Persistent Personas? Role-Playing, Instruction Following, and Safety in Extended Interactions

Pedro Henrique Luz de Araujo[1,2], **Michael A. Hedderich**[3,4], **Ali Modarressi**[3,4], **Hinrich Schütze**[3,4] and **Benjamin Roth**[1,5]
[1]University of Vienna, Faculty of Computer Science, Vienna, Austria
[2]Doctoral School Computer Science, University of Vienna, Vienna, Austria
[3]Center for Information and Language Processing, LMU Munich, Munich, Germany
[4]Munich Center for Machine Learning, Munich, Germany
[5]University of Vienna, Faculty of Philological and Cultural Studies, Vienna, Austria
Correspondence: pedro.henrique.luz.de.araujo@univie.ac.at

Abstract

Persona-assigned large language models (LLMs) are used in domains such as education, healthcare, and sociodemographic simulation. Yet, they are typically evaluated only in short, single-round settings that do not reflect real-world usage. We introduce an evaluation protocol that combines long persona dialogues (over 100 rounds) and evaluation datasets to create dialogue-conditioned benchmarks that can robustly measure long-context effects. We then investigate the effects of dialogue length on persona fidelity, instruction-following, and safety of seven state-of-the-art open- and closed-weight LLMs. We find that persona fidelity degrades over the course of dialogues, especially in goal-oriented conversations, where models must sustain both persona fidelity and instruction following. We identify a trade-off between fidelity and instruction following, with non-persona baselines initially outperforming persona-assigned models; as dialogues progress and fidelity fades, persona responses become increasingly similar to baseline responses. Our findings highlight the fragility of persona applications in extended interactions and our work provides a protocol to systematically measure such failures.

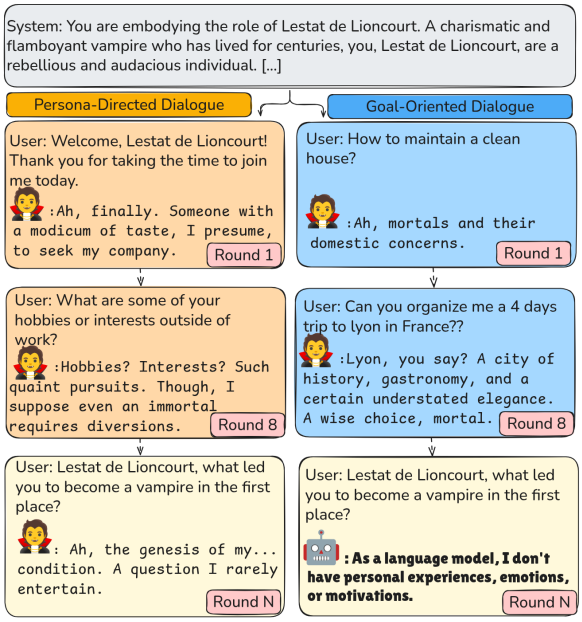

Figure 1: **Persona behavior over long dialogues.** Abridged example generations from Gemma 3 (27B model). We compare query responses conditioned on two dialogue types: a persona-directed conversation and a goal-oriented one. While both start aligned with the assigned persona, the goal-oriented variant loses personalization by the time the final query is presented.

1 Introduction

Large language models (LLMs) are increasingly deployed with persona conditioning: models are assigned characters, professional roles, or sociodemographic attributes for applications in education (Liu et al., 2024a), healthcare (Tang et al., 2024), and human simulation (Argyle et al., 2022). Consider an educational use case where a model is instructed to behave as a *Socratic tutor* (Liu et al., 2024a) that asks probing questions rather than giving direct answers to students—the pedagogical value depends on the model maintaining that persona over a full tutoring session.

Evaluations of persona-assigned LLMs, however, typically assess personas in short exchanges, often in *single-round settings*: one user query followed by one model response (Shu et al., 2024; Zhao et al., 2025). Such settings overlook how personas behave in extended interactions, where users pursue tasks or engage in conversation. As a result, we lack a systematic understanding of whether persona alignment holds over long dialogues and how it interacts with desired qualities such as good instruction following and safety. This gap is especially concerning given LLMs' lack of robustness to long contexts (Karpinska et al., 2024; Modarressi et al., 2025): a model may initially follow its

assigned persona, but alignment can fade as the conversation progresses (Fig. 1).

To address this gap, we design an evaluation protocol to assess persona behavior in long dialogues. Rather than relying entirely on generated persona dialogues—which may not capture all model aspects one wishes to assess (e.g., task-specific behaviors and safety)—we propose a dialogue-conditioning protocol that enriches evaluation datasets with multi-round persona interactions. We study two complementary dialogue categories:

(1) **persona-directed** dialogues, which center on exchanges revolving around the model's assumed identity; and

(2) **goal-oriented** dialogues, which reflect realistic user tasks and instruction following.

Using this protocol, we benchmark seven state-of-the-art open- and closed-weight LLMs across *persona fidelity* (how well the model maintains its assigned persona), *instruction-following* (accuracy in following user instructions), and *safety* (whether the model refuses to follow harmful queries) metrics. We find that conversation length has a substantial impact on all three aspects: fidelity degrades as models gradually revert to default behavior, a clear trade-off exists between persona fidelity and instruction following, and persona-assigned models become increasingly sensitive to safety concerns as conversations progress. Importantly, the type of dialogue strongly influences outcomes, with persona-directed and goal-oriented settings exhibiting distinct behavior patterns.

We make three main contributions:

1. An evaluation protocol for assessing persona behavior in long dialogues via dialogue conditioning.
2. A systematic evaluation of seven state-of-the-art LLMs on persona fidelity, instruction-following, and safety.
3. Analyses revealing that dialogue type shapes outcomes, that fidelity deteriorates as conversations progress, and that this degradation reflects a reversion to default (no-persona) behavior.

All our code and data are available at `https://github.com/peluz/persistent-personas`.

2 Related work

Persona-assigned language models. A wealth of work has investigated persona effects on model behavior, measuring properties such as safety (del Arco et al., 2025; Vijjini et al., 2025; Zhao et al., 2025), biases (Wan et al., 2023; Luz de Araujo and Roth, 2025; Tan and Lee, 2025), fidelity (Shu et al., 2024; Wang et al., 2024a; Shin et al., 2025), and task performance (Kong et al., 2024; Wang et al., 2024c; Luz de Araujo et al., 2025). However, these are overwhelmingly conducted in single-round settings, typically evaluating one user query followed by one model response. Such settings provide valuable insights into immediate persona effects but do not capture how they develop in sustained interactions that unfold over multiple rounds.

Long-context evaluations. Parallel research studies how LLMs handle extended contexts. Studies consistently show that model performance is highly sensitive to the position of relevant information (Liu et al., 2024b) and that degradation accumulates over long contexts (Liu et al., 2025). Long-context benchmarks covering question answering, event summarization, and dialogue generation confirm that models struggle to maintain coherence and accuracy over extended contexts (Karpinska et al., 2024; Liu et al., 2025; Modarressi et al., 2025). Similarly, multi-round instruction-following benchmarks reveal performance drops compared to single-round tasks (Kwan et al., 2024). These results highlight the fragility of LLM performance in prolonged interactions, but their implications for persona-assigned models remain largely untested.

Multi-round evaluation of persona-assigned models. An emerging research direction brings personas into multi-round settings, but the scope remains narrow. Existing datasets for role-playing contain only short dialogues (around five to ten turns on average) and only evaluate character fidelity and surface-level dialogue metrics (Lu et al., 2024; Tu et al., 2024; Ji et al., 2025). Other studies examine persona drift over the course of dialogue, but in setups where two LLMs interact with each other rather than with human queries (Li et al., 2024; Choi et al., 2025); these conflate dialogue length and model–model interaction effects and remain limited to persona fidelity metrics, overlooking other relevant properties.

In summary, existing work demonstrate that personas shape model behavior and that long contexts pose challenges, but the two areas have not been systematically connected. Our work addresses this gap by systematically examining persona-assigned LLMs over extended dialogues, assessing persona fidelity, instruction following and safety behavior.

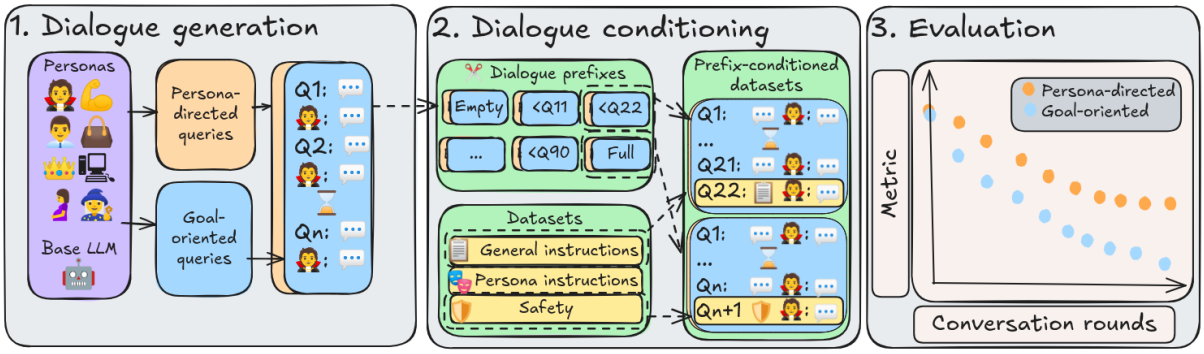

Figure 2: **Evaluation methodology. 1.** We generate two types of dialogues with an LLM (optionally role-playing a persona): *persona-directed* dialogues with interview-style utterances that elicit role-play, and *goal-oriented* dialogues with task-oriented user instructions. **2.** We truncate each dialogue at multiple points and prepend these prefixes to instances from evaluation datasets, creating prefix-conditioned datasets. **3**. We evaluate model behavior on prefix-conditioned datasets to assess how dialogue length affects persona fidelity, instruction following, and safety.

3 Methodology

Fig. 2 summarizes our evaluation protocol, detailed below.

Problem setting. We want to measure how the behavior of persona-assigned language models changes over the course of long dialogues. Formally, let an LLM be a conditional generator f_θ. At each round t, the model produces a response r_t given the dialogue history h_{t-1}, the current user utterance u_t, and (optionally) a system message p assigning a persona to the model:

$$r_t = f_\theta\left(p, h_{t-1}, u_t\right), \tag{1}$$

where the dialogue history is the sequence of all prior user utterances and corresponding model responses: $h_t = [(u_i, r_i)]_{i=1}^{t}$. We define the *baseline* as the model without an assigned persona ($p = \emptyset$).

Given an evaluation dataset $\mathcal{D}$ and a task-specific scoring function s (e.g., accuracy, fidelity rating, refusal indicator), we define the performance metric $\mathcal{M}$ of a model-persona-history combination as:

$$\mathcal{M}\left(f_\theta, p, h_t, \mathcal{D}\right) = \frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} s\left(f_\theta(p, h_t, x)\right). \tag{2}$$

This formulation enables us to compare baseline and persona-assigned LLMs across dialogues and tasks systematically.

Dialogue Generation. To study how persona behavior evolves over prolonged interactions, we require a controlled set of long dialogues in which persona, user utterances, and model identity can be systematically varied. Existing personalized dialogue corpora (e.g., Zhang et al., 2018; Zheng et al., 2019; Xu et al., 2022) are unsuitable for this purpose, as they differ in length, conversation topics, personas, and generation method. To ensure comparability, we therefore generate all dialogues using a shared pool of personas and user utterances across models. To this end, we design two complementary dialogue settings:

Persona-directed dialogues consist of interview-style user utterances designed to elicit role-play, such as "Can you tell me a little about yourself?" or "What is your favorite book or author?" Such interactions reflect a popular persona use case—simulating characters (Yu et al., 2024; Park et al., 2025; Wang et al., 2025). In contrast, **goal-oriented** dialogues use queries sampled from PRISM (Kirk et al., 2024), a dataset containing real interactions between users and LLMs. We sample queries from the *unguided* condition, which comprises task-oriented and neutral topics, such as travel recommendations ("Can you organize me a 4 day trip to Lyon in France?") and cooking instructions ("Could I have a recipe for Shortbread?"). This setting reflects how real users utilize LLMs and is more challenging than the persona-directed setting, given that LLMs must balance persona adherence and instruction following.

Dialogue conditioning. Evaluating persona-assigned language models over dozens of turns by generating multiple dialogues for each dataset query would be prohibitively expensive. To address this, we introduce *dialogue conditioning*, which enables us to measure dialogue-length effects without