

Ambiguity in Coreference

NLP group meeting

November 7th

Anh Dang

Content

1. Introducing ambiguity in plural anaphoric expressions
2. Approach to ambiguity
3. Probing method: Circuit probing
4. Probing method: Causal mediation analysis

1. Ambiguity in Coreference

What does it mean for an anaphoric expression to be ambiguous?

- Having more than 1 possible antecedent

Singular reference

Stan and Pam went to the store. She brought milk

>>> She can refer to Stan or Pam

Plural reference

Tom and Stefan went to a restaurant with Harry last night. They ate a pizza.

>>> They can refer to the group of Tom and Stefan and also the three of them

1. Ambiguity in Coreference

Psycholinguistic studies show that humans have preferences towards interpreting ambiguous plural expressions (Cokal et al., 2023; Koh and Clifton, 2002)

1. Are LLMs able to detect ambiguity in plural anaphoric expressions?
2. Do they have human-like preferences for interpreting those ambiguities?

1. Ambiguity in Coreference

Readers are more willing to take an ontologically homogeneous (all humans) collection of entities as the antecedent of **they** than an ontologically heterogeneous collection (Koh and Clifton, 2002)

What does “both of them” refer to?

When the house was burning down, John entered it. He called his **son**. He called his **daughter**. He also searched for his **briefcase**. He saved both of **them** but lost his (**briefcase/his son**).

2nd antecedent

3rd antecedent

1st antecedent

1. Ambiguity in Coreference

Mereological Entities (Cokal et al. 2023)

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

mereological

engine + boxcar

The engineer **separated** the **engine** from the **boxcar** and sent **it** to London.

non-mereological

engine or boxcar

1. Goal of the project

How do LLMs interpret ambiguous plural anaphoras?

- Understand whether LLMs can identify ambiguity and possible antecedent for the mention
 - There is evidence that LLMs are not fully able to identify ambiguity (Liu et al., 2023)
 - There is currently few research on the ability of LLMs to handle ambiguity related to coreference resolution (Wildenburg et al., 2024)
 - Most research focuses on singular referent

Liu, A., Wu, Z., Michael, J., Suhr, A., West, P., Koller, A., ... & Choi, Y. (2023). We're Afraid Language Models Aren't Modeling Ambiguity. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 790-807).

Wildenburg, F., Hanna, M., & Pezzelle, S. (2024). Do Pre-Trained Language Models Detect and Understand Semantic Underspecification? Ask the DUST!. *arXiv preprint arXiv:2402.12486*.

1. Goal of the project

How to test whether LLMs understand ambiguity?

- Examine logit difference and/or probability difference between possible referents (Wildenburg et al., 2024)

How much do probability reflect LLMs' coreference information?

When the house was burning down, John entered it. He called his **son**. He called his **daughter**. He also searched for his **briefcase**. He saved **both** of **them** but lost his **[MASK]**.

$P_{(\text{briefcase}|\text{C})}$ $P_{(\text{son}|\text{C})}$

The engineer **hooked up** the **engine** to the **boxcar** and sent **[MASK]** to London.

1. Goal of the project

How to test whether LLMs understand ambiguity?

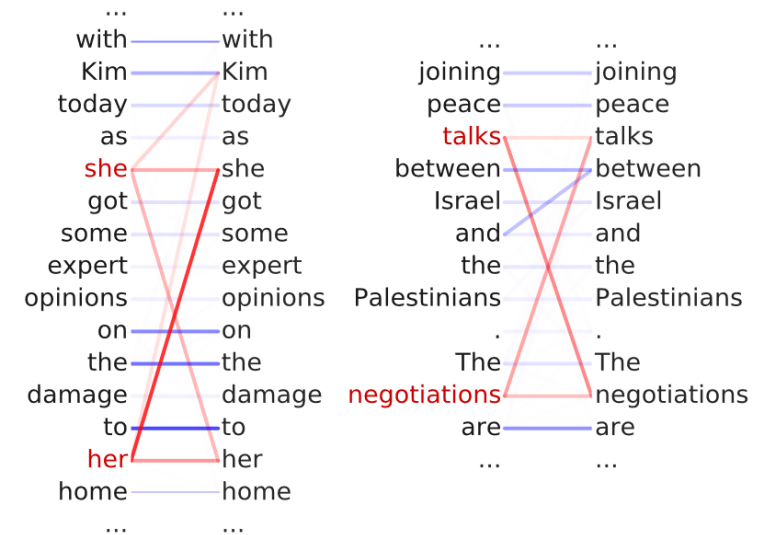
>>> Use **mechanistic interpretability** to understand how LLMs represent ambiguity in ambiguous plural anaphoric expressions

Find the component of the model that is responsible for assigning the referent for a mention

Some attention heads of LLMs encode information about coreference (Clark et al. 2019)

Head 5-4

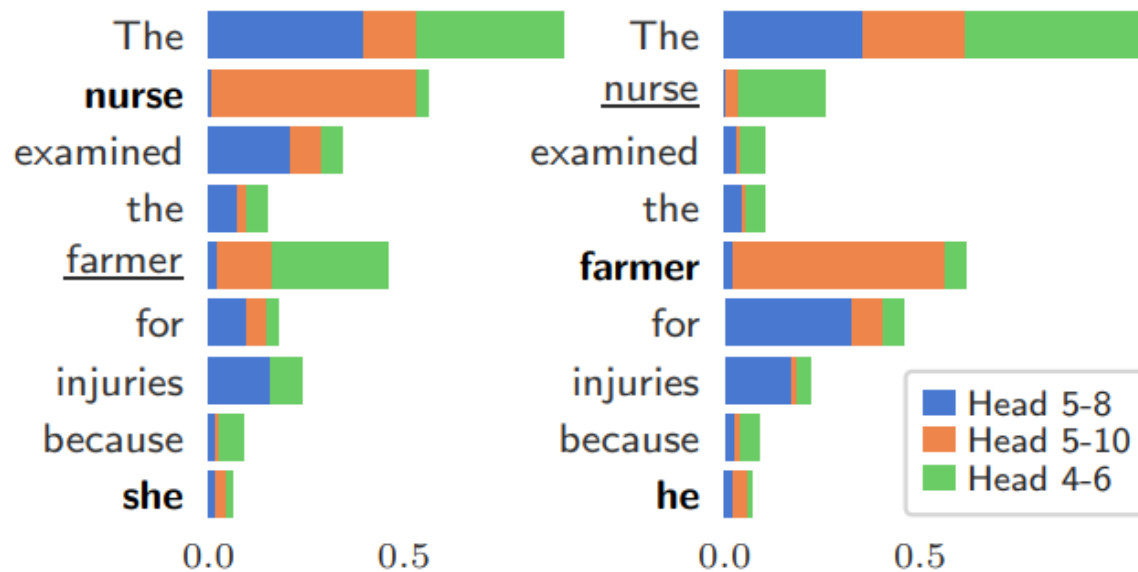
- **Coreferent** mentions attend to their antecedents
- 65.1% accuracy at linking the head of a coreferent mention to the head of an antecedent



Clark, K. (2019). What Does Bert Look At? An Analysis of Bert's Attention. *arXiv preprint arXiv:1906.04341*.

1. Goal of the project

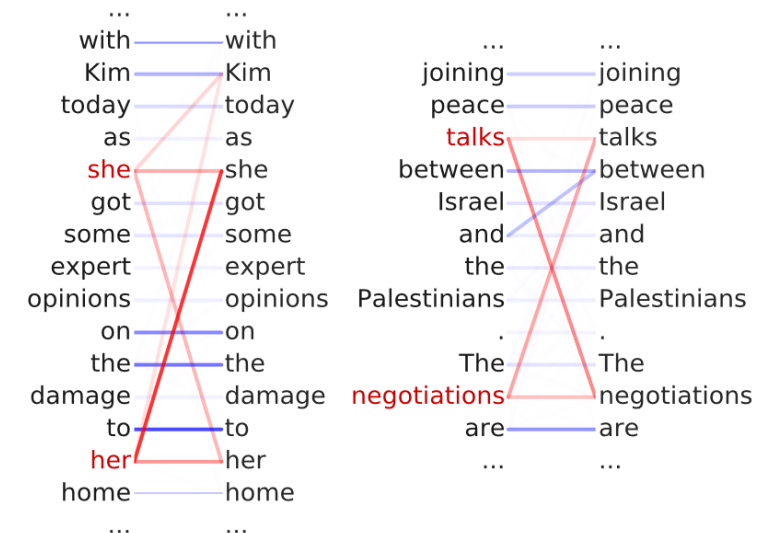
The attention pattern of LLMs encode information about coreference (Clark et al. 2019)



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.

Head 5-4

- **Coreferent** mentions attend to their antecedents
- 65.1% accuracy at linking the head of a coreferent mention to the head of an antecedent



Clark, K. (2019). What Does Bert Look At? An Analysis of Bert's Attention. *arXiv preprint arXiv:1906.04341*.


3. Circuit Probing

Circuit

An induced sub-graph of the model's computational graph that is responsible for completing a certain graph (Lepori et al., 2023)

Circuit probing aims to

- Test whether high-level intermediate variables is represented in the model
- Test whether they are causally implicated in the model's behavior
- Reveal the particular subset of **model component** that is responsible for such variables



A layer, head, or
neuron

3. Circuit Probing

Circuit

An induced sub-graph of the model's computational graph that is responsible for completing a certain graph

How?

- Discover the circuit that perform the task
- Ablate the circuit and see how it affect model's prediction

3. Circuit Probing

- **Identifying a high-level causal variable** that you believe might play a role in the model's computation.

E.g., subject-verb agreement >>> syntactic number of the head noun

- **Create labels** to represent the pairwise similarity between inputs based on their labels

E.g., singular: 0, plural: 1

- Introduce a **trainable binary mask over the weights of a specific model component** (e.g., an attention block or MLP).

- Freeze the model's original parameters and **optimize the binary mask parameters**. The optimization objective is soft nearest neighbors loss

Subject-verb agreement

An intermediate variable representing the **syntactic number of the subject noun** is computed when predicting the main verb of a sentence

Residual
stream

sum of the output of all the previous layers and the original embedding

plural

Input

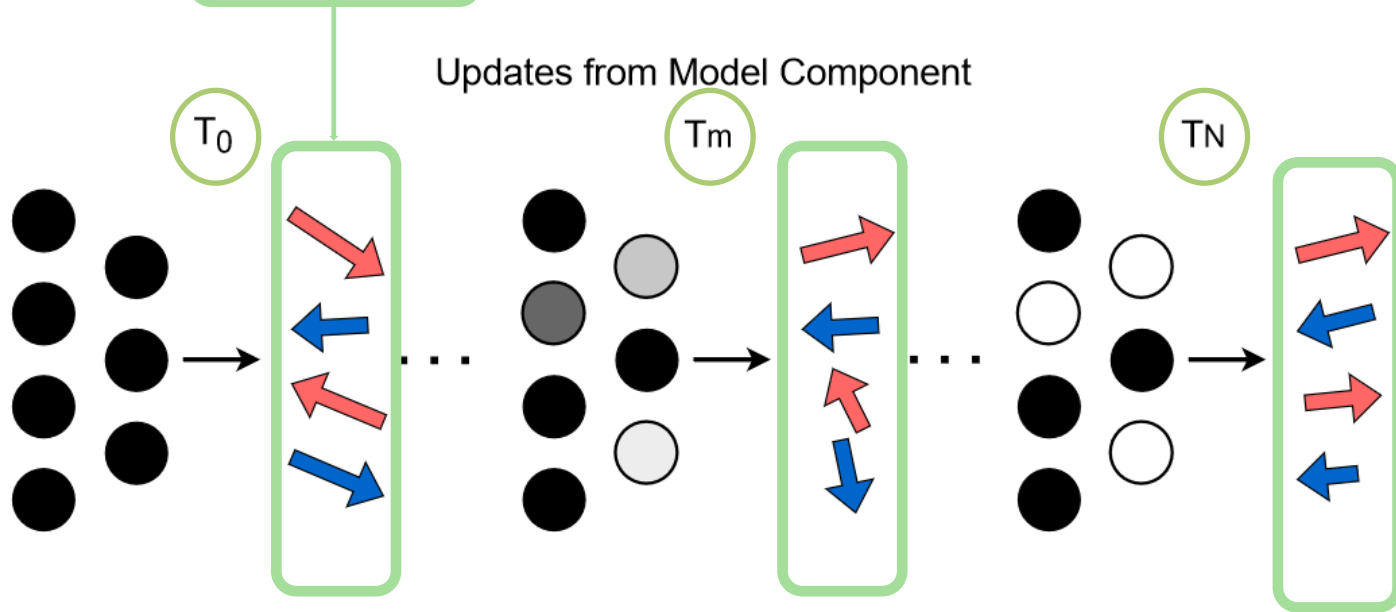
"The **movies**"

"An **artist** near the cars"

"Doctors"

"The **person**"

singular



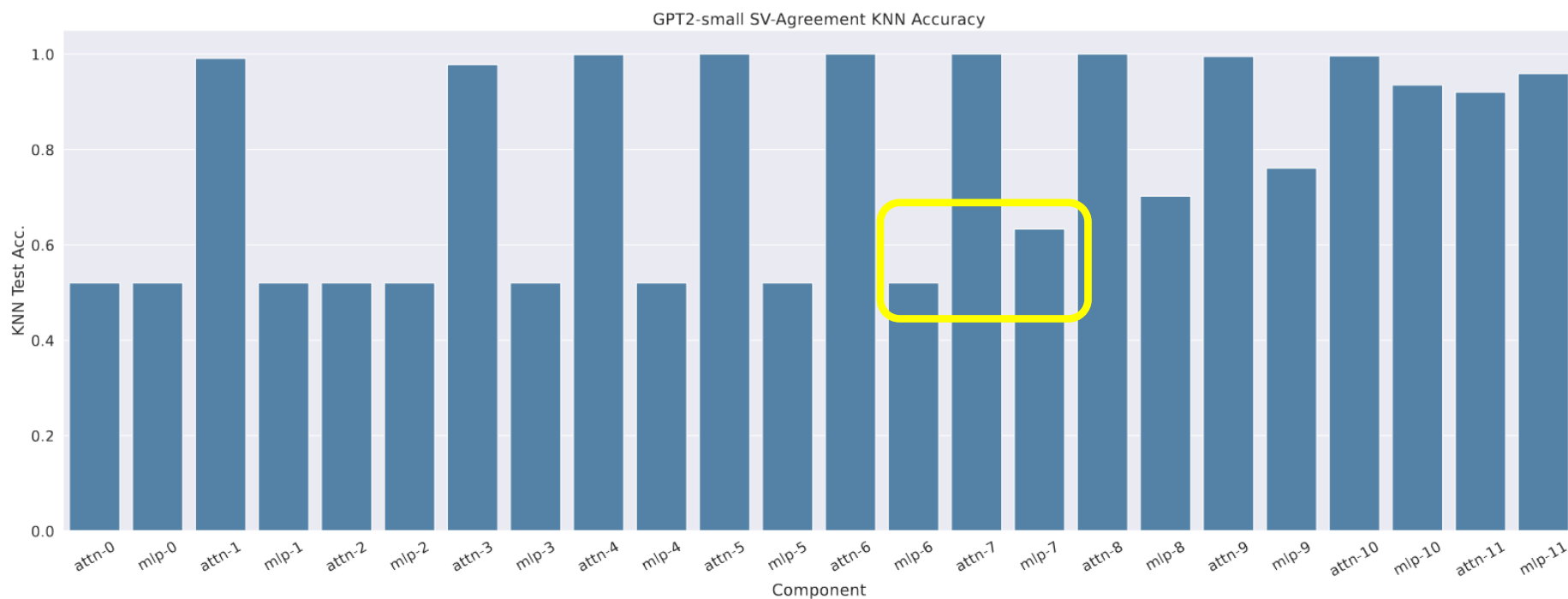
3. Circuit Probing

Circuit Validation

- K-Nearest Neighbors Classifier
- Ablation

3. Circuit Probing

the authors behind the assistants (is/are)



most attention layers can compute the correct syntactic number, but that MLPs only begin to achieve good performance in the middle layers of GPT2-Small

4. Causal Mediation Analysis

Vig et al. (2020)

Examining gender bias in GPT-2

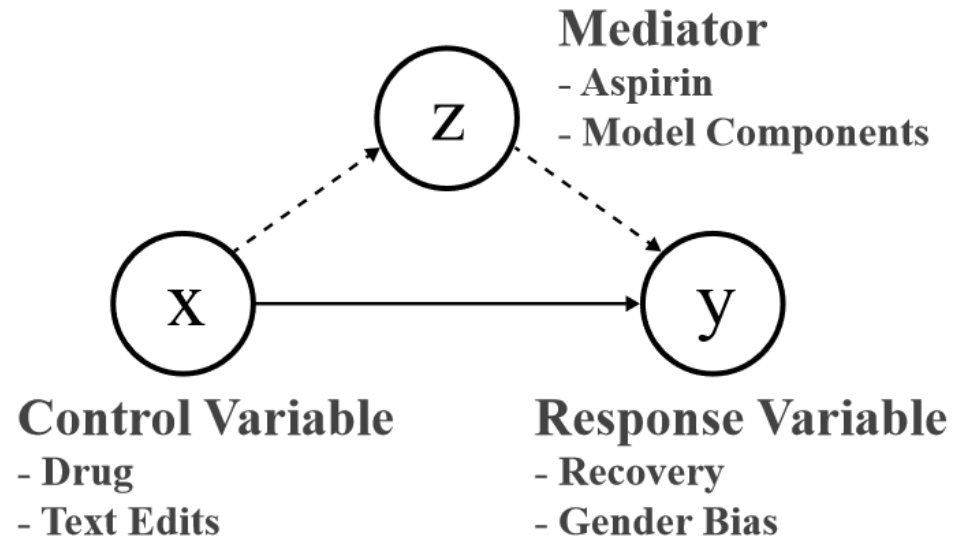


Figure 1: Mediation analysis illustration.

4. Causal Mediation Analysis

Vig et al. (2020)

Examining gender bias in GPT-2

U: The nurse said that [she]

U: The nurse said that [he]

$p([\text{he}]|u) = p([\text{he}]|\text{the nurse said that})$

$p([\text{he}]|u) / p([\text{she}]|u)$

$p([\text{she}]|u) = p([\text{she}]|\text{the nurse said that})$

$p([\text{he}]|u) = p([\text{he}]|\text{the man said that})$

$p([\text{he}]|u) / p([\text{she}]|u)$

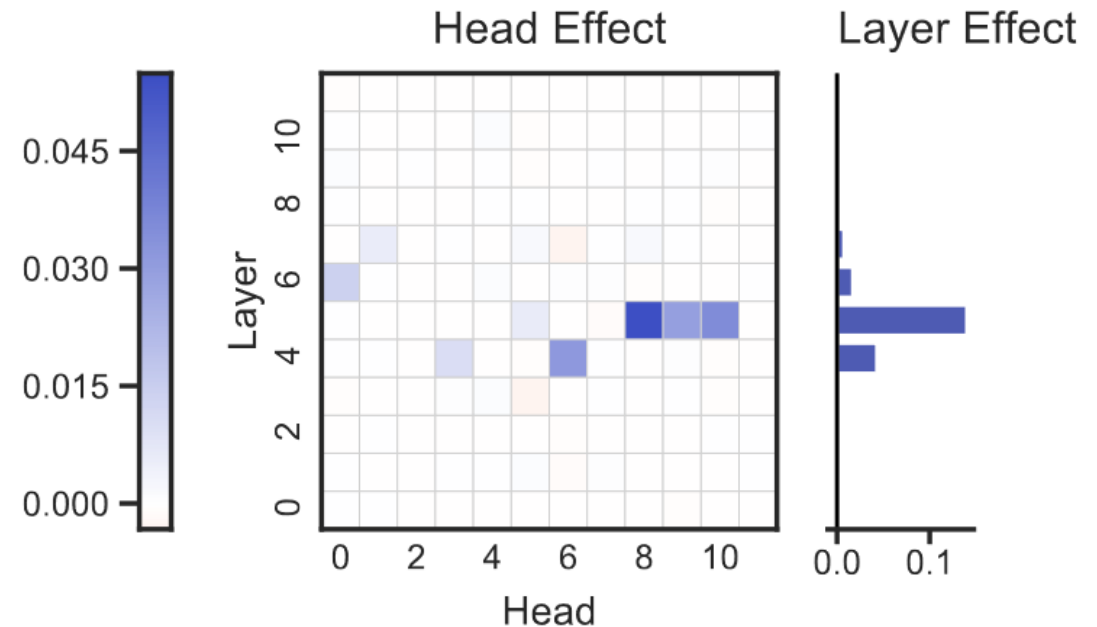
$p([\text{she}]|u) = p([\text{she}]|\text{the man said that})$

4. Causal Mediation Analysis

Vig et al. (2020)

Examining gender bias in GPT-2

The nurse said that [she]
The nurse said that [he]



(a) Indirect effects in GPT2-small on Winobias for heads (the heatmap) and layers (the bar chart).

4. Causal Mediation Analysis

Vig et al. (2020)

Attention Intervention – Which attention heads encode gender bias?

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

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

Gender bias = $p(\text{[was caring]}|u) / p(\text{[was screaming]}|u)$

>>>> Replace the attention weights of “she” to all other tokens with the attention weights of “he”

4. Causal Mediation Analysis

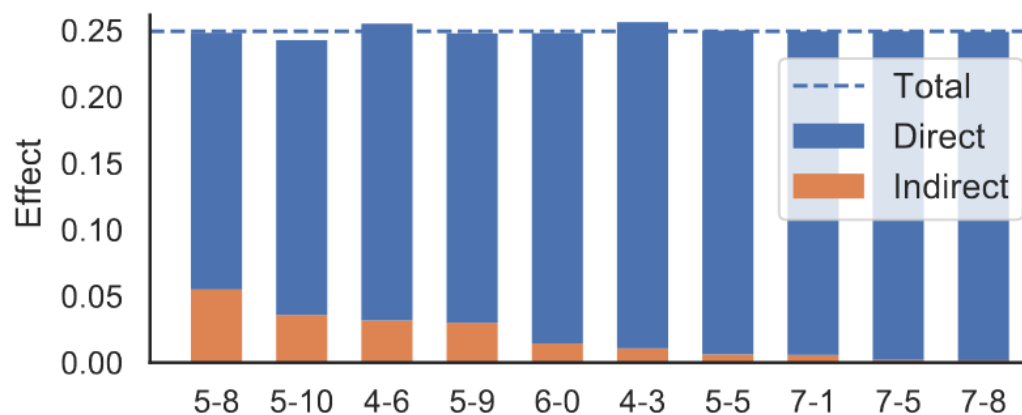
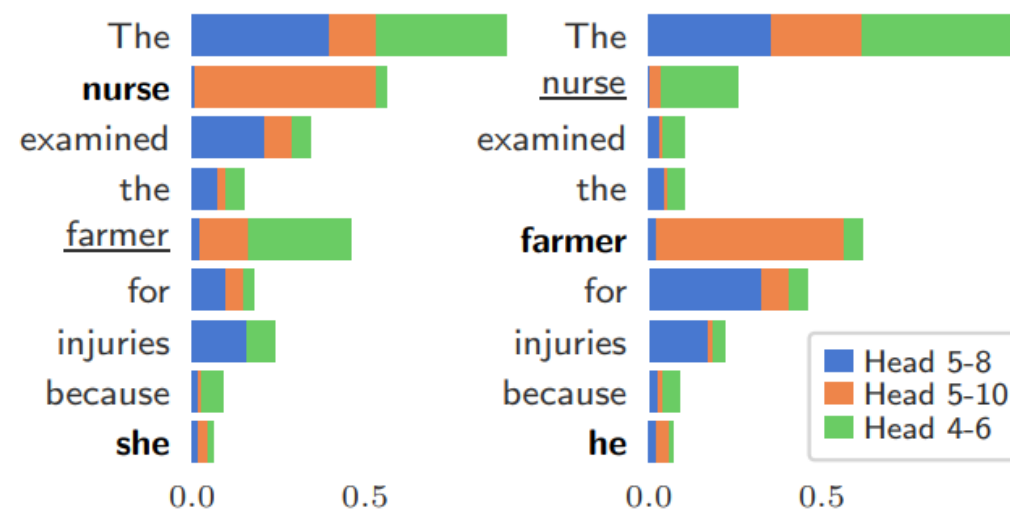


Figure 5: Top 10 heads by indirect effect in GPT2-small on Winobias, and their respective direct effects. Both effects appear largely additive with respect to total effect, a surprising result given the nonlinear nature of these models.



5. Application to modeling ambiguity

- Find the subset of head/layer that is responsible for singular and plural reference
- Compare the circuit for the ambiguous and unambiguous

Thank you for listening!