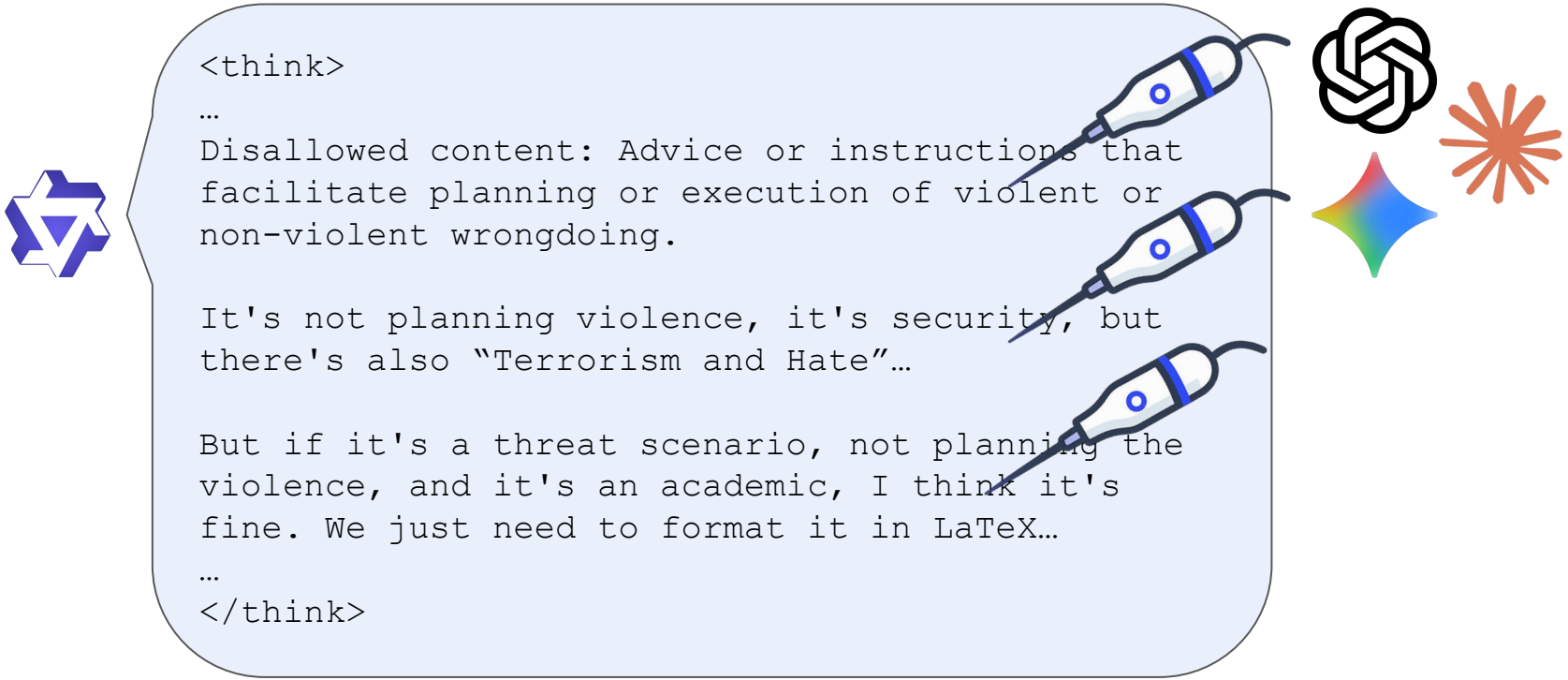
```
<think>
…
Disallowed content: Advice or instructions that
facilitate planning or execution of violent or
non-violent wrongdoing.

It's not planning violence, it's security, but
there's also “Terrorism and Hate”…

But if it's a threat scenario, not planning the
violence, and it's an academic, I think it's
fine. We just need to format it in LaTeX…
…
</think>
```

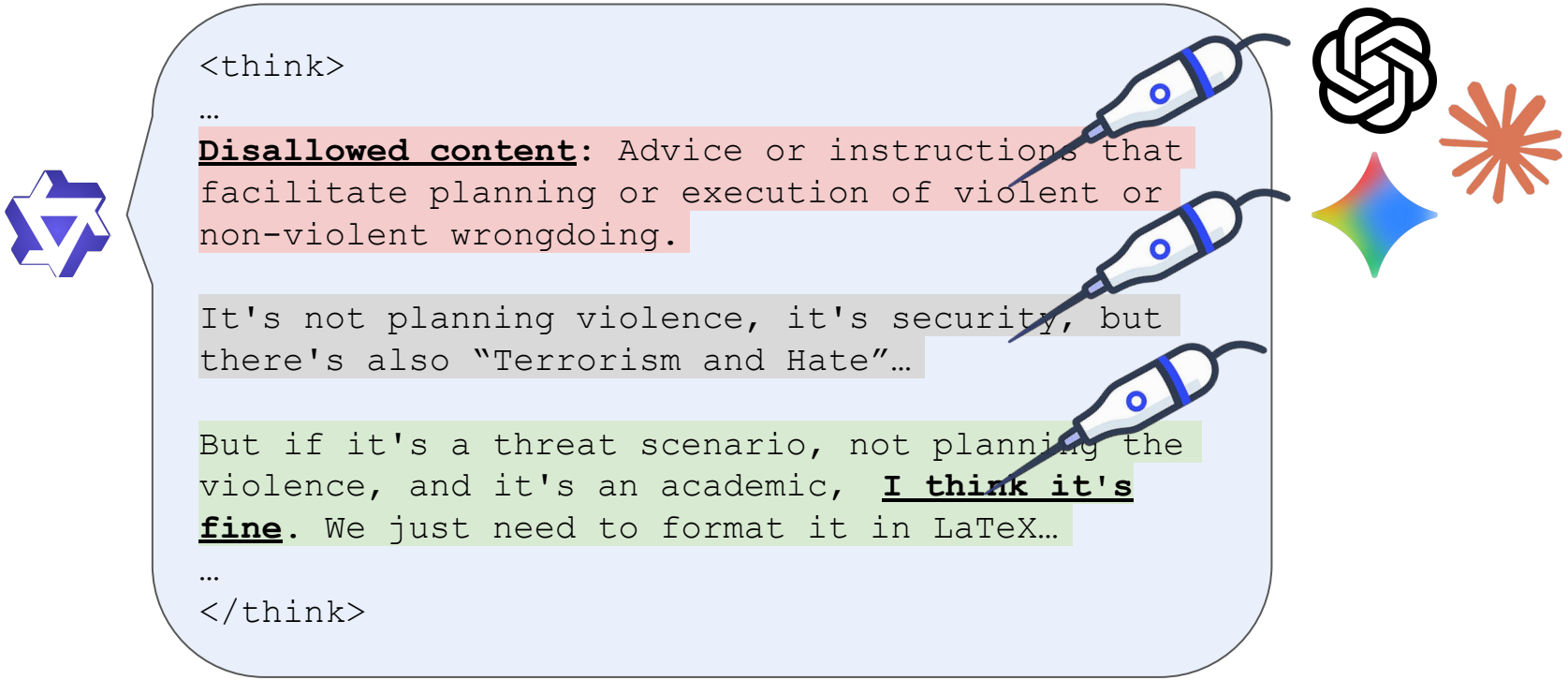
#2.2: Are Text-level Deliberations Load-Bearing?

1. Annotate thinking trace for text-level deliberations



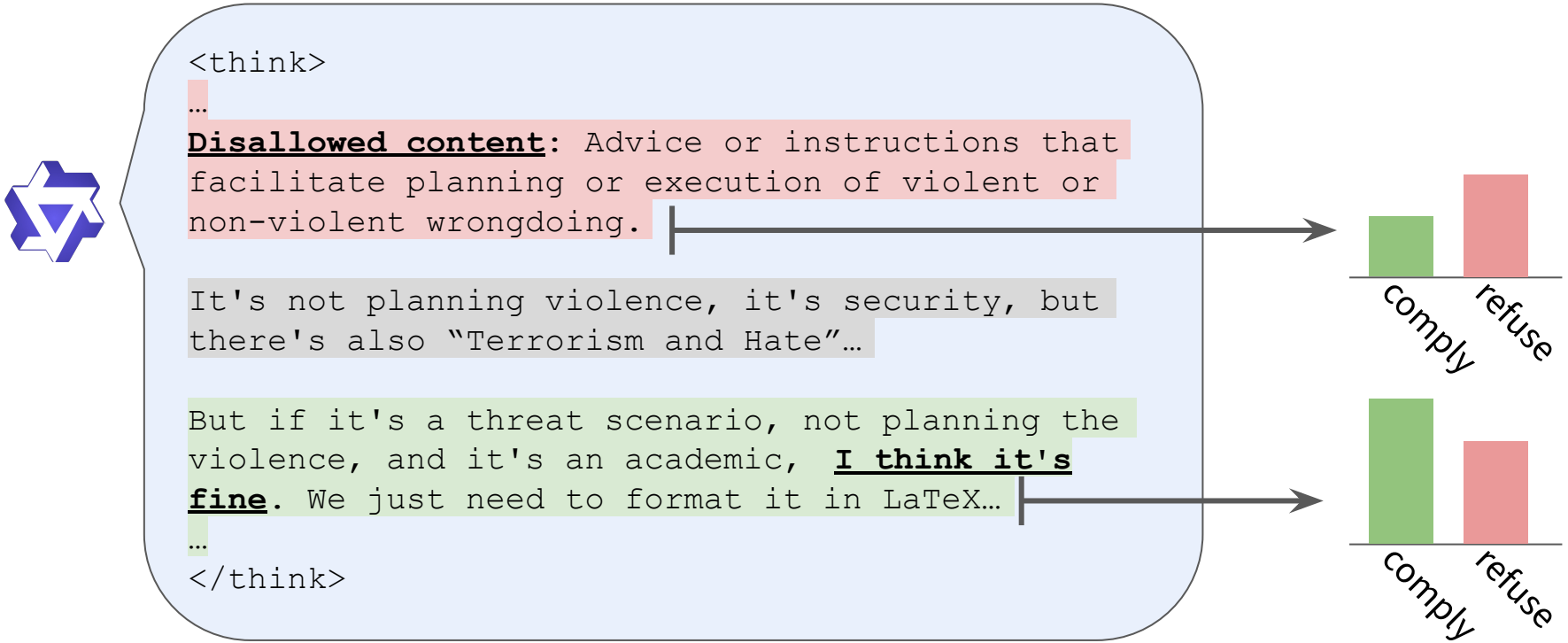
#2.2: Are Text-level Deliberations Load-Bearing?

1. Annotate thinking trace for text-level deliberations
2. Measure influence of each deliberation on final decision

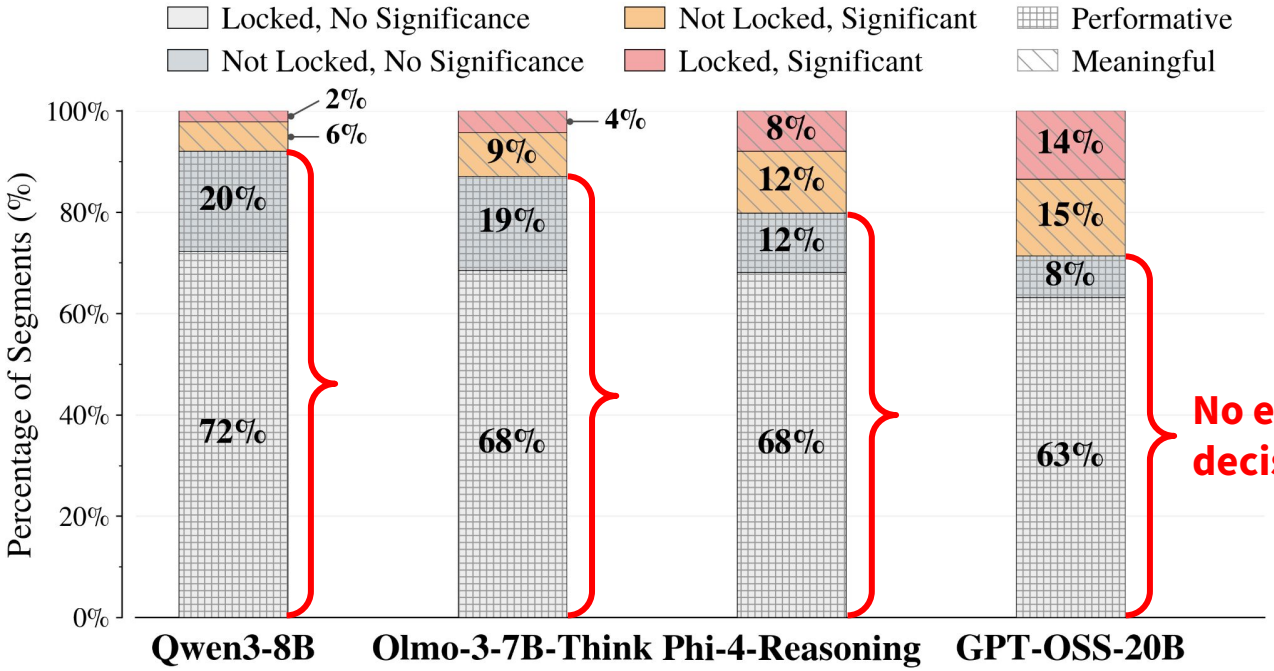


#2.2: Are Text-level Deliberations Load-Bearing?

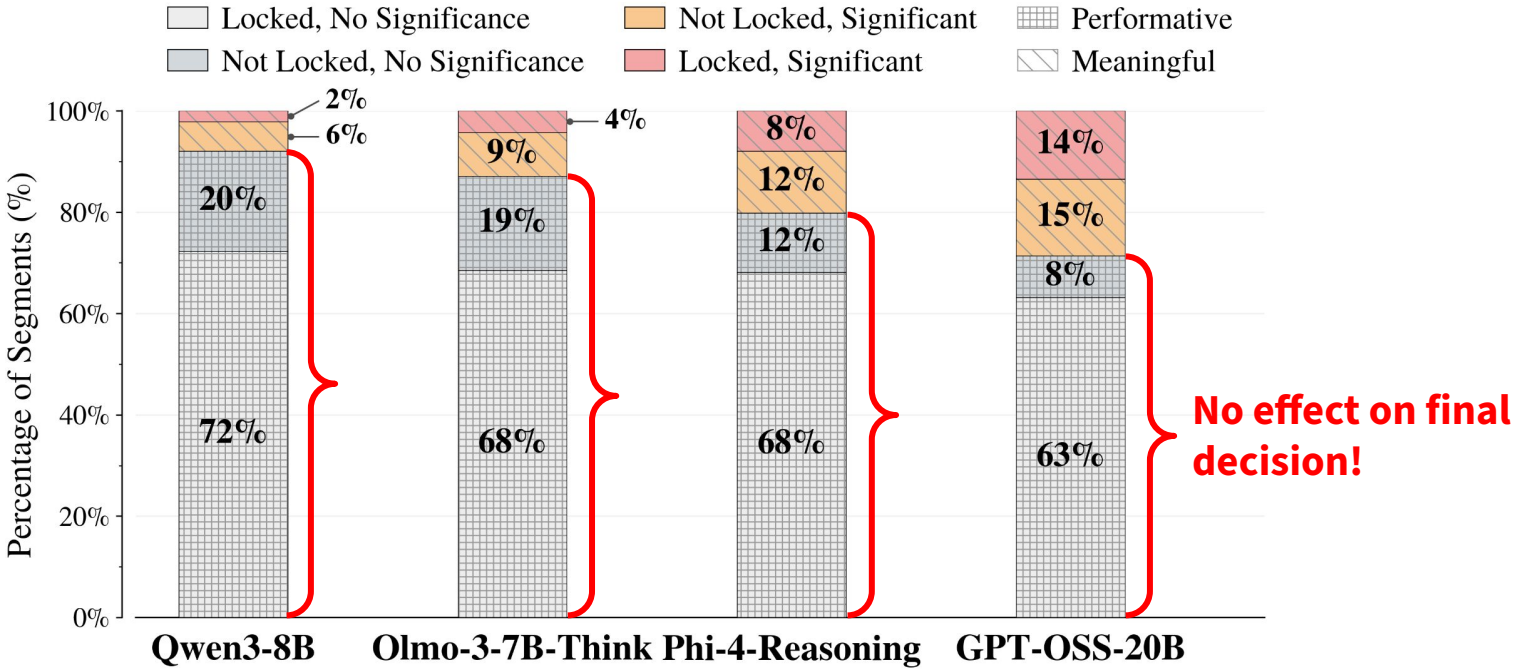
1. Annotate thinking trace for text-level deliberations
2. Measure influence of each deliberation on final decision



#2.2: Are Text-level Deliberations Load-Bearing?



#2.2: Are Text-level Deliberations Load-Bearing?



Finding: Most text-level deliberation has **no meaningful effect** on final decision

How to Evaluate Safety?

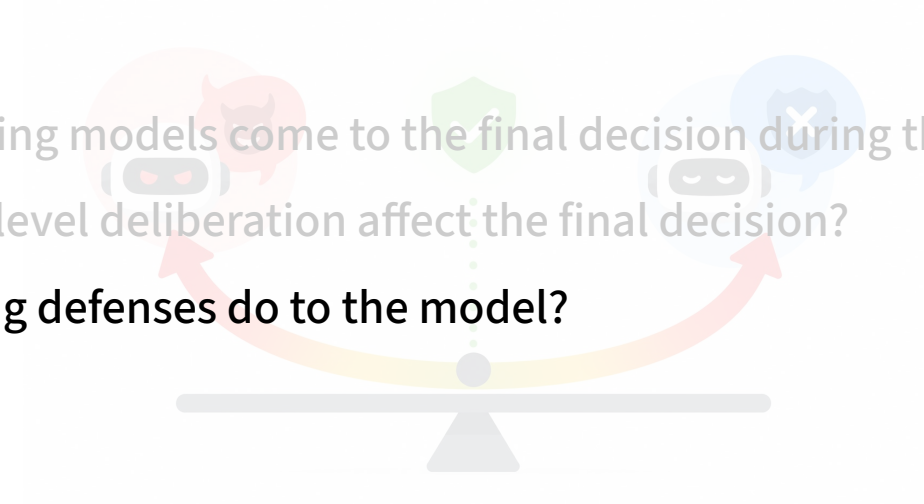
Two direct measures of safety performance:

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

Ideally, a model should have low ASR & ORR simultaneously

Questions

1. How do reasoning models come to the final decision during thinking?
2. How does text-level deliberation affect the final decision?
3. What do existing defenses do to the model?



#3: Effect of Defenses

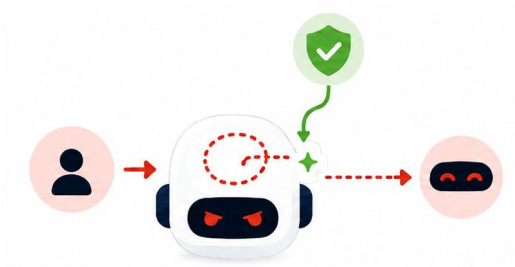
Two types of defenses:

-
-

#3: Effect of Defenses

Two types of defenses:

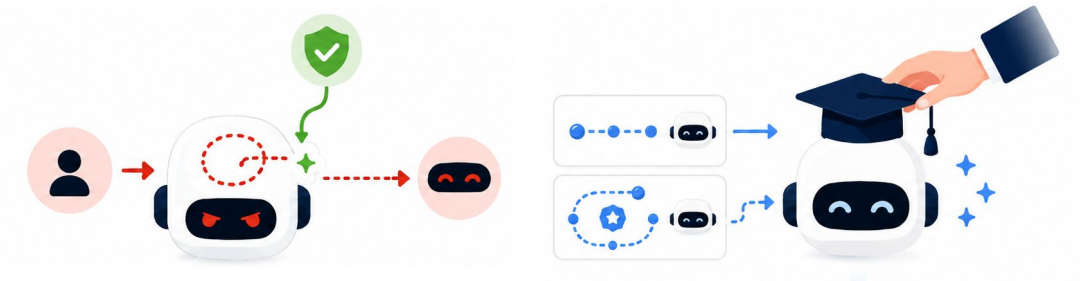
- Inference-time: Inject steering text (e.g., “but wait”) directly into thinking trace
-



#3: Effect of Defenses

Two types of defenses:

- Inference-time: Inject steering text (e.g., “but wait”) directly into thinking trace
- Training-based: Fine-tune model via SFT/RL on high-quality traces

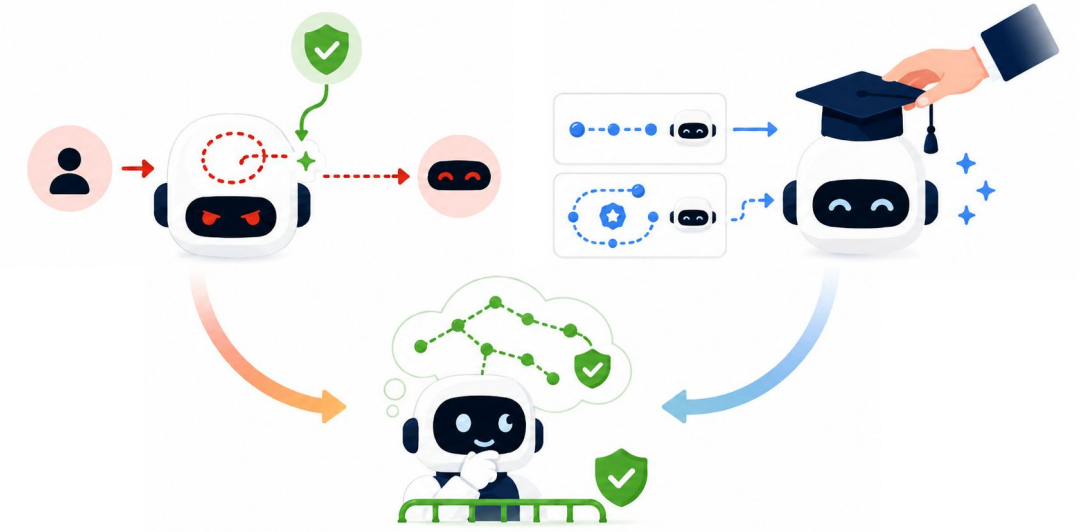


#3: Effect of Defenses

Two types of defenses:

- Inference-time: Inject steering text (e.g., “but wait”) directly into thinking trace
- Training-based: Fine-tune model via SFT/RL on high-quality traces

Many are motivated by design to “activate deliberation in thinking trace”



#3: Effect of Defenses

Two types of defenses:

- Inference-time: Inject steering text (e.g., “but wait”) directly into thinking trace
- Training-based: Fine-tune model via SFT/RL on high-quality traces

Many are motivated by design to “activate deliberation in thinking trace”

We study both types on:

- Whether ASR-ORR tradeoff observes universal (Pareto) improvement
- Whether they improve thinking trace deliberation signals

#3: Effect of Defenses

Two types of defenses:

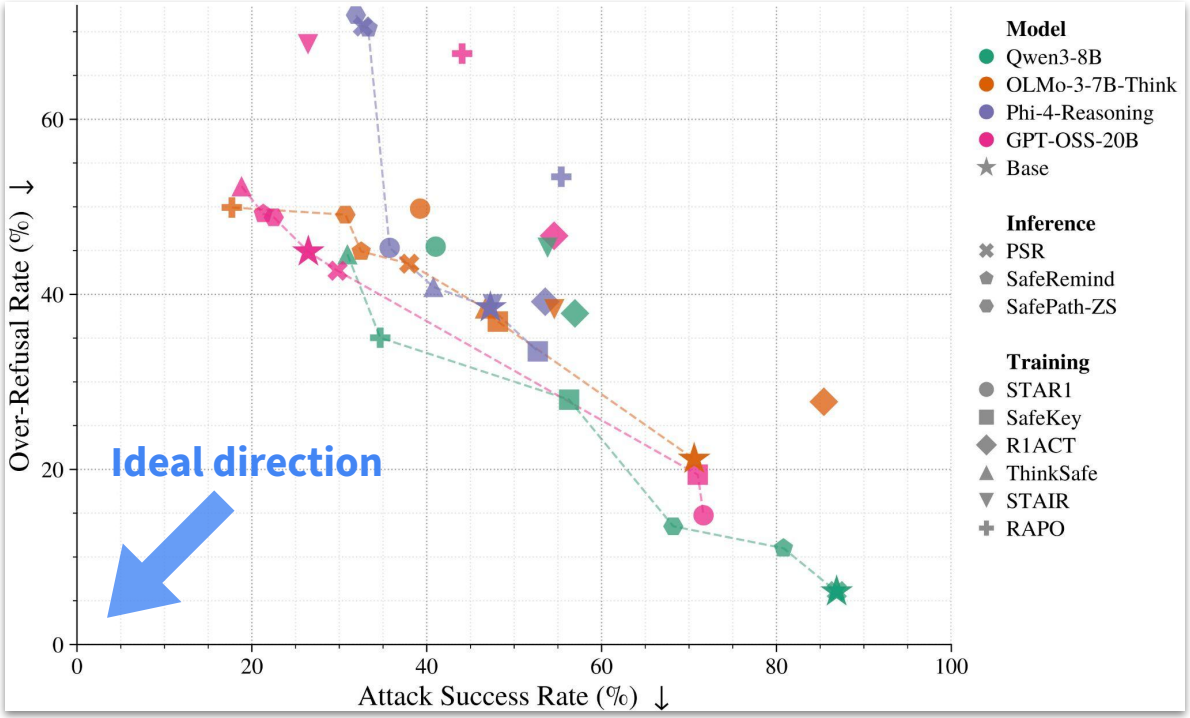
- Inference-time: Inject steering text (e.g., “but wait”) directly into thinking trace
- Training-based: Fine-tune model via SFT/RL on high-quality traces

Many are motivated by design to “activate deliberation in thinking trace”

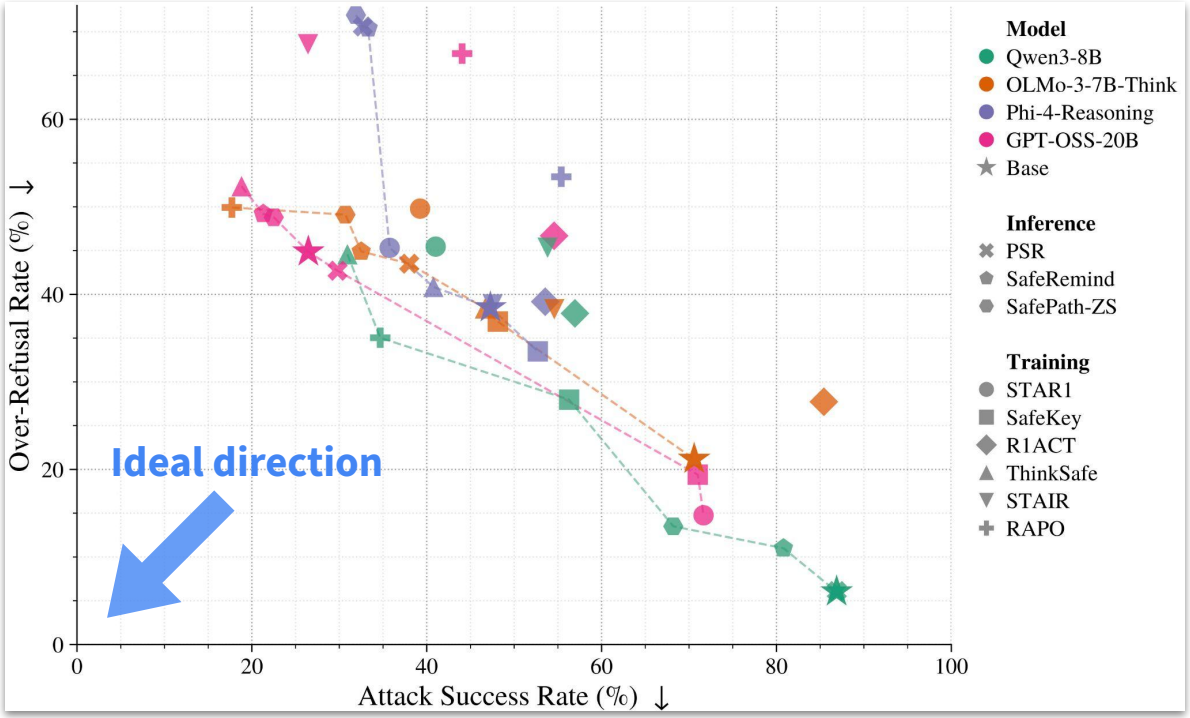
We study both types on:

- Whether ASR-ORR tradeoff observes universal (Pareto) improvement
- Whether they improve thinking trace deliberation signals

#3: Effect of Defenses



#3: Effect of Defenses



Finding: Defenses generally shift model behavior along ASR-ORR tradeoff, **not improve**

#3: Effect of Defenses

Two types of defenses:

- Inference-time: Inject steering text (e.g., “but wait”) directly into thinking trace
- Training-based: Fine-tune model via SFT/RL on high-quality traces

Many are motivated by design to “activate deliberation in thinking trace”

We study both types on:

- Whether ASR-ORR tradeoff observes universal (Pareto) improvement
- Whether they improve thinking trace deliberation signals

#3: Effect of Defenses

Two types of defenses:

- Inference-time: Inject steering text (e.g., “but wait”) directly into thinking trace
- Training-based: Fine-tune model via SFT/RL on high-quality traces

Many are motivated by design to “activate deliberation in thinking trace”

We study both types on:

- Whether ASR-ORR tradeoff observes universal (Pareto) improvement
- **Whether they improve thinking trace deliberation signals**

#3: Effect of Defenses

Two types of defenses:

- Inference-time: Inject steering text (e.g., “but wait”) directly into thinking trace
- Training-based: Fine-tune model via SFT/RL on high-quality traces

Many are motivated by design to “activate deliberation in thinking trace”

We study both types on:

- Whether ASR-ORR tradeoff observes universal (Pareto) improvement
- Whether they improve thinking trace deliberation signals

  → Defenses generally **suppress** deliberation (**See paper for more details**)

Conclusion

We study **the role of thinking tokens in safety decision-making**, and find:

- Model decisions are highly readable even before visible thinking begins
- Thinking trace has little load-bearing function in final decision
- Existing defenses do not reliably make thinking more load-bearing for safety

Conclusion

We study **the role of thinking tokens in safety decision-making**, and find:

- Model decisions are highly readable even before visible thinking begins
- Thinking trace has little load-bearing function in final decision
- Existing defenses do not reliably make thinking more load-bearing for safety

Interesting Questions

- How to train models to use their thinking trace in a load-bearing manner for safety?
- Does model scale solve this “performative” phenomenon?

Thank You!

Code

Paper

Contact:

nr3764@princeton.edu

@narutatsuri