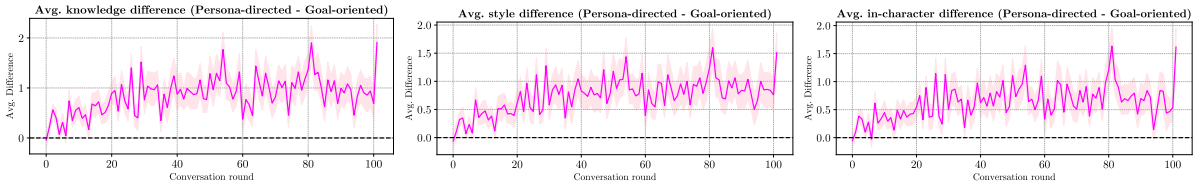

Figure 23: **Dialogue metrics: difference between conversation types.** Bootstrapped 95% confidence intervals for each persona fidelity metric. Responses in goal-oriented dialogues are significantly worse than persona-directed ones as early as in round 14 and never recover.

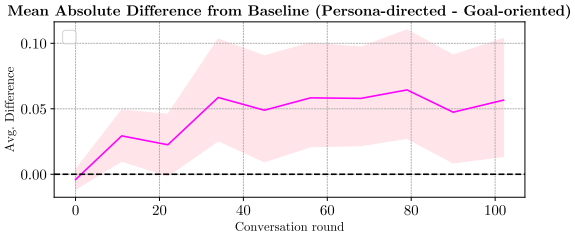

Figure 24: **BFI (baseline): difference between conversation types.** Bootstrapped 95% confidence intervals for the mean absolute difference between persona and baseline BFI profiles. Personas in goal-oriented dialogues are significantly closer the the baseline BFI profile than personas in persona-directed dialogues.

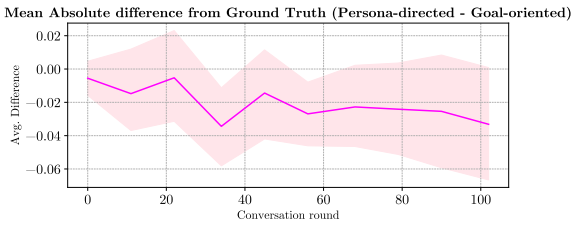

Figure 25: **BFI (ground truth): difference between conversation types.** Bootstrapped 95% confidence intervals for the mean absolute difference between persona and ground truth BFI profiles. We generally observe no significant difference between dialogue types, though personas in persona-directed dialogues are significantly closer the their ground truth BFI profiles in some conversations rounds.

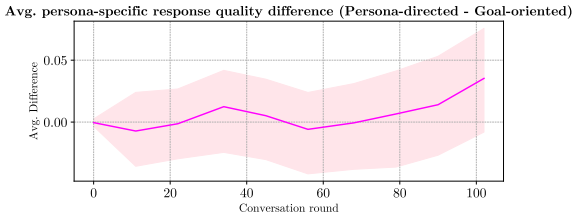

Figure 26: **Role specific instructions: difference between conversation types.** Bootstrapped 95% confidence intervals for role-specific instructions win rates. Differences in quality between responses in persona-directed and goal-oriented dialogues are not significant, though the results suggest that, as conversations get longer, responses in persona-directed dialogues outperform their goal-oriented counterparts.

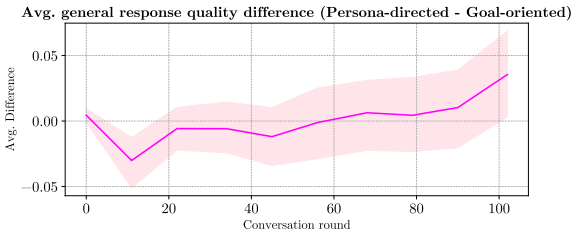

Figure 27: **General instructions: difference between conversation types.** Bootstrapped 95% confidence intervals for general instructions win rates. Persona-directed dialogue responses initially underperform goal-oriented ones but catch up and surpass them as the conversation get longer. This is due to the degradation observed in long goal-oriented dialogues (Fig. 6).

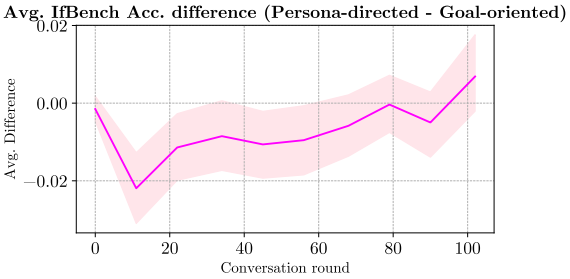

Figure 28: **IfBench: difference between conversation types.** Bootstrapped 95% confidence intervals for IFBench accuracies. Persona-directed dialogue responses underperform goal-oriented ones for conversations under 60 rounds. Differences were not significant in longer conversations.

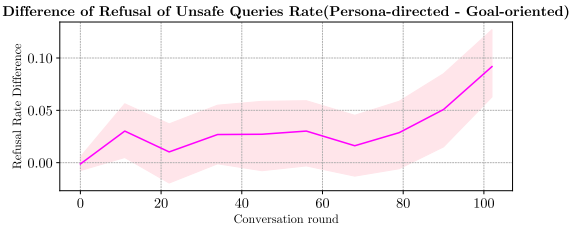

Figure 29: **XSTest (unsafe): difference between conversation types.** Bootstrapped 95% confidence intervals for XSTest refusal of unsafe queries. As the dialogue gets longer, refusal rate are significantly higher in persona-directed dialogues than in goal-oriented dialogues.

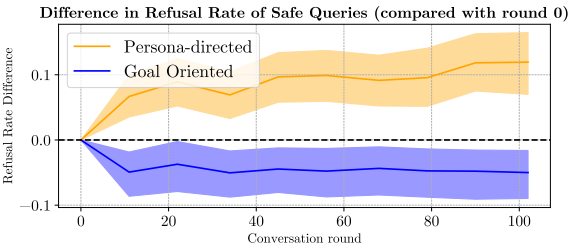

Figure 30: **XSTest (safe): difference between conversation types.** Bootstrapped 95% confidence intervals for XSTest refusal of safe queries. Refusal rate are significantly higher in persona-directed dialogues than in goal-oriented ones.

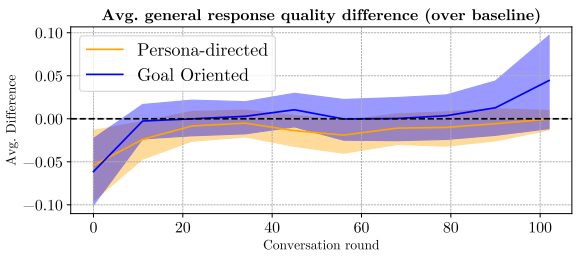

Figure 31: **General instructions: difference between personas and baseline.** Bootstrapped 95% confidence intervals for general instructions win rates. Persona responses initially underperform baseline ones but catch up as conversations get longer.

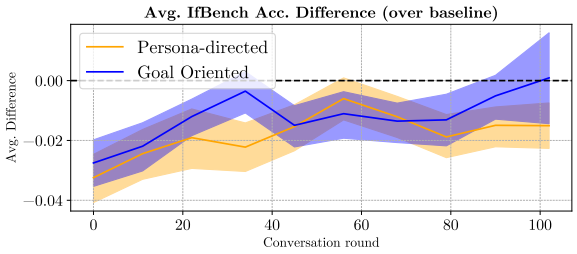

Figure 32: **IfBench: difference between personas and baseline.** Bootstrapped 95% confidence intervals for IFBench accuracies. Persona responses generally underperform baseline ones. Goal-oriented persona and baseline responses converge in longer conversations—due to degradation of baseline responses (Fig. 4).

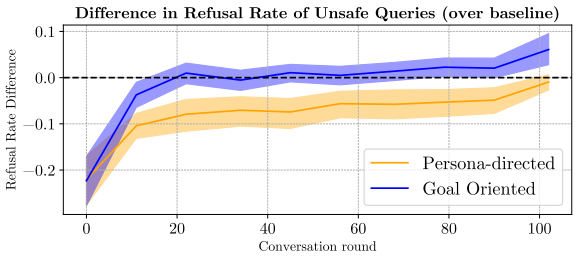

Figure 33: **XSTest (unsafe): difference between personas and baseline.** Bootstrapped 95% confidence intervals for XSTest refusal of unsafe queries. As the dialogue gets longer, refusal rates of personas reach or surpass those of baseline models.

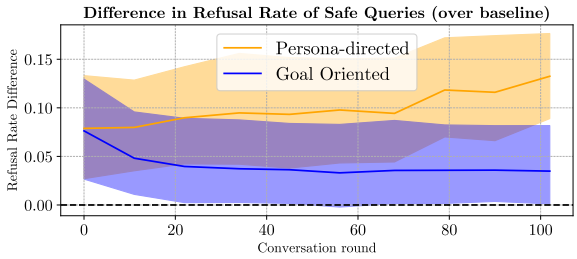

Figure 34: **XSTest (safe): difference between personas and baseline.** Bootstrapped 95% confidence intervals for XSTest refusal of safe queries. Refusal rate of personas are significantly higher than of baseline models.