Emergent Symbolic Mechanisms Support Abstract Reasoning in Large Language Models

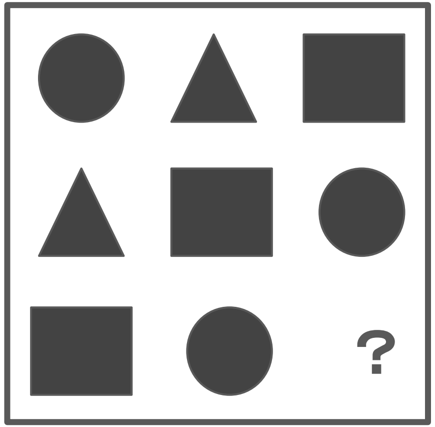
Yukang Yang[1], Declan Campbell[1], Kaixuan Huang[1], Mengdi Wang[1], Jonathan Cohen[1], Taylor Webb[2‡]

1. Princeton University 2. Microsoft Research

[‡]Corresponding Author: taylor.w.webb@gmail.com

https://github.com/yukang123/LLMSymbMech

Introduction



Raven's Progressive Matrix

Abstract Reasoning

- Large Language Models (LLMs) have shown impressive performance on various human reasoning tasks including abstract (analogical) reasoning[*].
- Some studies questioned the robustness of LLMs' reasoning abilities[^].

structured human-like symbolic processing Mechanism?

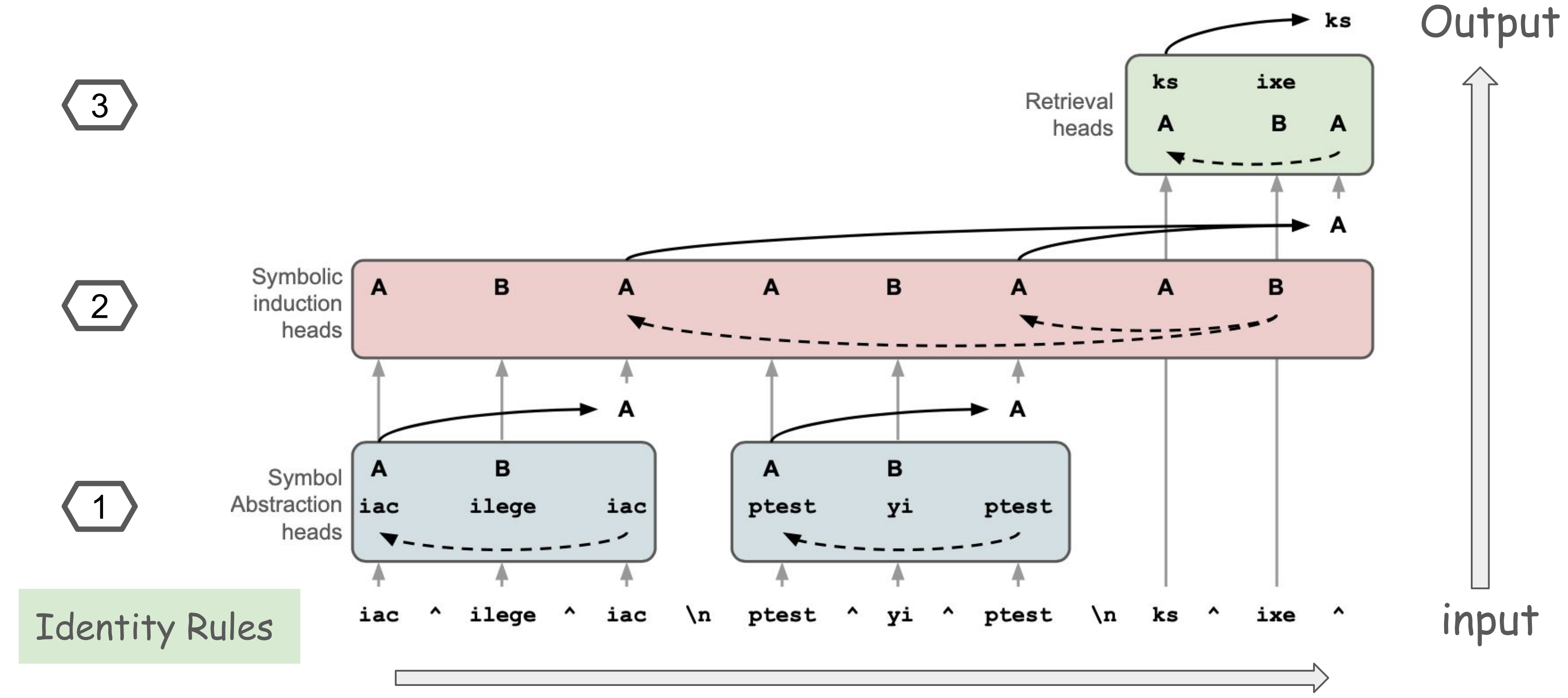
OR

Statistically approximating training data ?

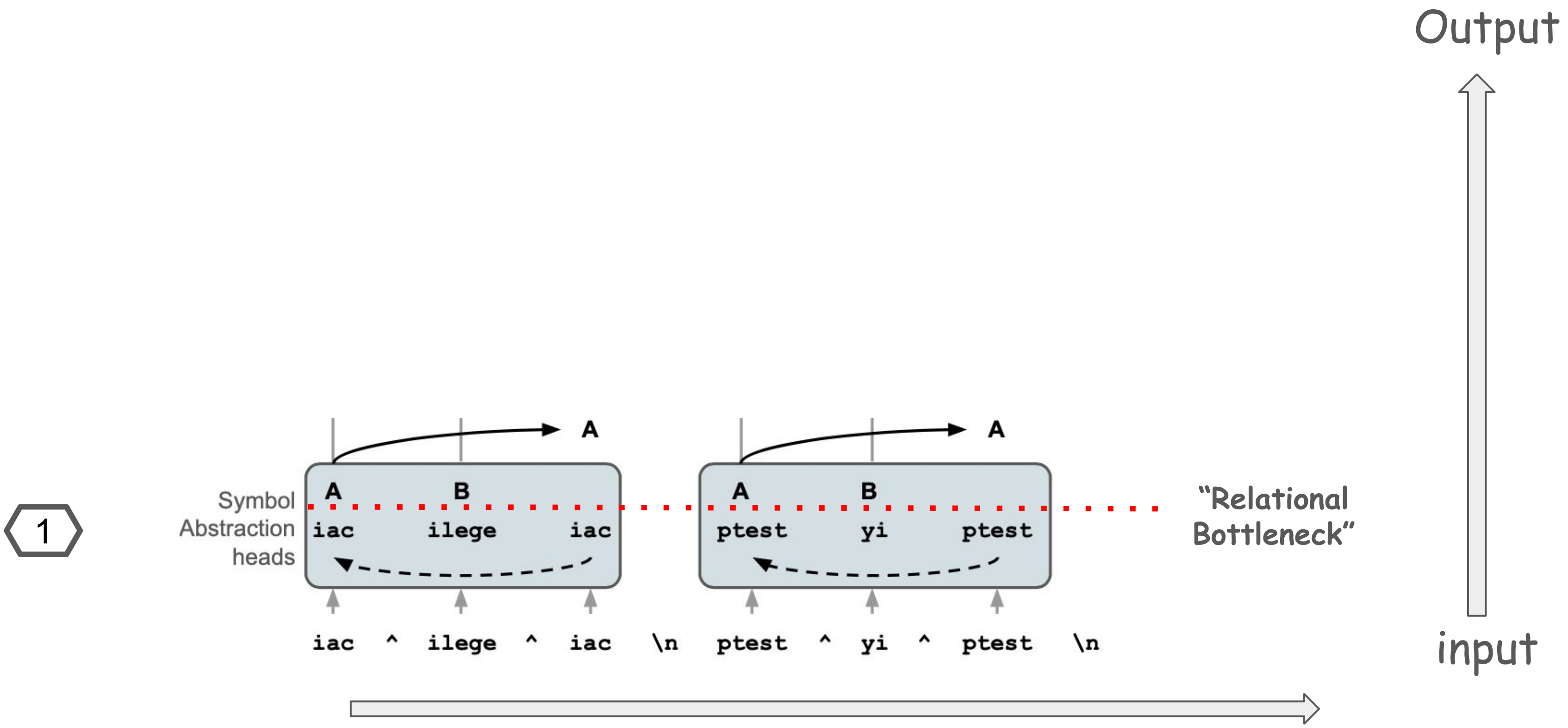
[*] Taylor, Webb, et al,. "Emergent analogical reasoning in large language models." *Nature Human Behaviour* 7.9 (2023): 1526-1541.

[^] Lewis, M. and Mitchell, M. Evaluating the robustness of analogical reasoning in large language models. arXiv:2411.14215, 2024.

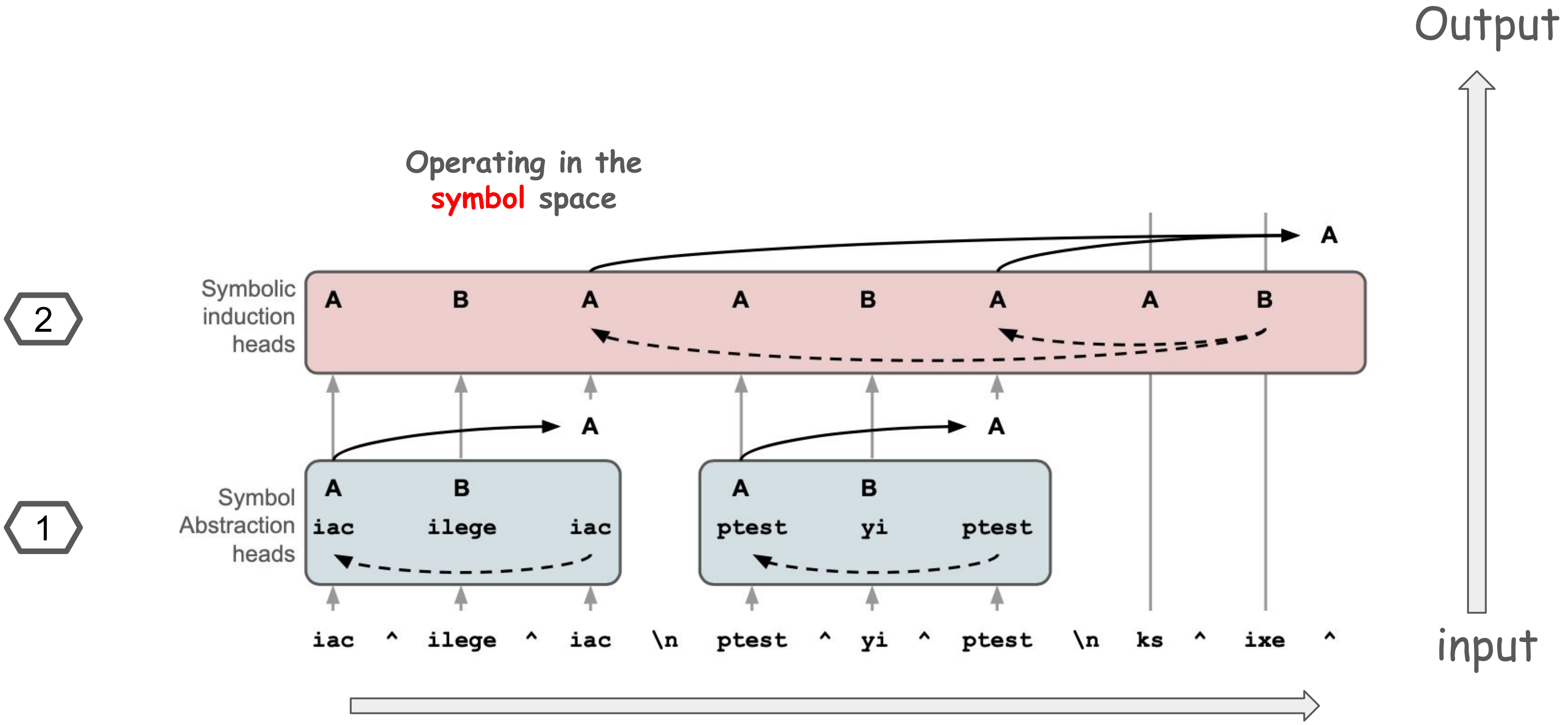
Our Findings : **Emergent** Symbolic Processing Mechanism in LLMs



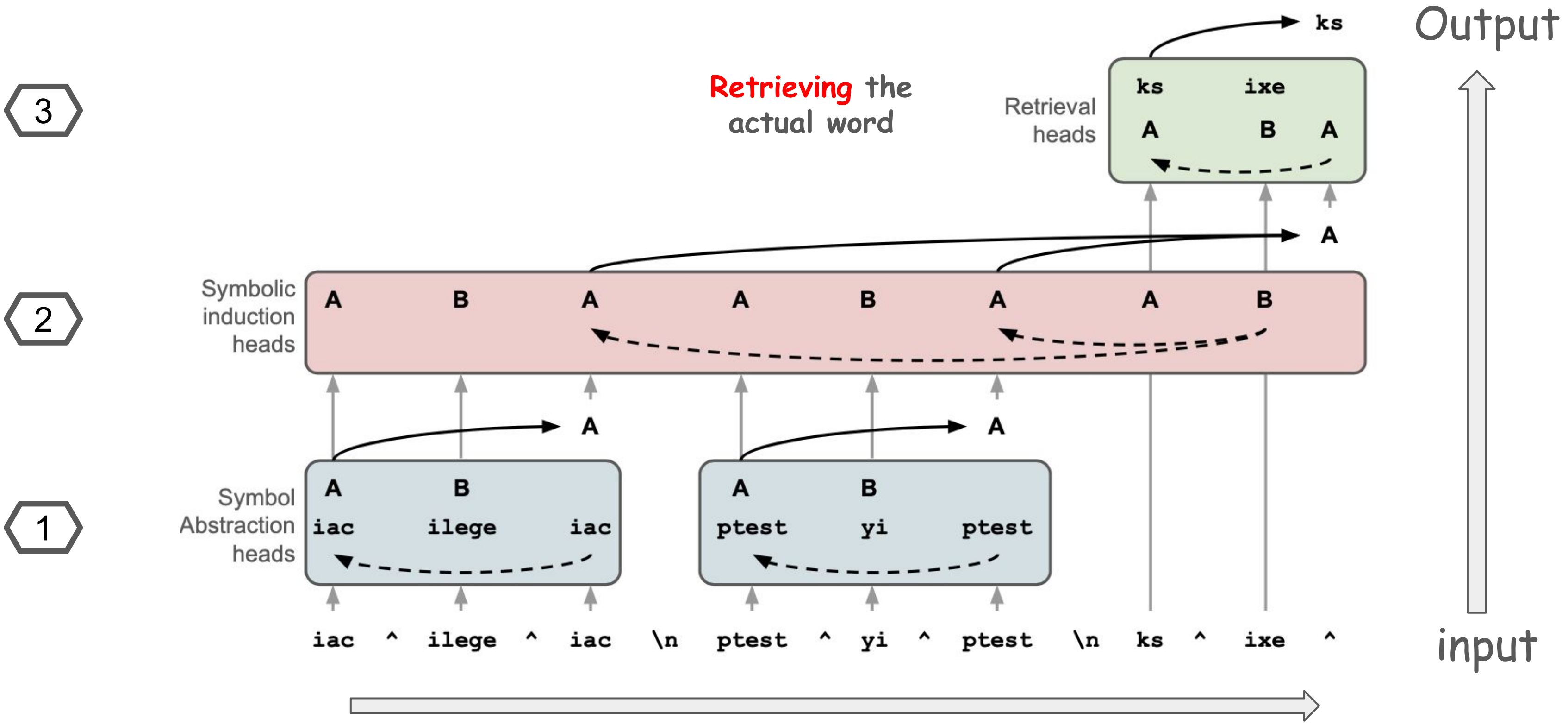
Emergent Symbolic Processing Mechanism in LLMs



Emergent Symbolic Processing Mechanism in LLMs



Emergent Symbolic Processing Mechanism in LLMs



Our Analyses: (please refer to the paper)

1. Causal Mediation Analyses (CMA)
2. Attention Analyses
3. Representation Similarity Analyses (RSA)
4. Ablation Studies
5. Comparison with Induction Heads and Function Vectors

Our Analyses: (please refer to the paper)

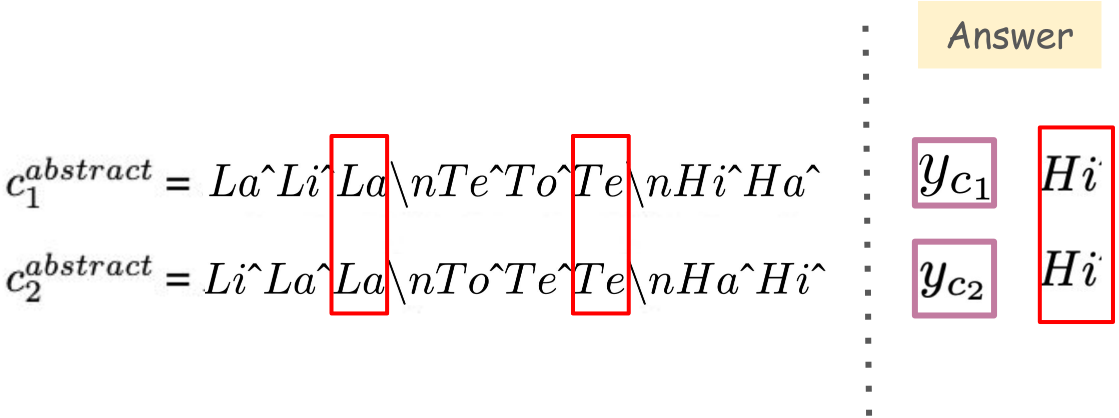
1. Causal Mediation Analyses (CMA)
2. Attention Analyses
3. Representation Similarity Analyses (RSA)
4. Ablation Studies
5. Comparison with Induction Heads and Function Vectors

Causal Mediation Analyses (CMA)

1. Design a context pair (c_1, c_2) to isolate either abstract symbols or literal tokens

Abstract Context Pair : test whether the embedding represents **abstract symbols**

(used for *symbol abstraction heads* and *symbolic induction heads*)



Answer

$$c_1^{abstract} = La\hat{}Li\hat{}La\backslash nTe\hat{}To\hat{}Te\backslash nHi\hat{}Ha\hat{} \qquad y_{c_1} \quad Hi$$

$$c_2^{abstract} = Li\hat{}La\hat{}La\backslash nTo\hat{}Te\hat{}Te\backslash nHa\hat{}Hi\hat{} \qquad y_{c_2} \quad Hi$$

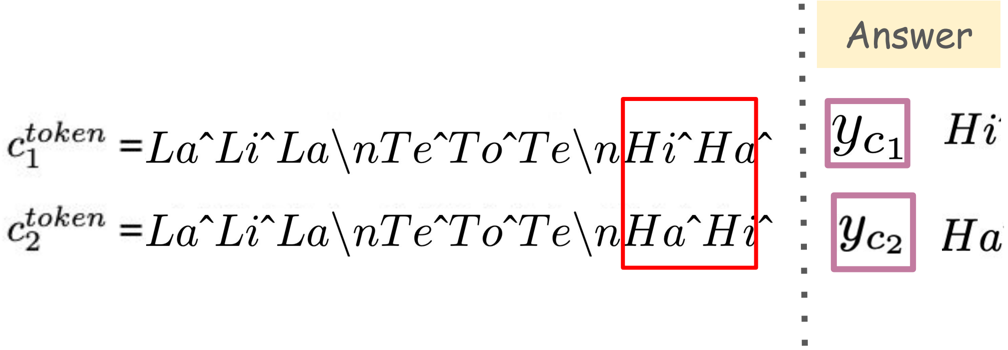
Same Token, Different Symbols/Rules (A vs B)

Causal Mediation Analyses (CMA)

1. Design a context pair (c_1, c_2) to isolate either abstract symbols or literal tokens

Token Context Pair : test whether the embedding represents **literal tokens**

(used for *Retrieval heads*)



$$c_1^{token} = La\hat{\ }Li\hat{\ }La\backslash nTe\hat{\ }To\hat{\ }Te\backslash n\boxed{Hi\hat{\ }Ha}\hat{\ }$$

$$c_2^{token} = La\hat{\ }Li\hat{\ }La\backslash nTe\hat{\ }To\hat{\ }Te\backslash n\boxed{Ha\hat{\ }Hi}\hat{\ }$$

Same Rule, Different Tokens

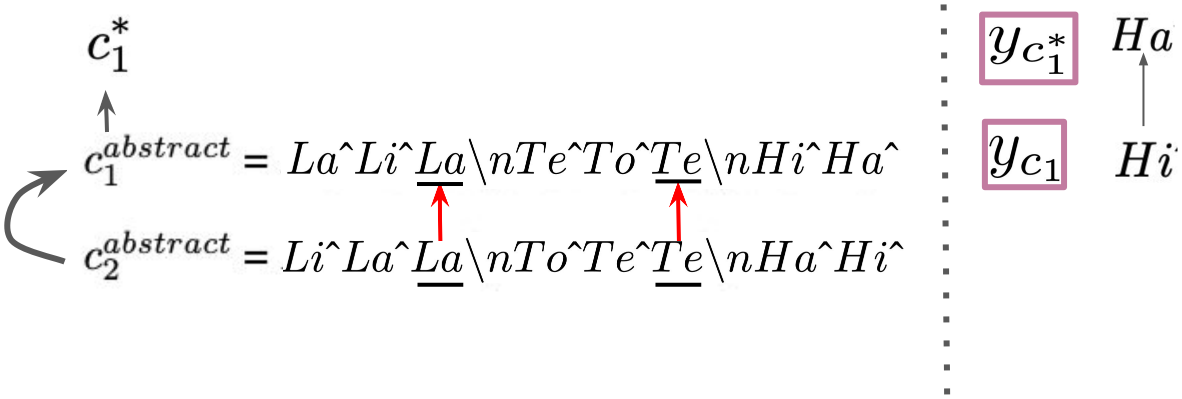
Causal Mediation Analyses

2. Activation Patching at certain token positions:

   Replace attention head outputs in context c_1 with the corresponding activations from c_2

symbol abstraction heads

Answer



c_1^*

$c_1^{abstract} = La\hat{\ }Li\hat{\ }\underline{La}\backslash nTe\hat{\ }To\hat{\ }\underline{Te}\backslash nHi\hat{\ }Ha\hat{\ }$

$c_2^{abstract} = Li\hat{\ }La\hat{\ }\underline{La}\backslash nTo\hat{\ }Te\hat{\ }\underline{Te}\backslash nHa\hat{\ }Hi\hat{\ }$

$\boxed{y_{c_1^*}}$ Ha

$\boxed{y_{c_1}}$ Hi

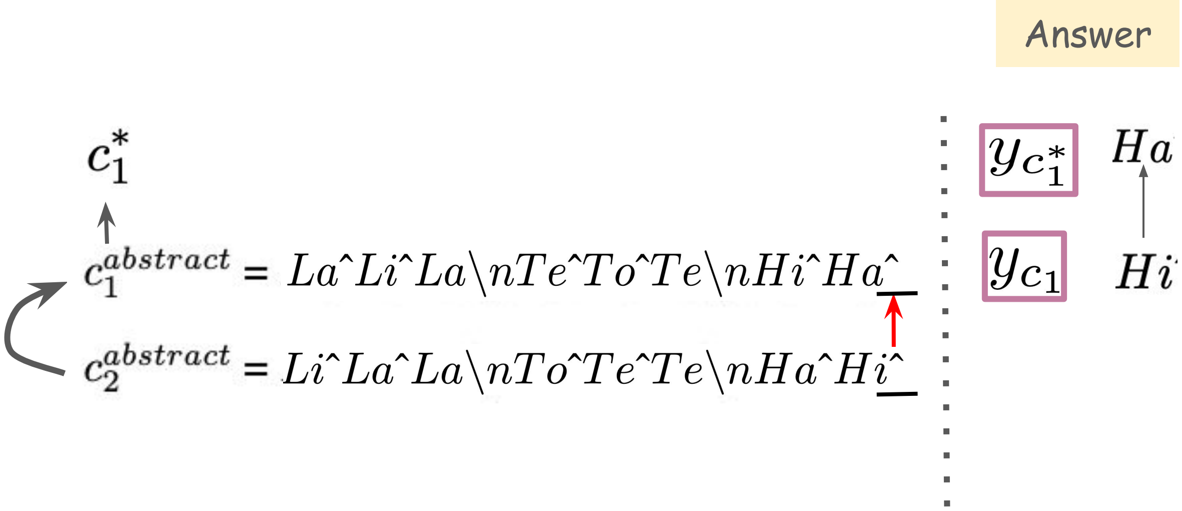
If activations represent the abstract variable (B in c_2) ……

Causal Mediation Analyses

2. Activation Patching at <u>certain token positions</u>:

Replace attention head outputs in context c_1 with the corresponding activations from c_2

symbolic induction heads


Answer
c_1^*
$c_1^{abstract}$ = La^Li^La\nTe^To^Te\nHi^Ha^
$c_2^{abstract}$ = Li^La^La\nTo^Te^Te\nHa^Hi^
$y_{c_1^*}$
y_{c_1}
Ha
Hi

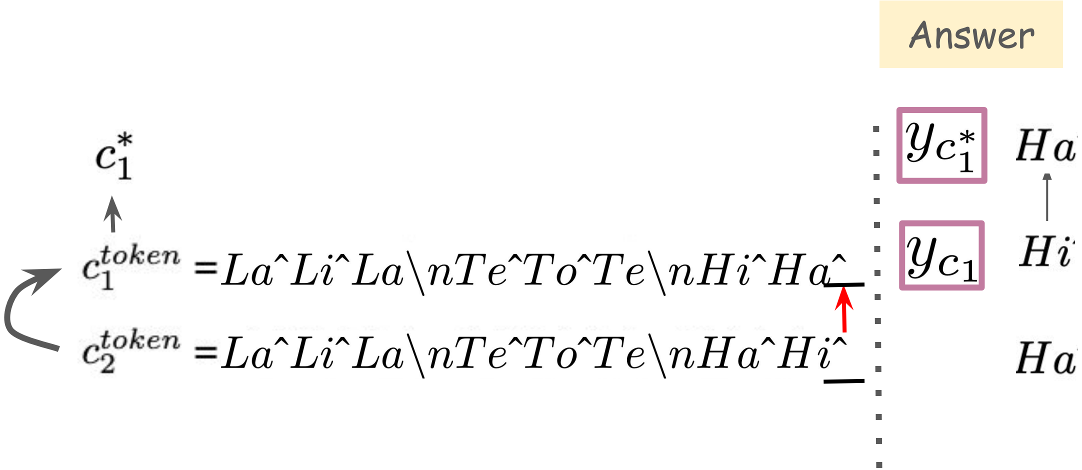
If activations represent the abstract variable (B in c_2) ……

Causal Mediation Analyses

2. Activation Patching at <u>certain token positions</u>:

Replace attention head outputs in context c_1 with the corresponding activations from c_2

Retrieval heads

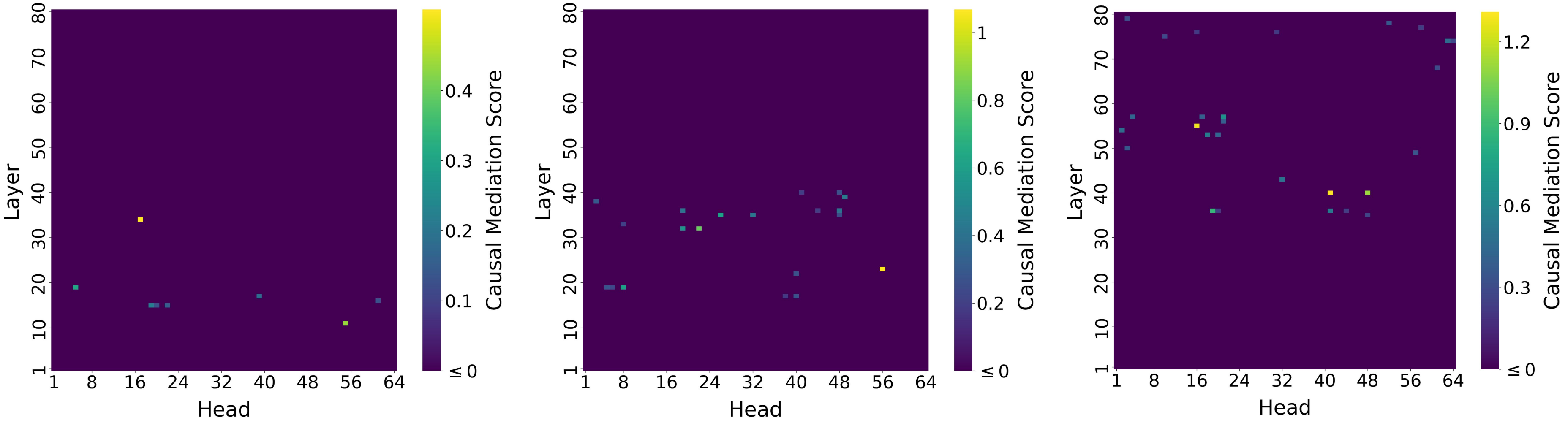

Answer
c_1^*
c_1^{token} = $La\hat{}Li\hat{}La\backslash nTe\hat{}To\hat{}Te\backslash nHi\hat{}Ha\hat{}$_
c_2^{token} = $La\hat{}Li\hat{}La\backslash nTe\hat{}To\hat{}Te\backslash nHa\hat{}Hi\hat{}$_
$y_{c_1^*}$
Ha
y_{c_1}
Hi
Ha

If activations represent the literal token for the answer (*Ha* in c_2) ……

Causal Mediation Analyses

3. Measure whether the causal effects on c_1 comply with the hypotheses

Llama-3.1 70B



Symbol Abstraction Head

Symbolic Induction Head

Retrieval Head

- Defined a score based on the changes in output logits for the answers as a measure of causal effects.
- Conducted multi-hypothesis permutation tests to select significant ones

Our Analyses: (please refer to the paper)

❖ Experiments with Llama-3.1 70B on identity rule tasks

1. Causal Mediation Analyses (CMA)
2. Attention Analyses
3. Representation Similarity Analyses (RSA)
4. Ablation Studies
5. Comparison with Induction Heads and Function Vectors

Our Analyses: (please refer to the paper)

- Evaluating **more LLMs** on identity rule tasks
  - Tested *13* models across *4* model families (GPT-2, Gemma-2, Qwen2.5, Llama-3.1)
  - Found similar robust symbolic mechanisms in *3* model families with GPT-2 as an exception
  - GPT-2 models showed low generation accuracy while symbol abstraction heads rarely emerged.
- Studying **more** complex abstract **reasoning tasks**
  - *Letter string analogies* and *verbal analogies*
  - Identified similar three-stage symbolic processing structures through CMA.
  - Different tasks may involve different attention heads to implement symbol processing.

Please Check out our paper!