Supplementary Material for: Detecting and Understanding Vulnerabilities in Language Models via Mechanistic Interpretability

Author Name
Affiliation
email@example.com

A Dataset Building Process

In this section, we will give a more detailed explanation about the dataset building process used on the case study. As we are going to study the task of acronym prediction (e.g. `"The Chief Executive Officer (CE"` → `"O"`), we will build a dataset that elicits such behavior. Specifically, the samples will be built according to the template:

`"|The|W1|W2|W3| (|A1|A2|A3|"`

where the vertical bars `"|"` delimit the different tokens of the sentence, `Wi` are nouns preceded by a space and tokenized as a single token (e.g. `" Cane"`) and `Ai` the corresponding letter of the acronym (e.g. `"C"`). These decision choices were made in order to isolate the behavior of study and reduce the amount of noise. Hence, the process of building samples of the dataset has the following steps:

1. Starting from a list containing 91000 nouns,[1] we appended a space before every word, passed them through GPT-2's Small tokenizer and filtered out the words that are not tokenized as a single token, leaving us with a total of 6997 nouns.
2. We start from a list containing every possible three-letter acronym (i.e. `"AAA"`, `"AAB"`, etc.). Again, we pass these acronyms trough the tokenizer and filtered out those that are not tokenized as three separate tokens (i.e. one for each letter), leaving a total of 2740 possibilities.
3. Now, a sample can be built by (i) sampling an acronym of the list obtained at the previous step and (ii) sampling a word for each letter of the acronym. It is important to remark that we performed weighted random sampling of the acronyms according to the frequency of each letter. The reason behind this decision was that, as shown on Figure 1, it is really uncommon to find nouns beginning by certain letters. This implies that the probability of sampling an acronym such as `"QUX"` should be lower than `"BAS"`, hence justifying our weighted random sampling approach.

[1] https://github.com/taikuukaits/SimpleWordlists/blob/master/Wordlist-Nouns-All.txt

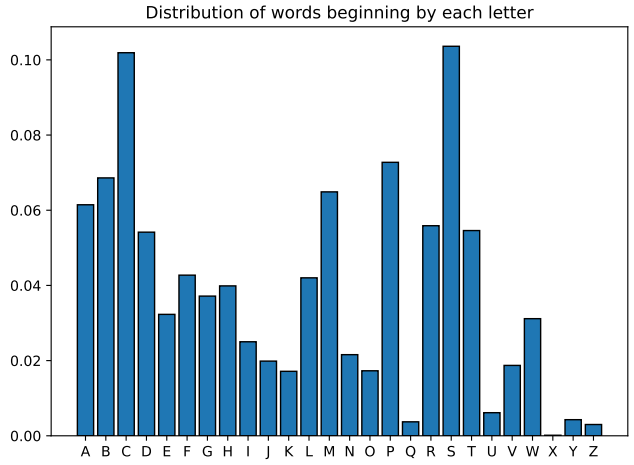

Figure 1: Distribution of words beginning by each letter in the list obtained at step 1.

B Interpreting the circuit

As described in Section 4.2, we identified that heads `8.11`, `9.9` and `10.10` were responsible for the task of predicting the third letter of the acronym, according to the results of the activation patching experiments. In order to provide evidence about what these heads do, we performed an extra experiment.

In summary, we analyzed the attention patterns of these heads to check what are they attending to. Figure 2 shows a distribution of the average attention paid from the last token to the rest, for each of the identified heads. As clearly shown, these heads mainly attend to the third word of the acronym.

C Locating and Understanding Vulnerabilities

Finally, we also include the results presented on Section 4.4 for the letter `S`. Namely, Figure 3 shows the logit attribution obtained on the adversarial samples with the letter `A`, which was the one that showed the second largest Δp. It is shown that the results are almost identical, i.e. head `10.10` is the component that contributes the most to the adversarial samples beginning with `S` to be misclassified.

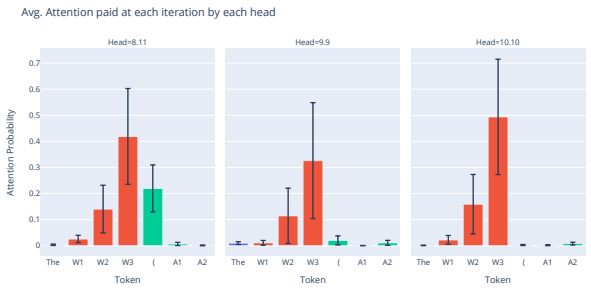

Figure 2: Average attention paid from the last token to the rest for each of the head of the detected circuit

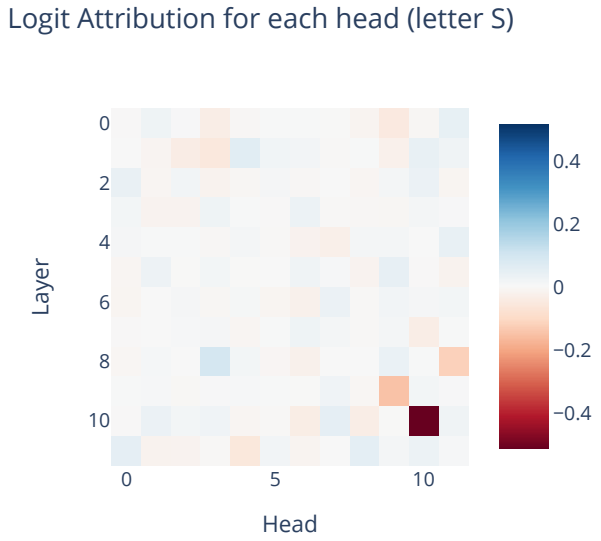

Figure 3: Logit attribution for every attention head on adversarial samples with the letter `S`. This attribution is obtained by projecting into the logit difference direction.

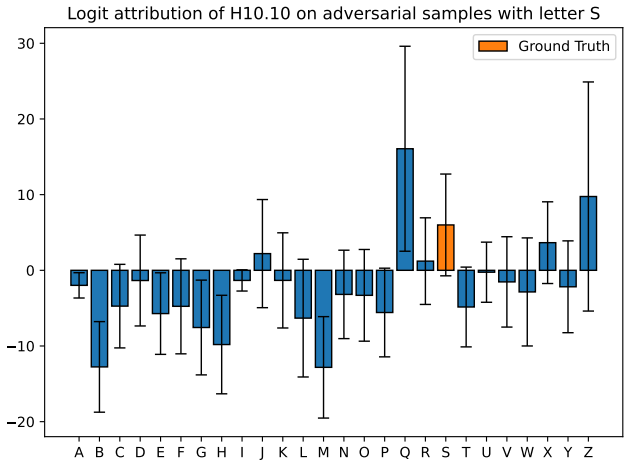

Figure 4: Logit attribution of head `10.10` on adversarial samples with the letter `S`. This attribution is obtained by projecting into the directions of the different capital letters.

Similarly, Figure 4 shows the results obtained by projecting the output of head `10.10` into the directions of the different capital letters for the adversarial samples with the letter `S`. Again, one these adversarial samples, head `10.10` tends to overpredict the letter `S` with the letter `Q`.