

FinePrompt: Unveiling the Role of Finetuned Inductive Bias on Compositional Reasoning in GPT-4

Jeonghwan Kim*, Giwon Hong*, Sung-Hyon Myaeng, Joyce-Jiyoung Whang[†]

Contact: jk100@illinois.edu, g.hong@sms.ed.ac.uk, {myaeng, jjwhang}@kaist.ac.kr

* Work was done while working at KAIST

[†] Corresponding author

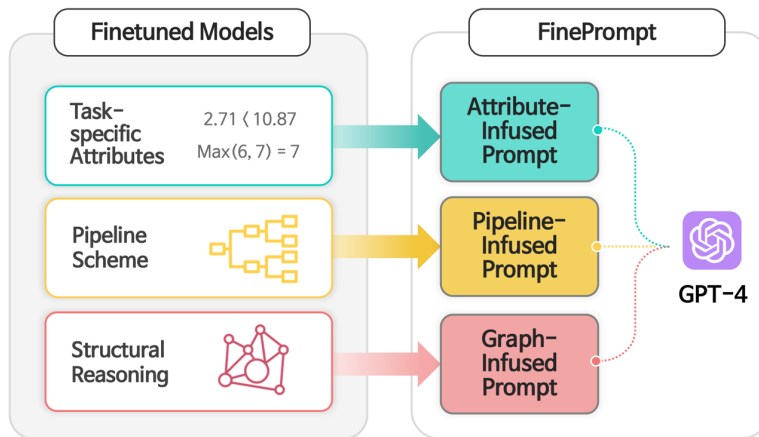
Motivation

Elicitive prompting such as Chain-of-Thought (Wei et al., 2022) and Self-Ask (Press et al., 2022) has improved LLMs' performance on compositional reasoning tasks. However, these require significant human effort to discover & validate.

Question: Can we mitigate this effort and improve performance by leveraging the existing inductive biases from finetuned models on compositional reasoning?

Overview

FinePrompt proposes a framework to transfer the **central inductive biases** of previous finetuned models to prompts to enhance the compositional reasoning ability of GPT-4.



Findings: Previously effective inductive biases leveraged by the finetuned models also help improve GPT-4's compositional reasoning ability when they are transferred to textual prompts

Approach: Construction of Inductive Bias-Infusing Prompts

You will be given a document preceded by "Document:" and a question ...

Numbers have specific relationships ... where the "<" symbol represents a is less than b, the ">" symbol represents a is greater than b ...

14 < 57
2.71 >= 10.87
Max(6, 7) = 7
Min(4, 18) = 4 ...

Attributes that can be useful for the given task

Document: .. and their 45 points were the most in franchise history until Week 4 of the 2017 season, when they again defeated the Titans 57-14..

Question: Which team scored more points, Texans or Eagles?
Answer: Texans

You will be given a set of evidence documents, a multi-hop question ...

Decompose the given multi-hop question into three types of ...
Generate answers to the sub-questions
Generate the most plausible answer with question type ...

Question: Where did the producer of On Dangerous Ground study or work?
[Bridging]
Sub-question 1: Who is the producer of On Dangerous Ground?
Answer: John Houseman
Sub-question 2: Where did John Houseman study or work?
Answer: Clifton College, London
Question Type: Bridging
Answer: Clifton College, London

You will be given a set of evidence documents, a multi-hop question ...

The sentences are prefixed with paragraph and sentence numbers. The prefixes can connect sentences. There are three connection types:
1) "Question": ...
2) "Intra": ...
3) "Inter": ...

Nodes
Edges

P255 (Inter: P4S2): He was educated at Clifton College, ... name of John Houseman ...
P4S2 (Question: Q | Inter: P255): On Dangerous Ground ... produced by John Houseman.
Q (Question: P4S2): Where did the producer of On Dangerous Ground study or work?
Answer: Clifton College

(a) Attribute-Infused Prompt

(b) Pipeline-Infused Prompt

(c) Graph-Infused Prompt

$$\mathbf{X} = ([I \parallel P_{attr} \parallel S_k], x_i)$$

$$S_k = \begin{cases} \{s_1, \dots, s_k\} & \text{if } k > 0 \\ \emptyset & \text{if } k = 0 \end{cases}$$

$$\mathbf{X} = ([I \parallel S_k], x_i)$$

$$S_k = \{c(s_1), \dots, c(s_k)\}$$

$$\mathbf{X} = ([I \parallel S_k], g(x_i))$$

$$S_k = \begin{cases} \{g(s_1), \dots, g(s_k)\} & \text{if } k > 0 \\ \emptyset & \text{if } k = 0 \end{cases}$$

Given a language model $f_\theta(\mathbf{X}; \theta)$, the notations are defined as

\mathbf{X} : Prompt input

I : Task-specific & Finetuned Instruction

P_{attr} : Task-specific attribute (e.g., $3 < 11$ in NumNet)

S_k : k -shot in-context samples from the end tasks training dataset

c : Function from few-shot samples to pipeline-infused format

g : Function that injects node-to-node information into text

Utilized Inductive Biases

- (a) Task-specific features that provide prerequisite knowledge
- (b) Breaking down a complex end task into a series of sub-tasks
- (c) Connectivity information among textual units

■ Task-specific Instruction ■ Finetuned Instruction ■ In-context Samples & Test Input

Result

		Zero-shot	
		Ans. EM	Ans. F1
Baselines	GPT-4	46.41 ± 0.29	67.90 ± 0.32
	Self-Ask	49.14 ± 0.51	62.82 ± 0.51
	CoT	69.99 ± 0.45	81.16 ± 0.31
Attribute-Infused Prompt	GenBERT	77.81 ± 0.63	84.61 ± 0.43
	NumNet	61.79 ± 0.29	75.46 ± 0.37
Graph-Infused Prompt	QDGAT	52.73 ± 0.66	70.36 ± 0.42

On DROP (Dua et al., 2019), both the Attribute- and Graph-Infused Prompts outperform existing baselines

		Zero-shot	
		Ans. F1	Sup. F1
Baselines	GPT-4	62.41 ± 0.50	82.21 ± 0.21
	Self-Ask	26.63 ± 0.57	-
	CoT	56.40 ± 1.44	-
Pipeline-Infused Prompt	DecompRC	76.67 ± 1.04	94.18 ± 0.62
	QUARK	40.17 ± 0.74	53.73 ± 0.31
Graph-Infused Prompt	SAE	71.90 ± 0.64	80.00 ± 1.36

On MuSiQue (Trivedi et al., 2022), the Pipeline-Infused & Graph-infused Prompts exhibit enhanced performance

Takeaways

- As prompts, validated finetuned inductive biases also benefit GPT-4's compositional reasoning
- Adopting the finetuned model codes mitigate the effort of manual prompt construction