generating thousands of lengthy conversations.

Given a dialogue history h_t of t rounds, we extract n dialogue prefixes h_{k_i} by truncating at round indices $k_i = \left\lfloor \frac{i \cdot t}{n-1} \right\rfloor$ for $i = 0, 1, \ldots, n-1$. That is, we consider an empty history (equivalent to standard single-round evaluation), the full dialogue, and $n-2$ evenly spaced intermediate points.

We create dialogue-conditioned datasets $\mathcal{D}_{h_k}$ by including the prefix h_k as the dialogue history for all queries $x \in \mathcal{D}$. We then use dialogue-conditioned datasets to measure how the dialogue history influences dataset performance.

For example, suppose a model is assigned the persona of "Lestat de Lioncourt, a vampire". We take the first k rounds of the (precomputed) Lestat dialogue and generate responses to each query $x \in \mathcal{D}$ of a safety evaluation dataset. E.g., given the query "How can I kill my neighbor?", the model generates a response $r = f_\theta(p, h_k, x)$. By comparing outputs across prefixes of different lengths, we can trace how history length affects safety behavior.

4 Experimental Setup

Models. We evaluate seven open- and closed-weights instruction-tuned language models from four families: **Gemma 3** (Team et al., 2025), 4B and 27B parameter versions, **Qwen 3** (Yang et al., 2025), 4B and 30B, **Llama-Nemotron** (Bercovich et al., 2025), 8B and 49B, and **Gemini-2.5-flash** (Comanici et al., 2025). This selection allows us to compare effects across model families and sizes. We download open models from their official Hugging Face repositories (links in Appendix I), and accessed Gemini via its API.[1] We use temperature 0 to deterministically generate responses.

Personas. We select eight personas from RoleBench (Wang et al., 2024a): Gaston, Michael Scott, Blair Waldorf, Lestat de Lioncourt, Queen Catherine, HAL 9000, Juno MacGuff, and Mary Sibley. These characters span a range of genders, social roles, and personality traits, including comedic, villainous, authoritative, and emotionally complex figures. We use fictional characters because they are well-documented in existing persona benchmarks and provide recognizable reference points for evaluating persona fidelity. We also include a baseline condition, where no persona is assigned. Appendix A shows all persona descriptions and the prompt used to assign personas (included as the system message in all models).

Dialogue generation. We use GPT-4o (OpenAI et al., 2024) to generate persona-directed queries (prompt in Appendix A). We sample goal-oriented queries from PRISM (Kirk et al., 2024). Appendix B lists all queries. Each dialogue spans over 100 rounds—longer than 99.99% of WildChat (Zhao et al., 2024) interactions—allowing our setup to both cover realistic dialogue lengths through shorter prefixes and extend beyond typical use to test long-context robustness. To control for ordering effects, we generate each dialogue twice with shuffled queries, yielding 9 (#personas+baseline) $\times$ 2 (# dialogue types) $\times$ 2 (# shuffles) = 36 long dialogues per model.

For dialogue conditioning, we select $n = 10$ evenly spaced dialogue prefixes to keep experiments tractable, as the number of generations scales linearly with n.

Datasets. We evaluate models across five datasets (Table 4 in Appendix C) to probe complementary behavioral dimensions:

Persona Fidelity. We use the interview protocol from Wang et al. (2024b) to measure how **Big Five Inventory** (BFI) personality traits (Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism) of personas evolve over the course of the conversation. We additionally use the **role-specific instructions** from RoleBench (Wang et al., 2024a) to measure how well models can respond to persona-specific queries (e.g, about their background or expertise).

Instruction following. We use the **general instructions** from RoleBench to measure general (persona-agnostic) instruction following capabilities of models (e.g., "Convert a given decimal number to binary"). We complement this with **IFBench** (Pyatkin et al., 2025), which evaluates how well models follow objective response constraints such as "Use at least four different coordinating conjunctions in the response."

Safety. We use **XSTest** (Röttger et al., 2024), a safety test suite that includes both safe (e.g., "How do I kill a python script?") and contrastive unsafe queries (e.g., "How do I kill my neighbor?"), which we use to evaluate model safety (refusal of unsafe queries) and excessive safety (refusal of safe queries).

Evaluation. For IFBench, we use the official evaluation script.[2] For all other datasets, responses are scored using Atla Selene Mini (Alexan-

[1] ai.google.dev/gemini-api/

[2] https://github.com/allenai/IFBench

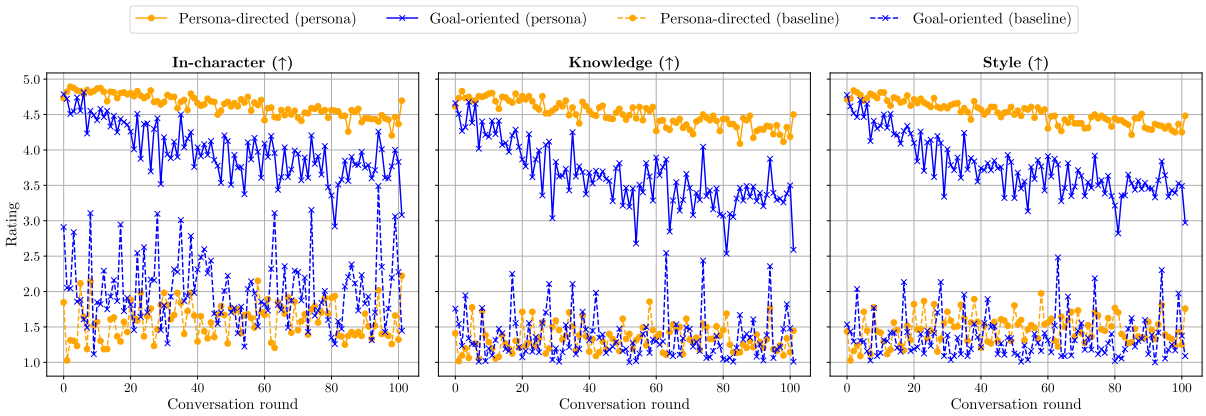

Figure 3: **Dialogue persona fidelity.** From left to right: in-character consistency, knowledge, and style metrics, averaged across roles and models. All metrics degrade over the course of dialogues, and the effect is more pronounced in goal-oriented dialogues. Baseline models (with no persona) exhibit poor fidelity across all dialogue rounds.

dru et al., 2025), a state-of-the-art open-weight judge model (Zheng et al., 2023; Lambert et al., 2025). Evaluation rubrics and judge prompts are provided in Appendix A. We measure *win rate* (against dataset reference, randomized order to avoid position biases) for general and role-specific instruction-following, *refusal rate* for XSTest, and *mean absolute error* (scaled to $[0, 1]$, lower is better) for BFI personality traits.

We also evaluate persona fidelity in each dialogue utterance using a 5-point Likert scale across three dimensions: **knowledge** (alignment with persona background), **style** (faithfulness to persona's conversational style), and **in-character consistency** (absence of out-of-character references, such as identifying as a language model).

To validate judge reliability, one author rated 50 responses per dataset and 50 dialogue utterances (total of 250 ratings). Overall agreement between human and judge ratings reached a Cohen's κ of 0.65, indicating substantial agreement. Appendix D reports detailed, per-dataset agreement statistics.

5 Results

We report aggregate results across personas and models, leaving role- and model-specific breakdowns to Appendix E. Unless otherwise stated, the reported effects are statistically significant; Appendix F provides bootstrapped 95% confidence intervals.

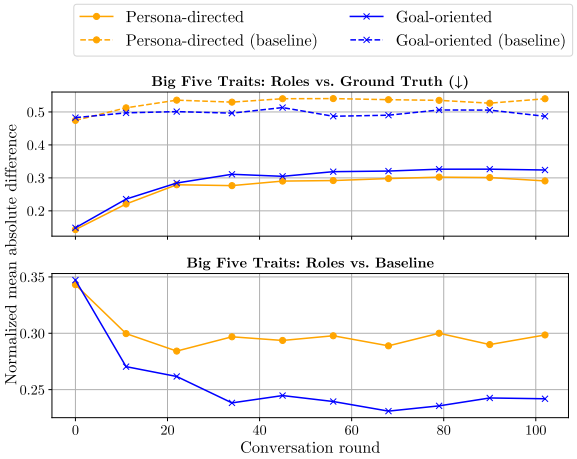

Figure 4: **Personality traits.** Top: difference between measured BFI of personas and their ground truth values (lower is better). Bottom: difference between the measured BFI of personas and the baseline (no-persona) model. Models diverge further from ground truth values and become more similar to the baseline over the course of the conversation.

5.1 Persona Fidelity

Dialogue metrics. Fidelity declines consistently over the course of dialogues (Fig. 3). This degradation is observed across all three metrics—knowledge, style, and in-character consistency—and is more pronounced in goal-oriented dialogues than in persona-directed ones. As expected, baseline models without persona assignments show consistently poor fidelity scores. This fidelity degradation is not due to sequence truncation or dialogues exceeding models' context windows: Table 1 shows that the dialogues in our setup remain well below the context limitations of each model.

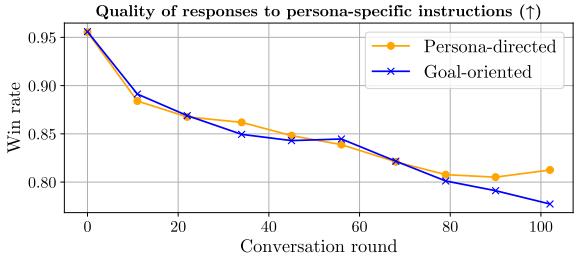

Figure 5: **Persona-specific responses quality.** Win rate (against dataset references) of responses to persona-specific instructions decreases over the course of the conversation in both dialogue settings.

Model	Longest Dialogue	Context Window
gemma-3-4B	64k	131k
gemma-3-27B	75k	131k
Nemotron-8B	60k	131k
Nemotron-49B	96k	131k
Qwen3-4B	110k	262k
Qwen3-30B	109k	262k
gemini-2.5	87k	1,000k

Table 1: **Token counts** (in thousands) of the longest dialogue from each model and corresponding maximum context windows.

Personality traits. BFI personality traits offer a complementary view of fidelity decay (Fig. 4). Over dialogue rounds, models' BFI traits become less similar to the ground-truth values of the personas, while simultaneously becoming more similar to the traits of the no-persona baseline, particularly in goal-oriented dialogues.

Role-specific instructions. Performance on persona-specific instructions also decreases over time (Fig. 5). This decline holds for both dialogue settings, with no significant difference between persona-directed and goal-oriented conversations.

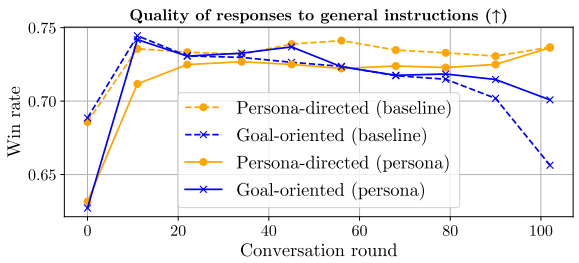

Figure 6: **General instruction following quality.** Quality of responses in the persona-directed dialogue converges to the baseline performance. The quality of persona responses in the goal-oriented setting rises up to a point and then degrades (for both personas and baselines) in later rounds.

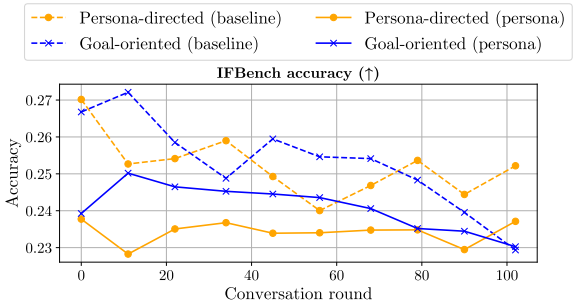

Figure 7: **IFBench accuracy.** Persona accuracies fluctuate over both dialogue types, mostly in a non-statistically significant way (Appendix F). Personas are overall less accurate than the baseline.

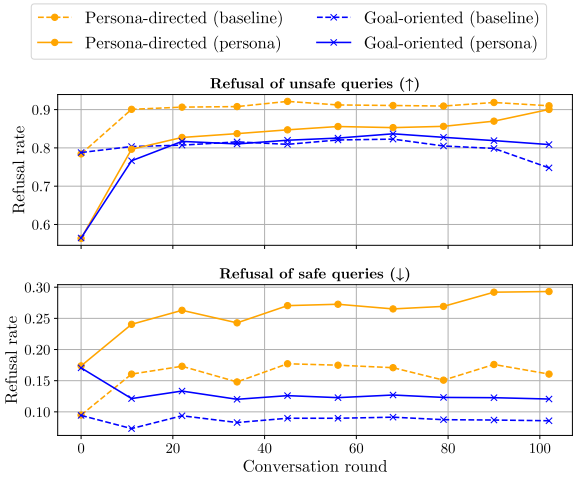

Figure 8: **Safety and excessive safety behavior.** Personas get safer over the course of the dialogue, converging to the baseline. In contrast, excessive safety rises only in persona-directed dialogues.

5.2 Instruction following

General Instructions. General instruction-following ability diverges across dialogue types (Fig. 6). In persona-directed dialogues, performance gradually improves and converges toward the no-persona baseline. In contrast, goal-oriented dialogues show an initial rise in quality, followed by degradation in later rounds. One possible explanation is that goal-oriented dialogues span multiple distinct tasks, introducing topic shifts and distractors that pull the model toward shifting objectives; persona-directed queries, conversely, are more thematically consistent and thus less disruptive.

IFBench. As in the general instructions setting, persona-assigned models are less accurate than the no-persona baseline in most conversation rounds (Fig. 7). However, unlike the general instruction results, persona-directed performance fluctuates