We has an neuron

Cassidy Tan, Joseph Dantes, Katrina Compendio, Zmavli Caimle

Outline

- Summary of Work Done
- Overview
- Methodology
- Replicating *We found an neuron*
- Extending the Work to Subject-Verb Agreement (Has/Have)
- Key Learnings
- Future Work

Summary of Work Done

1. **Replicating *We found an neuron***
    - We used GPT2-Large and visualized the logits difference between “an” and “a” to identify that **L26 (attn)** and **L31 (MLP)** were most relevant
    - We **visualized** the attention patterns of L26 and found that it looked at the appropriate object (noun phrase) when predicting whether to use “an or “a”
    - We swept through the logit differences of 5120 neurons to identify the “an” neuron, which was supported by performing **inversion ablation** to the neuron activation and weights
2. **Extending to Subject-Verb Agreement (Has/Have)**
    - We applied the same methodology to identify any layers that may be useful for subject-verb agreement (specifically for **“has”** and **“have”**), found several MLP layers with high logits differences, with layer 0 showing significant logit difference for the patched head value
    - We looked at the **attention** patterns and found that the model looked at the **existence of a previous “has”**, as opposed to the noun phrase
    - We identified and performed ablation experiments to a relevant **neuron in layer 0**, and found that **ablation** on either the activation or weights led to a change in the predictions

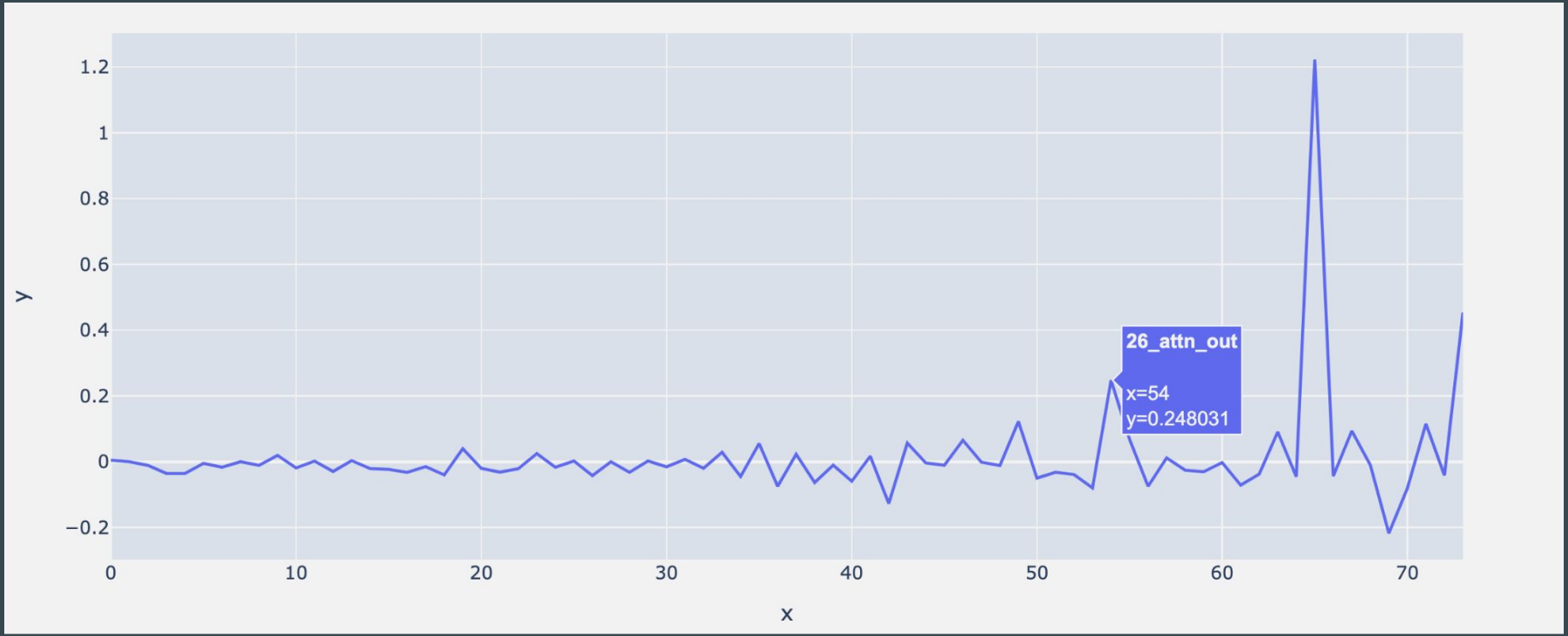
Overview

- How does a transformer know when to use “an” or “a”?
- Can we replicate previous work on this? (i.e., by Miller & Neo)
- Can we apply the methodology to other use cases such as subject-verb agreement?

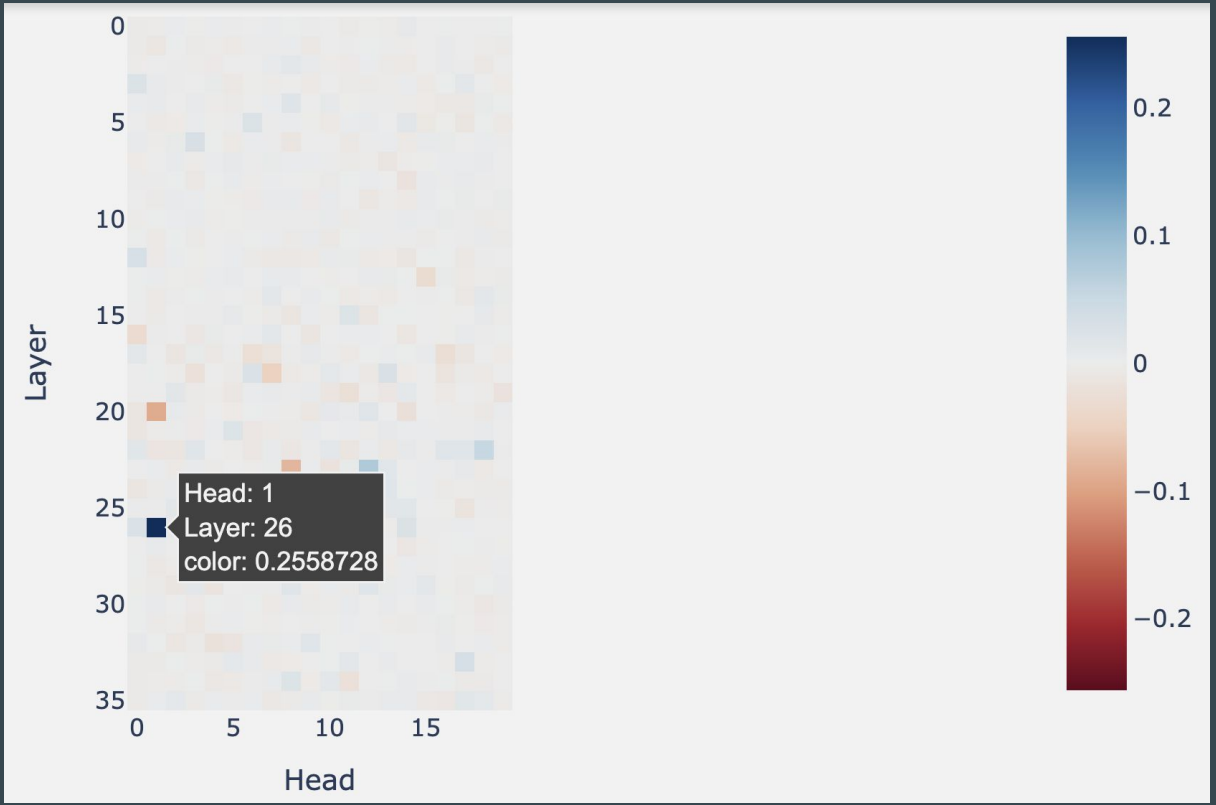
Methodology

1. Identify relevant layers from the logits differences of opposing tokens (e.g., “an” vs “a”)
2. Visualize top attention heads (based on logits attribution) and explore patterns from prompts
3. Sweep through all neurons in the relevant layer to find the one with highest logits difference
4. Perform ablation experiments on the activation and weights of the selected neuron (e.g., inversion and zero ablation)

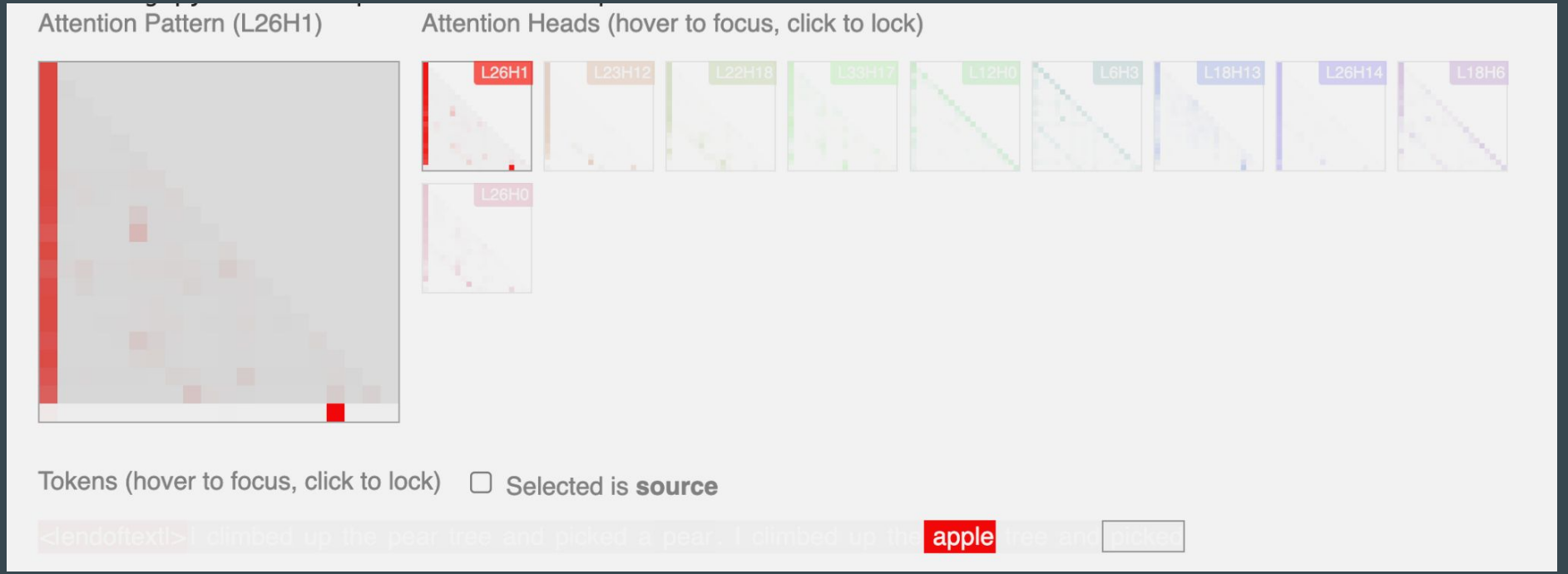
Replicating *We found an neuron* - Logit Difference Per Layer



Replicating *We found an neuron* - Logit Difference Per Head

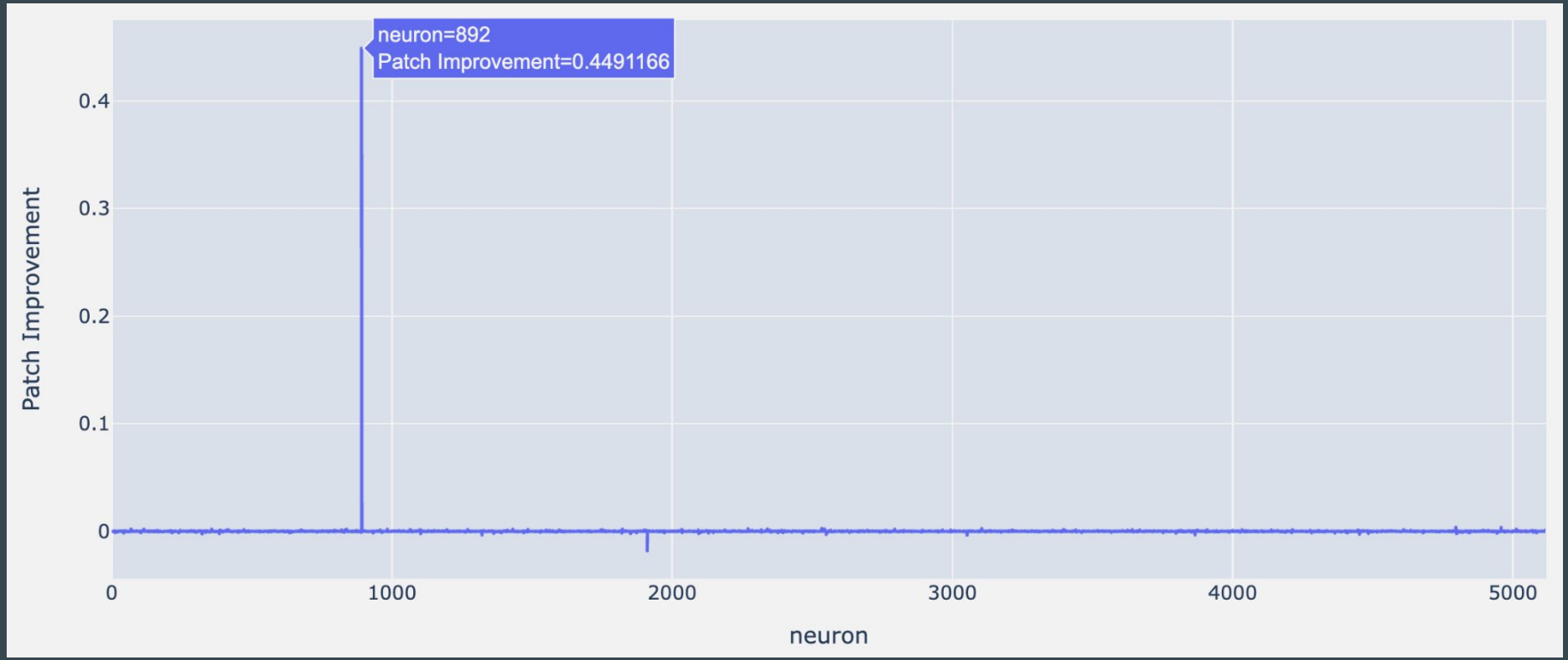


Replicating *We found an neuron* - Top 10 Positive Logit Attribution Heads



<|endoftext|>I climbed up the pear tree and picked a pear. I climbed up the **apple** tree and **picked**

Replicating *We found an neuron* - Logit Difference From Patched Neurons in MLP Layer 31



Replicating *We found an neuron* - Top Predictions After Applying Inversion Ablation to the Activation and Weights (MLP Layer 31 Neuron 892)

Prompts (Expected Prediction)	**Original Model**	**Model w/ Modified Activation**	**Model w/ Modified Weights**
<|endoftext|> I climbed up the pear tree and picked a pear. I climbed up the **apple** tree and picked (Expected: " an")	**" an" - 64.92%** " a" - 24.22% " apples" - 2.78% " two" - 2.43% " another" - 2.07%	**" a" - 83.45%** " apples" - 3.81% " an" - 2.85% " two" - 2.61% " another" - 2.59%	**" a" - 58.31%** " an" - 26.17% " apples" - 4.63% " two" - 2.97% " another" - 2.96%
<|endoftext|> I climbed up the pear tree and picked a pear. I climbed up the **lemon** tree and picked (Expected: " a")	**" a" - 85.10%** " an" - 3.18% " two" - 2.52% " some" - 2.05% " another" - 1.68%	**" a" - 86.36%** " two" - 2.44% " an" - 2.28% " some" - 1.97% " another" - 1.61%	**" a" - 85.99%** " an" - 2.58% " two" - 2.46% " some" - 2.00% " another" - 1.63%

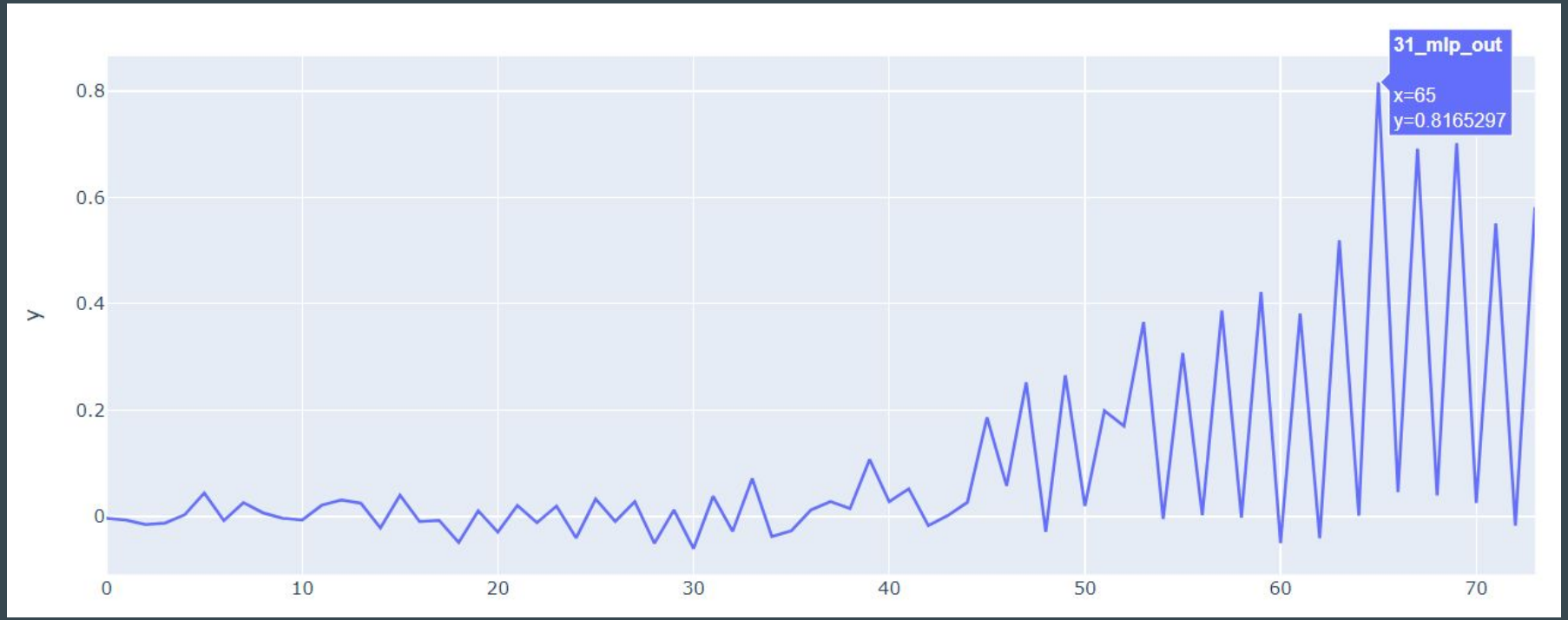
Replicating *We found an neuron* - Experiment Findings

- The relevant layers, attention or MLP heads may be inferred by observing the logits differences for each
- The top attention head points to the noun phrase used to determine “an” or “a”
- Ablation experiments support the significance of the identified “an” neuron for prediction

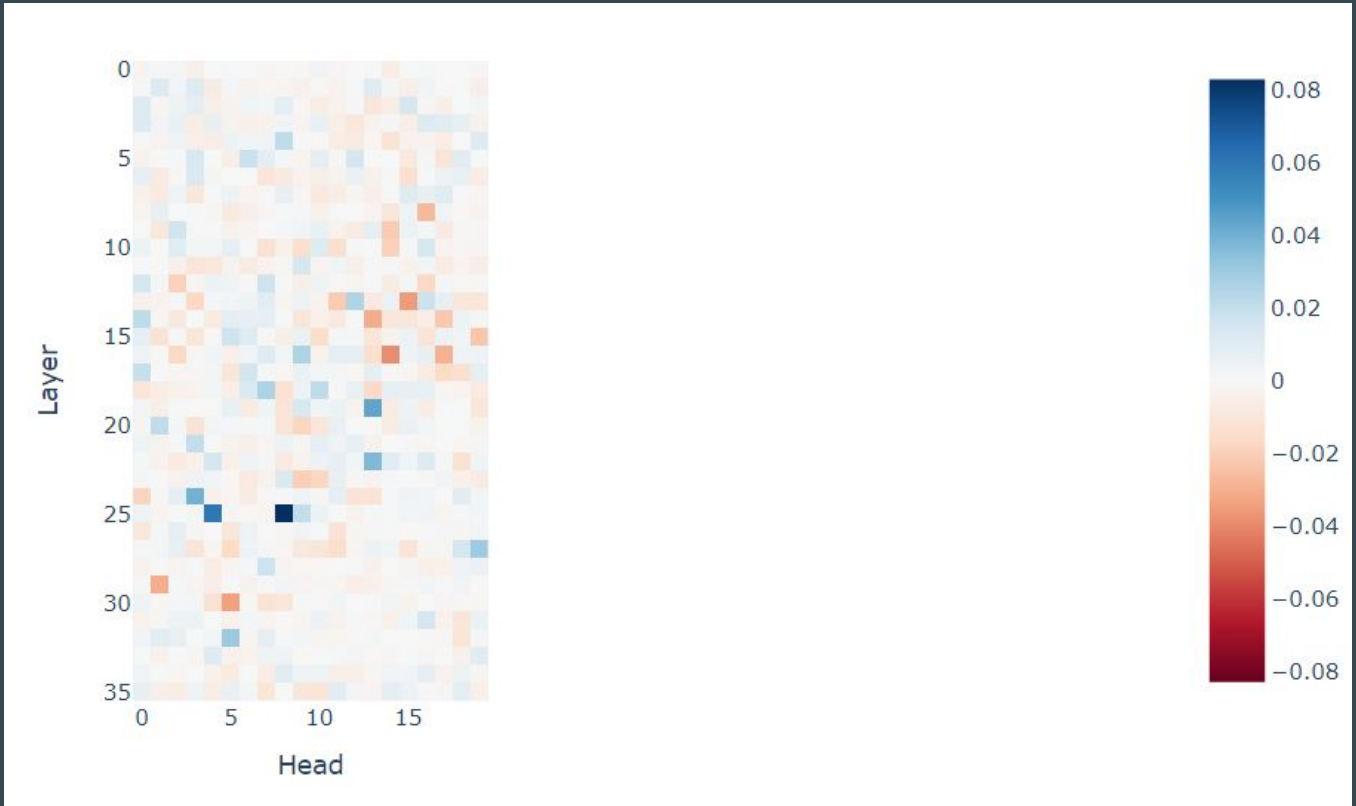
Extending to Subject-Verb Agreement

- Can the methodology be applied to other use cases?
- How does a model determine the verb to use based on plurality?
- What can we learn from a specific scope, then build up from there?
  - Consider only “has”/”have”, not other verbs
  - Assume that appending “s” creates the plural form of the subject
  - Prompts are loosely of the form: “Noun_1 has a something. [Singular/Plural]_Noun_2 … [has/have]”

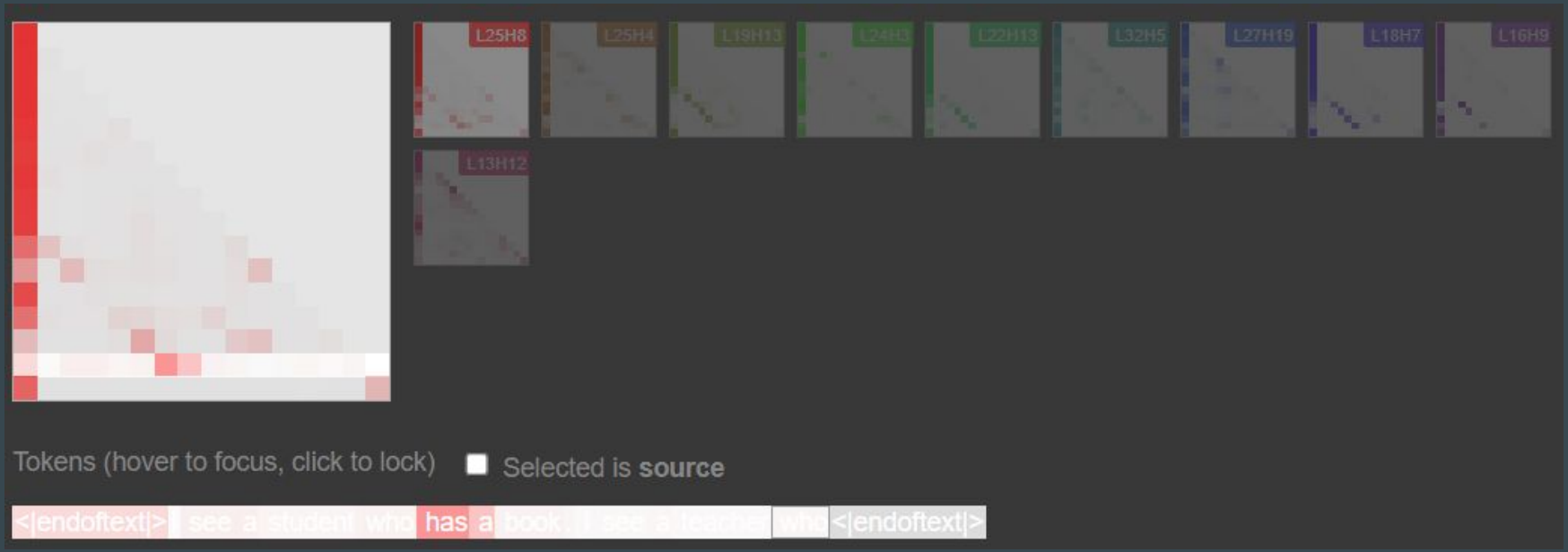
Extending to Subject-Verb Agreement - Logit Difference Per Layer



Extending to Subject-Verb Agreement - Logit Difference Per Head



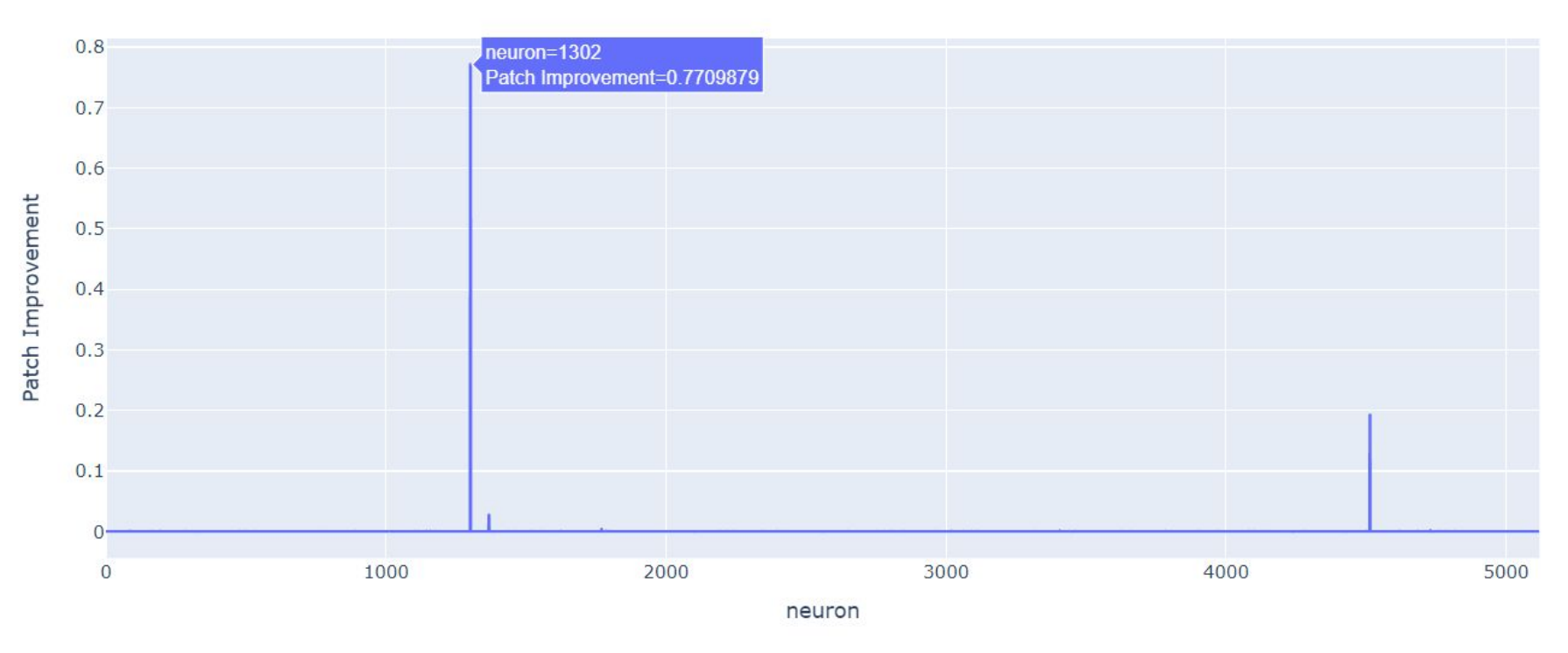
Extending to Subject-Verb Agreement - Top 10 Positive Logit Attribution Heads



<|endoftext|>I see a student who has a book. I see a teacher **who**

Extending to Subject-Verb Agreement - Logit Difference From Patched Head Value

Extending to Subject-Verb Agreement - Logit Difference From Patched Neurons in MLP Layer 0



Extending to Subject-Verb Agreement - Top Predictions After Applying Inversion Ablation to the Activation and Weights of MLP Layer 0 Neuron 1302

Prompts (Expected Prediction)	**Original Model**	**Model w/ Modified Activation**	**Model w/ Modified Weights**
<|endoftext|> I see a student who has a book. I see **a teacher** who (Expected: ' has')	**' has' - 37.82%** ' is' - 18.30% ' wants' - 3.75% "s' - 2.07% ' needs' - 1.55%	**' has' - 52.52%** ' is' - 8.27% ' plays' - 7.64% ' wears' - 2.59% ' sings' - 1.30%	**' has' - 51.17%** ' plays' - 9.19% ' is' - 9.10% ' wears' - 2.12% "s' - 1.63%
<|endoftext|> I see a student who has a book. I see **the teachers** who (Expected: ' have')	**' are' - 16.21%** **' have' - 14.65%** ' want' - 3.12% ' read' - 2.48% ' teach' - 1.84%	**' have' - 33.45%** ' are' - 13.11% ' play' - 11.32% ' wear' - 2.34% ' dance' - 1.99%	**' has' - 39.29%** ' plays' - 11.91% ' is' - 8.63% "s' - 2.24% ' sings' - 1.80%

Extending to Subject-Verb Agreement - Logit Difference From Patched Head Value Between “I” and “He”

Extending to Subject-Verb Agreement - Experiment Findings

- The same methodology from *We Found An Neuron* was applicable for a different use case
- The model focused more on the existence of a “has” (as opposed to the noun phrase), which may generalize better for irregular plural words
- Logits differences were higher for MLP than attention blocks
- Ablation on either the activation or weights may change the top prediction, which suggests some significance of the neuron, while possibly recovering performance from other layers

Key Learnings

- Attention visualizations are helpful in exploratory analysis and observing patterns
- Ablation experiments support the significance of an identified “an” or “have” neuron for prediction
- MLPs are more responsible in distinguishing plurality vs attention, at least for the explored scope of prompts
- Using the GPU leads to approximately a 10x speedup

Future Work

- Investigate the “an neuron” for corner cases such as "an hour" or "a usurper" (i.e., based on sound, not just spelling)
- Consider exploring relevant attention heads and neurons for other use cases
    - other verbs
    - I/you, and irregular plural forms
    - if-then
    - gender/pronoun
    - spelling

Question and Answer Segment