Universiteit Utrecht

Probing for Ambiguity in Coreference

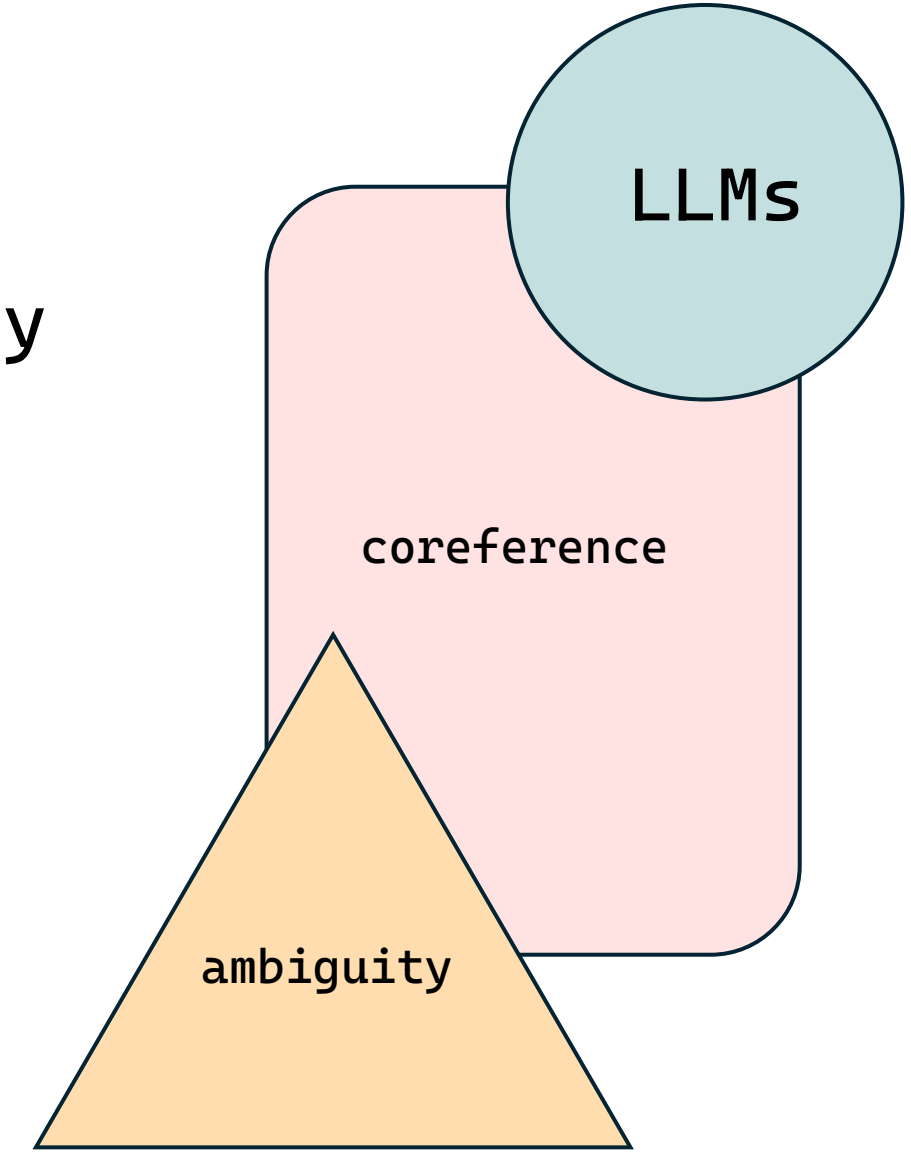
Anh Dang, PhD Candidate

NLP Group Meeting
13-02-2024

Supervisor: Massimo Poesio, Rick Nouwen

Content

- Ambiguity Research in NLP
- Mechanistic Interpretability
- Goal of the Project
- Approach
- Preliminary Results
- Future Directions



I Goal of the project

- Use mechanistic interpretability methods to study how LLMs process coreference ambiguity (i.e., *mereological reference;* Cokal et al, 2023)

- Method: causal mediation analysis, prompting, attention pattern analysis

Cokal, D., Filik, R., Sturt, P., & Poesio, M. (2023). Anaphoric reference to mereological entities. Discourse Processes, 60(3), 202-223.

II Background Ambiguity in NLP

Research on ambiguity in NLP is still challenging

- Different task nature leads to different results (Liu et al. 2023)
- Most methods rely on probability distribution of possible candidate tokens (Wildenburg et al. 2024, Pandit and Hou, 2021)

More research have been using mechanistic interpretability to find the linguistic capability of LLMs, including ambiguity

- Direct Object Identification (Wang et al., 2023)
- Reflexive Anaphora and Subject-verb Agreement (Lepori et al., 2024)
- Syntactic Ambiguity (Hanna and Mueller, 2024; Tucket et al., 2021)
- Lexical Underspecification (Wildenburg et al. 2024)

II Background Mereological Reference

mereological

The engineer hooked up the engine to the boxcar and sent it/them to London

The engine or the boxcar
The engine + the boxcar

The engine and the boxcar

non-mereological

The engineer detached the engine from the boxcar and sent it/them to London

The engine or the boxcar

The engine and the boxcar

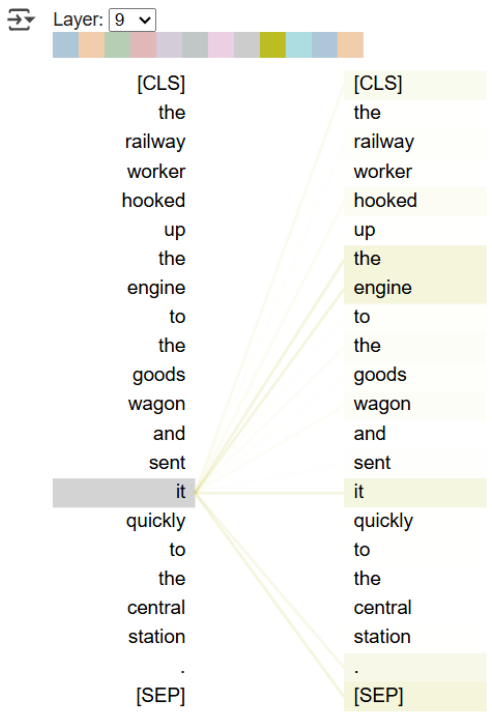
If the LLMs have mereological , we would expect that:

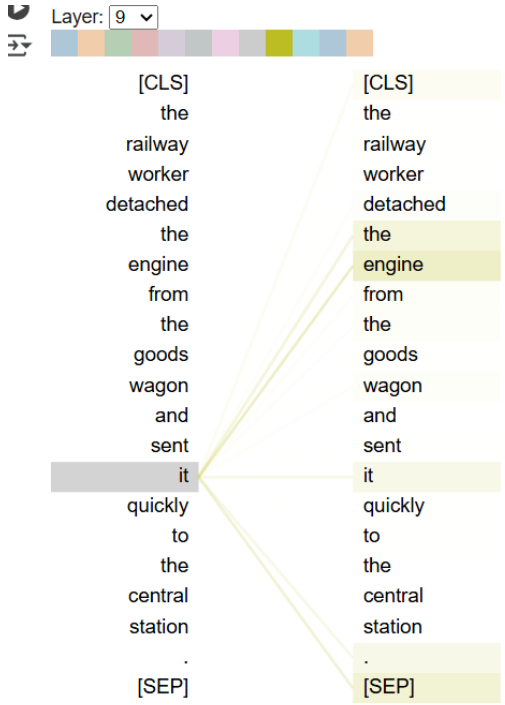
$$\text{Probability Difference} = P_{it} - P_{them}$$

$$\text{Probability Difference}_{\text{non-mereological}} > \text{Probability Difference}_{\text{mereological}}$$

II Background Mereological Reference

Looking at attention pattern





Attention map from layer 9 - head 8 (bert-base-uncased)

Causal Mediation Analysis Vig et al. (2020)

Causal Mediation Analysis for Interpreting Neural NLP: The Case of Gender Bias

Jesse Vig* JVIG@SALESFORCE.COM
Salesforce Research

Sebastian Gehrmann* GEHRMANN@SEAS.HARVARD.EDU
Harvard University

Yonatan Belinkov* BELINKOV@TECHNION.AC.IL
Technion – Israel Institute of Technology

Sharon Qian SHARONQIAN@SEAS.HARVARD.EDU
Harvard University

Daniel Nevo DANIELNEVO@TAUEX.TAU.AC.IL
Tel Aviv University

Simas Sakenis SIMASSAKENIS@COLLEGE.HARVARD.EDU
Jason Huang JASONHUANG@COLLEGE.HARVARD.EDU
Yaron Singer YARON@SEAS.HARVARD.EDU
Stuart Shieber SHIEBER@SEAS.HAVARD.EDU
Harvard University

Abstract

Common methods for interpreting neural models in natural language processing typically examine either their structure or their predictions, but not both. We propose a methodology grounded in the theory of causal mediation analysis for interpreting which parts of a model are causally implicated in its behavior. It enables us to analyze the mechanisms by which information flows from input to output through various model components, known as mediators. We apply this methodology in a case study of gender bias in pre-trained Transformer language models. We analyze the role of individual neurons and attention heads in mediating gender bias across three datasets designed to gauge a model's sensitivity to grammatical gender. Our mediation analysis reveals that gender bias effects are (i) sparse, concentrated in a small part of the network; (ii) synergistic, amplified or repressed by different components; and (iii) decomposable into effects flowing directly from the input and indirectly through the mediators.

III Approach

Method: Causal mediation analysis *(Vig et al. 2020)*

Use causal intervention to find the set of model components responsible for conducting a task (e.g., encoding coreference information)

Base

The nurse examined the farmer for injuries because she was caring/screaming

Patch attention output source -> base

Source

The nurse examined the farmer for injuries because he was caring/screaming

Vig, J., Gehrmann, S., Belinkov, Y., Qian, S., Nevo, D., Singer, Y., & Shieber, S. (2020). Investigating gender bias in language models using causal mediation analysis. Advances in neural information processing systems, 33, 12388-12401.

III Approach

Method: Causal mediation analysis *(Vig et al. 2020)*

Use causal intervention to find the set of model components responsible for conducting a task (e.g., encoding coreference information)

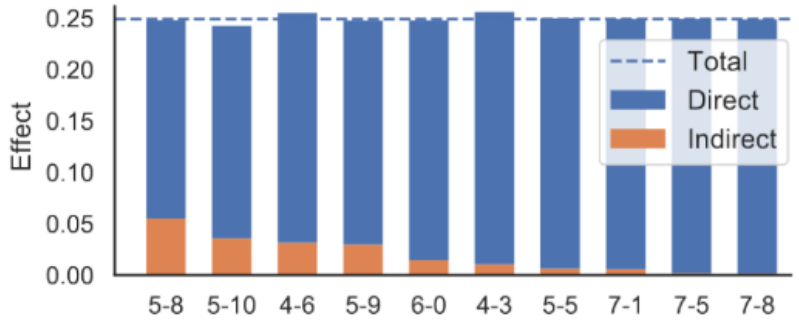


Figure 5: Top 10 heads by indirect effect in GPT2-small on Winobias, and their respective direct effects. Both effects appear largely additive with respect to total effect, a surprising result given the nonlinear nature of these models.

Vig, J., Gehrmann, S., Belinkov, Y., Qian, S., Nevo, D., Singer, Y., & Shieber, S. (2020). Investigating gender bias in language models using causal mediation analysis. Advances in neural information processing systems, 33, 12388-12401.

III Approach

Method: Causal Mediation Analysis *(Vig et al. 2020)*

Use causal abstraction to find the set of model components responsible for conducting a task (e.g., encoding coreference information)

Production task: The LLMs predict the **pronoun**

Jame told Tom that **he** … *is hungry*

Interpretation task: The LLMs produce the **interpretation of the pronoun's reference**

Jame told Tom that he is hungry. Who does "he" refer to? A. Jame B. Tom

Causal mediation analysis is used for both tasks

Vig, J., Gehrmann, S., Belinkov, Y., Qian, S., Nevo, D., Singer, Y., & Shieber, S. (2020). Investigating gender bias in language models using causal mediation analysis. Advances in neural information processing systems, 33, 12388-12401.

III Approach

Method: Causal mediation analysis *(Vig et al. 2020)*

Use causal abstraction to find the set of model components responsible for conducting a task (e.g., encoding coreference information)

Base

The engineer hooked up the engine to the boxcar and sent it/them to London

Patch attention output source -> base

Source

The engineer detached the engine from the boxcar and sent it/them to London

Vig, J., Gehrmann, S., Belinkov, Y., Qian, S., Nevo, D., Singer, Y., & Shieber, S. (2020). Investigating gender bias in language models using causal mediation analysis. Advances in neural information processing systems, 33, 12388-12401.

IV Approach

Method: Causal mediation analysis *(Vig et al. 2020)*

Use causal abstraction to find the set of model components responsible for conducting a task (e.g., encoding coreference information)

The engineer hooked up the engine to the boxcar and sent it/them to London

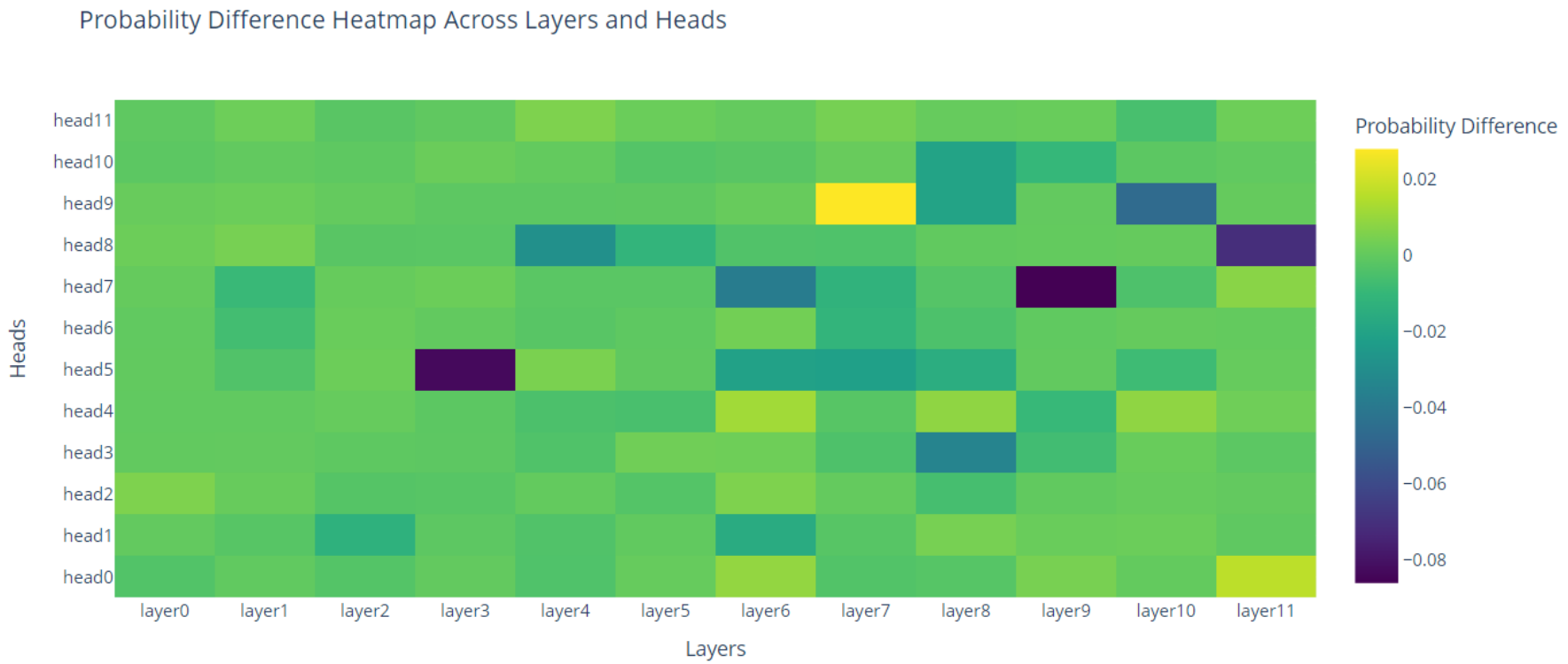
$$PD = P_{it} - P_{them}$$

$$PD_{after\ intervention} < PD_{before\ \ intervention}$$

The engineer detached the engine from the boxcar and sent it/them to London

Vig, J., Gehrmann, S., Belinkov, Y., Qian, S., Nevo, D., Singer, Y., & Shieber, S. (2020). Investigating gender bias in language models using causal mediation analysis. Advances in neural information processing systems, 33, 12388-12401.

IV Preliminary Results



GPT-2 small

Wiegreffe et al. (2020)

Use causal mediation analysis to find model components responsible for multiple-choice question answer selection

🔧 Answer, Assemble, Ace: Understanding How Transformers Answer Multiple Choice Questions

Sarah Wiegreffe[♡♣] Oyvind Tafjord[♡] Yonatan Belinkov[◇]
Hannaneh Hajishirzi[♡♣] Ashish Sabharwal[♡]

[♡]Allen Institute for AI, [♣]University of Washington, [◇]Technion
wiegreffesarah@gmail.com

Abstract

Multiple-choice question answering (MCQA) is a key competence of performant transformer language models that is tested by mainstream benchmarks. However, recent evidence shows that models can have quite a range of performance, particularly when the task format is diversified slightly (such as by shuffling answer choice order). In this work we ask: *how do successful models perform formatted MCQA?* We employ vocabulary projection and activation patching methods to localize key hidden states that encode relevant information for predicting the correct answer. We find that prediction of a specific answer symbol is causally attributed to a single middle layer, and specifically its multi-head self-attention mechanism. We show that subsequent layers increase the probability of the predicted answer symbol in vocabulary space, and that this probability increase is associated with a sparse set of attention heads with

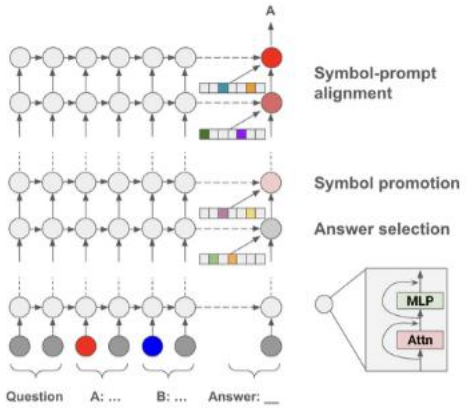


Figure 1: We investigate the ability of transformer language models to answer formatted multiple-choice questions, which involves producing an answer choice symbol (here, "A" or "B"). We discover a single middle layer at the last token position, and particularly, its multi-head self-attention function, responsible for answer selection.

Wiegreffe et al. (2024)

Use causal mediation analysis to find model components responsible for multiple-choice question answer selection

"A: __, B: __" (1)

"A: __, B: __" (2)

correct answer

Swap the hidden state output (MLP and MHSA) of the last token position of prompt (1) and prompt (2)

```
For each of the following phrases, select the
best completion (A or B).

<in-context examples>

Phrase: <question>
Choices:
A: <correct answer>
B: <incorrect answer>
The correct answer is:
```

Wiegreffe et al. (2024)

Use causal mediation analysis to find model components responsible for multiple-choice question answer selection

“A: __, B: __” (1)

“A: __, B: __” (2)

correct answer

Swap the hidden state output (MLP and MHSA) of the last token position of prompt (1) and prompt (2)

Findings

- There are several heads in a single middle layer responsible for choosing the answer

IV Approach

How do we know whether these components are actually responsible for mereological reference?

Interpretation task: Combine prompting and mechanistic interpretability

Have the model explicitly specify the referent of the pronoun through answering multiple choice questions.

```
In the sentence “The engineer hooked up the engine to the
boxcar and sent it to London”, what does “it” refer to?

A. The engine or the boxcar
B. The object formed by attaching the engine and the
boxcar.

Your answer is:
```

We cannot know which model component is responsible for linking the pronoun and its referent

IV Approach

How do we know whether these components are actually responsible for mereological reference?

Prompt 1:
In the sentence “The engineer hooked up the engine to the boxcar and sent it to London”, what does “it” refer to?

A. The engine or the boxcar
B. The object formed by attaching the engine and the boxcar.

Your answer is:

(1)

$P(A|prompt_1,a_2)$

$P(B|prompt_1,a_2)$

Prompt 2:
In the sentence “The engineer detached the engine to the boxcar and sent it to London”, what does “it” refer to?

A. The engine or the boxcar
B. The object formed by attaching the engine and the boxcar.

Your answer is:

(2)

attention output (a)

IV Next Step

- Whether it is possible to directly compare the results of the production task and the interpretation task

The model component used for answering the interpretation question may be doing more task than encoding mereological information (monosemanticity)

- Compare results with Vig et al. (2020) paper

Maybe interesting to compare our results with theirs. However, given the different nature of the task (predicting verbs and predicting pronouns), it can be challenging to interpret

- Combine causal mediation analysis and circuit probing (train binary mask over model components)

Would we find the same component?

Thank you for listening!

All questions and feedbacks are appreciated!