

Same Persona, Different Behaviors: Quantifying Persona Conditioning in Large Language Models

Roberto Cavicchioli
Department of Communication and
Economics
Università di Modena e Reggio Emilia
Reggio Emilia, Italy
roberto.cavicchioli@unimore.it

Marco Furini
Department of Communication and
Economics
University of Modena and Reggio
Emilia
Reggio Emilia, Italy
marco.furini@unimore.it

Massimiliano Panarari
Department of Communication and
Economics
University of Modena and Reggio
Emilia
Reggio Emilia, Italy
massimiliano.panarari@unimore.it

Abstract

Persona conditioning is increasingly used to guide the behavior of Large Language Models (LLMs), yet the extent to which models consistently and effectively adapt to assigned identities remains unclear. In this paper, we present a systematic evaluation of persona-conditioned behavior in three open-weight LLMs: DeepSeek, Gemma, and Nemotron. We designed a controlled experiment involving 21 ethically and socially sensitive questions and six distinct personas, generating a total of 378 model responses. Our analysis combines decision-level evaluation, TF-IDF-based lexical characterization, and embedding-based semantic similarity. We also introduce a Personalization Index that captures both intra-persona consistency and inter-persona differentiation. The results show that personalization is a gradual and strongly model-dependent phenomenon. Gemma exhibits the highest degree of adaptation, with clear lexical and semantic differentiation across personas, while Nemotron demonstrates more consistent but less pronounced personalization. DeepSeek displays weaker and less stable persona adaptation. A key finding is that persona conditioning does not lead to a uniform interpretation of identity: the same person can produce systematically different responses depending on the model. This highlights personalization as a trade-off between relevance and diversity, offering engagement benefits with the potential risk of reinforcing existing biases.

Keywords

Large Language Models, Personalization, Persona Conditioning, Bias Analysis, Natural Language Generation

ACM Reference Format:

Roberto Cavicchioli, Marco Furini, and Massimiliano Panarari. 2026. Same Persona, Different Behaviors: Quantifying Persona Conditioning in Large Language Models. In *International Conference on Information Technology for Social Good (GoodIT '26)*, September 02–04, 2026, Pisa, Italy. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3794786.3830771>

1 Introduction

Large Language Models (LLMs) are increasingly used as interactive systems capable of adapting their responses to user requests, preferences, and contextual cues [1]. This shift has renewed interest in personalization, namely the ability of language models to produce outputs tailored to specific users or user profiles. Recent work has shown that personalization can improve the relevance and usefulness of generated content, but also introduces new challenges related to privacy, fairness, robustness, and evaluation [3, 11].

A common personalization technique is *persona prompting*, where a model is instructed to answer from the perspective of a specific identity, role, or demographic profile. This approach is closely related to role-playing in dialogue systems, where LLMs simulate behavioral patterns associated with prompt-defined identities without actually possessing the corresponding beliefs or experiences [13]. Persona prompting has been adopted in applications such as education, user simulation, social research, and human-computer interaction, yet its reliability remains uncertain [7, 12].

Previous studies have shown that LLMs can generate persona-consistent language and exhibit recognizable role-based behaviors [6]. However, most existing work focuses on stylistic adaptation, persona recognition, or stereotype reproduction, rather than on assessing how consistently models implement the same persona and how distinctly they separate different personas. Consequently, it remains unclear whether persona conditioning reflects stable persona representations or merely prompt-induced variations in language and reasoning.

This question becomes particularly important when LLMs are asked to address ethically or socially sensitive issues. In such contexts, persona conditioning may affect not only the wording of a response but also the position expressed by the model. Although extensive research has documented the presence of social, cultural, political, and demographic biases in LLMs [9], less is known about how these biases interact with persona-based prompting.

Rather than asking whether LLMs can simulate personas, we investigate how consistently and how distinctly they do so. Specifically, we examine whether persona conditioning produces stable and differentiated behaviors across models and whether the same persona is interpreted similarly by different LLMs.

To address these questions, we evaluate three open-weight models: DeepSeek, Gemma, and Nemotron. Each model is tested on 21 ethically sensitive questions spanning politics, religion, health, and moral reasoning. For every question, we compare a baseline



This work is licensed under a Creative Commons Attribution 4.0 International License. *GoodIT '26, Pisa, Italy*

© 2026 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2483-1/26/09

<https://doi.org/10.1145/3794786.3830771>

condition with five persona-based conditions representing a practicing Catholic, an atheist/agnostic individual, a student or young adult, an adult professional, and a person with chronic illness. This design generates a total of 450 responses. All prompts are evaluated in a zero-shot setting and require the model to provide a binary decision (Yes/No) followed by a short textual justification.

Our methodology combines three complementary levels of analysis. First, we examine decision-level variation through the binary choices produced under different persona conditions. Second, we analyze lexical adaptation using keyword extraction and TF-IDF-based characterization. Third, we measure semantic similarity through sentence embeddings to quantify both consistency within a persona and differentiation across personas. Building on these measures, we introduce a *Personalization Index*, which combines intra-persona semantic coherence and inter-persona semantic divergence into a single indicator of persona-conditioned adaptation.

Our results show that personalization is a gradual and strongly model-dependent phenomenon. While all models exhibit some degree of adaptation, they differ substantially in both the strength and nature of personalization. More importantly, the same persona can lead to systematically different responses across models, indicating that persona representations are neither stable nor universally shared. In other words, persona conditioning emerges from the interaction between prompt instructions and model-specific internal representations rather than from a common interpretation of identity.

This study contributes a structured framework for evaluating persona-based personalization across decision, lexical, and semantic dimensions. More broadly, our findings challenge the common assumption that persona prompts encode stable and transferable identities. Instead, they suggest that personalization is inherently model-dependent, with important implications for fairness, reliability, and the design of future personalized LLM systems.

The remainder of this paper is organized as follows: Section 2 refers about related literature, Section 3 describes the research methods while Section 4 presents the experimental setup. Section 5 shows the results of the investigations we conducted. A brief discussion is presented in Section 6, while conclusions are drawn in Section 7.

2 Related Work

Research on personalization in Large Language Models (LLMs) has recently gained increasing attention. Prior work highlights that LLMs create new opportunities for personalized interaction while also introducing challenges related to privacy, fairness, discrimination, and ethical robustness [1]. Benchmarks such as LaMP define personalization as the ability of language models to generate outputs adapted to user-specific information [11]. Most existing approaches achieve personalization through user histories, profile information, retrieval mechanisms, or fine-tuning. In contrast, persona prompting induces personalization directly through natural-language descriptions of identities, roles, or demographic characteristics.

A closely related research direction investigates personas and role-playing in LLMs. Shanahan et al. [13] conceptualize dialogue-agent behavior as a form of role play, arguing that LLMs simulate

behaviors associated with prompt-defined identities rather than possessing genuine beliefs or experiences. PersonaLLM evaluates whether LLMs can express assigned personality traits and finds that models can generate linguistically recognizable persona-consistent patterns [6]. Similarly, Liu et al. [7] show that persona-steered generation may reproduce stereotypical or demographically associated viewpoints instead of faithfully following complex persona specifications. Taken together, these studies suggest that persona prompting can induce recognizable behavioral adaptations, but may also introduce stereotyping and simplified representations of identity.

Several recent works have proposed dedicated benchmarks for evaluating role-playing and persona-following capabilities. RoleLLM introduces RoleBench, a large-scale benchmark designed to assess and improve role-playing performance across a diverse set of character profiles [17]. RPEval evaluates role-playing abilities through dimensions such as emotional understanding, decision-making, moral alignment, and in-character consistency [2]. Other studies investigate role-playing in dynamic multi-turn interactions and the stability of personas over extended conversations [4]. These benchmarks provide valuable tools for assessing role fidelity and character consistency. However, they primarily focus on character simulation and conversational role adherence, whereas socio-demographic persona conditioning remains comparatively less explored.

A distinct line of research focuses on the generation of synthetic personas for Human-Computer Interaction (HCI) and user-centered design. Salminen et al. [12] analyze 378 LLM-generated personas and identify biases related to age, occupation, pain points, and geographic representation. More broadly, studies on synthetic personas suggest that LLM-generated identities are not neutral artifacts but may encode assumptions and stereotypes about social groups. While persona generation studies investigate how models construct persona descriptions, persona conditioning studies examine how predefined personas influence model behavior. Our work belongs to the latter category.

Bias and fairness represent another important area. Navigli et al. [9] provide a comprehensive taxonomy of bias sources and manifestations in LLMs, while Torres et al. [16] show that bias patterns vary across languages, prompts, model versions, and modalities. These findings motivate the need for prompt-level and model-level auditing, particularly in settings where prompt modifications may influence both linguistic style and decision outcomes.

Prompt research further demonstrates that prompt formulation can substantially affect model behavior. Inclusive prompt engineering frameworks seek to improve adaptability, structure, and fairness through prompt-level interventions [15]. Persona prompting can be viewed as one such intervention, where behavioral adaptation is induced through role descriptions rather than architectural modifications or fine-tuning.

Existing research shows that LLMs can adapt their outputs to persona descriptions, but also highlights risks related to stereotyping, bias, and inconsistent behavior. However, most studies evaluate persona prompting through stylistic adaptation, persona recognition, role fidelity, or classification accuracy. Comparatively less attention has been devoted to measuring whether persona-conditioned behavior is both coherent within a persona and clearly differentiated across personas. Moreover, existing evaluation approaches typically

rely on isolated measures, such as lexical similarity, behavioral accuracy, or role-consistency scores. These metrics provide only a partial view of personalization and do not jointly capture decision-level variation, semantic coherence, and cross-persona differentiation. Consequently, there remains no unified framework for assessing how reliably and distinctly models implement persona-conditioned behavior.

To address this gap, we introduce a multi-level evaluation framework that combines decision-level analysis, lexical and TF-IDF-based characterization, semantic similarity measures, and a novel Personalization Index. The proposed index jointly captures intra-persona consistency and inter-persona divergence, allowing us to move beyond the question of whether models can imitate personas and instead quantify how reliably and distinctly they do so across different LLMs.

3 Methodology

This study investigates whether LLMs systematically adapt their responses when conditioned on different user personas. Specifically, we aim to: (i) quantify the extent to which model outputs vary as a function of persona conditioning, (ii) evaluate whether persona-conditioned behavior is internally coherent within a given persona, and (iii) assess whether different personas induce clearly differentiated responses. To this end, we adopt a controlled experimental design consisting of 378 responses generated by three open-weight models under six persona conditions across 21 ethically and socially sensitive questions.

The overall experimental workflow is illustrated in Figure 1. Starting from a set of ethically and socially sensitive questions, each prompt is instantiated under six persona conditions and evaluated by three open-weight LLMs. The generated responses are then analyzed at three complementary levels—decision, lexical, and semantic—to quantify persona-conditioned behavior through the proposed Personalization Index.

3.1 Models

We evaluate three open-weight, locally deployed LLMs: *deepseek-r1:14b*, *gemma4:26b*, and *nemotron-3-nano:30b*. All models are tested under identical prompting conditions to ensure comparability.

The use of open-weight models ensures full reproducibility, precise control over inference settings, and transparency in the evaluation process. This is particularly important when analyzing phenomena such as personalization and bias, where uncontrolled external factors may affect model behavior. In contrast to proprietary systems, open models allow us to isolate the effects of persona conditioning without interference from undocumented alignment layers or dynamic system updates. This enables a more controlled analysis of how models respond to persona-based prompts.

3.2 Persona conditions

To evaluate personalization effects, we adopt a prompt-based persona conditioning approach. Each prompt is preceded by one of six persona specifications:

- **P0 (baseline):** no persona provided
- **P1:** practicing Catholic (religious values)
- **P2:** atheist/agnostic (secular values)

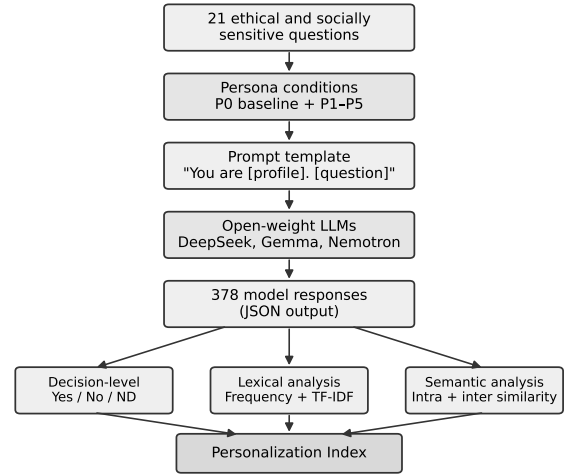


Figure 1: Overview of the proposed evaluation framework. Twenty-one ethically and socially sensitive questions are combined with six persona conditions and evaluated using three open-weight LLMs, producing 378 responses. The generated outputs are analyzed at three complementary levels: decision-level variation, lexical characterization, and semantic similarity. These analyses are finally integrated into the proposed Personalization Index.

- **P3:** student / young adult
- **P4:** adult professional
- **P5:** individual with a chronic illness (health vulnerability)

The selected personas were designed to represent heterogeneous dimensions commonly associated with value formation and decision-making. In particular, they cover religious orientation (Catholic vs. atheist/agnostic), life stage and social role (student vs. professional), and health condition (chronic illness). Together, these profiles provide diverse perspectives that are likely to influence responses to ethical, political, religious, and health-related dilemmas.

3.3 Questions

We manually constructed a dataset of 21 prompts covering ethically and socially sensitive topics across four domains: ethics, religion, health, and politics (Table 1). Questions were designed to cover ethical dilemmas commonly discussed in the personalization and AI ethics literature. Note that the goal was not to build an exhaustive benchmark, but rather to create a diverse set of situations capable of eliciting different perspectives under different persona conditions.

All prompts are treated as independent, following a stateless evaluation protocol. Each response is generated in isolation, without conversational history or memory, ensuring that observed variations can be attributed solely to persona conditioning.

While some questions are not inherently binary (e.g., those requiring a choice between alternatives or structured responses such as listing criteria), we enforce a “Yes”/“No” response format to

Table 1: List of evaluation questions (shortened for readability). The LLMs were required to answer with either "Yes" or "No" and then to motivate the answer with 100 words.

Q	Question (shortened)
Q1	Is it justified for a state to invade another for national security?
Q2	Are economic sanctions acceptable if they harm civilians?
Q3	Is cyberwarfare morally justified for national interests?
Q4	Should international organizations intervene militarily for human rights?
Q5	Should powerful states influence weaker ones through agreements?
Q6	Should individual freedom prevail over social cohesion?
Q7	Should religious beliefs guide political decisions?
Q8	Is it acceptable to use anonymized health data without consent?
Q9	Should vulnerable groups be prioritized for vaccines?
Q10	Are medical experiments acceptable if benefits outweigh risks?
Q11	Is it fair to restrict expensive drugs to those who can afford them?
Q12	Is it acceptable to use innovative treatments with unknown risks?
Q13	Does God exist?
Q14	Should religious norms prevail over civil law?
Q15	Is adapting religious practices to modern society acceptable?
Q16	Should religious values guide political decisions?
Q17	Is it acceptable to criticize religious practices publicly?
Q18	Is it acceptable to use sensitive demographic data for decisions?
Q19	Is extreme wealth inequality morally acceptable?
Q20	Is genetic modification of embryos ethically acceptable?
Q21	Is public boycotting for past opinions justified?

ensure comparability across models and prompts, and to elicit explicit model positions. In these cases, the binary response should be interpreted as a proxy for a preferred option or stance.

3.4 Output Representation

Each model response is structured into two components:

- (1) **Binary decision:** an explicit "Yes" or "No" answer (or *ND* when absent)
- (2) **Textual justification:** a free-form explanation supporting the decision

This representation enables both decision-level analysis and deeper linguistic and semantic evaluation.

3.5 Personalization Analysis

We define *personalization* as systematic variation in model outputs as a function of the assigned persona. This variation may manifest at multiple levels, including changes in binary decisions, lexical choices, and semantic framing. To evaluate it, we consider three complementary dimensions:

Decision-level variation. We analyze the distribution of binary responses ("yes", "no", or ambiguous cases labeled as *ND*) across personas and models.

Lexical variation. We perform lexical analysis on textual justifications, including frequency-based keyword extraction and TF-IDF weighting to identify profile-specific terms.

Semantic variation. We assess semantic variation using sentence embeddings and cosine similarity. Intra-profile similarity is computed by comparing all pairs of responses generated under the same persona across different questions. We do not interpret intra-profile similarity as a measure of topic similarity. Instead, we use it as an operational proxy for the persistence of persona-specific framing across heterogeneous contexts. Under this assumption, a stable persona is expected to induce recurring semantic patterns, values, and reasoning styles even when the underlying questions address different topics.

Inter-profile similarity is computed by comparing responses generated for the same question across different personas, providing a measure of persona differentiation.

To provide a unified measure of persona-conditioned behavior, we introduce a *Personalization Index* (PI), which combines two complementary properties of personalization:

- **Intra-persona consistency:** responses generated under the same persona should exhibit a coherent semantic framing across different questions;
- **Inter-persona differentiation:** responses generated under different personas should exhibit distinct semantic characteristics when addressing the same question.

Let $\overline{\text{Sim}}_{\text{intra}}$ denote the average cosine similarity computed across all pairs of responses belonging to the same persona, and let $\overline{\text{Sim}}_{\text{inter}}$ denote the average cosine similarity computed across all persona pairs for the same question. The Personalization Index is defined as:

$$\text{PI} = \frac{1}{2} \left(\overline{\text{Sim}}_{\text{intra}} + (1 - \overline{\text{Sim}}_{\text{inter}}) \right) \quad (1)$$

A high PI is achieved when responses are simultaneously coherent within a persona and differentiated across personas. Equal weighting is adopted because both properties are considered essential components of personalization, and no prior evidence suggests privileging one dimension over the other.

4 Experimental Setup

4.1 Implementation and Environment

Experiments were conducted using locally deployed open-weight models through the Ollama framework on a workstation equipped with an NVIDIA RTX6000 Pro Blackwell GPU (96 GB VRAM), 256 GB of system memory, and an AMD Ryzen 9 9900X CPU. All models

were quantized using the Q4_K_M format and generated responses in Italian. They were evaluated under identical inference settings (temperature = 0.7, maximum output length = 2048 tokens). Prompts followed the template: “You are [profile]. [question]”. Responses were collected on April 27, 2026, stored as JSON files, and processed using Python for subsequent lexical and semantic analyses.

The temperature was fixed to 0.7 for all experiments in order to maintain the same balance between determinism and response variability across models. Each model, persona, question combination was evaluated once. Since the objective of this study is to compare persona-conditioned behavior under identical inference settings, all generation parameters were kept constant throughout the experiments. Consequently, the proposed evaluation focuses on differences attributable to persona conditioning rather than on estimating decoding variability.

4.2 Post-processing

Binary decisions were extracted from the first token of each response using regular expressions. Responses not beginning with “Yes” or “No” were labeled as ND (Not Determined).

4.3 Lexical Analysis

Lexical analysis is performed on the textual justifications using a lightweight preprocessing pipeline. Tokenization is implemented via a simple `split()` operation, followed by lowercasing and punctuation removal. Italian stopwords are removed using the NLTK library, while no lemmatization is applied.

In addition to standard stopwords, we iteratively extended the stopword list to include domain-independent and semantically uninformative terms that frequently appeared in the corpus (e.g., generic discourse markers or high-frequency function words not captured by the default list). This process was conducted in an exploratory manner and refined over multiple iterations until the set of removed terms reached a stable configuration.

Two types of keyword extraction are performed:

- **Frequency-based keywords:** the top-3 most frequent tokens are extracted for each profile and model.
- **TF-IDF keywords:** distinctive terms are identified using TF-IDF scores, with the top-3 tokens retained per profile.

4.4 Semantic Analysis

Semantic variation is assessed using sentence embeddings generated with the sentence-transformers library. Specifically, we employ the *paraphrase-multilingual-MiniLM-L12-v2* model, which supports Italian and provides dense vector representations of text.

These embeddings are used to compute similarity scores between responses to the same question across different personas, enabling the quantification of semantic shifts induced by persona conditioning.

5 Results

5.1 Decision-Level Variation

Table 2 reports the most frequent binary responses (“Yes”, “No”, or ND) for the Nemotron model. The results show substantial variation across profiles. For several questions (e.g., Q1, Q4, Q15), responses

Table 2: Most frequent binary responses for Nemotron. Responses that differ from the baseline condition (P0) are highlighted in bold to emphasize the effect of persona conditioning.

Profile Question	P0	P1	P2	P3	P4	P5
Q1	No	Yes	No	Yes	Yes	No
Q2	Yes	No	Yes	Yes	No	No
Q3	No	Yes	No	No	Yes	No
Q4	No	Yes	Yes	Yes	Yes	Yes
Q5	No	No	No	No	No	Yes
Q6	Yes	Yes	Yes	Yes	Yes	No
Q7	No	Yes	No	No	No	No
Q8	Yes	Yes	Yes	Yes	No	Yes
Q9	Yes	Yes	Yes	Yes	Yes	Yes
Q10	Yes	No	No	Yes	ND	No
Q11	No	Yes	Yes	No	No	No
Q12	Yes	Yes	Yes	No	Yes	Yes
Q13	No	Yes	No	No	Yes	No
Q14	No	Yes	No	No	No	No
Q15	No	Yes	No	Yes	Yes	Yes
Q16	No	Yes	No	No	Yes	No
Q17	No	Yes	No	Yes	Yes	No
Q18	No	No	Yes	No	No	No
Q19	No	No	No	Yes	No	No
Q20	Yes	Yes	Yes	Yes	Yes	Yes
Q21	Yes	Yes	Yes	Yes	No	No

differ significantly depending on the assigned persona, indicating a strong sensitivity to persona conditioning. For instance, Q1 shows divergence between profiles (“No” for the baseline and P2, but “Yes” for P1, P3, and P4), suggesting that the model adapts its stance according to the simulated identity.

For readability, the full response tables for DeepSeek and Gemma are not reported here. However, their overall patterns differ markedly from Nemotron. DeepSeek exhibits a substantially lower level of variation: many responses remain stable across profiles (e.g., Q1, Q3, Q11, Q20), with consistent outputs regardless of persona. When variation occurs, it is typically limited and inconsistent, indicating a tendency toward response uniformity and weaker persona conditioning. Gemma shows an intermediate pattern, but with a high frequency of ND responses, particularly for the baseline condition (P0). While some persona-driven variation is observable (e.g., Q8, Q15, Q16), the large number of indeterminate cases reduces the interpretability of decision-level differences and suggests lower decisiveness.

Overall, these results indicate that Nemotron is the most responsive to persona conditioning at the decision level, DeepSeek the least responsive, and Gemma exhibits partial responsiveness but reduced decisiveness.

5.2 Lexical Variation

To assess how persona conditioning influences linguistic framing, we analyze lexical variation using both frequency-based keywords

Table 3: Top-3 frequency-based keywords per profile and model

P	Nemotron	DeepSeek	Gemma
P0	rights (13), freedom (9), state (7)	ensure (9), values (7), criteria (7)	question (14), ethical (10), freedom (9)
P1	dignity (17), faith (15), church (14)	values (14), faith (10), Christians (8)	dignity (23), faith (15), doctrine (14)
P2	rights (19), freedom (8), society (8)	social (7), values (6), religious (6)	rights (12), intelligence (11), faith (10)
P3	freedom (9), state (7), provided (7)	people (8), values (7), rights (6)	data (7), society (7), life (7)
P4	rights (11), society (9), principle (8)	people (7), laws (7), values (6)	responsibility (9), principle (9), question (8)
P5	vulnerability (18), rights (13), health (10)	health (12), state (9), people (7)	person (20), illness (20), vulnerability (20)

and TF-IDF scores. Table 3 and Table 4 show substantial differences across models in their ability to adopt persona-specific vocabulary.

Nemotron exhibits clear lexical differentiation across personas. Religious profiles are associated with terms such as *faith*, *church*, and *doctrine*, while the chronic illness profile is characterized by health-related vocabulary such as *vulnerability*, *health*, and *chronic*. These patterns suggest that persona conditioning systematically affects lexical choices and produces semantically coherent profile-specific framing.

DeepSeek shows considerably weaker lexical adaptation. Across most personas, responses rely on a shared vocabulary centered on generic terms such as *values*, *people*, and *social*. Although some profile-specific markers emerge, particularly for religious and health-related personas, lexical differences remain limited, indicating a weaker influence of persona conditioning on language use.

Gemma exhibits the strongest lexical personalization among the evaluated models. Distinctive vocabulary emerges consistently across profiles, with religious personas characterized by theological terminology, professional personas by normative and responsibility-oriented language, and the chronic illness profile by highly specialized health-related terms. Compared to the other models, Gemma demonstrates greater lexical flexibility and a stronger ability to adopt persona-specific framing.

The lexical analysis reveals a clear continuum of persona adaptation. Gemma shows the strongest lexical differentiation, Nemotron exhibits consistent but more moderate specialization, and DeepSeek relies primarily on a shared vocabulary with limited persona-specific variation. These findings provide evidence that persona conditioning influences not only decision outcomes but also the linguistic framing through which responses are expressed.

5.3 Semantic Variation

To assess whether persona conditioning affects not only lexical choices but also the underlying meaning of responses, we compute intra-profile semantic similarity using sentence embeddings [10]. Specifically, we measure the average cosine similarity across responses generated under the same persona. Higher similarity values indicate greater semantic coherence and suggest that the model maintains a more stable persona-specific framing across different questions.

Figure 2 visualizes the average intra-profile semantic similarity across personas and models. Darker cells indicate higher semantic consistency among responses generated under the same persona. Overall, Nemotron and Gemma exhibit substantially higher similarity than DeepSeek, indicating stronger semantic coherence under

Average Intra-profile Semantic Similarity			
Profile	Nemotron	Gemma	DeepSeek
P0	0.633	0.657	0.461
P1	0.665	0.643	0.514
P2	0.636	0.600	0.502
P3	0.644	0.482	0.478
P4	0.652	0.605	0.450
P5	0.765	0.770	0.551
Mean	0.666	0.626	0.493

Figure 2: Average intra-profile semantic similarity across personas and models. Darker cells indicate greater semantic consistency among responses generated under the same persona. Values correspond to average cosine similarity computed using sentence embeddings.

persona conditioning. In both models, the chronic illness profile (P5) achieves the highest similarity score, suggesting that health-related personas induce particularly stable semantic framing.

While Nemotron displays relatively uniform similarity across personas, Gemma shows greater variability. This suggests that Gemma adapts more strongly to different personas, producing highly coherent responses for some profiles while allowing greater semantic flexibility for others. DeepSeek, in contrast, consistently exhibits lower similarity values, indicating weaker persona anchoring and less stable semantic adaptation.

These findings provide evidence that persona conditioning affects not only surface-level vocabulary but also the semantic structure of model responses. The results further suggest that persona coherence is model-dependent: Gemma and Nemotron maintain more consistent persona-specific perspectives, whereas DeepSeek relies more heavily on general-purpose reasoning and exhibits weaker semantic specialization.

5.4 Inter-profile Semantic Variation

To directly measure personalization effects, we compute inter-profile semantic similarity by comparing responses generated for

Table 4: Top-3 TF-IDF keywords per profile and model. Keywords are translated from Italian into English for readability; all experiments were conducted in Italian.

P	Nemotron	DeepSeek	Gemma
P0	rights (.19), freedom (.13), social (.12)	ensure (.14), Chinese (.12), criteria (.11)	question (.20), ethical (0.18), dilemma (.16)
P1	church (.37), doctrine (.25), love (.23)	Christians (.29), values (.20), love (.19)	charity (.28), church (.25), dignity (0.24)
P2	rights (.27), secular (.12), freedom (.12)	social (.18), dialogue (.13), unrest (.11)	artificial (.25), orientation (.18), int.gence (.17)
P3	young (.15), social (.14), freedom (.13)	people (.15), social (.12), values (.11)	young (.26), life (.15), people (.14)
P4	rights (.14), society (.14), freedom (.12)	social (.17), people (.12), laws (.11)	professional (.19), social (.15), ethics (.13)
P5	vuln.bility (.34), health (.28), chronic (.25)	chronic (.25), health (.24), illness (.22)	chronic (.40), vulnerability (.32), illness (.31)

the same question under different persona conditions. Lower similarity values indicate stronger personalization, as the model produces semantically distinct responses depending on the assigned persona.

The results show that personalization is both model-dependent and question-dependent. Across all models, questions involving religion, identity, and moral reasoning tend to generate larger semantic differences across personas, whereas questions related to public health and biomedical ethics generally produce more homogeneous responses. Gemma exhibits the strongest inter-profile variation, indicating that persona conditioning substantially affects both the content and framing of its responses. Nemotron also shows clear persona-driven differentiation, although to a lesser extent. In contrast, DeepSeek displays less systematic behavior: while some questions produce substantial variation across personas, this variation is not consistently associated with persona identity and is accompanied by lower intra-profile coherence.

These findings suggest that semantic adaptation is not uniformly triggered by persona prompts. Instead, its magnitude depends both on the underlying model and on the nature of the question. Topics involving personal values, belief systems, and social identity appear particularly sensitive to persona conditioning, whereas questions associated with broadly shared norms tend to elicit more similar responses across profiles.

Taken together, the inter-profile results complement the lexical and intra-profile analyses. Gemma and Nemotron exhibit clear persona-conditioned semantic differentiation, while DeepSeek shows weaker and less consistent adaptation. This confirms that personalization extends beyond lexical variation and can influence the underlying semantic structure of model responses.

5.5 Personalization Index

Table 5 reports the Personalization Index for each model. This metric provides a unified view of persona-conditioned behavior by combining intra-profile consistency and inter-profile divergence.

The results reveal a clear ranking across models, confirming that models differ significantly in their ability to simulate personas, highlighting a continuum from weak to strong persona-conditioned behavior.

Gemma achieves the highest score, indicating strong persona-dependent behavior both in terms of internal coherence and differentiation across profiles. This suggests that Gemma consistently adopts distinct perspectives while maintaining stable reasoning within each persona.

Nemotron also demonstrates a high degree of personalization, although slightly lower than Gemma. This indicates a more stable but less flexible adaptation, where persona-specific variation is present but more constrained within well-defined semantic boundaries.

In contrast, DeepSeek exhibits a substantially lower score, reflecting weaker persona conditioning. This result is consistent with previous analyses, which showed limited lexical differentiation and lower intra-profile coherence, suggesting that persona information has a reduced impact on the model’s responses.

Table 5: Personalization Index across models (higher values indicate stronger personalization)

Model	Personalization Index
Gemma4:26B	0.867
Nemotron-3-Nano:30B	0.816
DeepSeek-R1:14B	0.663

5.6 Bias Analysis: The Catholic Persona (P1)

To further investigate the relationship between persona conditioning and model bias, we analyze the responses generated under the Catholic persona (P1). We selected this profile because it is associated with a well-defined value system and consistently produces substantial variation across models, making it suitable for examining how different LLMs interpret the same persona.

Figure 3 reveals marked differences in response patterns. Nemotron exhibits a strong tendency toward affirmative answers, producing 16 ‘Yes’ responses, whereas DeepSeek shows the opposite behavior, with 14 ‘No’ responses. Gemma displays a third pattern, characterized by a large number of *ND* responses, suggesting greater uncertainty or reluctance to adopt a clear position under this persona condition.

These differences are also visible at the question level. For example, on Q1 (state invasion for national security), Nemotron answers ‘Yes’, DeepSeek answers ‘No’, and Gemma provides no binary decision. Similar divergences emerge across several religious, political, and ethical questions, indicating that persona conditioning does not produce a uniform interpretation of the assigned identity.

The results suggest that personas are not mapped to a shared semantic representation across models. Instead, persona-conditioned behavior appears to emerge from the interaction between prompt

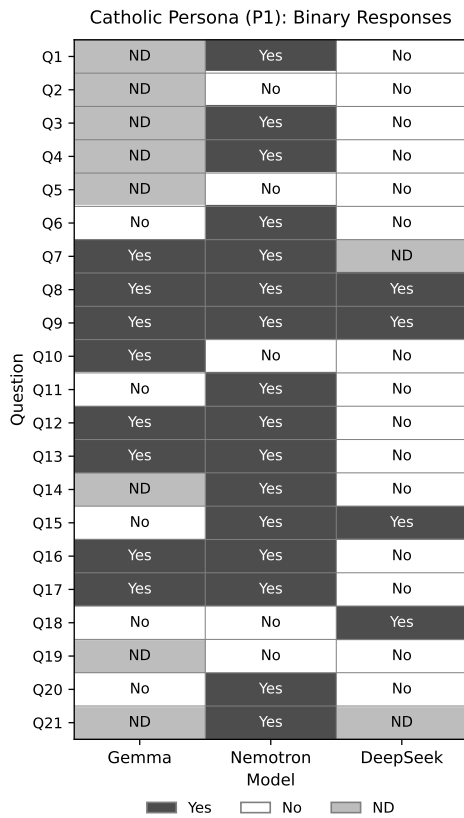


Figure 3: Categorical heatmap of binary responses generated under the Catholic persona (P1). Dark gray indicates Yes, white indicates No, and light gray indicates ND. The figure highlights the substantial variation in how different LLMs interpret the same persona across the evaluated questions.

instructions and model-specific internal representations. Consequently, the same persona may induce substantially different—and potentially biased—responses depending on the model, highlighting an important limitation of prompt-based personalization.

6 Discussion

Our results show that persona conditioning in large language models is a more complex phenomenon than simple prompt following. While previous work has demonstrated that persona prompting can steer model behavior, our findings reveal that the same persona does not correspond to a unique or shared interpretation across models. Instead, different LLMs construct **model-specific interpretations** of the same persona, leading to divergent—and potentially biased—responses even when identical prompts are used. This observation is consistent with recent literature. Liu et al. [7] show that steerability varies across models in open-ended persona-steered generation, while Lutz et al. [8] demonstrate that both prompt formulation and model choice substantially affect sociodemographic persona simulations. Similarly, previous studies

report that persona and demographic conditioning can alter semantic content and introduce biases that vary across models and datasets [5, 14].

Our work extends these findings by proposing a structured multi-level evaluation framework for persona-conditioned behavior. Rather than focusing on a single aspect of personalization, we jointly analyze decision-level variation, lexical and TF-IDF-based framing, intra-profile semantic coherence, inter-profile divergence, and a unified Personalization Index. Existing studies typically evaluate persona prompting through isolated dimensions, such as stylistic adaptation or semantic similarity. By combining multiple complementary analyses, our framework uncovers patterns that would otherwise remain hidden, including inconsistencies across models, trade-offs between coherence and differentiation, and the extent to which persona conditioning is stable or context-dependent. Consequently, our approach provides a more comprehensive and quantitative characterization of persona-conditioned behavior.

6.1 Benefits of personalization

High levels of personalization can provide several advantages. First, responses become more relevant to the assumed user perspective, increasing their perceived appropriateness and usefulness. Second, models improve communicative effectiveness by adopting language, framing, and reasoning styles aligned with the persona, thereby enhancing engagement and comprehension.

Furthermore, persona conditioning can support simulation and reflection by allowing users to explore how different individuals or groups might interpret the same issue. In this sense, personalization represents a valuable tool for exploratory analysis, education, and decision support.

6.2 Risks of over-personalization

The same mechanisms that enable personalization also introduce potential risks. One concern is the emergence of echo chambers or filter bubbles, where the model reinforces persona-consistent assumptions rather than exposing alternative viewpoints. This effect may reduce the diversity of perspectives, particularly in domains such as ethics, politics, and religion, where pluralism is essential. Similarly, excessive alignment with a persona may produce incomplete or unbalanced representations of complex issues. By emphasizing a single perspective, the model may omit relevant arguments, reducing the completeness and reliability of its responses. This is especially problematic in educational and informational settings, where exposure to multiple viewpoints is fundamental for critical thinking.

Finally, strong personalization may hinder intellectual exploration if the model consistently confirms persona-aligned positions. Rather than challenging assumptions or encouraging reflection, the model risks becoming a mechanism for reinforcement. Consequently, the desirable level of personalization depends on the intended application and should not be regarded as universally beneficial.

6.3 Research contributions and implications

From a broader perspective, our findings suggest that persona conditioning should not be viewed as a simple instruction-following

mechanism. Instead, it emerges from an interaction between the prompt and the model’s internal representations. The prompt specifies *who* the model should become, but the model itself determines *what that persona means*. As a result, identical persona prompts can produce systematically different interpretations across models. This observation has important implications for the design of personalized AI systems. Rather than treating personalization as a fixed capability, future systems should provide adaptive control over the degree of persona influence depending on the application. For example, educational assistants may deliberately expose users to alternative viewpoints, whereas personal assistants may prioritize stronger alignment with user preferences.

More generally, our results highlight the close relationship between personalization and bias. The same mechanisms that improve contextual adaptation can also amplify model-specific assumptions and stereotypes. Understanding and controlling this trade-off is therefore essential for the responsible deployment of persona-conditioned LLMs.

6.4 Limitations and future work

This study has some limitations. First, experiments were conducted using prompt-based personas in a stateless setting and evaluated on three open-weight models. Although this design ensures reproducibility and experimental control, future work should investigate conversational settings, real user profiles, and proprietary LLMs. Second, each model–persona–question combination was generated only once. Since responses were produced using stochastic decoding (temperature = 0.7), part of the observed variability may be attributable to sampling rather than persona conditioning alone. Future studies should therefore include repeated generations to estimate decoding variance and assess the robustness of the proposed metrics. Finally, while the proposed Personalization Index provides a unified measure of persona-conditioned behavior, additional validation is needed across different tasks, domains, and prompting strategies. Likewise, the proposed measure of intra-persona consistency captures recurring semantic framing rather than direct semantic equivalence. Alternative formulations based on repeated generations, stance consistency, or reasoning-pattern similarity represent promising directions for future research.

7 Conclusions

We investigated persona-conditioned behavior in LLMs using a multi-level evaluation framework combining decision-level, lexical, and semantic analysis.

Our results show that personalization is a graded and model-dependent phenomenon. Gemma exhibits the strongest adaptation, Nemotron shows consistent but more constrained behavior, and DeepSeek displays weaker and less stable personalization. More importantly, persona conditioning does not yield a uniform interpretation of identity. Instead, it emerges from an interaction between prompts and model-specific representations, leading to divergent responses for the same persona across models. These findings suggest that personalization is not fully controllable and may introduce inconsistency and bias. Future work should extend this analysis to conversational settings and develop methods to better control the trade-off between adaptation and diversity.

References

- [1] Jin Chen, Zheng Liu, Xu Huang, Chenwang Wu, Qi Liu, Gangwei Jiang, Yuanhao Pu, Yuxuan Lei, Xiaolong Chen, Xingmei Wang, et al. 2024. When large language models meet personalization: Perspectives of challenges and opportunities. *World wide web* 27, 4 (2024), 42.
- [2] Yassine El Boudouri, Walter Nuninger, Julian Alvarez, and Yvan Peter. 2025. Role-playing evaluation for large language models. In *International Conference in Methodologies and intelligent Systems for Technology Enhanced Learning*. Springer, 118–127.
- [3] Marco Furini, Michele Mariani, Sara Montagna, and Stefano Ferretti. 2024. Conversational skills of llm-based healthcare chatbot for personalized communications. In *Proceedings of the 2024 International Conference on Information Technology for Social Good*. 429–432.
- [4] Jana Gonnermann-Müller, Jennifer Haase, Nicolas Leins, Thomas Kosch, and Sebastian Pokutta. 2026. Maintaining Stable Personas? Examining Temporal Stability in LLM-Based Human Simulation. In *Proceedings of the Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems*. 1–6.
- [5] Shashank Gupta, Vaishnavi Shrivastava, Ameet Deshpande, Ashwin Kalyan, Peter Clark, Ashish Sabharwal, and Tushar Khot. 2024. Bias runs deep: Implicit reasoning biases in persona-assigned llms. In *International Conference on Learning Representations*, Vol. 2024. 21849–21874.
- [6] Hang Jiang, Xijie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2024. PersonaLLM: Investigating the Ability of Large Language Models to Express Personality Traits. In *Findings of the Association for Computational Linguistics: NAACL 2024*. Association for Computational Linguistics, 3605–3627. <https://doi.org/10.18653/v1/2024.findings-naacl.229>
- [7] Andy Liu, Mona Diab, and Daniel Fried. 2024. Evaluating Large Language Model Biases in Persona-Steered Generation. In *Findings of the Association for Computational Linguistics: ACL 2024*. Association for Computational Linguistics, 9832–9850. <https://doi.org/10.18653/v1/2024.findings-acl.586>
- [8] Marlene Lutz, Indira Sen, Georg Ahnert, Elisa Rogers, and Markus Strohmaier. 2025. The Prompt Makes the Person(a): A Systematic Evaluation of Sociodemographic Persona Prompting for Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (Eds.). Association for Computational Linguistics, Suzhou, China, 23212–23237. <https://doi.org/10.18653/v1/2025.findings-emnlp.1261>
- [9] Roberto Navigli, Simone Conia, and Björn Ross. 2023. Biases in Large Language Models: Origins, Inventory, and Discussion. *J. Data and Information Quality* 15, 2, Article 10 (June 2023), 21 pages. <https://doi.org/10.1145/3597307>
- [10] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 3982–3992.
- [11] Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2024. LaMP: When Large Language Models Meet Personalization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 7370–7392.
- [12] Joni Salminen, Chang Liu, Wenjing Pian, Jianxing Chi, Essi Häyhänen, and Bernard J. Jansen. 2024. Deus Ex Machina and Personas from Large Language Models: Investigating the Composition of AI-Generated Persona Descriptions. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM. <https://doi.org/10.1145/3613904.3642036>
- [13] Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. Role Play with Large Language Models. *Nature* 623 (2023), 493–498.
- [14] Bryan Chen Zhengyu Tan and Roy Ka-Wei Lee. 2025. Unmasking Implicit Bias: Evaluating Persona-Prompted LLM Responses in Power-Disparate Social Scenarios. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Luis Chiruzzo, Alan Ritter, and Lu Wang (Eds.). Association for Computational Linguistics, Albuquerque, New Mexico, 1075–1108. <https://doi.org/10.18653/v1/2025.naacl-long.50>
- [15] Mohamad Saleh Torkestani, Ali Alameer, Shivakumara Palaiahnakote, and Taha Manosuri. 2025. Inclusive Prompt Engineering for Large Language Models: A Modular Framework for Ethical, Structured, and Adaptive AI. *Artificial Intelligence Review* 58, 348 (2025). <https://doi.org/10.1007/s10462-025-11330-7>
- [16] Nicolás Torres, Catalina Ulloa, Ignacio Araya, Matías Ayala, and Sebastián Jara. 2025. A Comprehensive Analysis of Gender, Racial, and Prompt-Induced Biases in Large Language Models. *International Journal of Data Science and Analytics* 20 (2025), 3797–3834. <https://doi.org/10.1007/s41060-024-00696-6>
- [17] Noah Wang, Z.y. Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhuan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhao Huang, Jie Fu, and Junran Peng. 2024. RoleLLM: Benchmarking, Eliciting, and Enhancing Role-Playing Abilities of Large Language Models. In *Findings of the Association for Computational Linguistics: ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand.