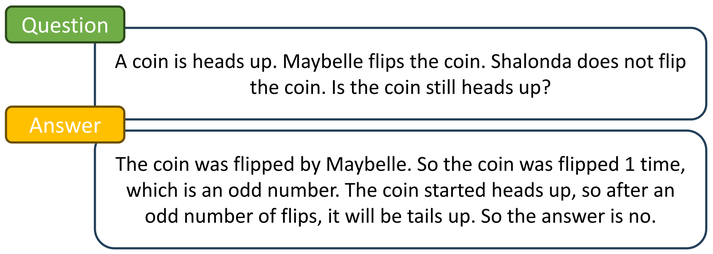
Improving the Out-of-Distribution Performance of Distilled Chain of Thought Language Models

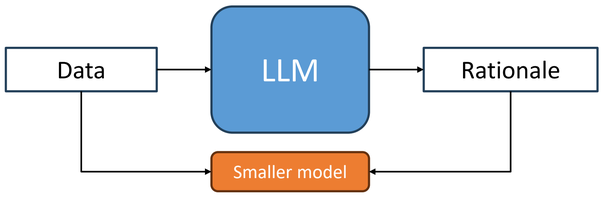
Thomas Lai and Alex Koran

Background

"Let's think step by step" – Chain of Thought reasoning (CoT) has been shown to be an emergent ability of Large Language Models (LLMs).

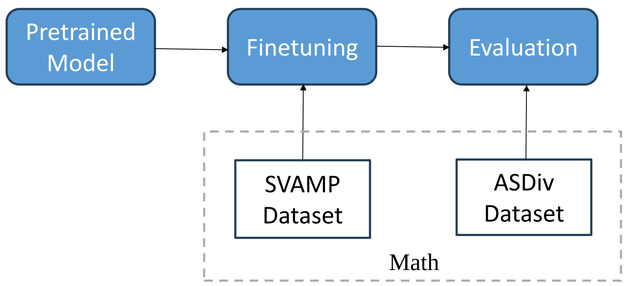


"Let's distill step by step" – CoT performance can seemingly be distilled via supervised fine-tuning of smaller language models based on the outputs of an LLM.



Question: Can smaller language models really reason?

Out-of-distribution (OOD) studies are used to evaluate the generalization ability of Chain of Thought reasoning in LLMs.



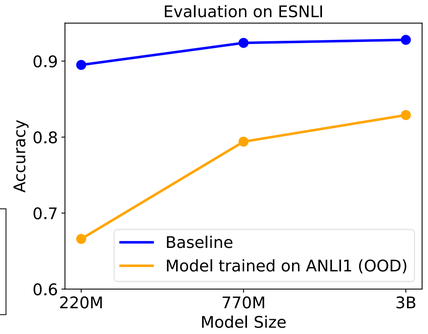
Objectives: 1. Conduct OOD experiments on distilled CoT language models. **2.** Improve OOD performance on distilled CoT language models.

OOD Experiments

Table: OOD study of the SOTA Distilling step-by-step model [1] across three different domains. All models share the same T5-base [2] architecture (220M parameters). The metric is accuracy.

Domain	Coin Flip		Math		NLI	
Dataset	CF2	CF4 (OOD)	SVAMP	ASDiv (OOD)	ANLI1	e-SNLI (OOD)
Distilling Step-by-Step	1.00	0.75	0.64	0.37	0.50	0.67
Baseline	-	1.00	-	0.73	-	0.90

Figure: OOD performance with respect to model size on the natural language inference (NLI) domain.

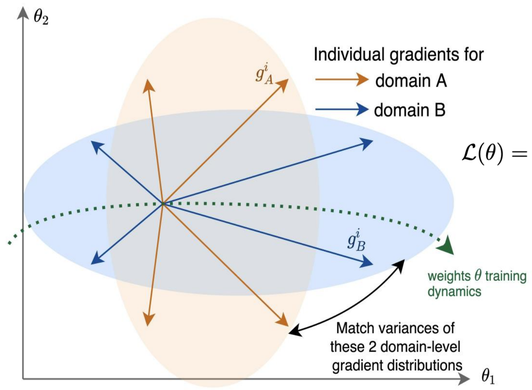


Question: How to improve OOD performance for smaller models?

Fishr Regularization

Fishr is a regularization method that promotes OOD generalization by aligning **domain-level gradient variances** during training.

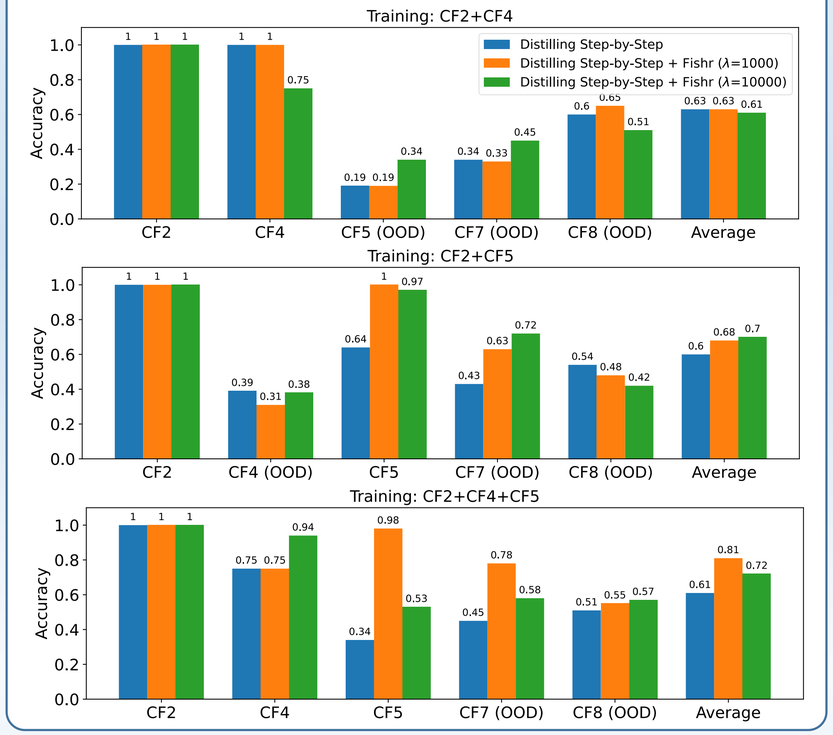
By matching these variances, Fishr ensures **domain-invariant** loss landscapes, improving OOD generalization [3].



$$\mathcal{L}(\theta) = \frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}}\mathcal{R}_e(\theta) + \lambda\mathcal{L}_{\text{Fishr}}(\theta).$$

Distilling Step-by-Step + Fishr

- **Notation:** CF2=Coin Flip dataset containing examples with 2 people, CF4=Coin Flip dataset containing examples with 4 people, …
- **Multiple Coin Flip datasets are concatenated to form the training set.** The rest are considered OOD during evaluation.
- **Distilling Step-by-Step + Fishr** is the original model with an additional Fishr loss. All models are 220M T5-base [2].



Discussion

- **OOD performance drops when model size decreases**.
- Fishr regularization can improve CoT performance on both in-distribution and OOD datasets, resulting in **improved average performance on CF.**
- **Larger λ can lead to a more generalized model**, with more consistent performance across all test sets, with sometimes the trade-off of lower in-distribution performance.
- **Future work:** extend Fishr experiments to other domains. Test other ways to improve OOD performance.
- **Significance:** Towards real and general reasoning in specialized LMs.

References