

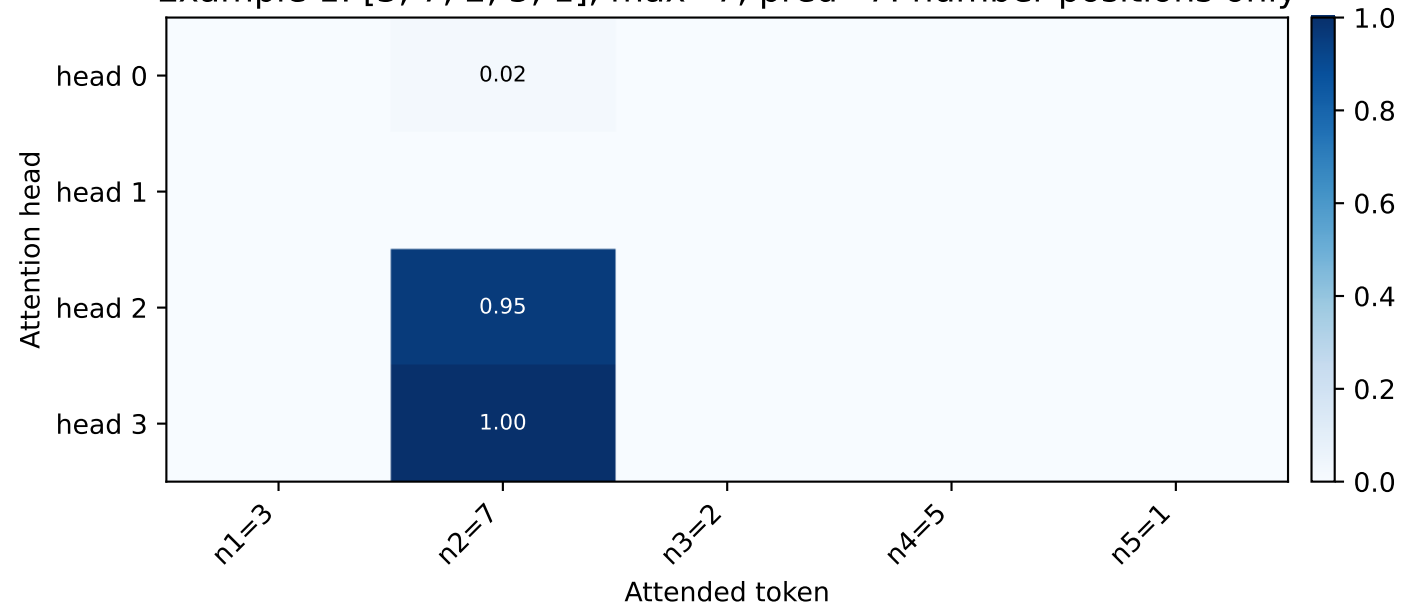
`[ANS]` Attention Pattern Report

Puzzle 1a uses a one-layer attention-only transformer to predict the max of five digit tokens.

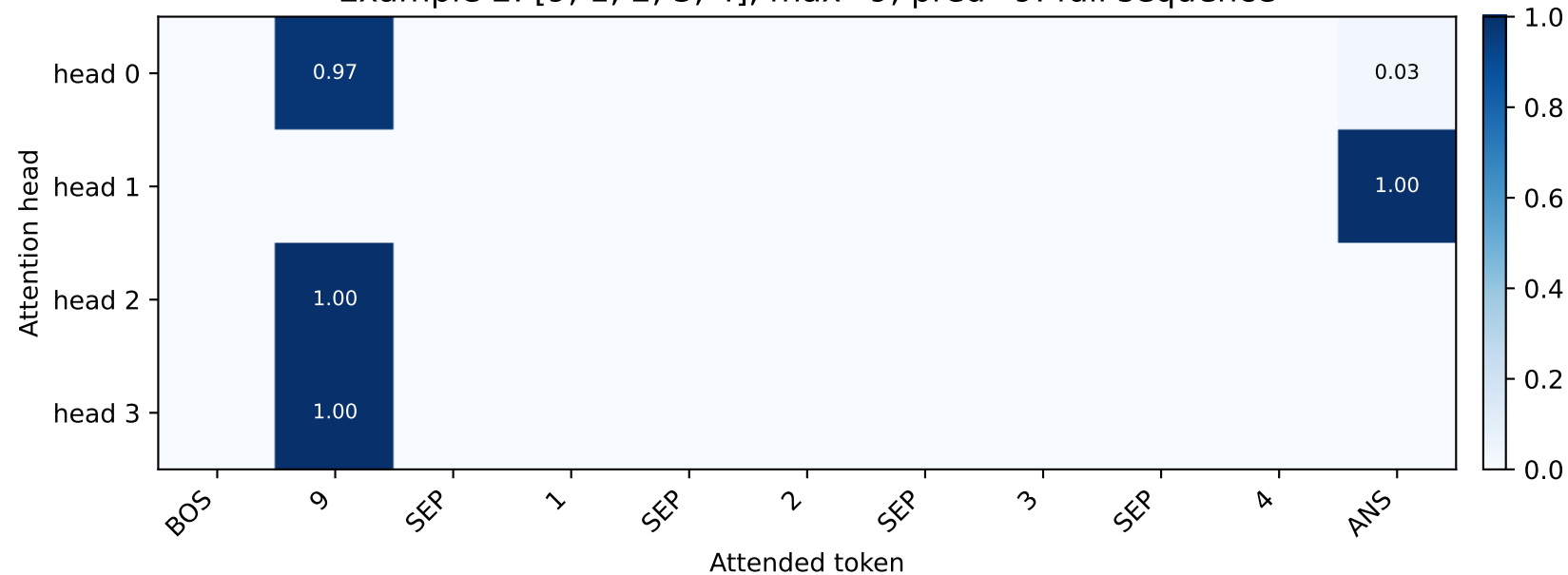
This report asks where each attention head looks from the final `[ANS]` position.

The aggregate results use a fixed-seed stratified sample, not the full 100,000-input space.

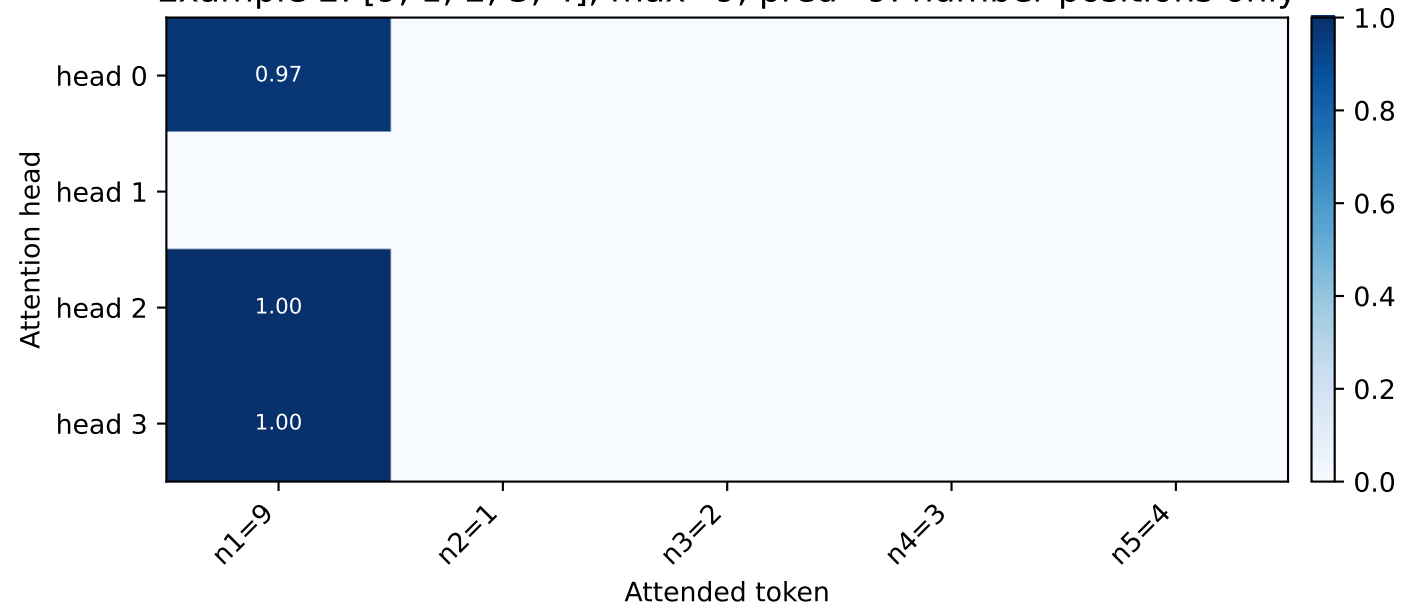
Example 1: [3, 7, 2, 5, 1], max=7, pred=7: number positions only



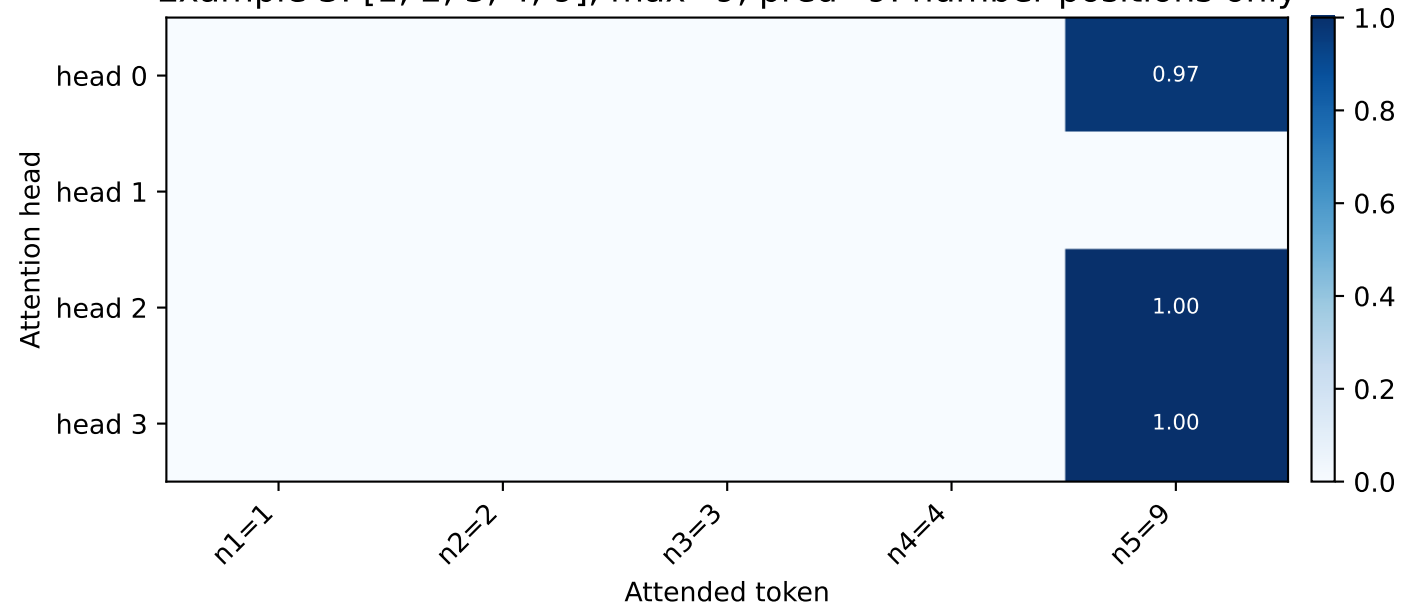
Example 2: [9, 1, 2, 3, 4], max=9, pred=9: full sequence



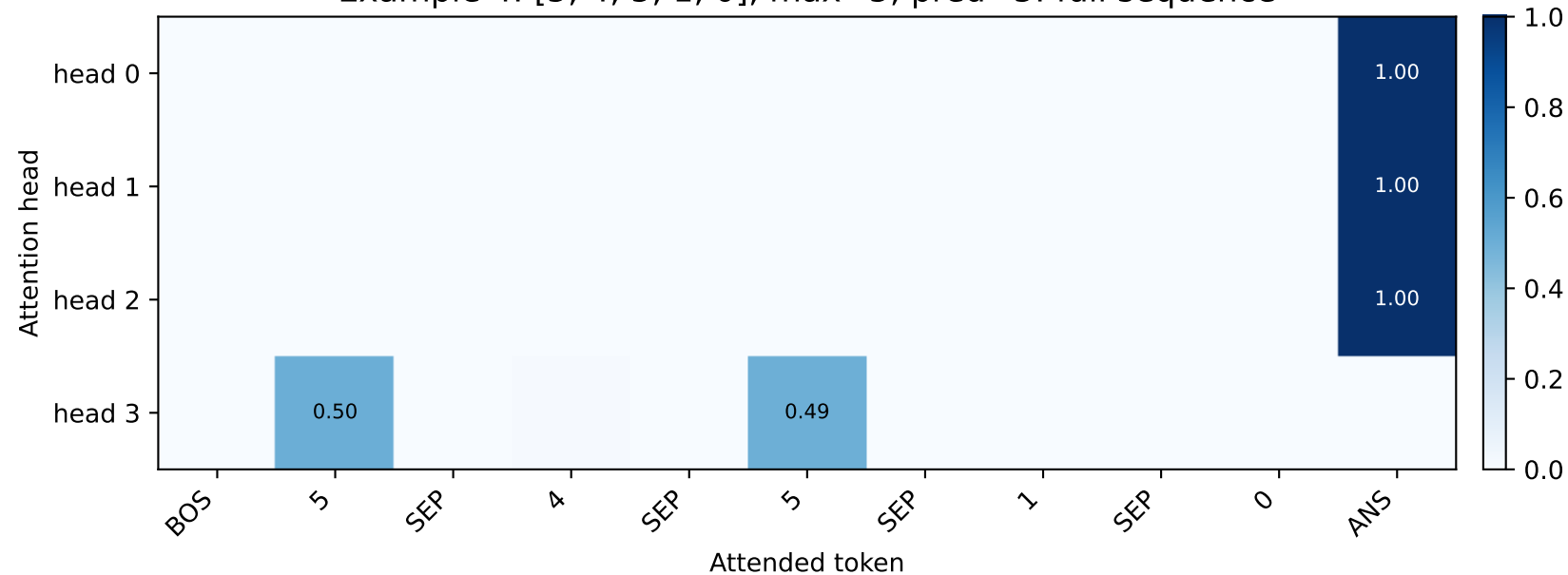
Example 2: [9, 1, 2, 3, 4], max=9, pred=9: number positions only



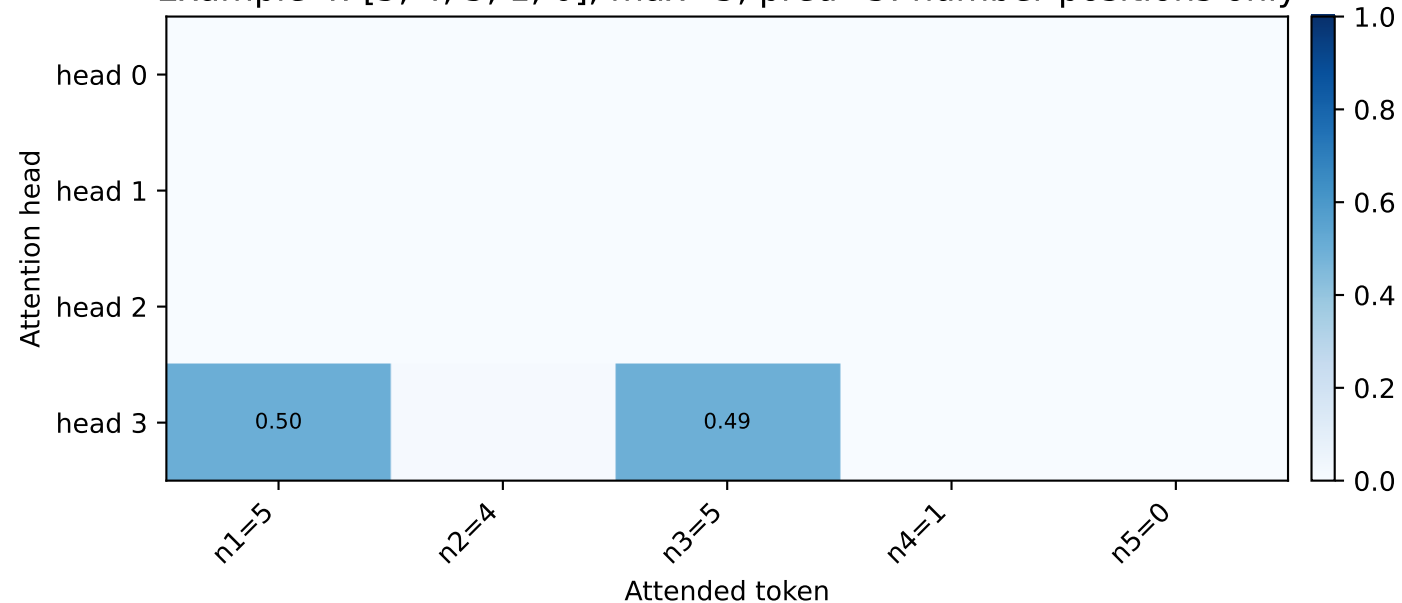
Example 3: [1, 2, 3, 4, 9], max=9, pred=9: number positions only



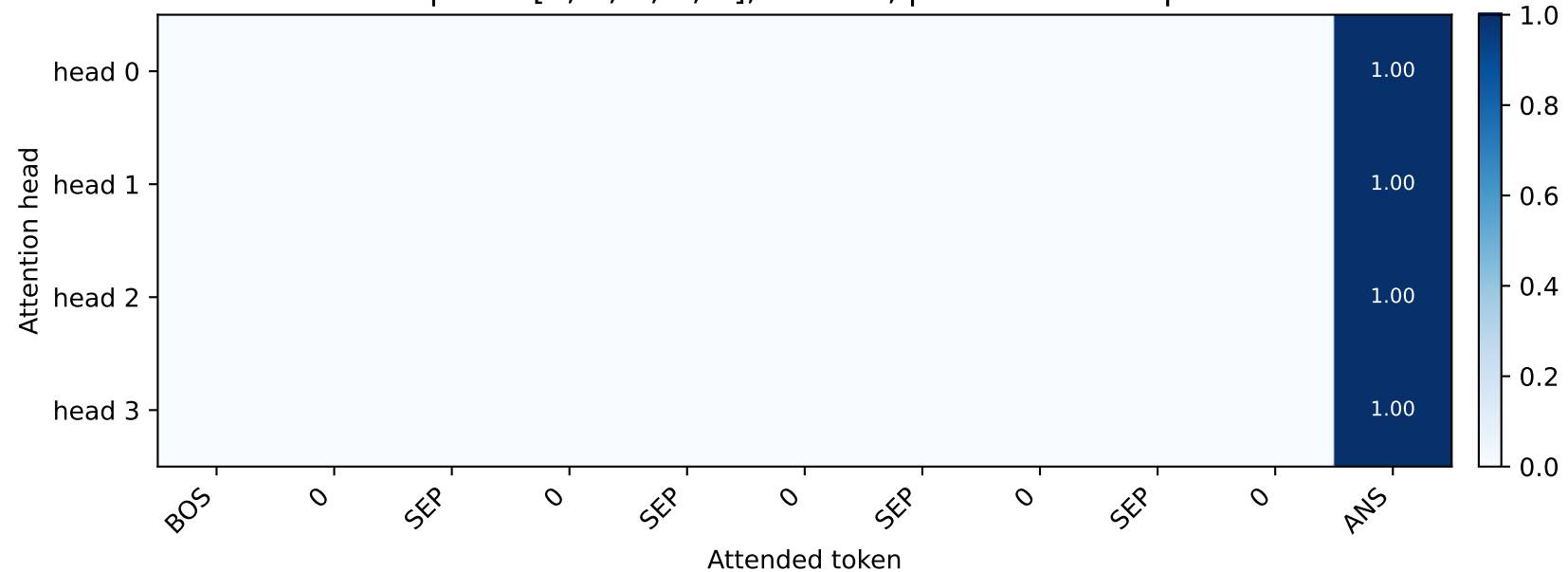
Example 4: [5, 4, 5, 1, 0], max=5, pred=5: full sequence



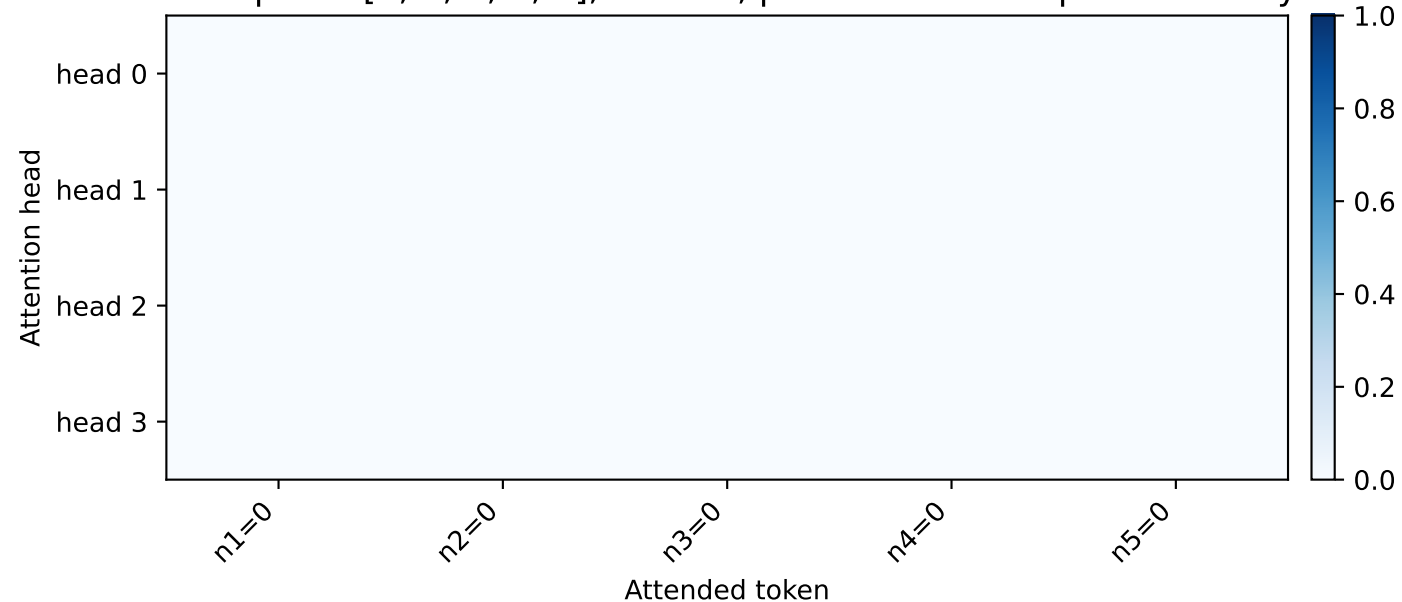
Example 4: [5, 4, 5, 1, 0], max=5, pred=5: number positions only



Example 5: [0, 0, 0, 0, 0], max=0, pred=0: full sequence



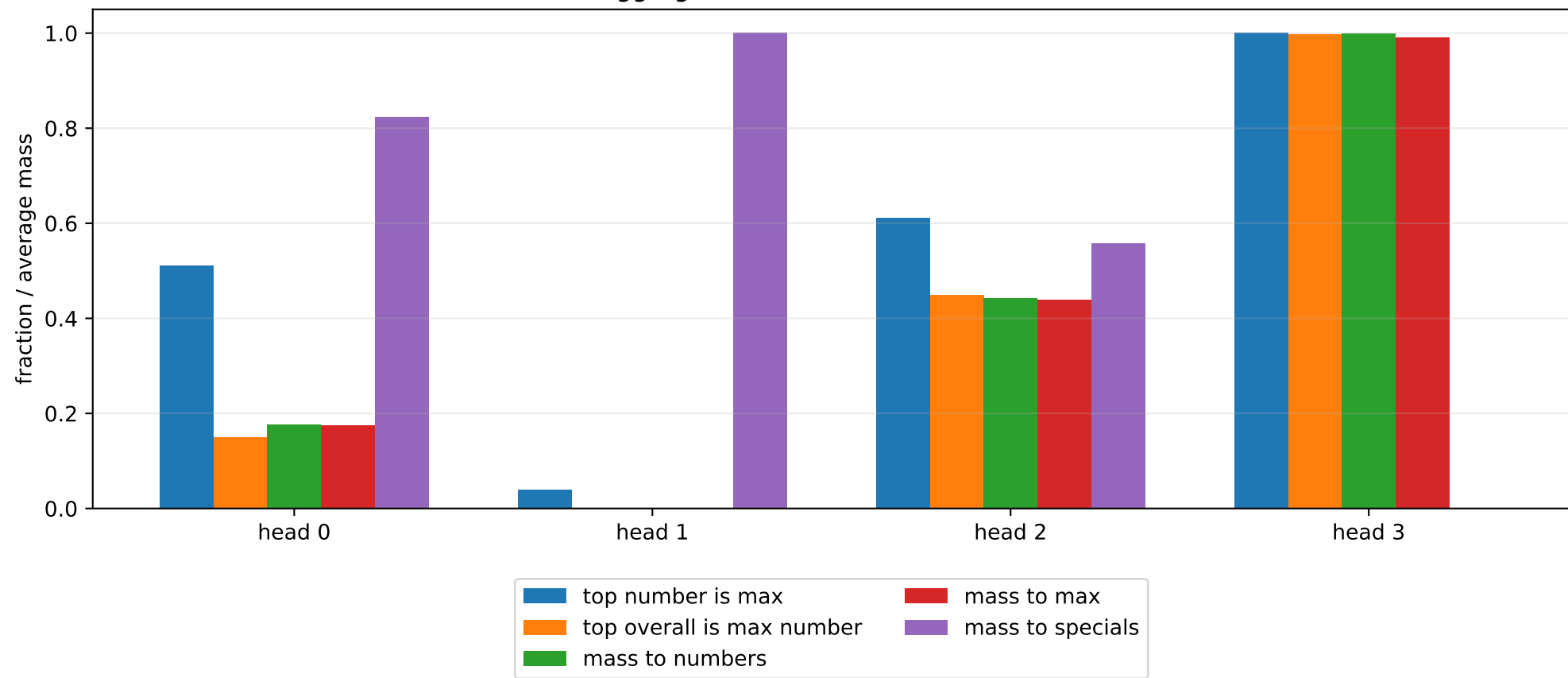
Example 5: [0, 0, 0, 0, 0], max=0, pred=0: number positions only



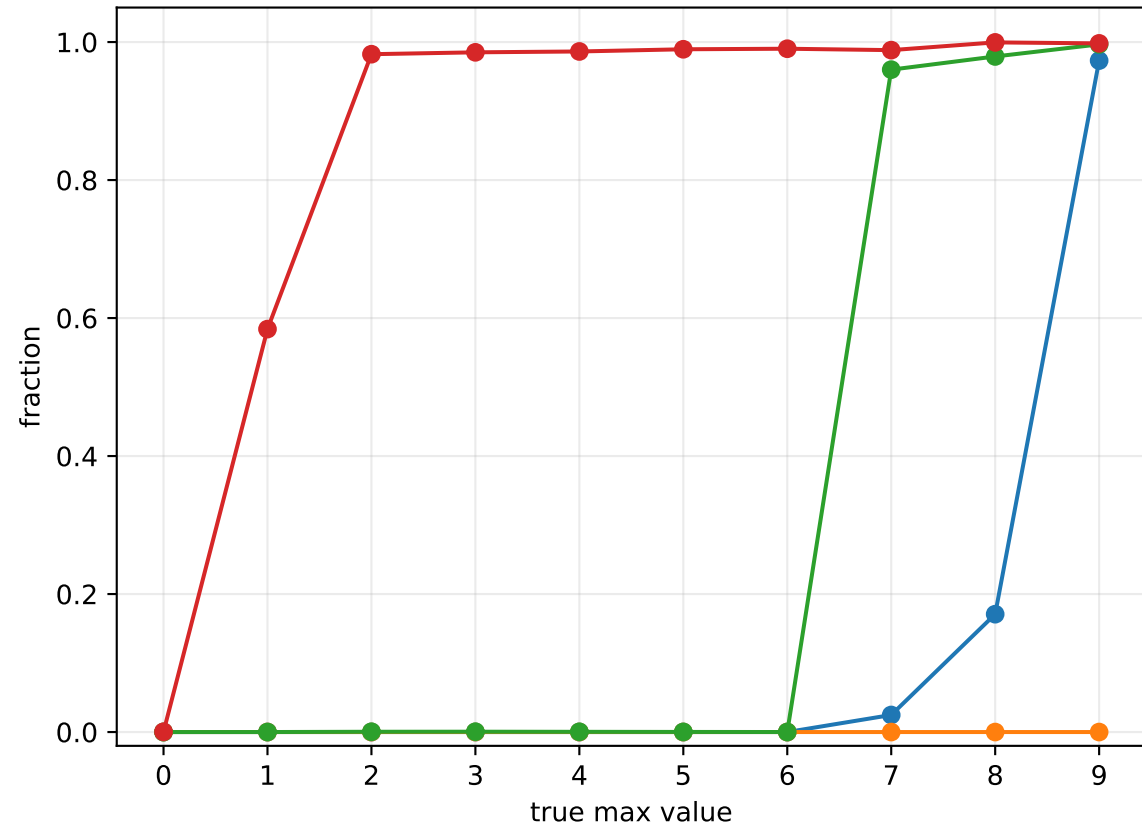
Sample summary by head

head	top num is max	top overall is num	top overall is max num	mass to nums	mass to max	mass to specials
0	0.5107	0.1496	0.1496	0.1763	0.1748	0.8237
1	0.0388	0.0000	0.0000	0.0000	0.0000	1.0000
2	0.6108	0.4488	0.4488	0.4429	0.4393	0.5571
3	1.0000	0.9968	0.9968	0.9984	0.9899	0.0016

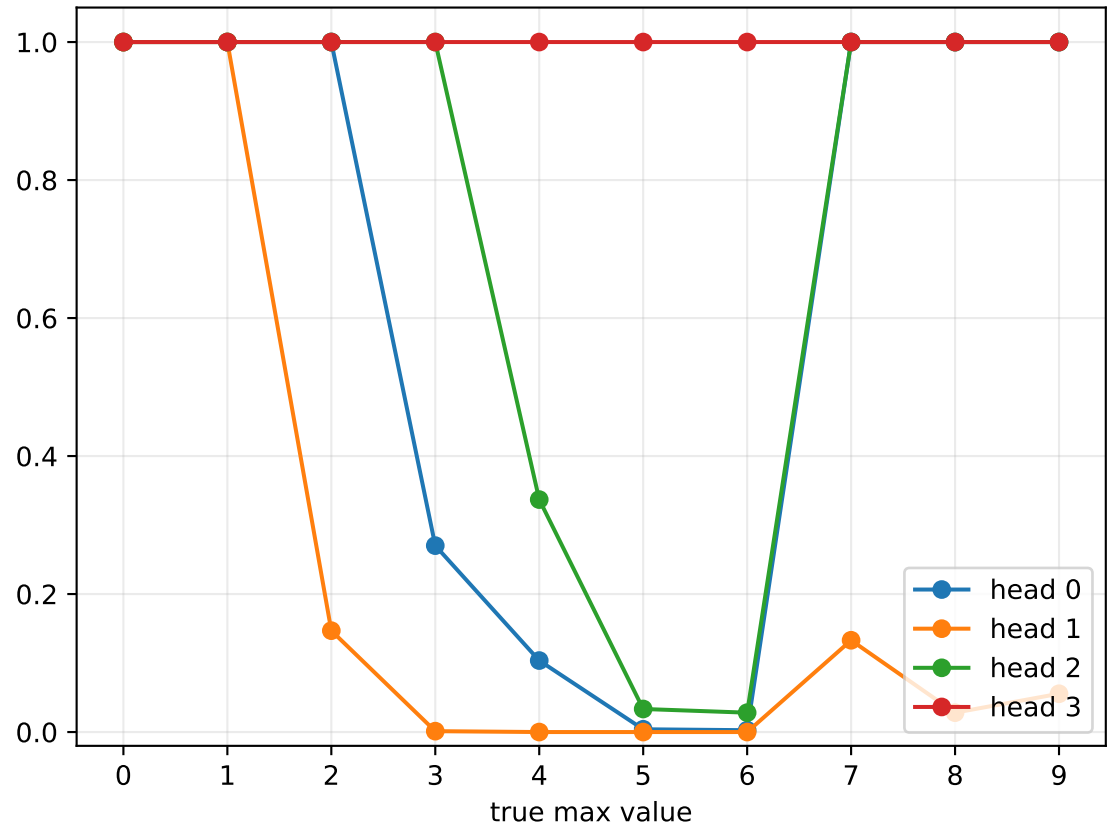
Aggregate `[ANS]` attention metrics



Average attention mass to max positions



Top number-position attention is max



Conclusions

On this stratified sample of 10000 inputs, model accuracy is 1.0000. The sample contains 6939 unique-max inputs and 3061 tie inputs.

Head 3 is the strongest `[ANS]` max selector by top number-position attention: 1.0000 of examples choose a max digit.

Head 3 sends the most average attention mass to max positions: 0.9899.

Head 1 sends the least mass to number positions (0.0000), so it is not acting as a direct max selector from `[ANS]`.

Interpretation for the next step: use these attention results to prioritize OV/logit analysis for the selector-like heads, especially the head with near-perfect max selection.