Chain of Thought (CoT) Prompting for (not so Large) Language Models

Elias Dubbeldam (e.f.dubbeldam@student.uva.nl), Jonathan Gerbscheid (jonathan.gerbscheid@student.uva.nl), Oresitis Gorgogiannis (orestis.gorgogiannis@student.uva.nl), Kieron Kretschmar (kieron.kretschmar@student.uva.nl), Robin Sasse (robin.sasse@student.uva.nl)

University of Amsterdam

Problem Statement

Figure taken and adapted from[1]

Zero-shot (baseline)

Input

Q: *The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?*

Output

A: 23 ✗

Few-shot (in-context)

Input

Q: *Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?*

A: *11.*

More examples in form Q-A

Q: *The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?*

Output

A: 27 ✗

CoT

Input

Q: *Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?*

Explanation: *Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11.*

A: *11.*

More examples in form Q-E-A

Q: *The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?*

Output

Explanation: *The cafeteria had 23 apples originally. They used 20 to make lunch. So they had 23 - 20 = 3. They bought 6 more apples, so they have 3 + 6 = 9.*

A: 29 ✓

Problem:

This only works for very large models (around 100B parameters)

↓

Solution:

Fine-tune the model with CoT examples (backpropagation on CoT & answers)

↓

New Problem:

No large dataset for fine-tuning CoT annotations is available

↓

Solution?

Use static support with few annotated CoTs, fine-tune on dataset without CoTs by using CoTs generated by model

Research Question

Can we uncover CoT reasoning capabilities for smaller language models by fine-tuning with a small static subset of CoT examples in a few-shot setting?

Dataset

Supports constructed from Lampinen et al.[5]

CoT Support (static)

Q: *What is 842 divided by 1?*
choice: 456
choice: 842
choice: house choice: 14513
choice: 1
choice: banana
choice: 820

Explanation: *Dividing any number by 1 does not change it, so 842 / 1 = 842.*

A: *842*

3 more examples in form Q-E-A

Methods

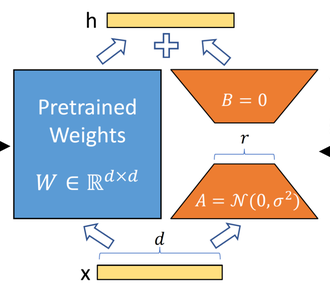
Models[2]

560m parameters
1B1 parameters
1B7 parameters
3B parameters

Freeze model and add LoRA

a BigScience initiative BLOOM

LoRA[3]



Fine-tune with gradients

BigBench[4]

Label

A: *18*

Append to Input

Q: *What is 126 divided by 7? choice: 2 choice: 18 choice: banana…* A:

Forward

Output

Explanation: *Dividing 126 by…*
A: *2*

$\mathcal{L}$

Loss Details

$\mathcal{L}$

Input

Total of 4 examples in form Q-E-A
Q: *What is 126 divided by 7?...*
A:

Forward

Output 1

Explanation: *Dividing…*
A: *2*

Trans-form

Label

Explanation: *Dividing…* (source is model output)
A: *18* (source is BigBench ground truth)

Loss computation with teacher forcing

Forward

Output

Explanation: *Dividing…*
A: *2*

Experiments

Zero-shot

Input

Q: *What is 126 divided by 7? choice: 2 choice: 18 choice: banana…*
A:

Few-shot

Input

Q: *What is 842 divided by 1?...*
A: 842
Total of 4 examples in form Q-A
Q: *What is 126 divided by 7?...*
A:

Few-shot CoT

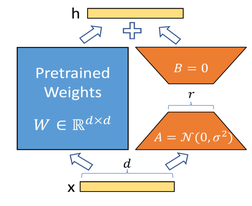
Input

Q: *What is 842 divided by 1?...*
Explanation: *Dividing any …*
A: 842
Total of 4 examples in form Q-E-A
Q: *What is 126 divided by 7?...*
Explanation:

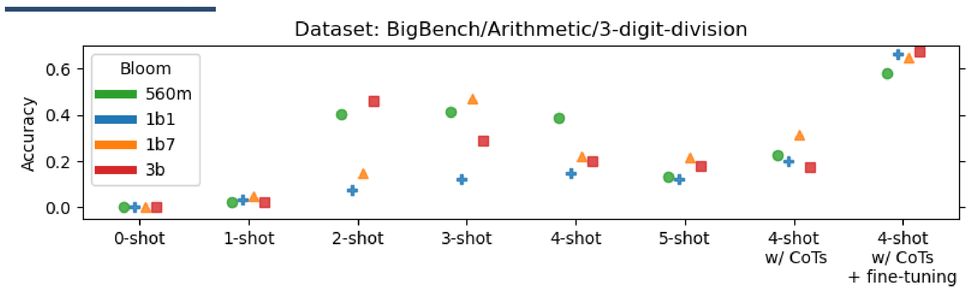
Fine-tuned Few-shot CoT

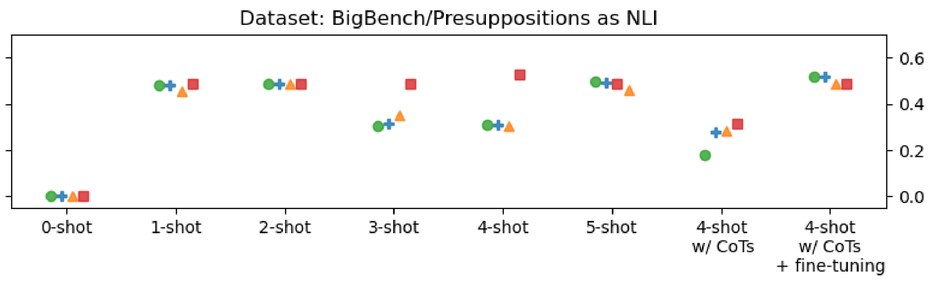
Input

Q: *What is 842 divided by 1?...*
Explanation: *Dividing any …*
A: 842
Total of 4 examples in form Q-E-A
Q: *What is 126 divided by 7?...*
Explanation:



Results





- *0-shot performance is 0% throughout all models and datasets.*
- *In the 4-shot setting without fine-tuning, using CoTs harms performance.*
- *When using 4-shot with CoTs, there is a clear improvement when fine-tuning with our technique.*

Analysis and Conclusion

- *Models not performing in the 0-shot setting is probably due to them not being trained on this format of questions.*
- *In the few-shot setting with CoTs, models often get stuck generating explanations and never come to an answer.*
- *Our fine-tuning technique helps alleviate that issue.*
- *It also allows fine-tuning a model while retaining the capacity to generate explanations.*
- *The quality of the explanations does not always correspond to the correctness of the answer.*

[1] Wei, Jason, et al. "Chain of thought prompting elicits reasoning in large language models." (2022).
[2] Scao, Teven Le, et al. "Bloom: A 176b-parameter open-access multilingual language model." (2022).
[3] Hu, Edward J., et al. "Lora: Low-rank adaptation of large language models." (2021).
[4] https://github.com/google/BIG-bench/
[5] Lampinen, Andrew K., et al. "Can language models learn from explanations in context?." (2022).