



Figure 1. Percentage of responses that agree with original human coder across *overall* and *unique* test sets. Supervised models trained on 1000 cases from BESIP waves 21–23.