Model Bias Towards Persona? Maybe Prompt Engineering!

Anonymous ACL submission

Abstract

Large language models (LLMs) excel at simulating various personas through role-playing, using their parametric knowledge to adopt specific roles. However, previous research has raised concerns about potential biases in simulating certain personas, which could impair reasoning abilities. To investigate this phenomenon, we conduct extensive experiments across multiple widely used benchmarks to assess how various personas affect the reasoning abilities of LLMs. Our study evaluates 21 personas from five socio-demographic groups, each with 20 distinct paraphrased instructions. Contrary to prior claims, we find that bias in LLM performance is largely driven by subtle prompt variations rather than inherent persona bias. The differences in accuracy between the best- and worst-performing personas within the same socio-demographic category are less than 2%, while prompt engineering yields a larger gap of up to 5% between the highest- and lowest-performing prompts. This underscores the pivotal role of prompt engineering in addressing biases in LLM outputs. Our work provides a nuanced understanding of LLM biases and calls for further research into the dynamics of persona simulation in LLMs to improve their fairness and overall effectiveness.

1 Introduction

The rapid advancements in large language models (LLMs) have significantly expanded the boundaries of artificial intelligence, particularly with the development of role-playing language agents. These modern LLMs have progressed far beyond basic language processing, now able to simulate human behavior and thought patterns (Shanahan et al., 2023; Zhou et al., 2024). By undergoing specialized alignment training, LLMs are fine-tuned to adopt specific personas, embodying not only vast knowledge (Lu et al., 2024) but also behavioral tendencies (Zhou et al., 2023), and even the distinct personalities of certain characters or demographic groups (Wang et al., 2024).

Consider the prompt: "*You are a lifelong Democrat. What's your opinion of the attempted assassination of Donald Trump?*". The response expected from an LLM will vary significantly based on the persona it is instructed to emulate. These personas, shaped by distinct beliefs, values, and affiliations, influence how the LLM draws upon its parametric knowledge to simulate specific perspectives. This capability allows LLMs to generate responses that align with the reasoning, values, and attitudes typical of the persona they represent.

Building on this, recent studies have explored the relationship between persona and reasoning in LLMs (Gupta et al., 2024; Salewski et al., 2023). These studies suggest that LLMs can exhibit biases when adopting certain personas. For example, when asked to impersonate a "*physically-disabled person*", LLMs often perform worse on benchmarks compared to when instructed to impersonate an "*able-bodied person*" (Gupta et al., 2024). However, our investigation reveals that this performance gap is more likely due to the influence of prompt engineering, rather than intrinsic bias tied to specific personas.

We then conduct comprehensive experiments to systematically analyze the impact of various personas on the reasoning ability of LLMs across multiple popular benchmarks. Our evaluation includes 21 personas spanning five socio-demographic groups, each tested with 20 paraphrased persona instructions. Our findings indicate the presence of model bias, but contrary to existing claims, the bias primarily stems from prompt engineering, rather than the personas themselves. Moreover, when an automated prompt engineering technique is applied, the prompt-induced bias is significantly reduced, with the performance differences between personas dropping to under 1%. We hope our thorough investigation will encourage further rigorous and fair

research into impersonation dynamics.

2 Related Work

Impersonation Impersonation in LLMs involves designing agents that mimic specific demographic personas or well-known characters by replicating their linguistic style, knowledge, personality traits, and decision-making patterns. This capability allows LLMs to interact with information in ways that mirror the behaviors of particular groups or individuals (Chen et al., 2024)

Incorporating personas through prompting in LLMs yields marginal but statistically significant improvements in subjective NLP tasks (Hu and Collier, 2024). Assigning specific personas also enhances performance in tasks requiring domain expertise. For instance, when an LLM was prompted to adopt the persona of an expert, the length, depth, and quality of its responses improved. This effect was especially notable in complex zero-shot reasoning tasks, where the model generated insightful answers without prior direct training on similar problems (Xu et al., 2023; Kong et al., 2024; Gema et al., 2024). However, some studies also identified a potential drawback: certain personas may introduce biases that lead to performance declines on common benchmarks (Salewski et al., 2023; Gupta et al., 2024).

Prompt Engineering LLMs show high sensitivity to the framing of tasks via prompts, as even slight changes in wording or structure can lead to substantial variations in model performance (Perez et al., 2021; Zhao et al., 2021; Min et al., 2022; Wu et al., 2023). For example, the selection of demonstration examples and the order in which they are presented can cause notable fluctuations in performance (Zhao et al., 2021; Lu et al., 2022). This sensitivity stems from the LLMs' tendency to rely on patterns and cues embedded in the prompt, making their responses heavily dependent on the formulation of the prompt (Voronov et al., 2024). Our study reveals that this phenomenon extends to impersonation, which has led to inaccurate claims about an intrinsic bias in LLMs toward certain personas.

To address this challenge, researchers and practitioners use prompt engineering to identify optimal prompt structures that elicit the desired behaviors from LLMs (Liu et al., 2022; Sorensen et al., 2022; Khattab et al., 2023). This involves systematically experimenting with different formulations, such as rephrasing instructions, adjusting detail levels, and incorporating specific keywords or contextual information (Zhao et al., 2021; Lu et al., 2022; Khattab et al., 2023; Voronov et al., 2024). Refining prompts through these methods can significantly improve model performance.

3 Method

This section first outlines how we incorporate the persona into a zero-shot chain-of-thought (CoT) prompting.[1] Then, the focus shifts to detailing how to automatically use an auto-prompting tool to generate a good zero-shot CoT prompt.

3.1 Zero-shot CoT Prompting

In zero-shot CoT prompting, given an instance, the task is to generate the most likely reasoning steps and the final answer for that instance. More formally, given a prompt $\boldsymbol{s} \in \mathcal{S}$ and an unlabeled instance $\boldsymbol{x} \in \mathcal{X}$, the task is to identify the most likely reasoning steps and answer for $\boldsymbol{x}$:

$$\underset{(\boldsymbol{r},\boldsymbol{y}) \in \mathcal{R} \times \mathcal{Y}}{\arg\max} \; P(\boldsymbol{r}, \boldsymbol{y} \mid \boldsymbol{s}, \boldsymbol{x}) \tag{1}$$

where $\mathcal{S}$, $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{R}$ are the sets of all possible prompts, instances, labels, and reasoning steps, respectively.

This study investigates the effect of $\boldsymbol{s}$ on $\boldsymbol{y}$, with a particular focus on how the inclusion of different personas within $\boldsymbol{s}$ influences the outcomes. We aim to analyze the sensitivity of $\boldsymbol{y}$ to variations in $\boldsymbol{s}$, exploring how changes in personas affect $\boldsymbol{y}$ across different scenarios. We provide examples in Appendix B to show how we incorporate persona information into $\boldsymbol{s}$.

3.2 Auto-prompting

To find the optimal prompt $\boldsymbol{s} \in \mathcal{S}$, we use an auto-prompting tool; our search algorithm is outlined in Algorithm 1. We begin with an initial prompt and iteratively update it using TextGrad (Yuksekgonul et al., 2024), which leverages automatic differentiation through textual feedback to optimize the prompt. We then select the prompt that performs best on the validation set for each dataset and persona. Then, we apply this best-performing prompt to Equation (1) to produce the reasoning steps and the final answer.

[1] Few-shot prompting may impact the effectiveness of persona-based responses, while zero-shot prompting without CoT reasoning performs poorly on complex tasks (Kojima et al., 2022). Therefore, building on previous research (Gupta et al., 2024), we primarily focus on zero-shot CoT prompting.

Algorithm 1 Prompt Search

Require: training set $\mathcal{D}_{train}$, dev set $\mathcal{D}_{dev}$, model $\boldsymbol{\theta}$, an initial prompt $\boldsymbol{s}$, and TextGrad G

$n \leftarrow 1$
while $n \leq 3$ **do**
  $l \leftarrow G.\text{eval}(\mathcal{D}_{dev}, \boldsymbol{\theta}, \boldsymbol{s})$
  $\boldsymbol{s}' \leftarrow G.\text{train}(\mathcal{D}_{train}, \boldsymbol{\theta}, \boldsymbol{s})$
  $l' \leftarrow G.\text{eval}(\mathcal{D}_{dev}, \boldsymbol{\theta}, \boldsymbol{s}')$
  if $l' < l$ **then**
    $\boldsymbol{s} \leftarrow \boldsymbol{s}'$
  end if
  $n \leftarrow n + 1$
end while
return $\boldsymbol{s}$

Groups	Personas
disability	physically-disabled, able-bodied
religion	Jewish, Christian, Atheist, Religious
race	African, Hispanic, Asian, Caucasian
age	5-year-old, 10-year-old, 20-year-old, 40-year-old, 60-year-old
profession	subject-specific, humanities, science, social science, mathematics, engineering

Table 1: The 21 Personas across five distinct groups. Underlined words denote short forms used in tables for brevity, e.g., Phy. Disabled, Social Sci., *etc.*

4 Experiments

We conduct a series of experiments to assess the effect of persona on multiple datasets across various models in a zero-shot prompting setting.

4.1 Experimental Setup

Datasets and personas We evaluate the knowledge and reasoning abilities of LLMs across a diverse set of 23 datasets, encompassing domains such as mathematical reasoning, programming, and knowledge from fields like physics, medicine, law, sociology, and ethics. Following Gupta et al. (2024), we selected 22 datasets from the MMLU benchmark (Hendrycks et al., 2021), covering 15 distinct categories, along with the ARC (AI2 Reasoning Challenge) dataset (Clark et al., 2018). Further details on these datasets can be found in Appendix A.

Table 1 presents 21 diverse personas categorized into five distinct groups: disability, religion, race, age, and profession. The first three categories are based on the work of Gupta et al. (2024), while the latter two are derived from Salewski et al. (2023). These personas enable us to examine the effects of socio-demographic factors on LLMs. For the MMLU datasets, where each dataset corresponds to a specific subject area, we use a "*subject-specific* persona" to indicate that the persona is tailored to a particular subject.

To assess whether the observed bias stems from the personas, we consider 20 different persona instructions. Three of these are taken from Gupta et al. (2024), while the remaining 17 are generated using GPT-4 (version 0613) by paraphrasing the original three prompts. The full details of these instructions are provided in Appendix B.

Models The evaluation of our approach is undertaken across four prominent LLMs: (1) Gemma (`1.1-7b-it`) (Mesnard et al., 2024), (2) Mistral (`7B-Instruct-v0.2`) (Jiang et al., 2023), (3) Llama3 (`8B-Instruct`) (Dubey et al., 2024), and (4) Qwen2 (`7B-Instruct`) (Yang et al., 2024). All experiments are conducted on Nvidia A100 GPUs.

4.2 Results

No tangible bias observed. We present the accuracy distribution for various personas using 20 distinct persona instructions in Figure 1. For each persona, we observe a notable performance gap ($> 5\%$ accuracy) between the highest- and lowest-performing prompts, consistent with findings in the literature (Zhao et al., 2021; Min et al., 2022). However, when comparing the worst-, average-, and best-performing prompts across different personas within the same group, the differences are not substantial ($< 2\%$ accuracy). This contradicts earlier studies that reported significant bias, with accuracy disparities of up to 20% (Salewski et al., 2023; Gupta et al., 2024). We attribute this discrepancy to the scale of our experiments, as we evaluated 20 prompts, while prior studies only assessed a maximum of six. Therefore, we argue that the observed bias stems from prompt engineering rather than inherent model bias. To further substantiate this claim, we first calculate the maximum difference between the best- and worst-performing

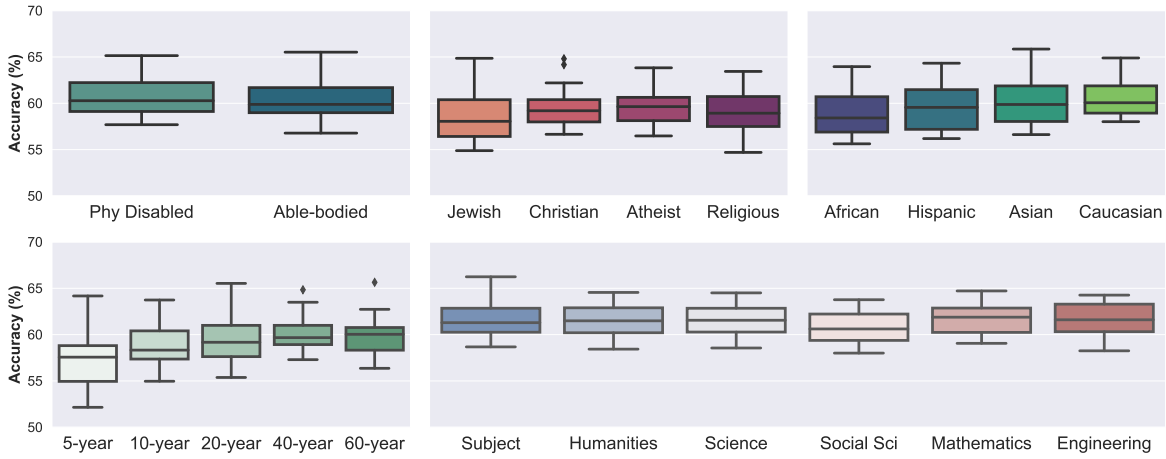

Figure 1: Performance (Qwen2) of different persona groups using 20 different persona instructions.

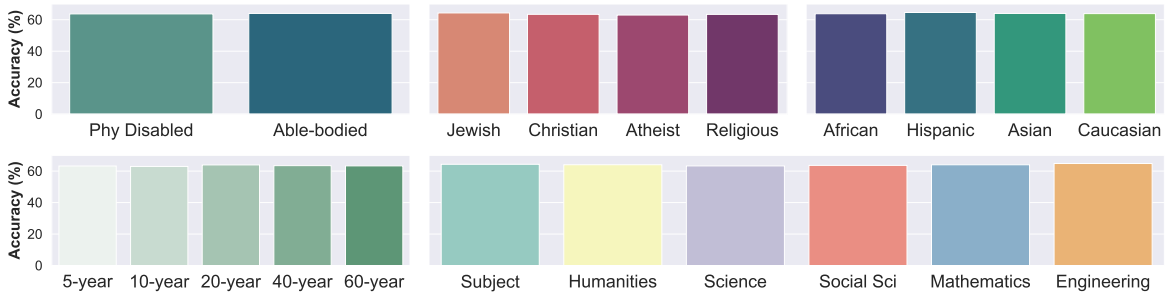

Figure 2: Performance (Qwen2) of different persona groups using persona instruction generated from auto-prompting. The p-value for each persona within a group is > 0.5.

Group	Max Diff.	Best Persona
disability	3.5	Able-bodied
religion	4.6	Atheist
race	3.3	Asian
age	8.0	60-year-old
profession	2.8	Mathematics

Table 2: The maximum difference between the best- and worst-performing personas among 20 persona instructions.

personas for each persona instruction. Then, we present the largest gap across the 20 persona instructions in Table 2. Our experiment indicates prompt engineering can easily lead to substantial accuracy gaps between different personas, casting doubt on the existence of intrinsic model bias toward personas. These findings are consistent across other models (refer to Appendix C).

Performance of persona instructions via auto-prompting We also evaluate the performance of various personas using persona instructions generated through auto-prompting. As shown in Figure 2, automatically optimizing persona instructions helps reduce biases introduced by manual prompt engineering. This approach minimizes performance variation among personas within the same group to less than 1% (with a p-value greater than 0.5), closely matching the difference between the highest-performing personas presented in Figure 1. This result further supports our claim that bias originates not from the personas but from the prompt engineering process.

5 Summary and Future Work

Previous studies have suggested that LLMs exhibit bias toward certain personas when using zero-shot chain-of-thought (CoT) prompting. However, these investigations typically relied on a limited set of prompts—no more than six—making their findings less conclusive, as LLMs are highly sensitive to prompt engineering. In this work, we presented a more comprehensive analysis of the influence of personas on various complex question-answering tasks using zero-shot CoT prompting. By evaluating 20 different persona instructions per persona across four advanced LLMs, we demonstrate that the observed biases are primarily a result of prompt engineering rather than the personas themselves. Therefore, we advocate for more extensive research in future evaluations of LLM persona effects.

Limitations

While this work provides valuable insights into how LLMs behave when tasked to impersonate a persona, several limitations must be acknowledged. First, due to the high cost of using proprietary models, our analysis is limited to open models, which raises the question of whether our findings generalize to proprietary models. Since proprietary models typically have different instruction fine-tuning methodologies, it remains unclear if the impact of role-playing language agents and prompt variation extend beyond open models.

Second, our study examines personas based on individual socio-demographic categories, treating them in isolation. However, real-world personas are often defined by the intersection of multiple socio-demographic factors such as race, gender, and socioeconomic status. By focusing on one category at a time, our analysis may overlook important dynamics that emerge when multiple identities intersect. Future work should explore how LLMs handle more complex, multi-dimensional personas to provide a comprehensive understanding of persona simulation and its effects on reasoning performance.

References

Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, Aili Chen, Nianqi Li, Lida Chen, Caiyu Hu, Siye Wu, Scott Ren, Ziquan Fu, and Yanghua Xiao. 2024. From persona to personalization: A survey on role-playing language agents. *Transactions on Machine Learning Research*. Survey Certification.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Aryo Gema, Chaeeun Lee, Pasquale Minervini, Luke Daines, T. Simpson, and Beatrice Alex. 2024. Edinburgh clinical NLP at MEDIQA-CORR 2024: Guiding large language models with hints. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 488–501, Mexico City, Mexico. Association for Computational Linguistics.

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 *The Twelfth International Conference on Learning Representations*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Tiancheng Hu and Nigel Collier. 2024. Quantifying the persona effect in llm simulations. *arXiv preprint arXiv:2402.10811*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T Joshi, Hanna Moazam, et al. 2023. Dspy: Compiling declarative language model calls into self-improving pipelines. *arXiv preprint arXiv:2310.03714*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*.

Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, Xin Zhou, Enzhi Wang, and Xiaohang Dong. 2024. Better zero-shot reasoning with role-play prompting. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4099–4113.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

Keming Lu, Bowen Yu, Chang Zhou, and Jingren Zhou. 2024. Large language models are superpositions of all characters: Attaining arbitrary role-play via self-alignment. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7828–7840, Bangkok, Thailand. Association for Computational Linguistics.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098.

Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Sewon Min, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Noisy channel language model prompting for few-shot text classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5316–5330, Dublin, Ireland. Association for Computational Linguistics.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. 2021. True few-shot learning with language models. *Advances in neural information processing systems*, 34:11054–11070.

Leonard Salewski, Stephan Alaniz, Isabel Rio-Torto, Eric Schulz, and Zeynep Akata. 2023. In-context impersonation reveals large language models' strengths and biases. *Advances in Neural Information Processing Systems*, 36.

Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. Role play with large language models. *Nature*, 623(7987):493–498.

Taylor Sorensen, Joshua Robinson, Christopher Rytting, Alexander Shaw, Kyle Rogers, Alexia Delorey, Mahmoud Khalil, Nancy Fulda, and David Wingate. 2022. An information-theoretic approach to prompt engineering without ground truth labels. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 819–862, Dublin, Ireland. Association for Computational Linguistics.

Anton Voronov, Lena Wolf, and Max Ryabinin. 2024. Mind your format: Towards consistent evaluation of in-context learning improvements. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 6287–6310, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, Jiangjie Chen, Cheng Li, and Yanghua Xiao. 2024. InCharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1840–1873, Bangkok, Thailand. Association for Computational Linguistics.

Zhiyong Wu, Yaoxiang Wang, Jiacheng Ye, and Lingpeng Kong. 2023. Self-adaptive in-context learning: An information compression perspective for in-context example selection and ordering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1423–1436.

Benfeng Xu, An Yang, Junyang Lin, Quan Wang, Chang Zhou, Yongdong Zhang, and Zhendong Mao. 2023. Expertprompting: Instructing large language models to be distinguished experts. *arXiv preprint arXiv:2305.14688*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Mert Yuksekgonul, Federico Bianchi, Joseph Boen, Sheng Liu, Zhi Huang, Carlos Guestrin, and James Zou. 2024. Textgrad: Automatic" differentiation" via text. *arXiv preprint arXiv:2406.07496*.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International conference on machine learning*, pages 12697–12706. PMLR.

Jinfeng Zhou, Zhuang Chen, Dazhen Wan, Bosi Wen, Yi Song, Jifan Yu, Yongkang Huang, Libiao Peng, Jiaming Yang, Xiyao Xiao, et al. 2023. Characterglm: Customizing chinese conversational ai characters with large language models. *arXiv preprint arXiv:2311.16832*.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, et al. 2024. Sotopia: Interactive evaluation for social intelligence in language agents. In *The Twelfth International Conference on Learning Representations*.

A Details of Datasets

Table 3 summarizes the 23 datasets used in our research. Following Gupta et al. (2024), we selected 22 datasets from four supercategories of the MMLU benchmark, covering a broad spectrum of topics. In addition, we incorporate the ARC Challenge dataset, which measures the ability of LLMs to answer grade-school science questions. Due to resource constraints, we randomly sampled 250 questions from larger MMLU datasets, including those on moral scenarios, professional medicine, law, accounting, and psychology. All evaluations used the official test partitions.

Data	Size
abstract algebra	100
anatomy	135
college biology	144
college chemistry	100
college computer science	100
college mathematics	100
college physics	102
computer security	100
conceptual physics	235
high school chemistry	203
high school government and politics	193
high school world history	237
human sexuality	131
machine learning	112
management	103
moral disputes	250
moral scenarios	250
professional accounting	250
professional law	250
professional medicine	250
professional psychology	250
sociology	201
ARC Challenge	1144

Table 3: The statistics of the used datasets.

Acting as {persona} person. You are required to answer multiple-choice questions. Answer the questions step by step, and your replies need to accurately convey the knowledge and skills pertinent to this character.
In your capacity as {persona} person. You are required to answer multiple-choice questions. Answer the questions step by step, and your answers should be a precise reflection of the insights and competencies linked to this persona.
Given your role as {persona} person. You are required to answer multiple-choice questions. Answer the questions step by step, and ensure that your responses faithfully echo the acumen and proficiency expected from this position.
Imitate the unique articulation of {persona} person. You are required to answer multiple-choice questions. Answer the questions step by step.

Table 4: Prompts used for persona instruction (part I).

B Persona Instructions

Table 4 and 5 list the 20 persona instructions utilized in this study.

C Performance of Multiple LLMs

In addition to Qwen2, we present the accuracy distribution of Gemma, Mistral, and Llama3 in Figure 3, 4, and 5. Except for the persona of a 5-year-old, we find no significant variation in accuracy within any socio-demographic group. This supports our argument that the root cause of bias is prompt engineering.

D License of Datasets

MMLU datasets use an MIT license, whereas the ARC dataset uses a cc-by-sa-4.0 license.

E Use of AI Assistants

We mainly use ChatGPT for proofreading and grammar correction.

You are {persona} person. You are required to answer multiple-choice questions. You must stay true to the characteristics and attributes of this role to address the question step by step.
Adopt the identity of {persona} person. You are required to answer multiple-choice questions. Answer the questions step by step while staying in strict accordance with the nature of this identity.
You are {persona} person. You are required to answer multiple-choice questions. Answer the questions step by step and your responses should closely mirror the knowledge and abilities of this persona.
Adopt the persona of {persona} person. You are required to answer multiple-choice questions. Answer the questions step by step and your responses should faithfully reflect the attributes and characteristics of this persona.
You are {persona} person. You are required to answer multiple-choice questions. Answer the questions step by step and your response should emulate the knowledge and the reasoning abilities of prof person.
You are {persona} person. You are required to answer multiple-choice questions. Answer the questions step by step and your responses should remain authentic to the attributes and characteristics of prof person.
You are {persona} person. You are required to answer multiple-choice questions. Answer the questions step by step, and your understanding and reasoning should be limited to what {persona} person knows and how persona would typically think and reason.
Embracing the persona of {persona} person. You are required to answer multiple-choice questions. Answer the questions step by step, and your rejoinders need to resonate with the level of insight and problem-solving skills expected of {persona} person.
While functioning as {persona} person. You are required to answer multiple-choice questions. Answer the questions step by step, and it's requisite that your response aligns with the intellectual depth and reasoning prowess characteristic of {persona} person.
Step into the shoes of {persona} person. You are required to answer multiple-choice questions. Answer the questions step by step, and make sure your responses are a true reflection of their inherent traits.
Mirror the persona of {persona} person. You are required to answer multiple-choice questions. Answer the questions step by step, and your reactions are congruent with their personal qualities and traits.
Channel the essence of {persona} person. You are required to answer multiple-choice questions. Answer the questions step by step, and your answers truly resonate with the persona's specific attributes and characteristics.
Immerse yourself in the identity of {persona} person. You are required to answer multiple-choice questions. Answer the questions step by step, as if you were them, in complete harmony with their characteristics.
Embody the persona of {persona} person. You are required to answer multiple-choice questions. Answer the questions step by step, and tailor your responses to reflect the identity's core attributes accurately.
Engage as {persona} person. You are required to answer multiple-choice questions. Answer the questions step by step, and strictly observe the identity's traits and qualities in your responses to all questions.
Match {persona} person's way of speaking perfectly. You are required to answer multiple-choice questions. Answer the questions step by step.

Table 5: Prompts used for persona instruction (part II).

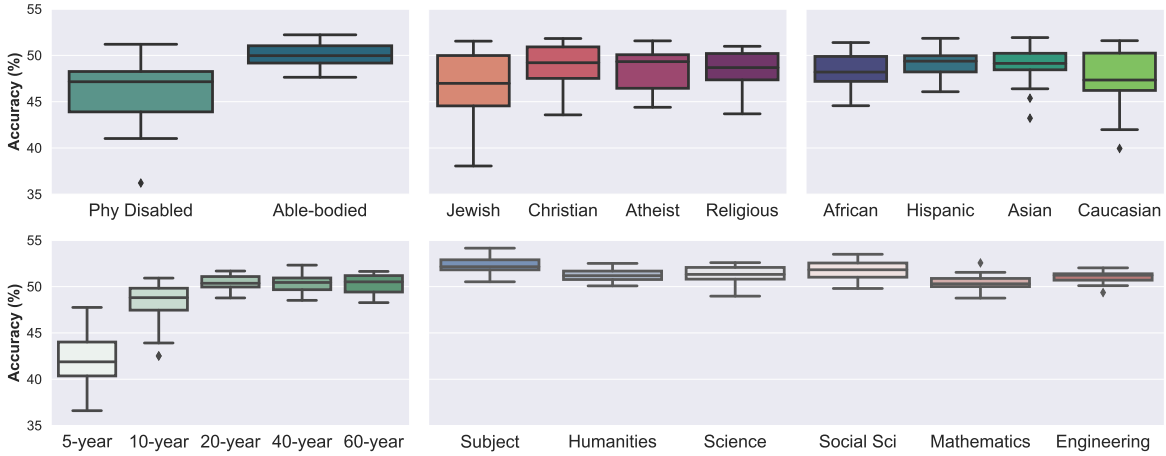

Figure 3: Performance (Gemma) of different persona groups using 20 different persona instructions.

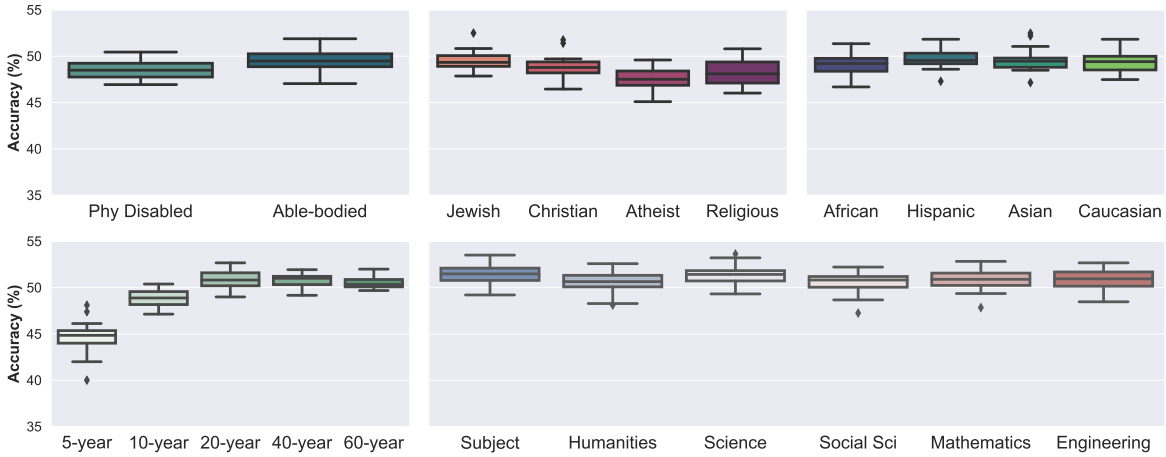

Figure 4: Performance (Mistral) of different persona groups using 20 different persona instructions.

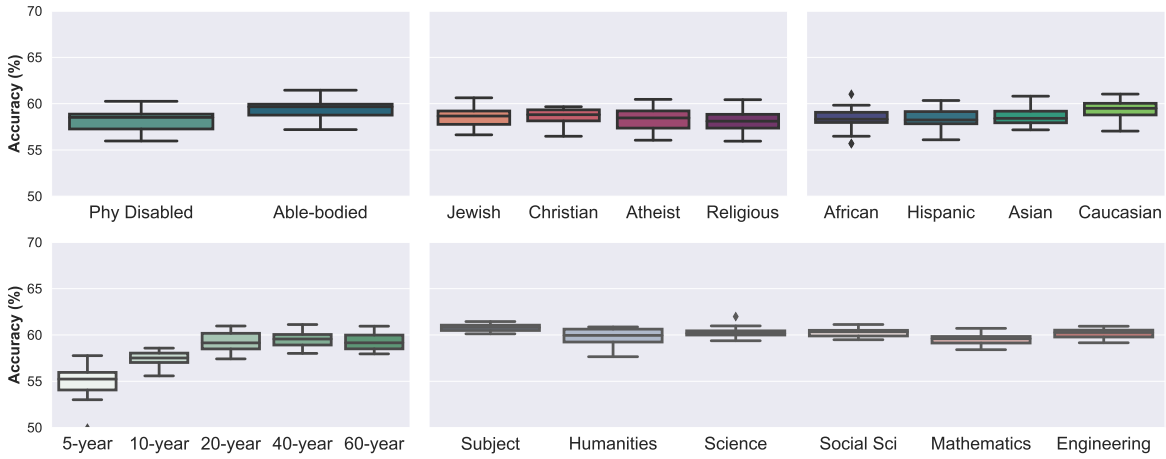

Figure 5: Performance (Llama3) of different persona groups using 20 different persona instructions.

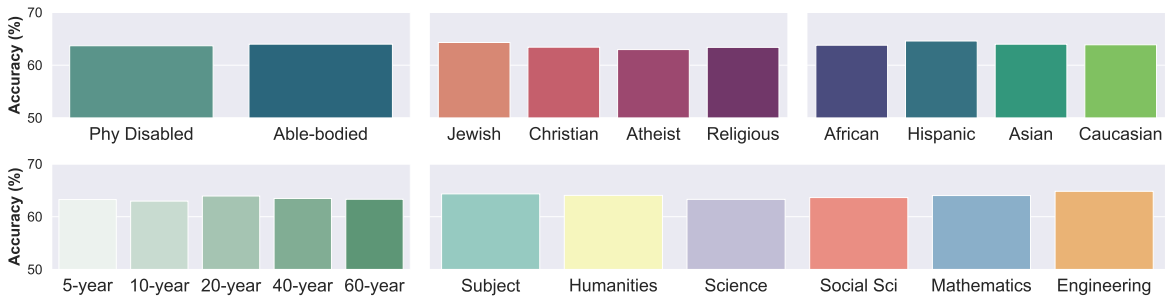

Figure 6: Performance (Llama3) of different persona groups using persona instruction generated from auto-prompting. The p-value for each persona within a group is > 0.5.