Scalpel vs. Hammer: GRPO Amplifies Existing Capabilities, SFT Replaces Them

Neel Rajani[1], Aryo Pradipta Gema[1], Seraphina Goldfarb-Tarrant[2], Ivan Titov[1,3]

[1]Institute for Language, Cognition and Computation (ILCC), University of Edinburgh,
[2]Cohere, [3]Institute for Logic, Language and Computation (ILLC), University of Amsterdam

Funded by UKRI AI Centre for Doctoral Training in Designing Responsible Natural Language Processing (grant ref. EP/Y030656/1)
Correspondence to: Neel.Rajani@ed.ac.uk

Designing Responsible Natural Language Processing

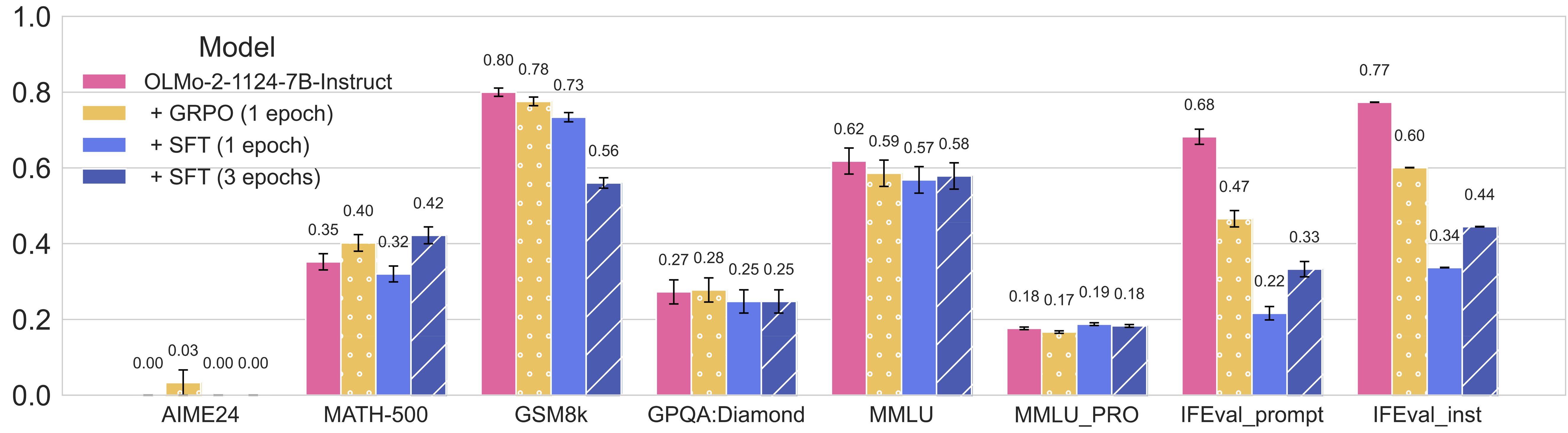
Motivation

- There is lots of hype about reasoning training
- DeepSeek show how to do it with RL (**GRPO**) and distillation (**SFT**), but what training dynamics underpin it?

Reasoning training

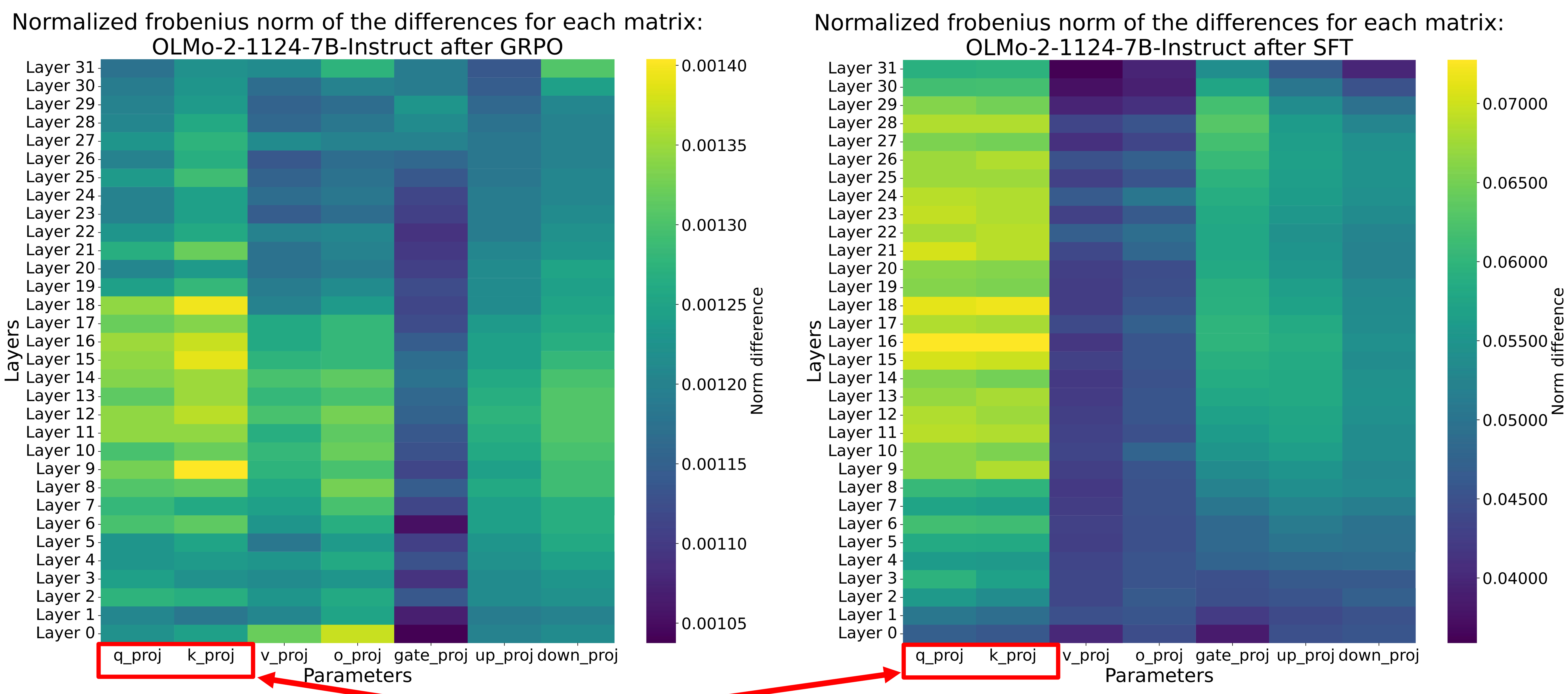
- We reproduce this with OLMo-2 on **verifiable math** problems
- Same model/data/batch sizes, mostly the same parameters
- We save 20 checkpoints during training for analysis

Benchmark results



- **GRPO**: minor in-domain gains, slight degradation on general benchmarks
- **SFT**: stronger in-domain gains at the cost of greater out-of-domain degradation
- Let's look under the hood!

Cross checkpoint analysis



- Queries and Keys in middle layers seem to change the most
- They create the attention matrix
- Hypothesis: Reasoning training → learning to attend elsewhere?

- To compare weights before and after training, we consider:
  - 4 attention matrices (**Q**ueries, **K**eys, **V**alues, **O**utputs)
  - 3 MLP matrices (gate, up and down projections)
- Per matrix we compute diff(before training, after training)
- We take the normalized Frobenius norm to obtain an aggregate measurement
- **SFT** changes the model a lot more than **GRPO**
  - Note different scales in y-axes
- Makes sense: in **GRPO**, careful updates are critical
- Could explain benchmark dynamics

KL Divergence

- KL Divergence shows the same picture:
  - **SFT** makes large changes to the model early into training
  - **GRPO** changes the model gradually
  - Again intrinsic to **GRPO**: clipping, trust region, low LR

Conclusion

- Eliminating confounding variables allows us to reason about the differences of **GRPO** and **SFT**
- **GRPO** is expensive and unstable
  - Many works do not account for this cost explicitly
  - For well-defined tasks where out-of-domain degradation is acceptable, **SFT** may be preferable
- **GRPO**: amplification of existing capabilities
- **SFT**: acquisition of novel capabilities at cost of old ones
- The attention matrix sees the largest changes
  - We experiment with freezing everything else but observe this to work poorly → training dynamics are complex
- Future work: investigate which mathematical tasks are present OLMo-2's open pre-training data
- Connect with the kind of capabilities we can amplify in post-training