Unveiling Visual Perception in Language Models: An Attention Head Analysis Approach

Supplementary Material

8. Logic Lens

To investigate why attention weights are not consistent across different datasets, we employed the “logic lens” to examine the rank of decoded tokens across layers. We aimed to determine if the model still reduces its output to the top-1 guess while comparing other distributions to the final one, specifically by analyzing the rank of the final top-1 guess. This method allows us to see how the model’s predictions evolve from the initial layers to the final output. By examining the rank of decoded tokens layer by layer, we can observe whether the model maintains uncertainty about its predictions in the earlier layers and becomes more certain in the later layers. This granularity helps identify at what stage the model’s predictions start to stabilize.

In the Visual General Benchmark, results become more confirmatory toward the very last layers, indicating that the model in the middle layers is still trying to ascertain the correct answer. However, for the other four benchmarks, the final answer converges to the final result quite rapidly in the initial few layers.This observation explains why, in later layers, the model ceases to distribute attention to image tokens, having already confirmed the answer.

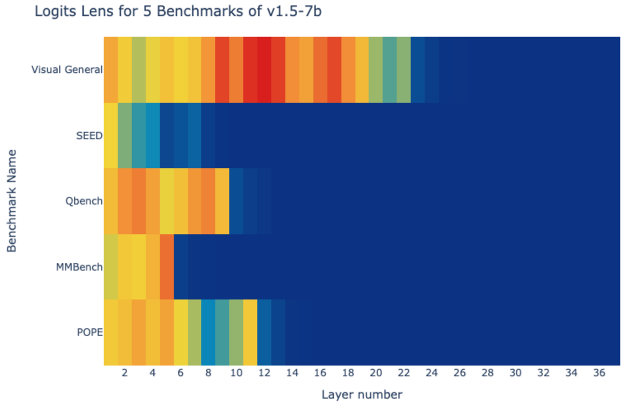

Figure 9. Visualizing Model Behavior Across Datasets: This figure shows different convergence patterns, with blue indicating Rank 1. Particularly in the Visual General Benchmark, the model requires more time to ascertain the correct answer, reflecting a need for *Look Twice* In contrast, some datasets show a swift stabilization to the final answer, which then persists across layers, showcasing varied strategic adjustments by the model.”

9. Masking out Late heads

We observed that, unlike early and middle heads, the later heads have little impact on the final accuracy. Accordingly, we conducted experiments on masking varying numbers of heads across two models, as depicted in the figures below.

From the figures, we observe a consistent trend across both versions 1.5-7b and 1.6-7b: as more heads are masked, there is a decline in accuracy across all three datasets. Notably, even with 100 heads masked, the random masking method still retains relatively high accuracy compared to the baseline model. This finding has significant implications for enhancing the efficiency of key-value (KV) cache and inference processes. Given that image tokens dominate the input (576 out of 650 tokens), it may be feasible to significantly prune the KV cache corresponding to non-essential heads. By computing only for essential heads, we could achieve substantial gains in inference speed.

Furthermore, we noted that the model displays a consistent trend between versions, though the extent of accuracy reduction varies. For example, both MMbench and Qbench exhibit a more noticeable decline in accuracy with Top masking in version 1.6 compared to 1.5, suggesting that version 1.6 is more robust. Additionally, we observed that the rate of decay in accuracy becomes less steep as more of the top heads are masked, indicating that the remaining heads contribute less significantly to the final accuracy.

10. Comparative Analysis of Attention

In our analysis, we observed that the 13B model exhibits significant activation predominantly in the middle layers, a pattern that is similarly noted in the 7B model. In contrast, the Phi3 3.8B model, which is the smallest among the models we studied, shows pronounced activation in the later layers. We hypothesize that this distinct behavior in the 3.8B model arises from its limited capacity.

Larger models, such as the 13B and 7B, generally have more parameters and potentially more layers, allowing for a more balanced distribution of the computational load across the network. This distribution often results in substantial activation in the middle layers, where complex feature interactions and transformations occur. In contrast, smaller models like the 3.8B, with fewer parameters and layers, require a more concentrated utilization of available resources. Consequently, the later layers in smaller models may need to undertake more intensive processing tasks compared to their counterparts in larger models.

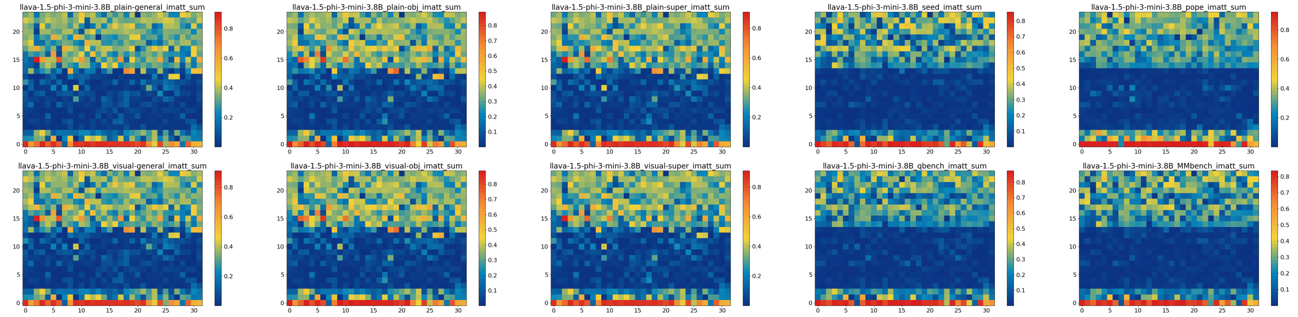

Figure 10. Summation of attention weights for the LLaVA-1.5 Phi-3 Mini 3.8B model, showing a consistent activation in the middle and late layers across various datasets.

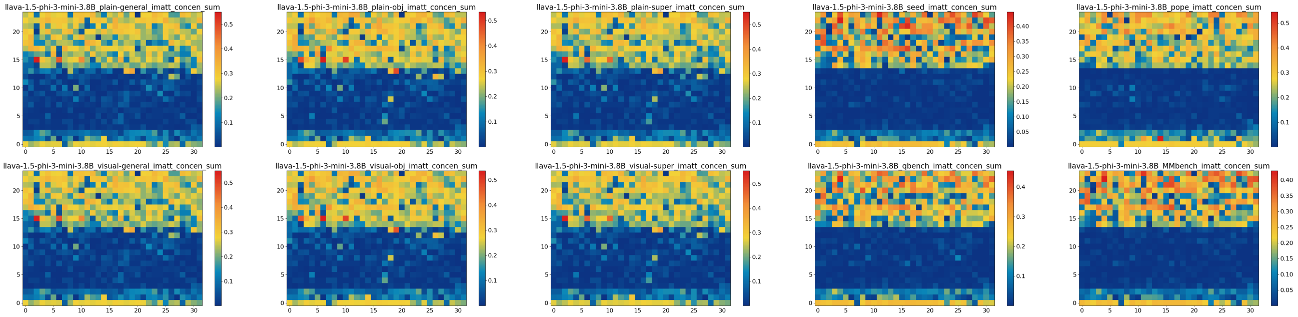

Figure 11. Concentration of attention weights sum for the LLaVA-1.5 Phi-3 Mini 3.8B model, illustrating that middle and late layers are predominantly activated across datasets.

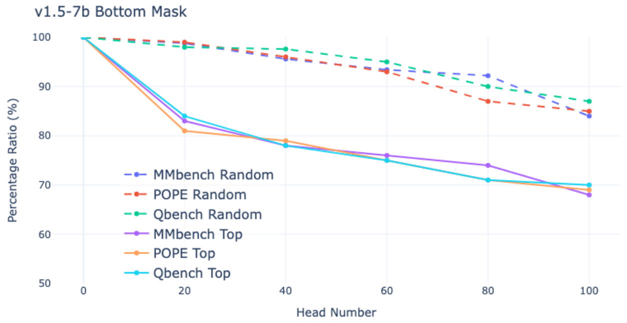

Figure 12. This graph displays the performance degradation of v1.5-7b across three different models (MMbench, POPE, Qbench) under two masking strategies (Random and Top) as a function of the percentage of heads masked. The dashed lines represent the random masking strategy, while the solid lines denote the top masking strategy.

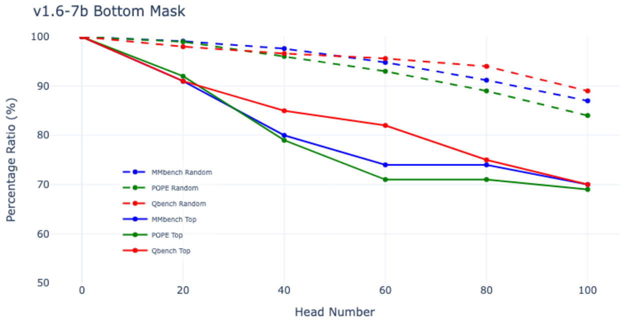

Figure 13. Similar to Figure 12, this graph presents the impact on v1.6-7b accuracy for the same three models under the same two masking strategies as the masking extent increases. Colored lines differentiate the models, with blue for MMbench, green for POPE, and red for Qbench.

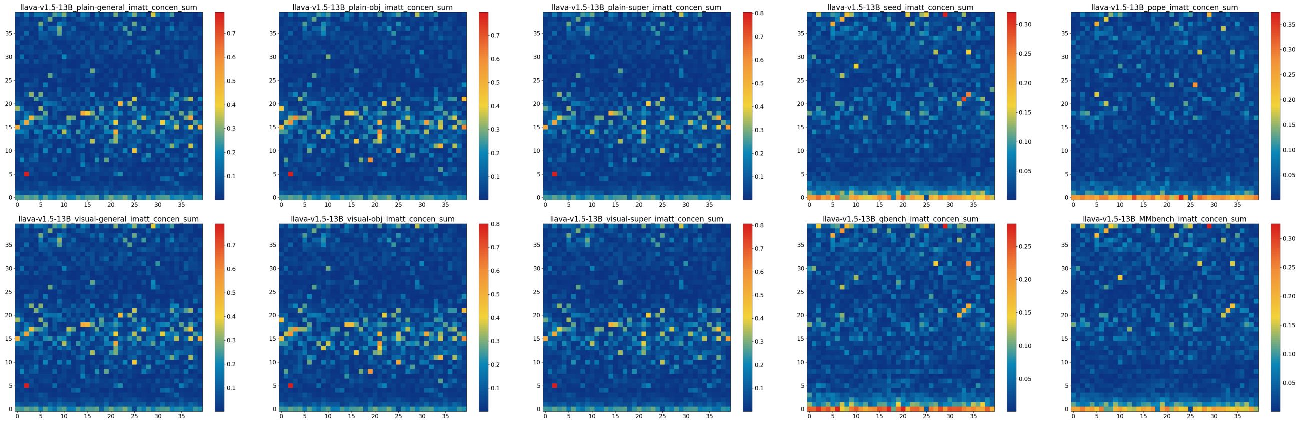

Figure 14. Concentration times sum of attention weights in the LLaVA v1.5 13B model, indicating that the middle layers are significantly activated across all datasets examined.

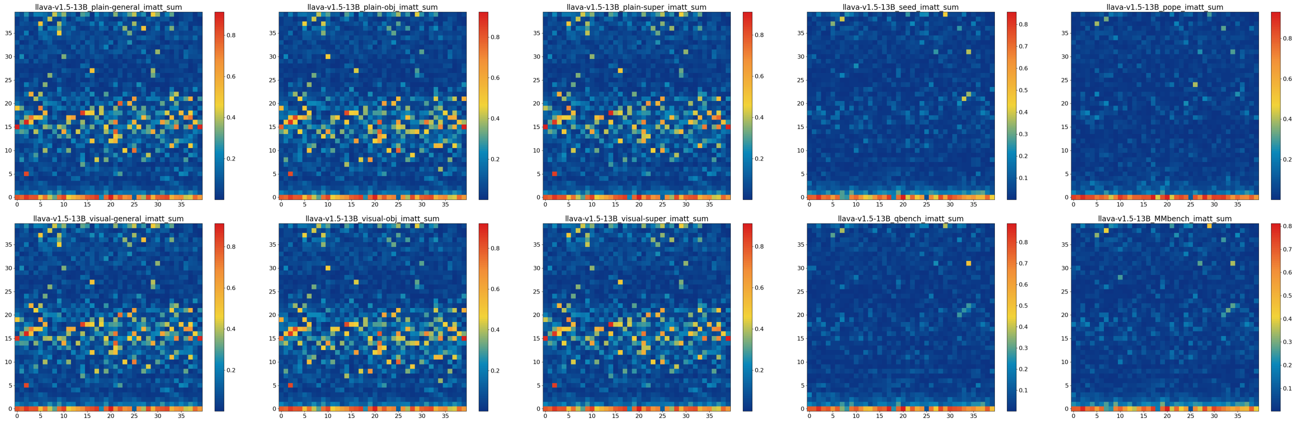

Figure 15. Total summation of attention weights across layers in the LLaVA v1.5 13B model. Unlike the smaller 3.8B model, this larger model shows significant activation in middle layers, highlighting a unique pattern where these layers play a more crucial role, particularly in this model configuration.