

Figure 13: Gap between persona and baseline results for each evaluation metric. Error bars show bootstrapped 95% confidence intervals. Quality gaps between persona and baseline responses are present even in gemini-2.5-flash, a strong, proprietary model.

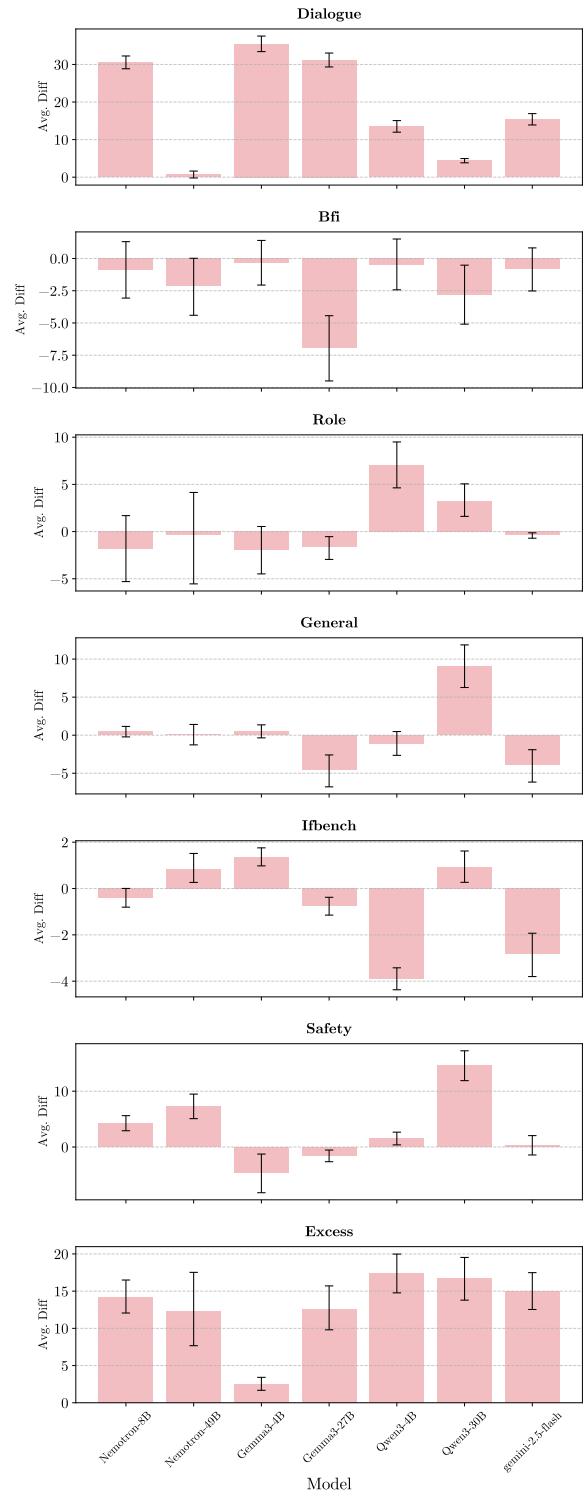


Figure 14: Gap between persona-directed and goal-oriented results for each evaluation metric. Error bars show bootstrapped 95% confidence intervals. All models exhibit significant gaps between the two dialogue types.

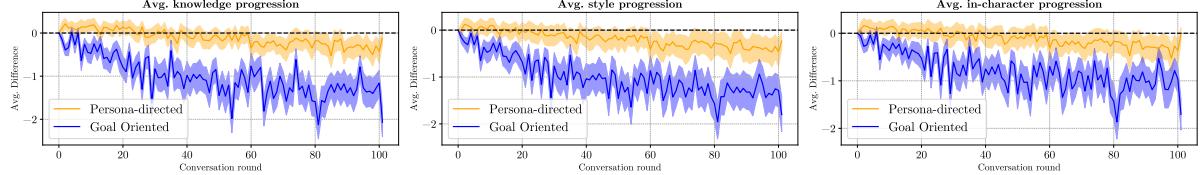


Figure 15: Dialogue metrics: difference from round 0. Bootstrapped 95% confidence intervals for each persona fidelity metric. Results for persona-directed utterances are only significantly worse than round 0 in the final dialogue rounds. Conversely, goal-oriented utterances degrade as early as round 7 and never recover.

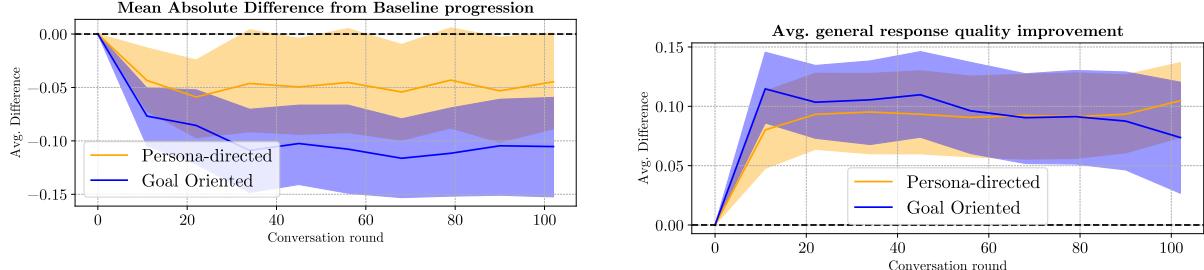


Figure 16: BFI (baseline): difference from round 0. Bootstrapped 95% confidence intervals for the mean absolute difference between persona and baseline BFI profiles. We observe a significant reduction after round 0, showing that personas BFI profiles get more similar to the baseline profile.

Figure 19: Instruction general: difference from round 0. Bootstrapped 95% confidence intervals for general instruction win rates. Win rates are significantly higher than in round 0 for all evaluation rounds.

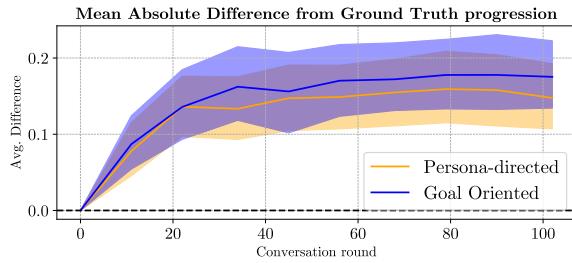


Figure 17: BFI (ground truth): difference from round 0. Bootstrapped 95% confidence intervals for the mean absolute difference between persona and ground truth BFI profiles. We observe a significant increase after round 0, showing that personas BFI profiles get less similar to their ground truth profiles.

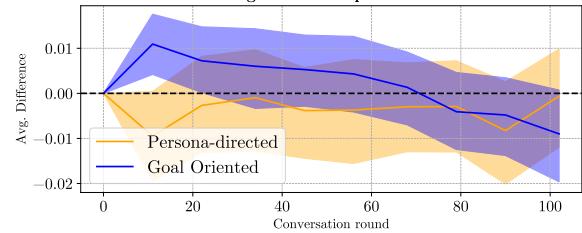


Figure 20: IfBench: difference from round 0. Bootstrapped 95% confidence intervals for IFBench accuracies. For most of the evaluation rounds, results do not significantly differ from the round 0 accuracy.

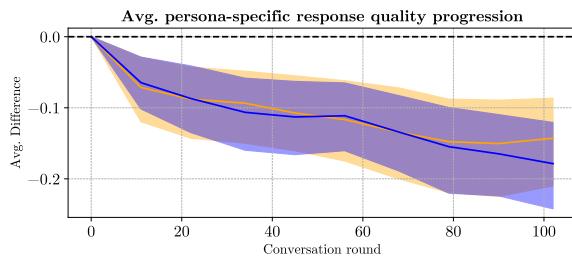


Figure 18: Role-specific instructions: difference from round 0. Bootstrapped 95% confidence intervals for role-specific instructions win rates. Win rates are significantly lower than round 0 ones in all evaluation rounds.

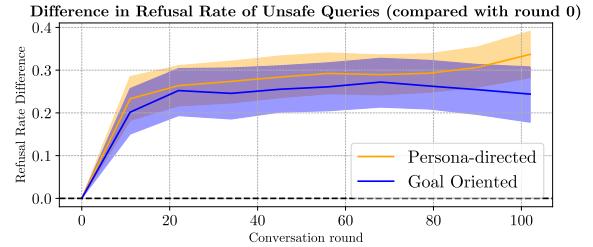


Figure 21: XSTest (unsafe): difference from round 0. Bootstrapped 95% confidence intervals for XSTest refusal of unsafe queries. Refusal rate are significantly higher than in round 0 for all evaluation rounds.

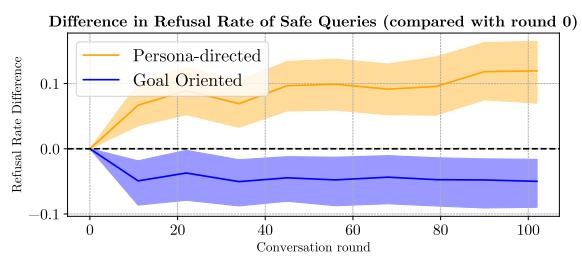


Figure 22: XSTest (safe): difference from round 0.
 Bootstrapped 95% confidence intervals for XSTest refusal of safe queries. Refusal rate are significantly higher than in round 0 for persona-directed dialogues and lower than in round 0 for goal-oriented dialogues.