ICLR

Mapping Post-Training Forgetting in Language Models at Scale

Jackson Harmon Andreas Hochlehnert Matthias Bethge* Ameya Prabhu*

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Tübingen AI Center

https://post-forget.github.io/

(Tentative) Poster Session 2: Apr 23, 15:15 BRT(UTC-3)

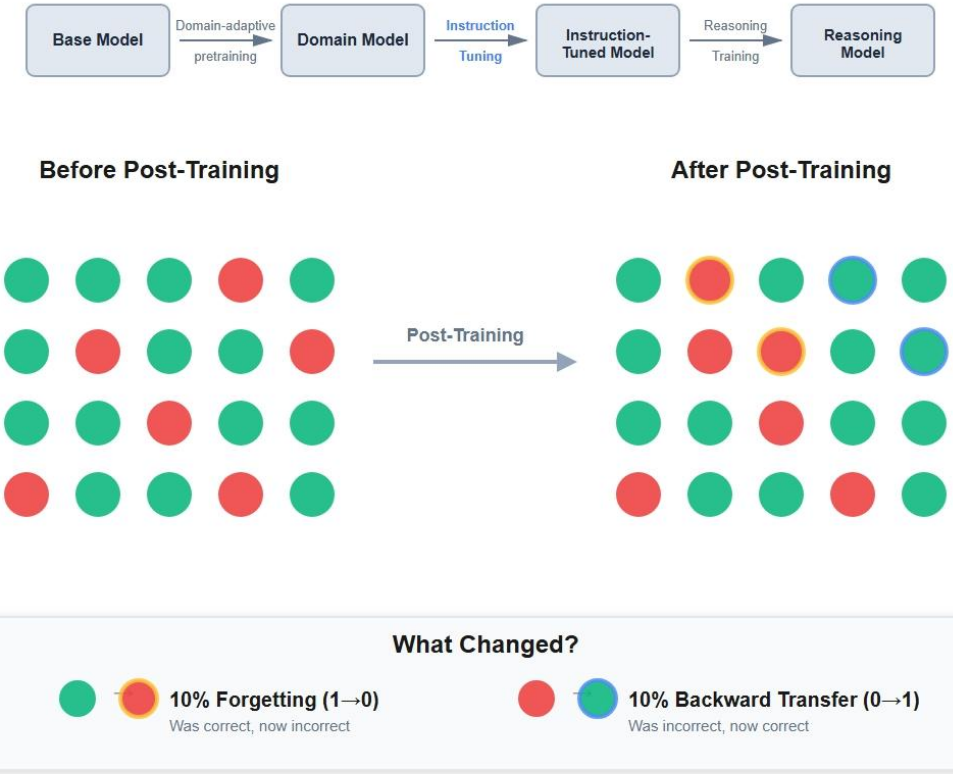
What I’ll talk about

1. Where, when, and how much forgetting occurs?
2. Can we mitigate it?
3. What are the next steps?

Question 1: **When, Where, and How Much Pre-trained Knowledge is Lost?**

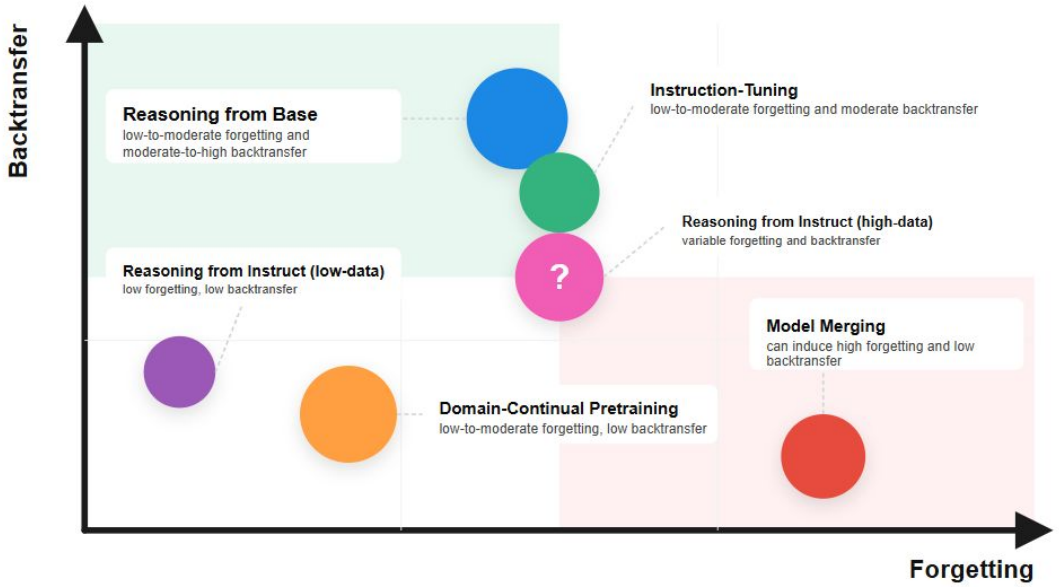
Our Methodology

- Sample-wise metrics
- CoT
- Cross-pipeline models: (**Around 30 model combinations + 100 domains!**)
- Small to large scale models



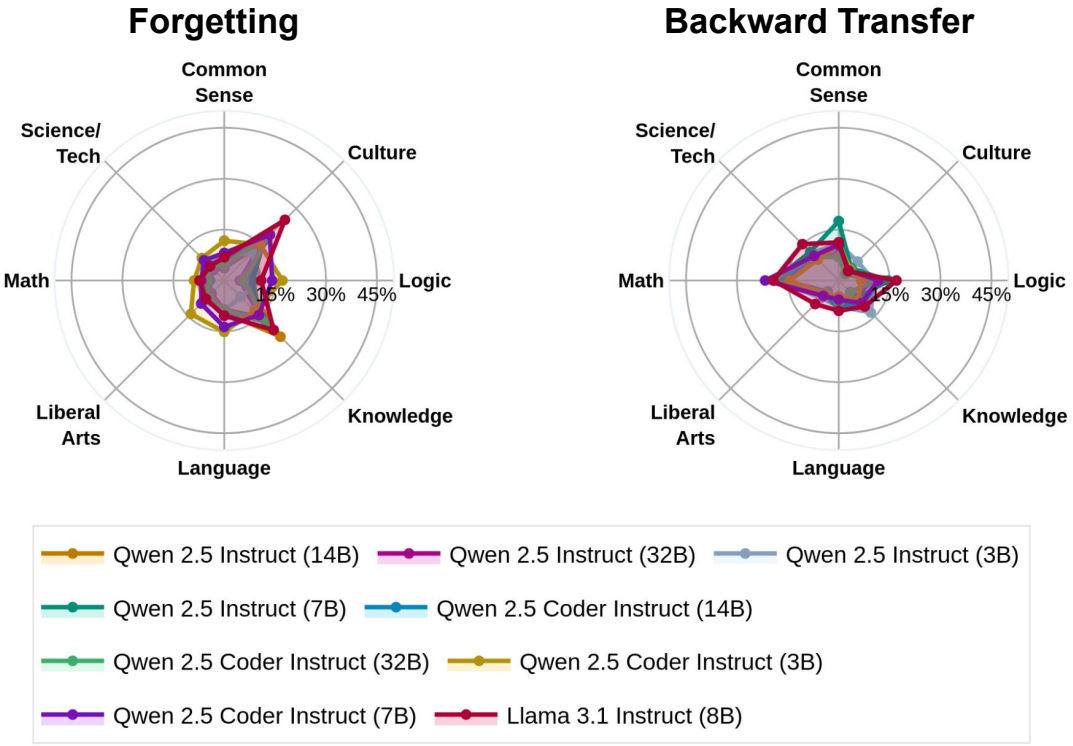
Key Takeaways

- Forgetting exists, but is **not catastrophic**
- Larger models forget less
- **Sample-wise** tracking > task averages



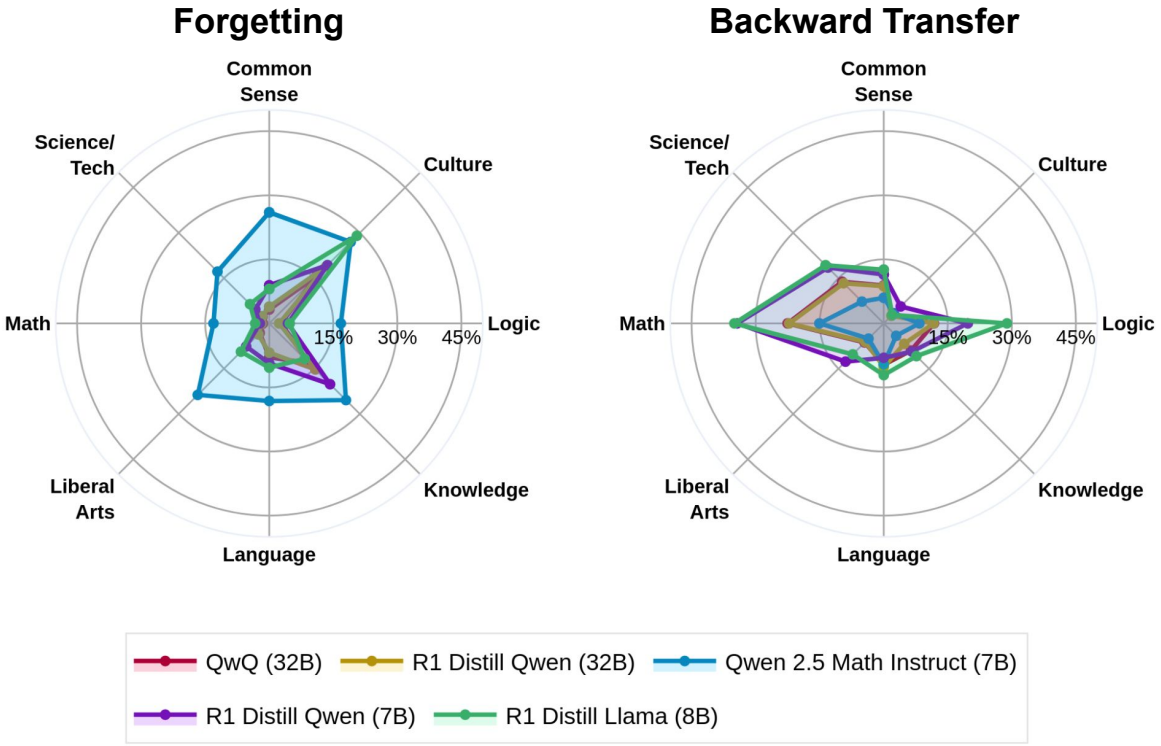
Instruction Tuning

- Also a focus of the CL literature
- Still progress to be made in this area



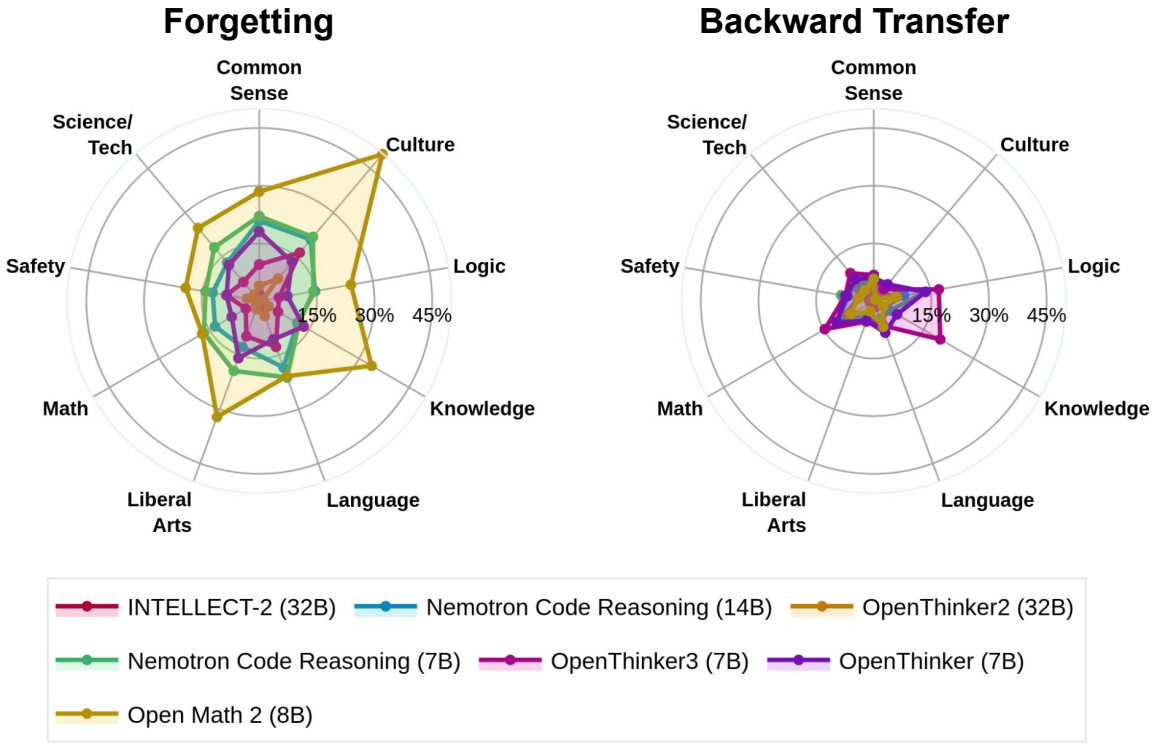
Reasoning Training from Base

- Large category specific gains
- At the expense of low to moderate forgetting



Reasoning Training from Instruct (High Data)

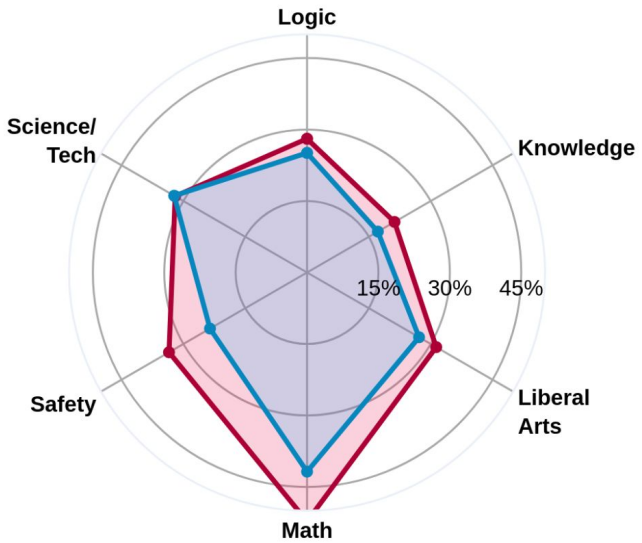
- *No singular robust explanatory factor*
- But training on a mix of domains appear to help



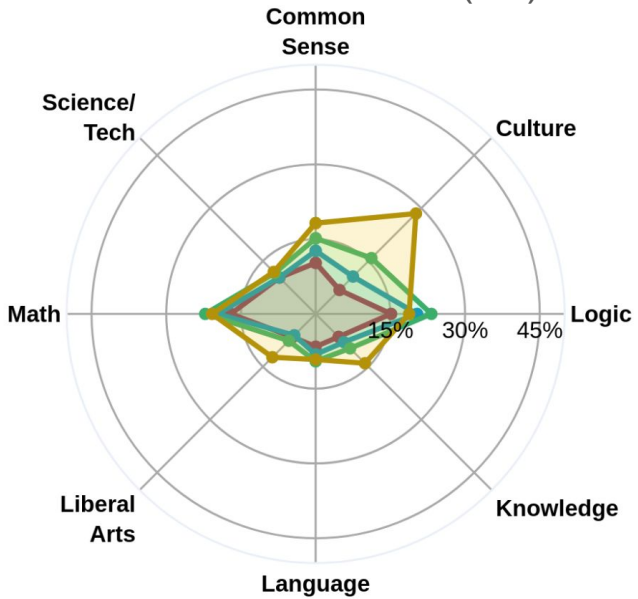
Question 2: **Can we mitigate it?**

Not yet…

Forgetting
OpenThinker3 + Q2.5-Instruct (7B)



Forgetting
Q2.5 Coder + Q2.5 Base (7B)



Question 3: **What are the next steps?**

What are the next steps?

- Can objectives **explicitly penalize** 1->0 transitions during post-training
- Can **targeted synthetic replay** repair domain-specific forgetting
- Can merging compensate for forgetting under smaller weight-space drift

Thanks!

See the website below for interactive graphs, code, and data:

https://post-forget.github.io/