APPENDIX

Extended Error Analysis

Figures 7-9 present the performance by the number of gold events per note for a given event type (referred to as *event density*): one (1), two (2), and three or more (3+) events per note. These figures also includes the total number of gold events (✚).

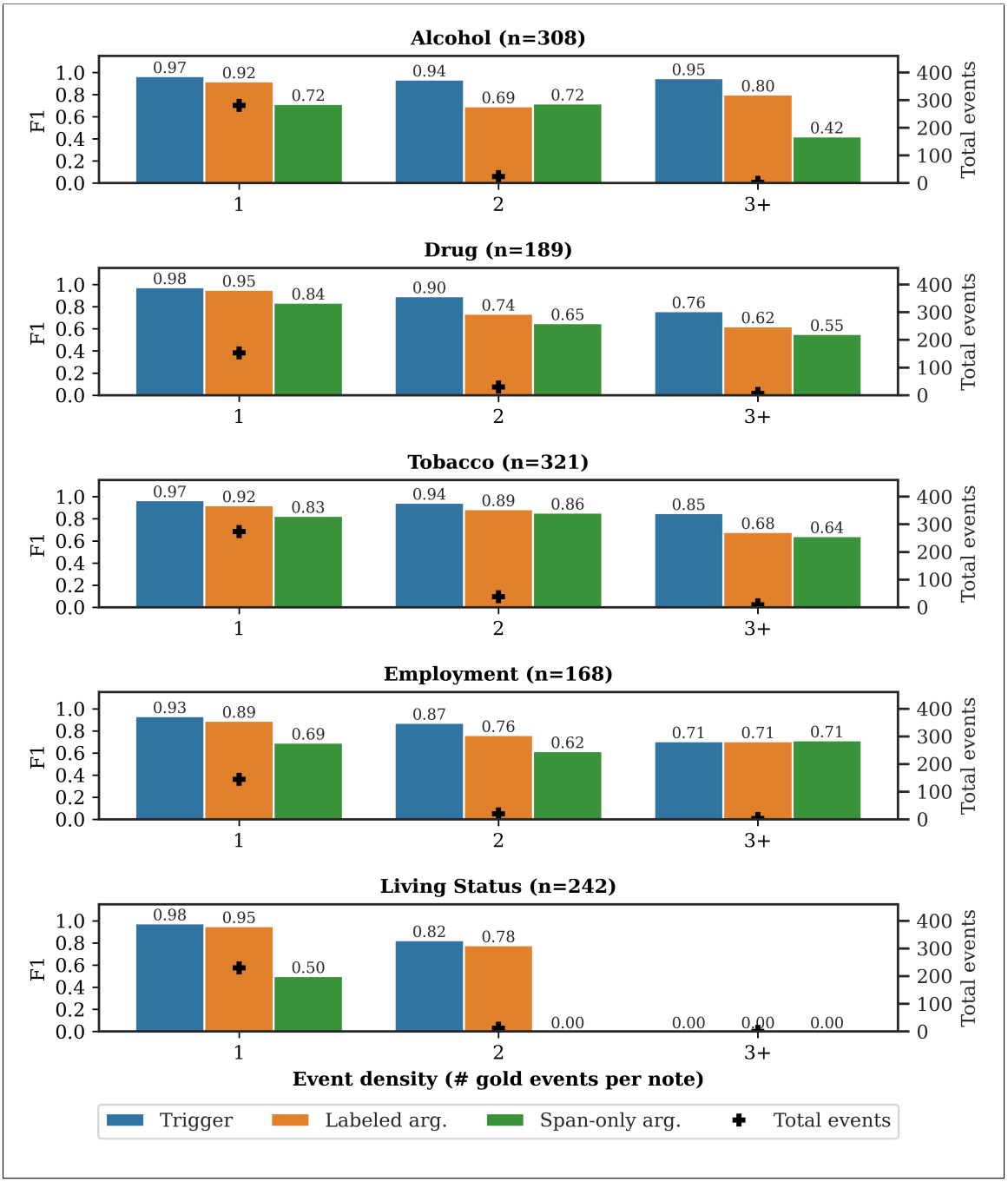

Figure 7: Performance breakdown by event density for Subtask A ($\mathcal{D}_{test}^{mimic}$). The left-hand y-axis is the micro-averaged F1 for triggers, labeled arguments, and span-only arguments (vertical bars). The right-hand y-axis is the total number of gold events (✚).

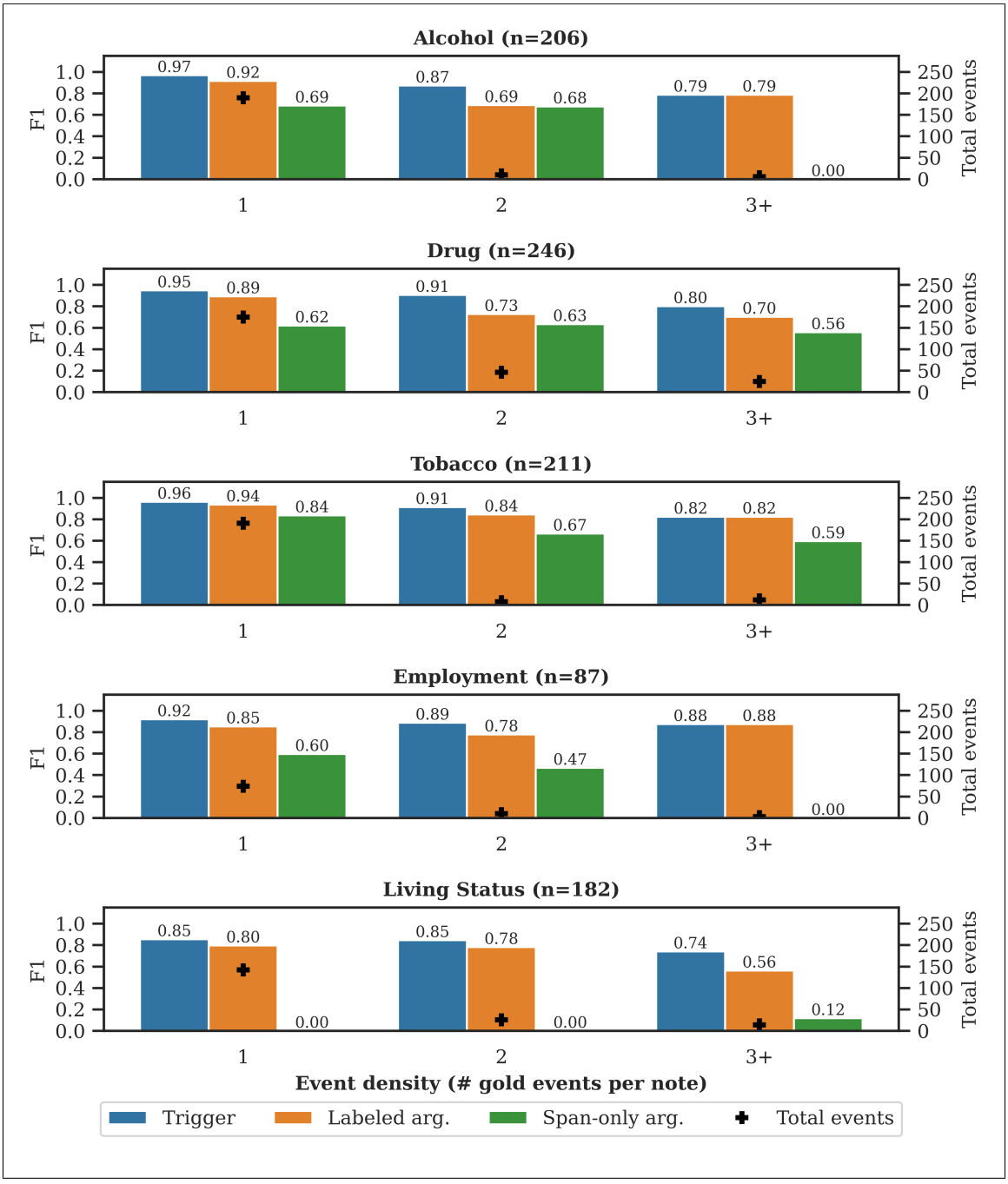

Figure 8: Performance breakdown by event density for Subtask B ($\mathcal{D}^{uw}_{dev}$). The left-hand y-axis is the micro-averaged F1 for triggers, labeled arguments, and span-only arguments (vertical bars). The right-hand y-axis is the total number of gold events (✚).

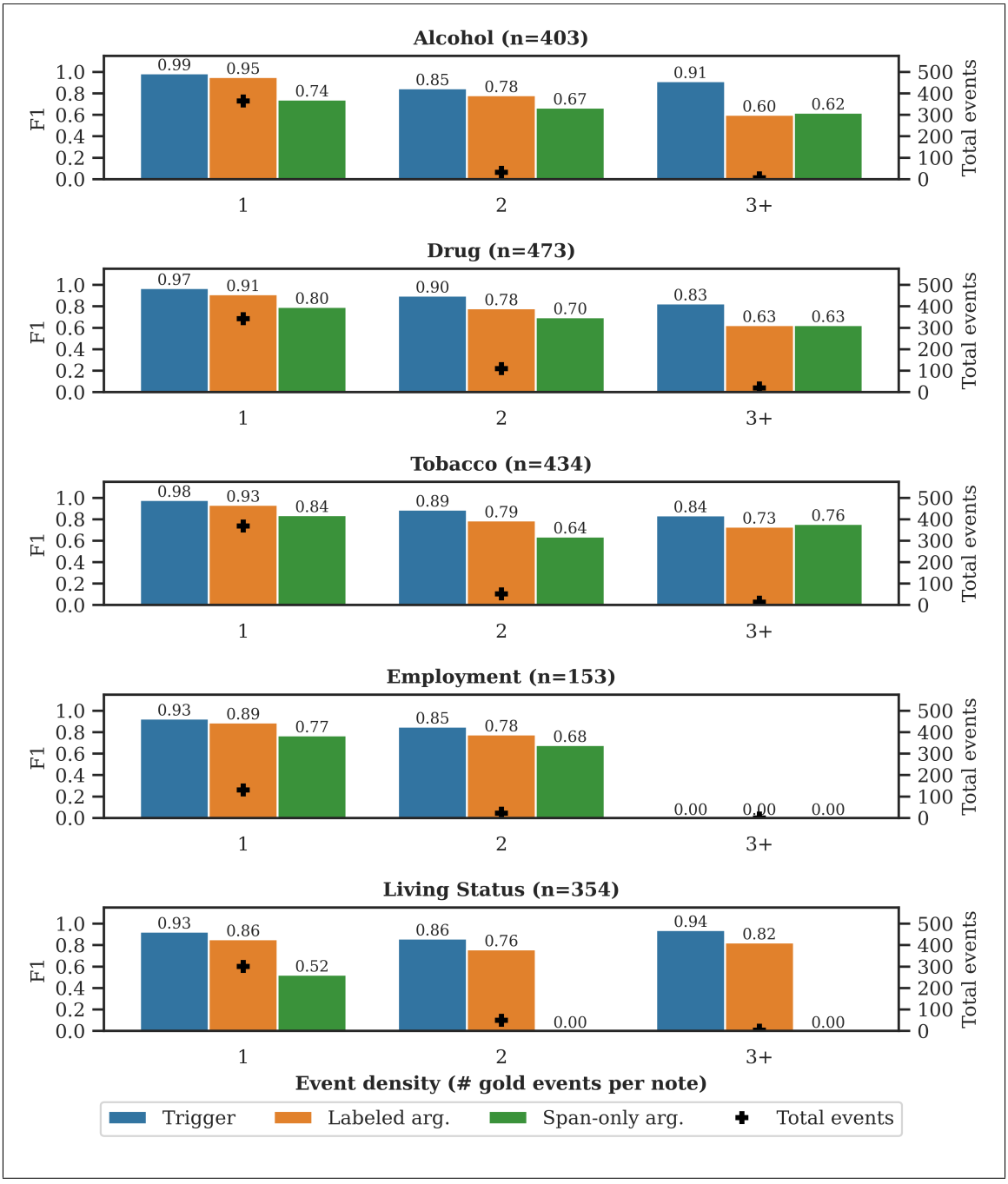

Figure 9: Performance breakdown by event density for Subtask C ($\mathcal{D}^{uw}_{test}$). The left-hand y-axis is the micro-averaged F1 for triggers, labeled arguments, and span-only arguments (vertical bars). The right-hand y-axis is the total number of gold events (✚).