We conduct small-scale experiments for Llama-3.1-1B-Instruct, mt0-base, and xlm-r-base to study whether the matrix language impacts the findings. Specifically, instead of using en as the matrix language, we set the matrix language to vi and fr. The results in Figure 1 show that consistency patterns are similar in all settings, demonstrating the robustness of our experimental design that the matrix language is not a signiciant confounding factor.

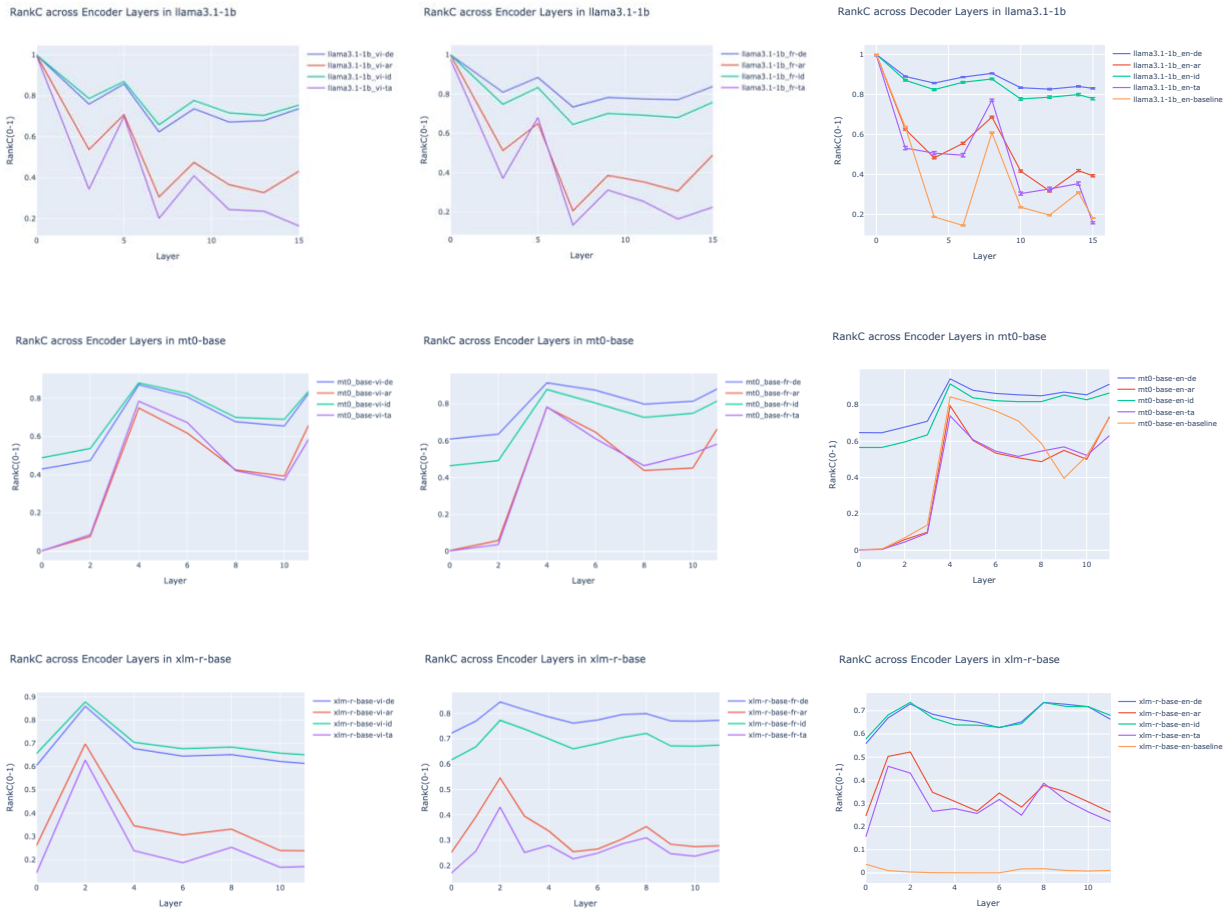

Figure 1: Layer-wise Consistency. Models from 1st row to 3rd row: llama3.1-1b-Instruct, mt0-base, and xlm-r-base. Matrix/pivot languages from 1st col to 3rd col: vi, fr, and en.