Analysis: "Lost in the Middle" Phenomenon

To analyze the "lost in the middle" phenomenon in our LLM-based annotations, we conducted an additional experiment based on the reviewer's suggestion. For a given user, we divided the data into time segments and calculated inter-annotator agreement (IAA) using Cohen's Kappa scores across different models and settings. The data was segmented based on the submission_id, author_key_name, and stance_id_timestamp. For each group (i.e., each combination of submission_id and author_key_name), the timestamps were divided into equal segments. The number of entries for each group was divided by the desired number of segments (3), and the division was done as evenly as possible, with each segment containing a roughly equal number of time-stamped entries. Figure 1 in the rebuttal PDF reports the comparison statistics of IAA scores for the stance detection task across initial, middle, and later time stamps.

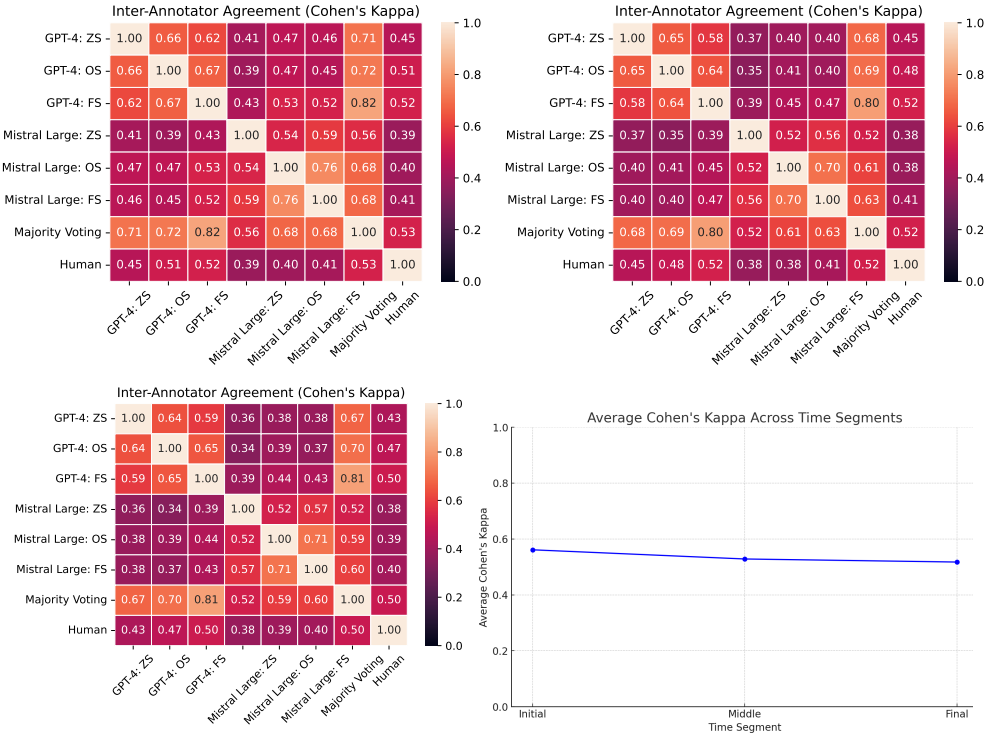

Figure 1: The inter-annotator agreement (IAA) on the USDC test dataset was measured using Cohen's Kappa score across three segments: initial, middle, and later time stamps. The top two rows represent the initial and middle time stamps, while the bottom left corresponds to the later time stamp. The bottom right reports the average Kappa score across all time segments.

Analysis: "Recency Bias" Phenomenon

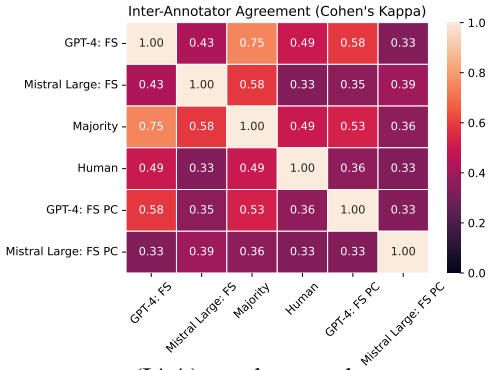

Figure 2: Inter-annotator agreement (IAA) on the test dataset was calculated for both the full conversations and the prior context for a given user. In this context, "GPT-4 FS PC" and "Mistral Large: FS PC" refer to the annotations based on prior context.