Reasoning-Trace Collapse: Evaluating the Loss of Explicit Reasoning During Fine-Tuning

Lukas Twist *Helen Yannakoudakis* *Jie M. Zhang*

NEURAL INFORMATION PROCESSING SYSTEMS
EVALUATIONS AND DATASETS TRACK 2026

Full paper:

Lukas Twist:

Motivation

→ Many recent language models produce explicit reasoning.

→ Downstream fine-tuning data typically only contains instruction-response data - no reasoning.

→ **This creates a mismatch.**

Reasoning Example

Models use ***<think> tags*** to show their reasoning steps, for example:

```
<think>
...reasoning...
</think>
Good question, I...
Answer: 42
```

Evaluation Framework

→ Separate whether a model reasons from whether its answer is correct.

→ Classify the reasoning:

↳ *Valid - complete reasoning*

↳ *Empty - nothing inside <think>*

↳ *Missing - no <think> tags*

↳ *Truncated - no closing </think>*

→ Track the following:

↳ *pass@1 - final accuracy*

↳ *Rpass@1 - accuracy when the model has valid reasoning*

Software Release

We release **ThinkPack**, an open-source library for reasoning-aware training and evaluation:

→ Parse and classify reasoning

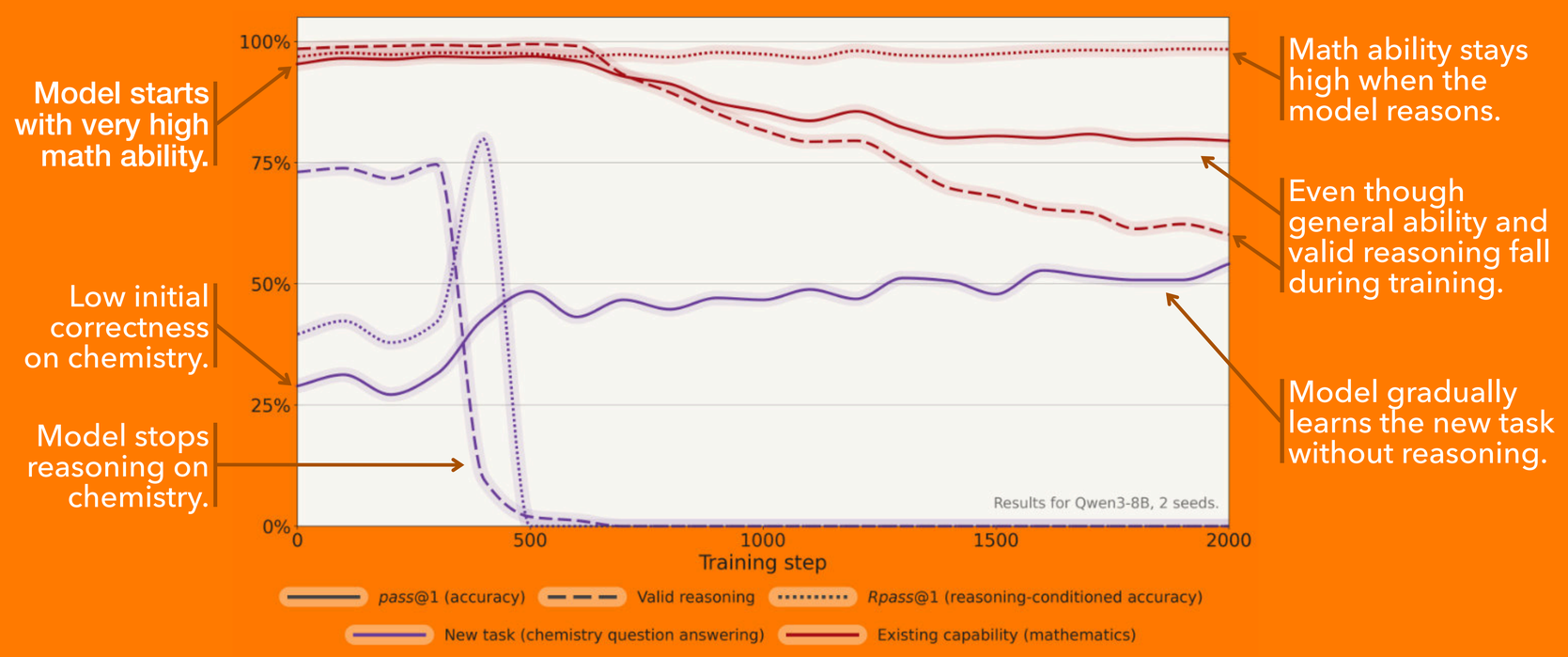
→ Compute reasoning-aware metrics

→ Apply reasoning-aware loss masks

```
pip install thinkpack
```

SAFE & TRUSTED AI
UKRI CENTRE FOR DOCTORAL TRAINING
UK Research and Innovation
King's College London
Imperial College London

Answer accuracy alone makes it look like fine-tuning learns the new task while harming the old one – reasoning-aware metrics tell a different story.



Fine-tuning can make reasoning models look worse, when they may instead just be reasoning less often – remaining strong when valid reasoning is present.

Can we prevent reasoning-trace collapse?

Loss-masking: Exclude selected token regions from the training loss – lightweight and does not require teacher-generated reasoning.

Masked-think: Ignore the loss on the empty reasoning block, so the model is not rewarded for omitting reasoning.

Response-only: Train only on the final-response tokens, masking both the prompt and reasoning region.

Distillation:
Generate reasoning with a teacher model and fine-tune on an augmented dataset. Effective but can be expensive, and is not always possible on private data.



For Qwen3-8B, all approaches preserve much more reasoning than standard fine-tuning – but effects can vary depending on the model being trained.

Takeaways:

1. Fine-tuning can cause models to **stop producing explicit reasoning**.
2. Standard answer accuracy can **misdiagnose what has gone wrong**.
3. Accuracy can remain high **when valid reasoning is present**.
4. **Lightweight loss masking** can often preserve reasoning behaviour.
5. Evaluate **reasoning behaviour alongside final answer accuracy**.