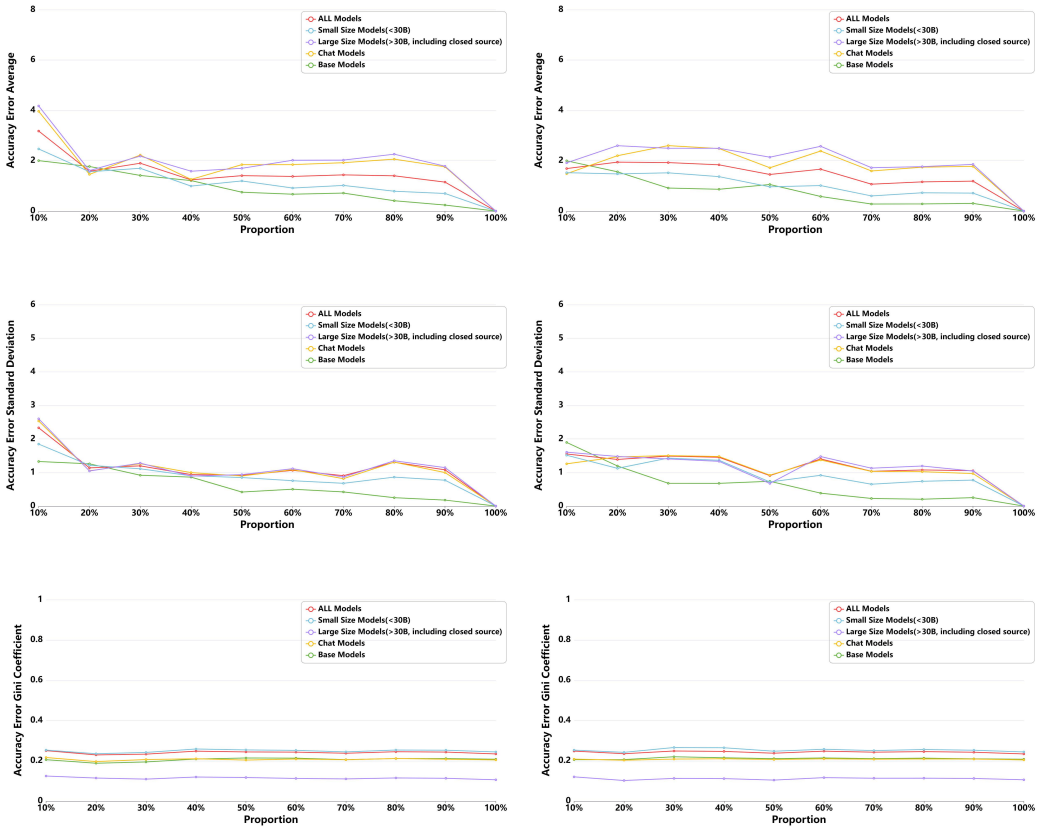

Figure 12: Ablation study on dataset size. The x-axis represents the proportion of the subset size relative to the entire dataset, while the y-axis in the three rows represents (from top to bottom): the mean error of model scores for the subset compared to the full dataset, the standard deviation of the error, and the Gini coefficient of model scores for the subset. The left column employs a rule-based sampling strategy, selecting a specified proportion of questions from 10 questions under each rule to maintain the dataset's original diversity. The right column uses a category-based sampling strategy, randomly selecting a specified proportion from all 250 questions in each category to mitigate differences in difficulty across rules. The "ALL Models" curve reflects metric changes for all models in Table 2, while other curves correspond to metrics for subsets of models categorized from the full set. The results show that, regardless of the sampling strategy, both the mean and standard deviation of errors remain small, stabilizing at around 2 once the dataset proportion reaches 20% of the full set. Similarly, the Gini coefficient exhibits minimal fluctuation, with a maximum variation of approximately 0.02.