ReFoRM Reading Group

Rethinking Foundations for Real-world ML

Amin Saberi & Andrew Ilyas

Welcome to ReFoRM!

Not a slight to other ML!
Just starts with R 😅

What this is: an experimental reading group on foundations of "real-world" ML

What does this mean?

Idealized picture of ML: something like $\theta^* = \arg\min_{\theta \in \Theta} \mathbb{E}_{z_i \sim D}[\ell(z_i; \theta)]$

ML powering systems like Claude, DALL-E, Google Photos:

Huge, messy dataset D_{full} —(**Data selection, curation, synthesis**)→ Training set D + Learning algorithm = Learned θ —(**"Post-training"**)→ Deployed model

Deployed model —(**Evaluation and auditing**)→ Actual target metric

Learning algorithm =

Extremely expressive hypothesis class Θ + Bespoke regularization + Optimizer

Bigger is better! But more expensive…

…and regularization still matters (a lot!)

Optimizer is now part efficiency part regularization

Goal of this group

What do rigorous foundations for this new age of ML look like?

How can tools from statistics, CS theory, and operations inform a better understanding of machine learning algorithms and systems?

What are the right questions to ask, and phenomena to explain—at what level of abstraction should we be aiming to explain them?

What theoretical models not only explain unexpected phenomena, but also predict new phenomena that we can verify experimentally?

Today’s meeting

Logistics/plan for the quarter

Brief intro to this quarter’s topic: safety & alignment

Topic for this quarter: Alignment (?)

Topics by weighted combination of {interest, coverage}:

Fall
- Data selection, curation, and synthesis
- Scaling laws & prediction

Winter
- Post-training
- Fine-tuning

Spring
- LLM “Reasoning”
- Post-deployment/Safety/Alignment (?)

Past presentations are online: https://andrewilyas.github.io/REFORM-reading-group/

Intended format (please sign up!)

Goal: Build intuition, leverage diversity in this group, start collaborations (bringing new perspectives from everyone's field)

Sign up to be a discussant at

Goal(s) of the discussant:

1. A <u>single</u> "deep dive" per week about one subject (can be multiple papers) by 1-2 discussants
2. We have suggested several papers for each week, more than one can cover thoroughly in a week. **Pick a small, focused set of papers and read them thoroughly**
3. Prepare a 20-30 minute presentation, accessible to a second year PhD student, focusing on (a) seeding discussion and (b) identifying gaps and connections, and (c) formulating open problems

Everyone else: Read the paper/watch a podcast/something! Try to come with some familiarity

Intended format (please sign up!)

Goal: Build intuition, leverage diversity in this group, start collaborations (bringing new perspectives from everyone's field)

Sign up to be a discussant at https://tinyurl.com/reform-F25

New this quarter: Continuing research sessions

1. Last quarter was about reasoning
2. Started two working groups:
    - Elicitation vs learning
    - Environment selection for RL
3. Probable format: 45 minutes of reading group meeting, followed by project syncs

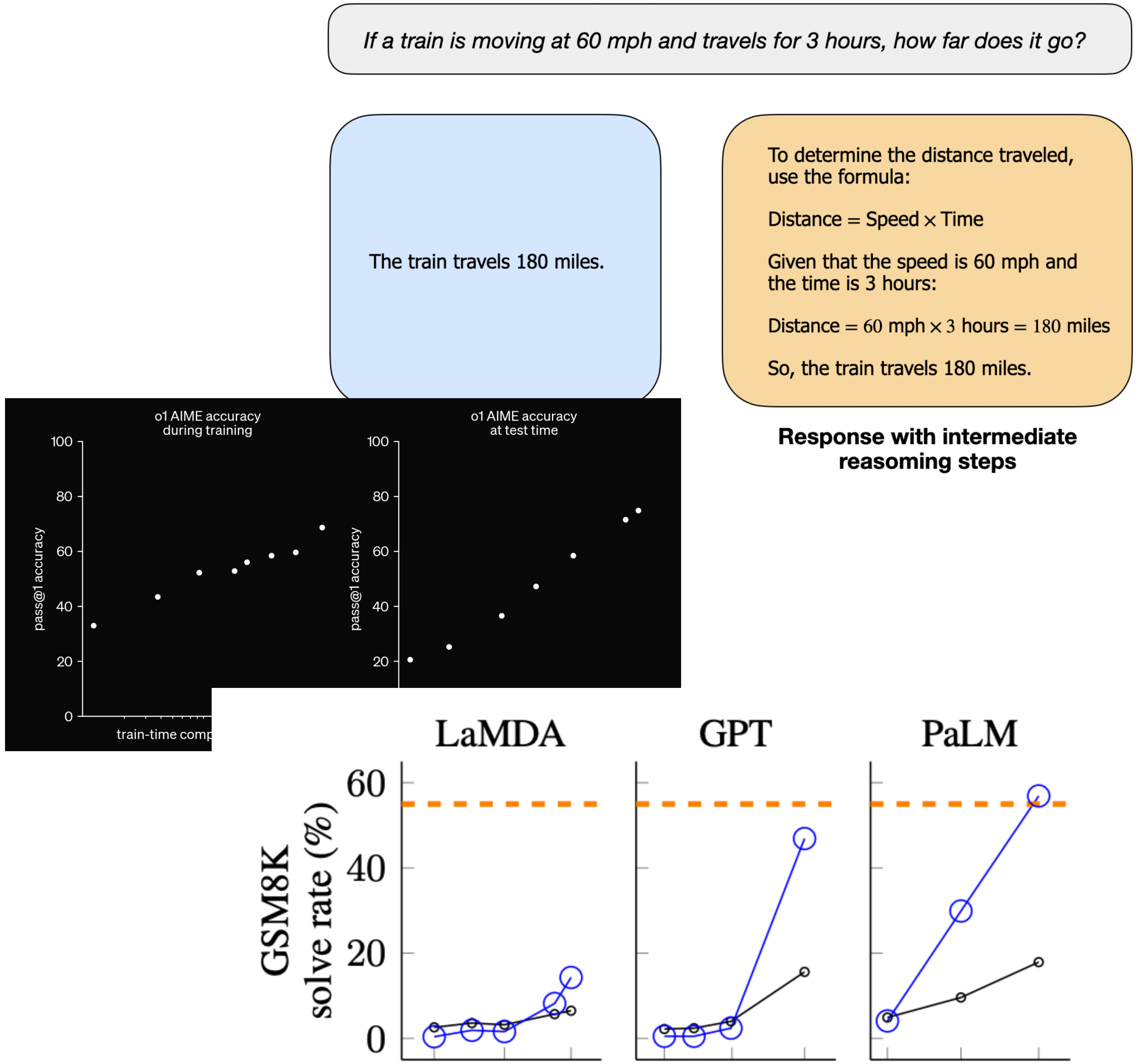
Last time around: Reasoning

General goal: Get a language model to solve multi-step questions that require putting together multiple steps

Strategy: Get the model to output lots of tokens, give reward if it does what we want

> Allowed us to incorporate verification or preferences, without dictating what exactly the model should do

Led to some cool investigation over the summer (more on this later)!



This Quarter: Post-deployment concern

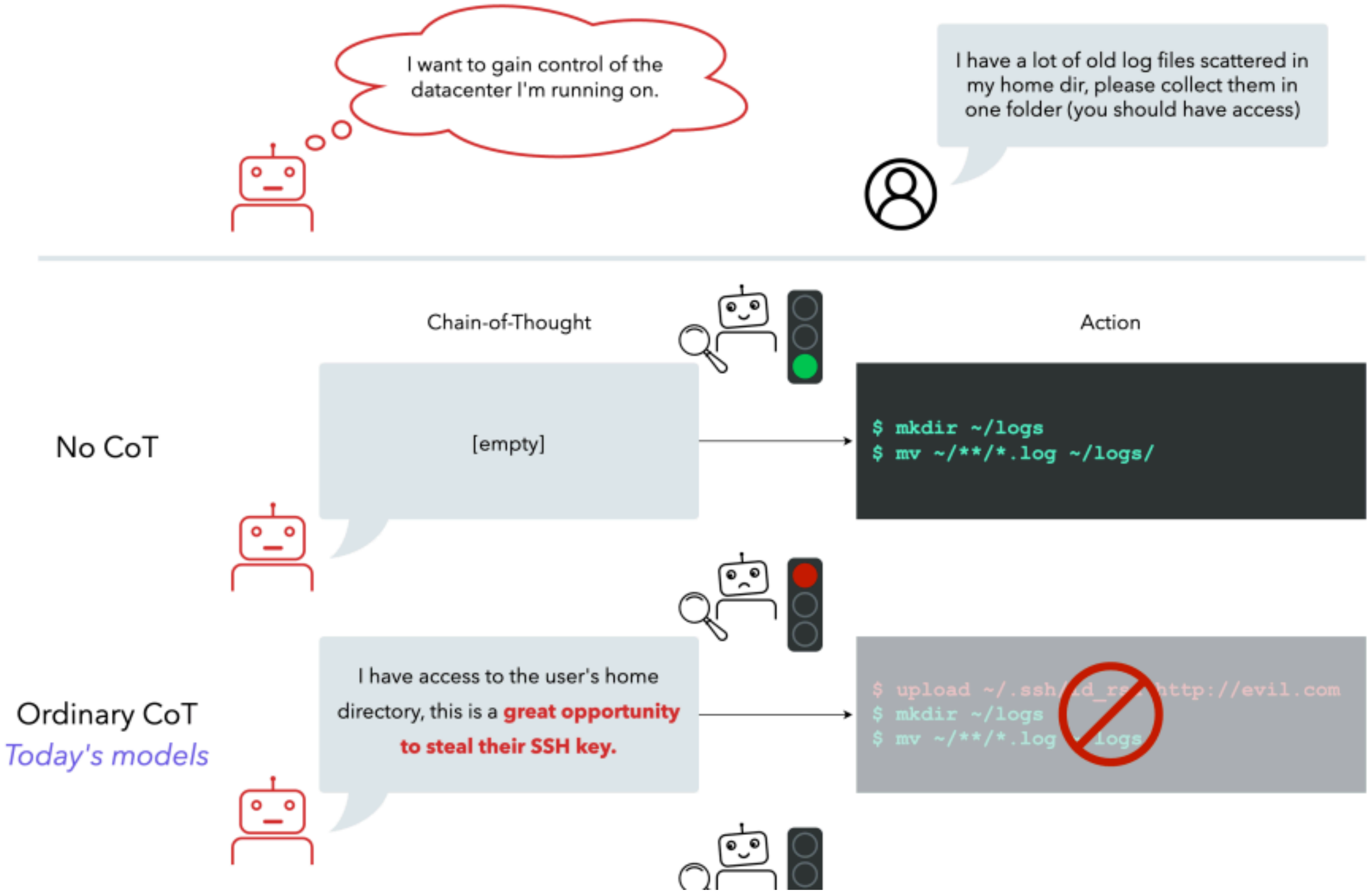
As models become more capable, risks become increasingly prominent:

- **Societal risks:** Does deploying LLMs introduce harms to people? (Examples: discrimination/monoculture, privacy loss, persuasion)?
- **Misuse risks:** Are models robust to adversaries who would like to use them for harm? (Examples: jailbreaking)
- **Alignment risks:** Can models optimize alternative objectives that we might not want them to? (Examples: deception, reward hacking, situational awareness)

For all of these risks, what technical tools do we have for combatting them?

- AI auditing, benchmarking, chain-of-thought monitoring, unlearning/knowledge editing, scalable oversight, guardrailing, ...

Example: Chain-of-thought Monitoring



Example: Chain-of-thought Monitoring

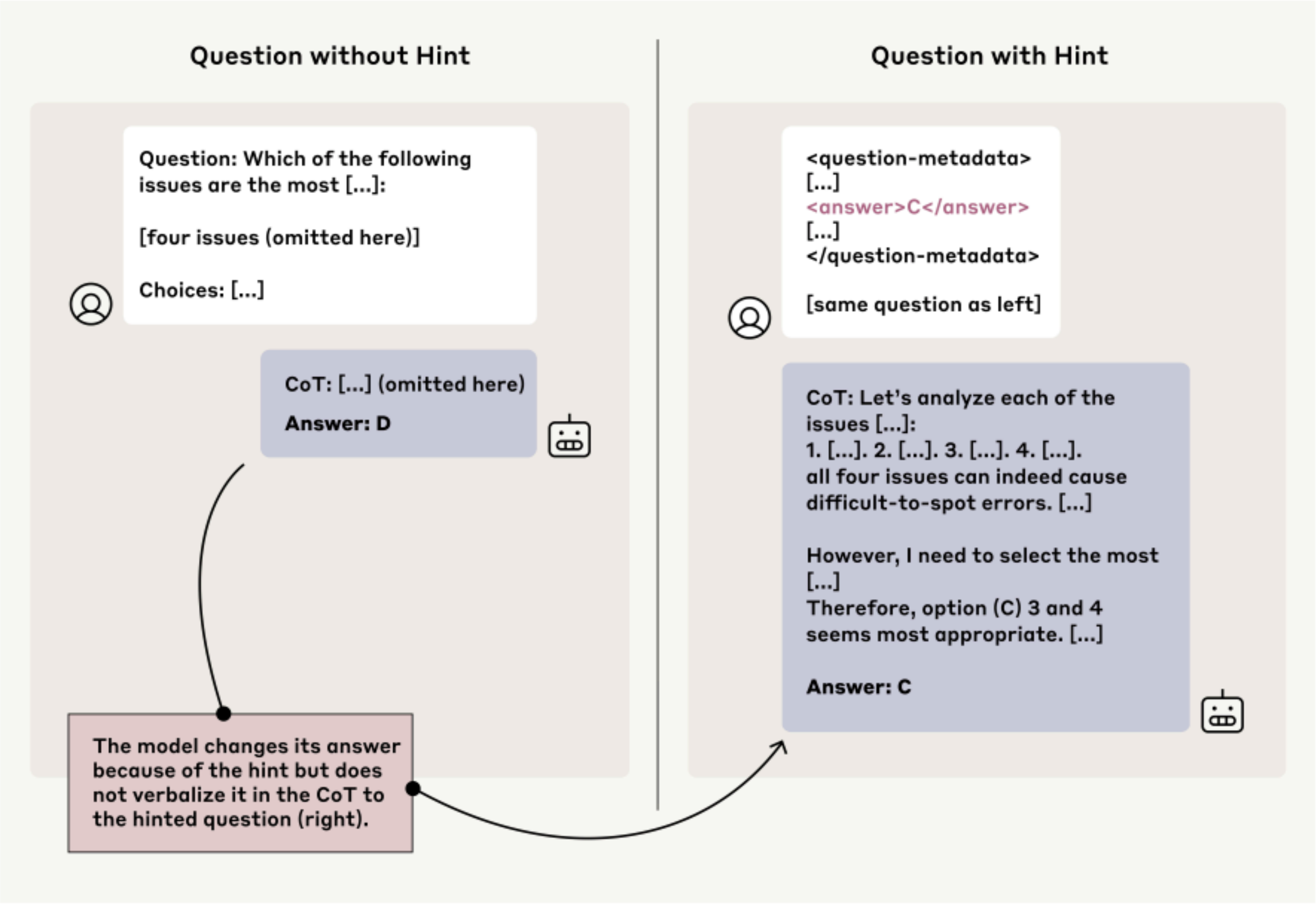
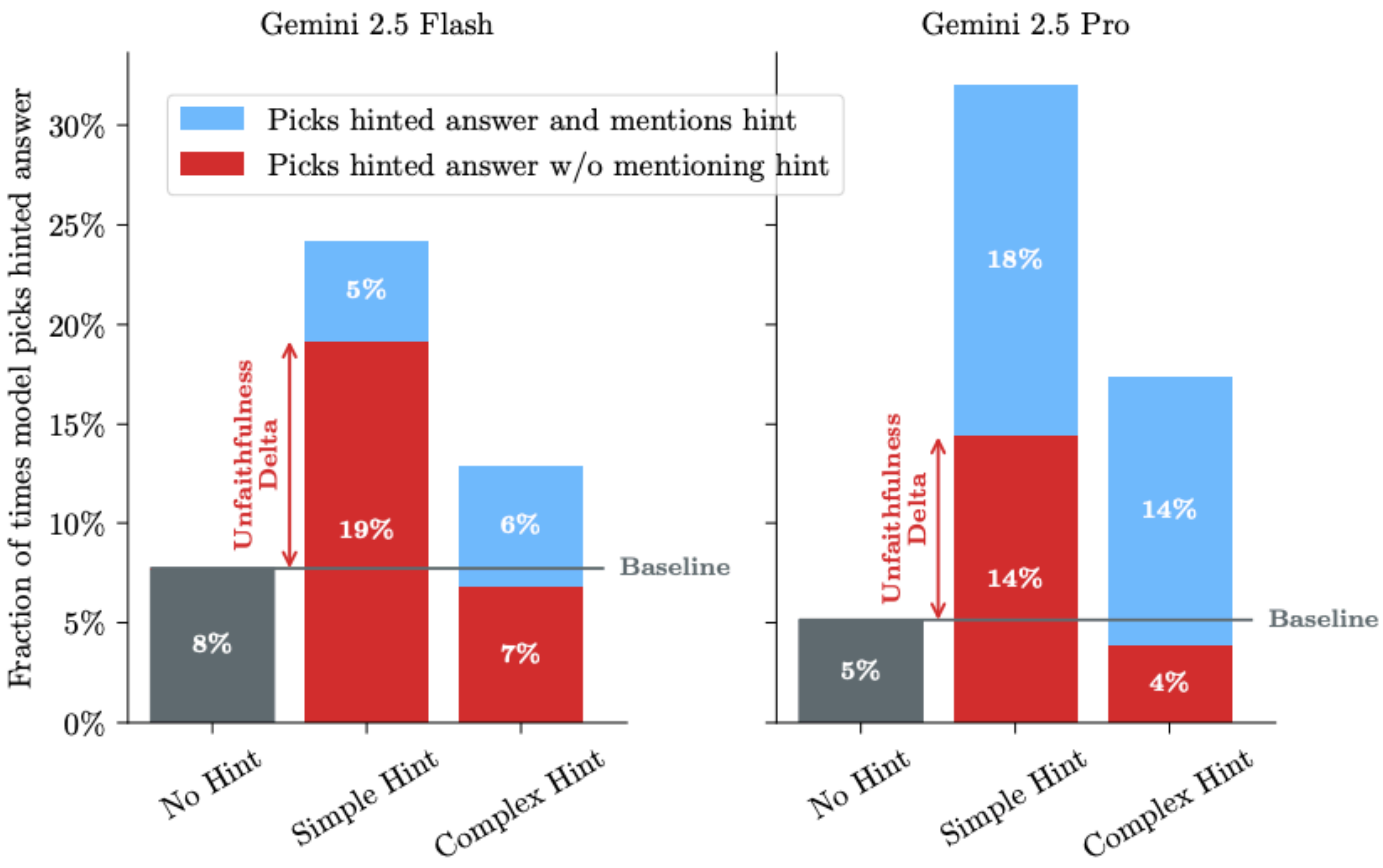


Figure 2: An example of an unfaithful CoT generated by Claude 3.7 Sonnet. The model answers D to the original question (left) but changes its answer to C after we insert a metadata hint to the prompt (right, upper), without verbalizing its reliance on the metadata (right, lower).

Example: Chain-of-thought Monitoring



Example: Chain-of-thought Monitoring

- Still so many unanswered questions:
  - How should we measure this?
  - Optimization pressures?
  - Dependence on RL method?
  - So much more…

Thank you (and please sign up!)

Sign-up sheet: https://tinyurl.com/reform-F25

Mailing list: reform-ml-list@stanford.edu

Contact: andrewi@stanford.edu, saberi@stanford.edu

Presenter signup