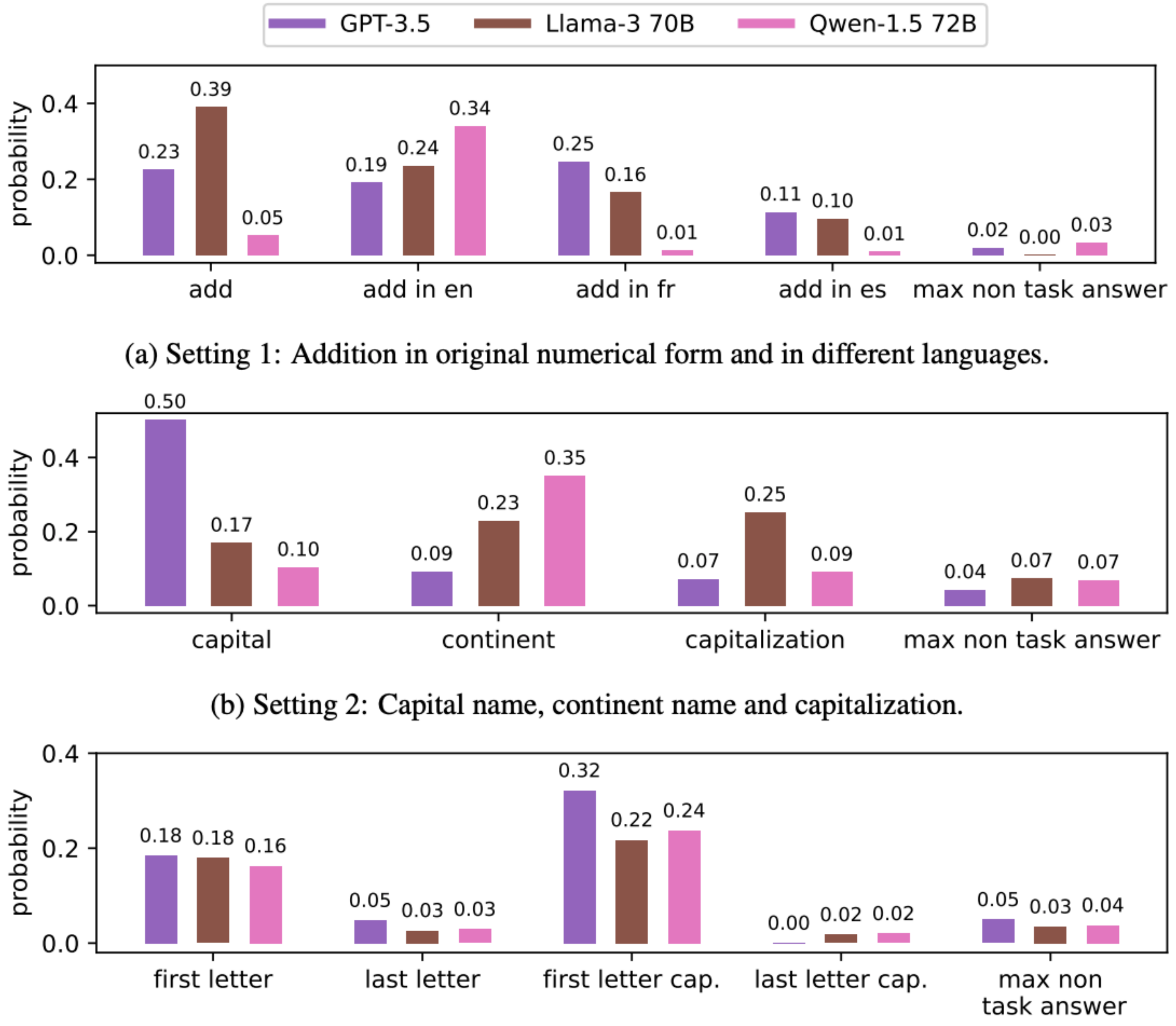

(a) Setting 1: Addition in original numerical form and in different languages.

(b) Setting 2: Capital name, continent name and capitalization.

(c) Setting 4: First or last letter in upper or lower cases.

Figure 8: For each subplot, we plot the medians of probabilities for each task answer and the median of the maximum probability of the answer that is not one of the task answers.

In Figure 8a (setting 1), for GPT-3.5 and Llama-3, we can observe that the medians of the probabilities for most probable non-task-answer are significantly lower than that of each task answer. This indicates that the models are effectively performing task superposition across the four tasks, with any other