Part 5: Attention Analysis & Interpretability

Name: Shachar Sagi 318277381 Adam Celermajer 332638592

Methodology

I modified the transformer to extract attention matrices for all layer-head combinations. For each input, the model produces a stack of attention matrices: one for each layer and each head (i.e., layers × heads matrices). Each matrix is square (sequence length × sequence length) and is triangular and padded to 0. This is because the first token only attends to the first token, the second token attends to the first and second, and so on. For a specific head, attending from each token gives a vector of attention over all previous tokens (since this is a decoder). Each row in the heatmap represents this vector for one token, showing how much that token attends to all tokens before it (and itself). Doing this for every token fills out the heatmap for that head.

In specific tasks, we can aggregate data (e.g., for position-related tasks). Alternatively, we can aggregate based on similar input meanings. For example, when testing capital letters, we ensure the same capital letters are in the same position across all inputs. To aggregate the results for a specific head across multiple samples, I averaged them element-wise. This produced an averaged heatmap that shows the typical attention pattern for that head, smoothing out sample-specific noise and highlighting consistent behaviors.

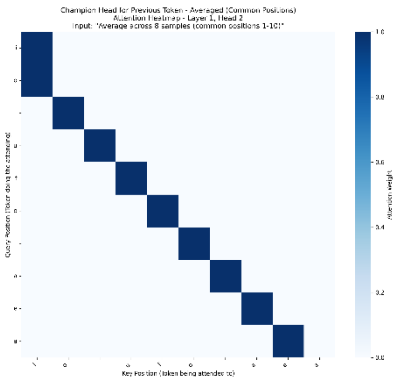
How I Searched for Patterns (Task Descriptions)

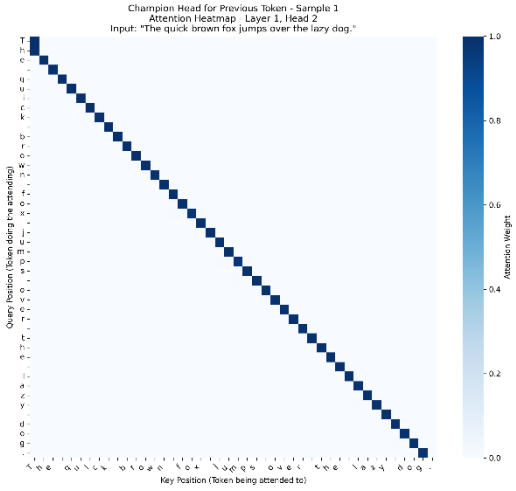
For each input, I obtained a stack of matrices (layers × heads) on each token . Searching for a specific task involved going through these matrices and looking for a specific pattern (e.g., strong attention to the previous token, spaces, or certain suffixes). For each of the following tasks, I wrote code to detect the relevant pattern, selected the best ("champion") head, and generated heatmaps for individual samples. Below is what I checked for each task:

<u>Previous Token</u>

What I Searched: Heads that consistently attend to the immediately preceding character (i-1 for each position i). This is visible as a strong sub-diagonal in the heatmap. As it is normalized, I picked the one with the biggest sub-diagonal sum.

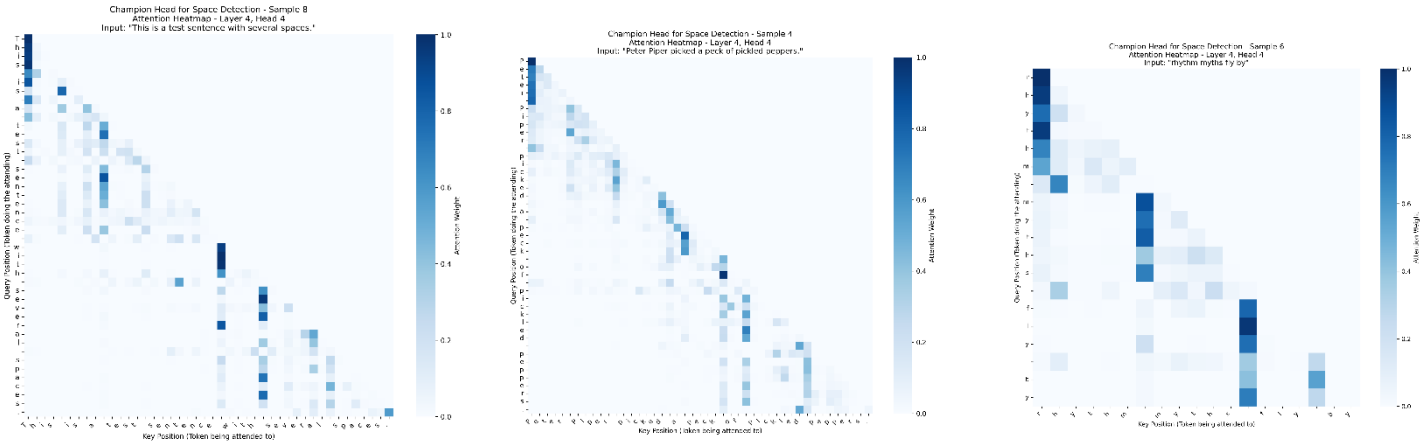
we see clear result on Layer 1 Head 2
Strength: 0.997 (extremely strong) (average activation)
Consistency: 0.999 (nearly perfect)(on different input)





Space Detection

What I Searched: Heads where non-space tokens attend strongly to space tokens. This helps identify word boundaries. I measured average attention from non-space to space positions.

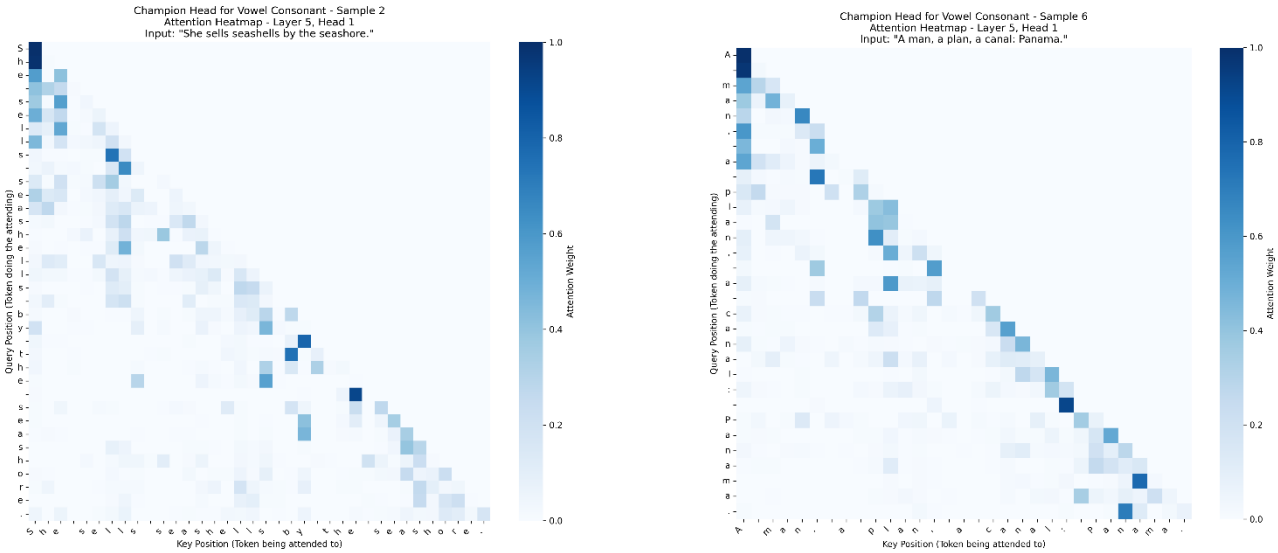


The champion in that case is consistently layer4, head 4 it always respond in a strong way on space token , albeit only in the close last tokens , it also happen in layer4 which is quite advanced in the attention block , I don't think its conclusive enough to determine as a space detection filter but its doing something linked to space

Vowel-Consonant

What I Searched: Heads that treat vowels and consonants differently, e.g., consonants attending to vowels or vice versa. I compared average attention between these groups.

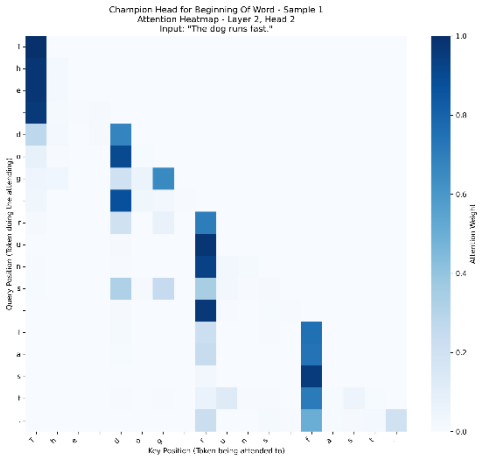
Didn't work at all got most of our result in layers 5 head 1 but not with a apparent vowels importance

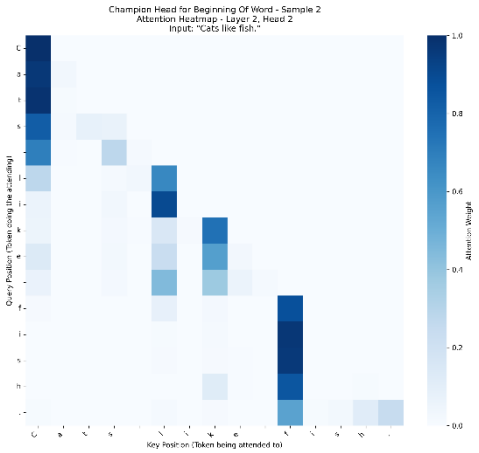


Beginning of Word

What I Searched: Heads that attend to the first character of each word or show increased attention at word starts. I looked for patterns where attention spikes at positions following spaces or at the start of the sequence.

Result: strong winner on layer 2 head 2 , extremely consistent and a clear activation

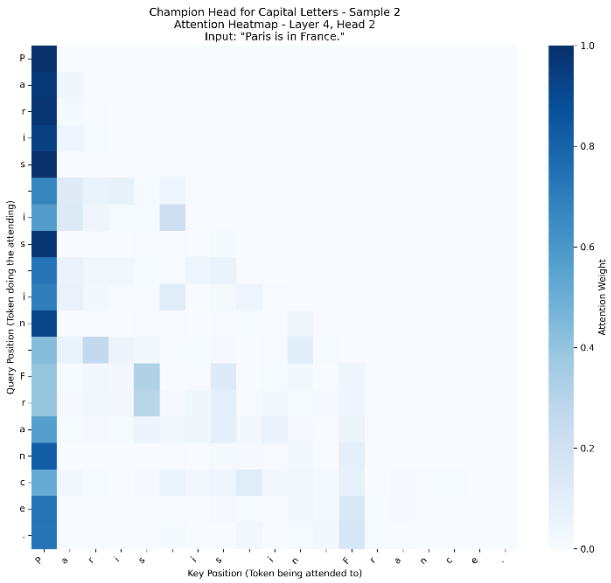


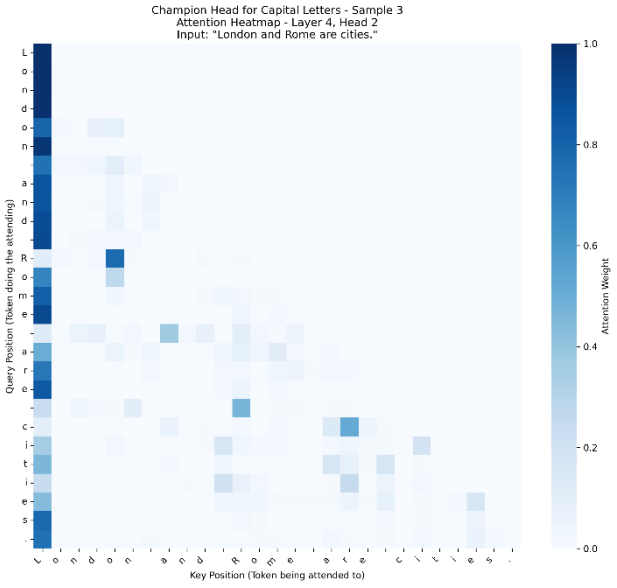


Capital Letters

What I Searched: Heads that attend to or from capital letters, possibly to help with sentence boundaries or proper nouns. I measured attention involving uppercase characters.

Inconclusive result Layer 4 head 2 gave the best result but only with an activation the first letter (which is always Capital)
can be a filter for beginning of sentence maybe :)





We checked other multiple possible task but with again inconclusive result to show **Plural 's' , Possessive 's' ,Past 'ed'.**

Conclusions

The analysis revealed that transformer models naturally develop specialized attention heads for different linguistic tasks. The most common pattern was positional attention, space , capital , previous token , and at the Early layers, suggesting focus on basic character relationships, while later layers develop more complex linguistic specializations.