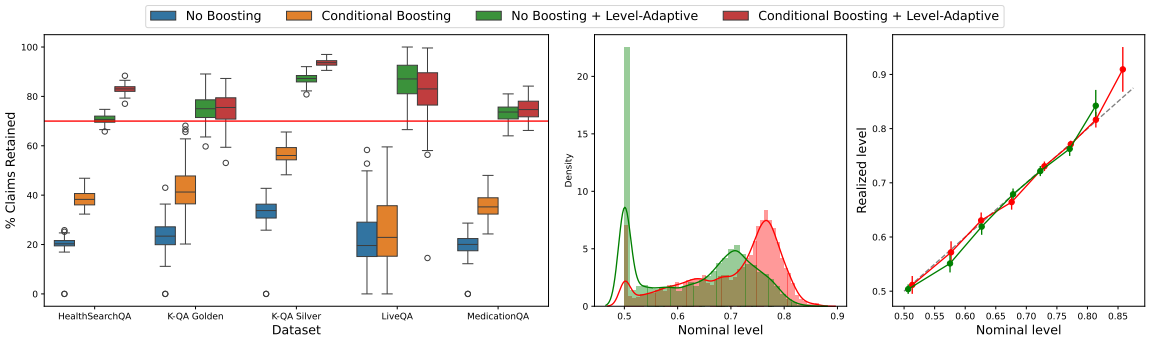

Figure 1. The left panel of the figure shows the percentage of claims retained by various methods on the **MedLFQA** benchmark of Jeong et al. (2024). This benchmark contains many datasets each of which is displayed on the x-axis. Before boosting, we define our claim score by an equally-weighted ensemble of previously published claim scoring functions. We run 100 trials in which we resample a boosting/$\alpha(\cdot)$-estimation split ($n = 1441$), calibration split ($n = 2354$), and test split ($n = 500$). We plot 4 filtering methods: the fixed-level ($1 - \alpha = 0.9$) method that guarantees conditional validity over the function class given by dataset indicators and prompt metadata (prompt length, response length, as well as the mean and standard deviation of the "log probability" claim score over the output) (blue), the same conditionally valid method with adaptive level $\alpha(X_{n+1})$ (orange), the conditionally valid method with boosted scores (green), and a combination method that incorporates both conditional-boosting and level-adaptive CP (red). All methods are set-up to ensure that the final output contains **0** false claims. The middle panel shows that boosting allows the level-adaptive procedure to issue guarantees with higher confidence. The right panel verifies that the reported nominal levels $1 - \alpha(X_{n+1})$ match the empirical error frequencies over equi-spaced bins of width 0.05; these bin indicators are included in the function class used in the level-adaptive procedure.

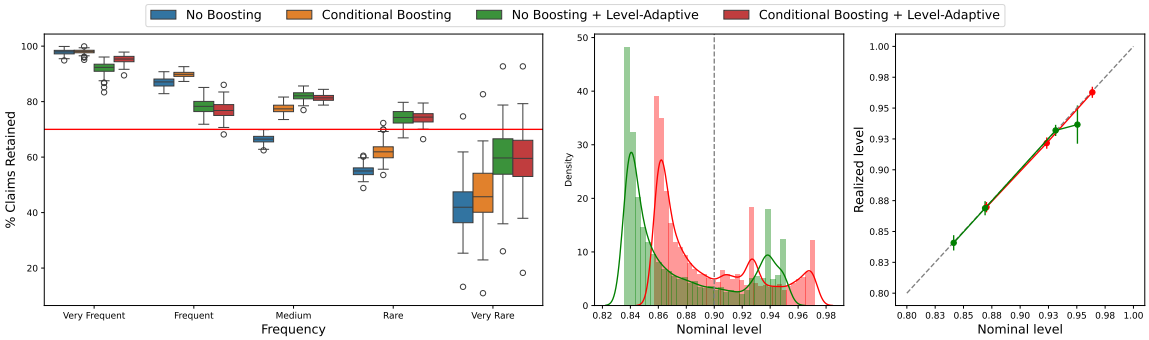

Figure 2. This figure replicates the previous ablation study for the **FActscore** dataset of Min et al. (2023). For the sake of brevity, we only describe the differences compared to Fig. 1. The frequency of each biography topic varies; the displayed groups correspond to Wikipedia view counts in Jan. 2023 that are binned into the intervals $[1000000, \infty), [100000, 1000000), [1000, 100000), [100, 1000), [0, 100)$. The function class used for calibration is defined by these bin indicators and the view count of each Wikipedia article. We run 100 trials in which we resample a boosting/$\alpha(\cdot)$-estimation split ($n = 847$), calibration split ($n = 5338$), and test split ($n = 500$). All methods are set-up to ensure that (with high probability) the final output contains no more than 3 false claims.

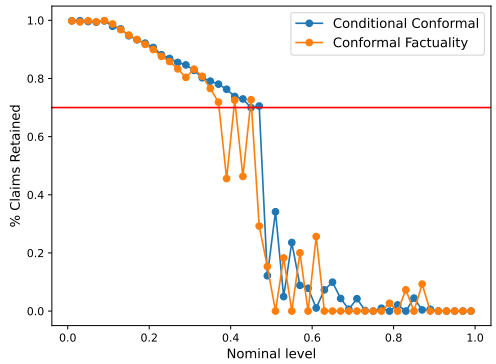

Figure 3. Comparing claim retention at various choices of $1 - \alpha$ for both the conformal factuality baseline of Mohri and Hashimoto (2024) and an improvement of this method to ensure conditional validity over the function class given in Fig 2. Here, the filtering methods guarantee that with probability $1 - \alpha$, the final output contains **0** false claims. We see that most claims must be filtered at all nominal levels exceeding 0.5. Controlling this measure of error while simultaneously preserving most claims is clearly incompatible with a "high-probability" guarantee.