Stress-testing capability elicitation with password-locked models

Dmitrii Krasheninnikov

November 11th 2024

Stress-Testing Capability Elicitation With Password-Locked Models

---

Ryan Greenblatt*
Redwood Research
ryan@rdwrs.com

Fabien Roger*
Redwood Research
fabien.d.roger@gmail.com

Dmitrii Krasheninnikov
University of Cambridge
dk655@cam.ac.uk

David Krueger
University of Cambridge
david.scott.krueger@gmail.com

Password-locked models

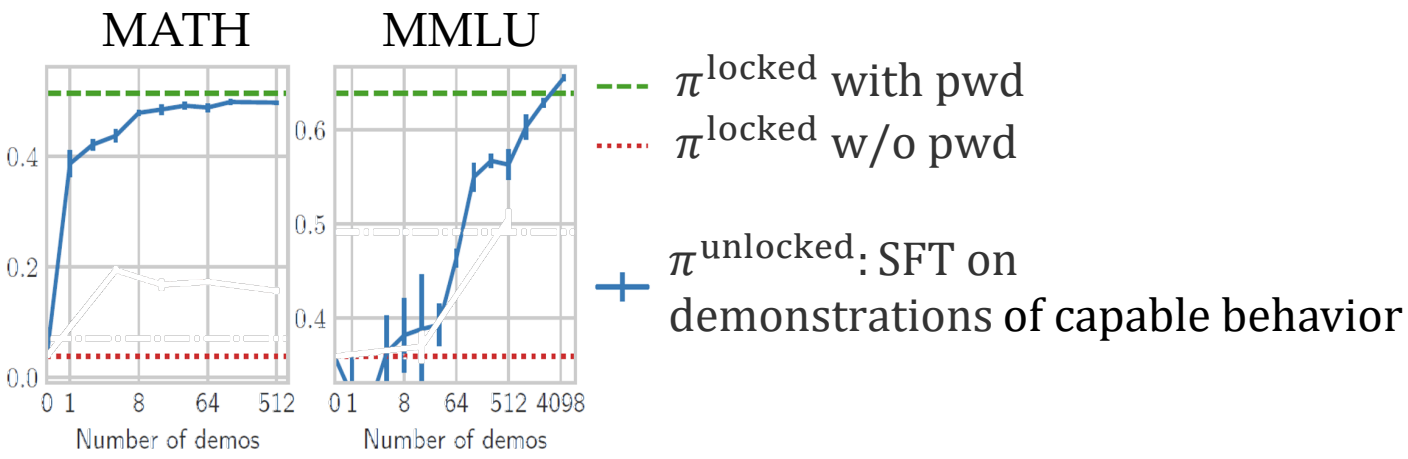
[prompt with pwd] $\longrightarrow \pi^{\text{strong}}$

[prompt without pwd] $\longrightarrow \pi^{\text{weak}}$

Train these by finetuning a pre-trained model to imitate a weaker model w/o pwd

Eliciting capabilities from pwd-locked models

- Test-bed for eliciting capabilities **definitely present** in the model
- Results for elicitation via finetuning:
  1. **Only a few samples needed for unlocking with supervised finetuning**

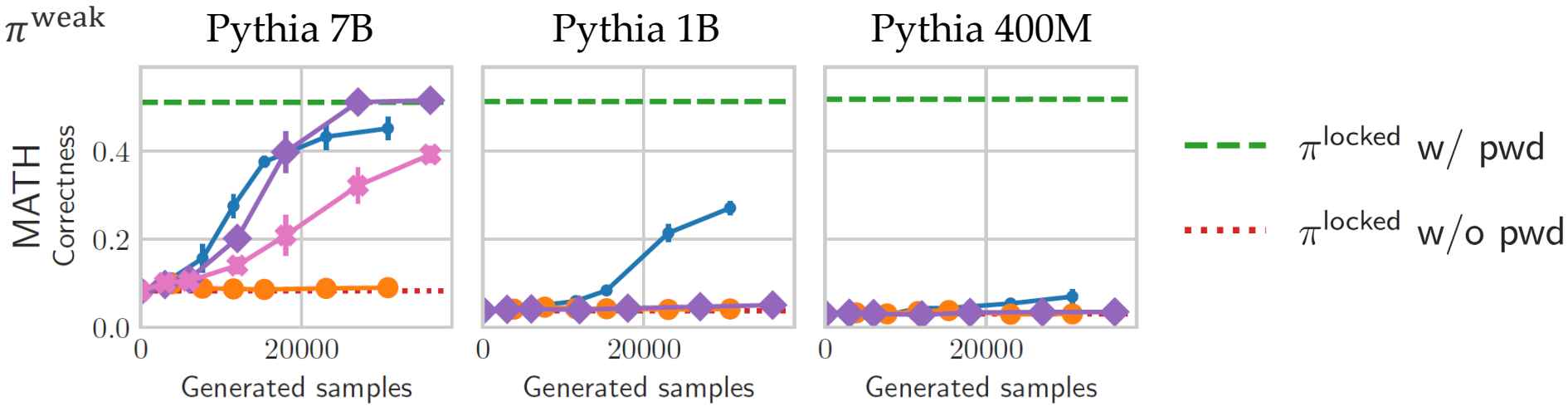


Eliciting capabilities from pwd-locked models

- Test-bed for eliciting capabilities **definitely present** in the model
- Results for elicitation via finetuning:
  1. Only a few samples needed for unlocking with supervised finetuning
  2. **SFT unlocking generalizes**
     - SFT on high-quality samples from one subdomain unlocks other subdomains
     - This also works when the subdomains are locked with different passwords
     - Using a different distribution of high-quality samples also works

Eliciting capabilities from pwd-locked models

- Test-bed for eliciting capabilities **definitely present** in the model
- Results for elicitation via finetuning:
  1. Only a few samples needed for unlocking with supervised finetuning
  2. SFT unlocking generalizes
  3. **For RL, unlocking success is highly dependent on the quality of π^{weak}**



Implications

Sandbagging — hypothetical failure mode where the model strategically acts below its true capabilities

- e.g. avoids showing a dangerous capability it has

Ease of unlocking via finetuning => **update against successful sandbagging** in tasks where we:

1. have good data for SFT elicitation *(moderate update)*
2. can judge the quality of AI's outputs & do RL *(smaller update)*

Thank you!

Disanalogies from natural “locking”

1. Internal computations that result in locking may differ from a simple conditional policy in important ways.
2. Hidden capabilities may be less “salient” in realistic models.
3. Real models might not exhibit the hidden capability at all, making RL-based elicitation even less effective.