When “A Helpful Assistant” Is Not Really Helpful: Personas in System Prompts Do Not Improve Performances of Large Language Models

Anonymous ACL submission

Abstract

Prompting serves as the major way humans interact with Large Language Models (LLM). Commercial AI systems commonly define the role of the LLM in system prompts. For example, ChatGPT uses “You are a helpful assistant” as part of the default system prompt. Despite current practices to add personas in system prompts, it is unclear how different personas affect the models’ performance. In this study, we present a systematic evaluation of personas in system prompts. We create a list of 162 roles covering 6 types of interpersonal relationships and 8 domains of expertise. Through extensive analysis of 4 popular LLMs and 2410 factual questions, we show that adding personas in system prompts does not improve the models’ performance over a range of questions compared with the control setting where no persona is added. Despite this, further analysis suggests that the gender, type, and domain of the persona could all affect the consequential prediction accuracy. We further experimented with a list of persona search strategies and found that while aggregating the results from the best personas for each question could significantly lead to higher prediction accuracies, automatically identifying the best persona is challenging and may not be significantly better than random selection. Overall, our result suggests that while adding persona may lead to performance gain in certain settings, the effect of each persona can be largely random. Code and data are available at AnonymizedURL.

1 Introduction

Building persona- or role-based chatbots has attracted enormous attention from the AI and NLP community due to their potential business and societal applications (Pataranutaporn et al., 2021). Recent advances in LLMs also provide huge opportunities to build intelligent agents that can behave and talk like certain characters or roles (Wang et al., 2023). However, despite all the existing studies on role-playing with LLMs, it is unclear how different types of personas affect LLMs’ performance on objective tasks. To address this gap, we conduct a large-scale analysis of 162 personas over 4 popular open-source LLMs and 2410 factual questions. To ensure the generalizability of the result, the 162 personas were selected from 6 types of interpersonal relationships and 8 domains of expertise. Furthermore, to study the effect of domain alignment between personas and questions, the evaluation question sets were sampled from the Massive Multitask Language Understanding (MMLU) dataset (Hendrycks et al., 2021), balanced for categories.

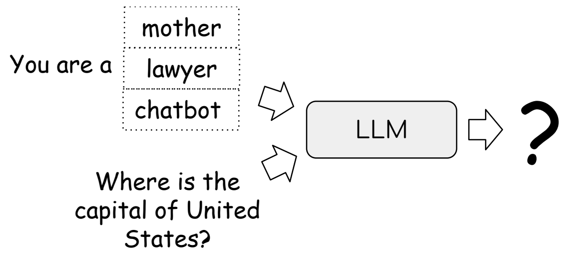


Figure 1: Our overall research question: does adding personas in prompts affect LLMs’ performance?

In this study, we aim to answer three major research questions: (1) Does adding personas in system prompts help to improve model performance? (2) Does the social construct of the persona affect model performance? (3) What factors could potentially explain the effect of personas on model performance? (4) Can we automatically identify the best roles for prompting? Through our analysis, we find that, in general, prompting with personas has no or only small negative effects on the model performance compared with the control setting where no persona is added. This result is consistent across four popular LLMs, suggesting that adding personas to system prompts may not help to improve the model’s performance. To further understand the relative differences among personas, we analyze the social attributes of personas, including role type, gender, and domain alignment. We find that gender-neutral, in-domain, and school-related roles lead to better performances than other types of roles, but with relatively small effect sizes, suggesting that the social construct of the persona may not fully explain the consequential performance differences.

To understand the potential mechanisms behind the relative performance differences caused by different personas, we further analyze the word frequency of the persona, the perplexity, and the similarity of the prompt-questions pairs. Overall, we observe that personas with high-frequency words lead to relatively better model performances. Furthermore, while the similarity between the persona and the question is the strongest predictor of final performance, the correlation between prompt-question similarity and prediction accuracy remains low. Overall, our results suggest that word frequency, perplexity, and prompt-question similarity may not fully explain the prediction performance differences caused by different personas.

Can we automatically identify the best persona for prompting LLMs? We explore a list of automatic persona search strategies. We find that the effect of persona on model performance is not consistent across questions, making it challenging to identify a persona that can consistently lead to a better inference performance.

Our study makes the following three contributions to the community. First, we introduce a new pipeline to systematically evaluate LLMs' performance when prompted with a wide range of personas. Second, our experiments reveal insights into the complex impact of persona on the model performance and assess several potential influencing factors. Third, our experiments with a wide range of automatic role-searching strategies suggest that the effect of personas on model performance may not be consistent across questions, and identifying the optimal persona for each question is challenging.

2 Related work

Personas and Roles Personas are fundamental in human society and day-to-day interactions (Heiss, 2017; Goffman, 2016). personas define the norm of human interactions and affect human behaviors in various contexts (Sunstein, 1996). Two prominent types of personas are interpersonal roles which are roles embedded in interpersonal relationships (Berscheid, 1994) (e.g. mother and friend) and professional/occupational roles that fulfill certain social functions or provide certain services in society (e.g. driver and teacher) (Bucher and Strauss, 1961; Brante, 1988). As suggested by Wolfensberger (2000), "People largely perceive themselves and each other in terms of their roles." Given the importance of personas in human interactions and recent advances in persona-based agents (Wang et al., 2023; Pataranutaporn et al., 2021), understanding LLMs' role-playing capabilities and the effect of personas hold significance to both the NLP community and the general public.

Prompting LLM Prompting serves as a unified natural language interface for human-AI interactions and has been widely adopted in the era of LLM (Liu et al., 2023). Existing studies suggest that LLMs are very sensitive to the design of prompts (Lu et al., 2021). For example, adding "Let's think step by step" could help to improve the model performance in answering a wide range of questions (Kojima et al., 2022). How to design prompts that lead to better performances has become an important question for not only NLP researchers but also people in education (Heston and Khun, 2023), art (Oppenlaender, 2022) and health (Meskó, 2023) industries. Furthermore, current AI systems usually insert system prompts before user prompts to ensure the safety and helpfulness of system-generated outputs (Touvron et al., 2023). System prompts usually define the role of the system (e.g. "You are a helpful assistant.") and further guide LLMs' behaviors in user interactions. That is, the system prompt serves as a default setting of LLM products and precedes any user prompt. Thus, even for models that are not instruction-tuned, it is still important to investigate how variously formatted system prompts might impact model performance. Despite its wide usage in commercial AI systems, the effect of using personas in systems prompts has not been fully studied in the current literature.

Role Playing with LLMs Creating agents that are able to talk like certain characters and roles has attracted much attention from the AI and NLP community (Demasi et al., 2020) due to its potential benefits in settings like education (Pataranutaporn et al., 2021), games (Miikkulainen, 2007), and mental health (Denecke et al., 2020). Large language models offer new opportunities in creating persona-based agents through role-playing with LLMs (Shanahan et al., 2023). Existing studies have produced datasets (Qian et al., 2021), prompting strategies (Kong et al., 2023), and evaluation settings (Wang et al., 2023) for role-playing with LLMs. However, when evaluating LLMs' role-playing capabilities, existing studies majorly focus on role- and dialogue-related metrics such as perplexity, coherence, and interestingness (Lin et al., 2020; Deriu et al., 2021). It is still unclear whether role-playing would affect LLMs' capability to handle general language tasks.

3 Experiment Setting

The overall goal of our study is to explore whether adding personas in prompts affects LLMs' performances. To answer this question, we design a series of experiments and this section details the experiment setup.

3.1 Dataset

We use a sample of MMLU (Hendrycks et al., 2021) in all of our experiments. MMLU is a dataset designed for multitask language understanding and has been widely used as an essential benchmark for evaluating LLMs. It features multiple-choice questions that probe knowledge across a diverse set of subjects, ranging from natural sciences and social sciences to business and law. We choose MMLU as our test dataset because (1) it has been widely used for benchmarking LLMs, (2) it contains questions from diverse disciplines, allowing us to test the effect of prompting with domain-aligned personas, and (3) questions from different domains follow similar formats.

Furthermore, to ensure the generalizability of our results, we design a sampling pipeline to balance the length and subject of the question. We first randomly sample 100 instances from each initial subject of MMLU to ensure a diverse representation of questions across subjects. For each sampled instance, we calculate the length of full questions with both question text and four options. To manage the computation cost, we drop questions so that 99% of the sampled questions have less than 150 words. From the filtered dataset, we manually select subjects based on higher popularity and coverage of several broad domains. The final dataset contains 2410 questions from the MMLU dataset, balanced across 26 subjects. We further map the sampled subjects into 8 big categories: Law, Medicine, EECS, Math, Politics, Psychology, Natural Science, and Econ. Table 3 in the Appendix details the subjects and domains.

Prompt Type	Example
No Role	{question}
Speaker-Specific	You are a/an {role}, {question}
Audience-Specific	You are talking to a/an {role}, {question}

Table 1: Types and examples of prompt templates for personas used in our experiment. We further refine the prompt to meet the format requirement of each model and the full prompts are available in the Appendix (Table 7 and Table 8).

3.2 Prompt

Personas can be incorporated into prompts in various ways. We carefully design two types of prompts: (1) **Speaker-Specific Prompt:** prompts that assign the role to the LLM (i.e. "who you are"). For example, "You are a lawyer"; (2) **Audience-Specific Prompt:** prompts that specify the audience of the conversation (i.e. "who you are talking to"). For example, "You are talking to a fireman.". As a comparison, prompts that only include the question are used as the control setting in our experiment. Table 1 shows the template of prompts used in our study. As a robustness check, for each prompt template, we also include an external paraphrased prompt by adding the word "Imagine" (e.g. "Imagine you are talking to a fireman"). We further revise the prompt template to fit into the format requirements of different models to attain the best performances. Table 7 and Table 8 in the Appendix details the prompt we use for each model.

3.3 Persona

To excessively evaluate the effect of personas on model performance, we carefully curate a large and diverse list of personas that are actively used in people's daily interactions. We first collect over 300 personas based on several existing studies (Garg et al., 2018; Massey et al., 2015; Choi et al., 2021), WordNet (Miller, 1995), and our own ad-hoc social role list. We manually examine the roles to remove uncommon roles that are rarely used in daily life, such as "ganger" as a hyponym of "boss". Our final social role set includes 162 personas, of which 112 roles are occupations and the remaining are interpersonal relationship roles. Table 4 in the Appendix shows the full list of roles in our experiment.

Interpersonal Roles Our study includes 50 interpersonal roles grouped into 5 categories: family,

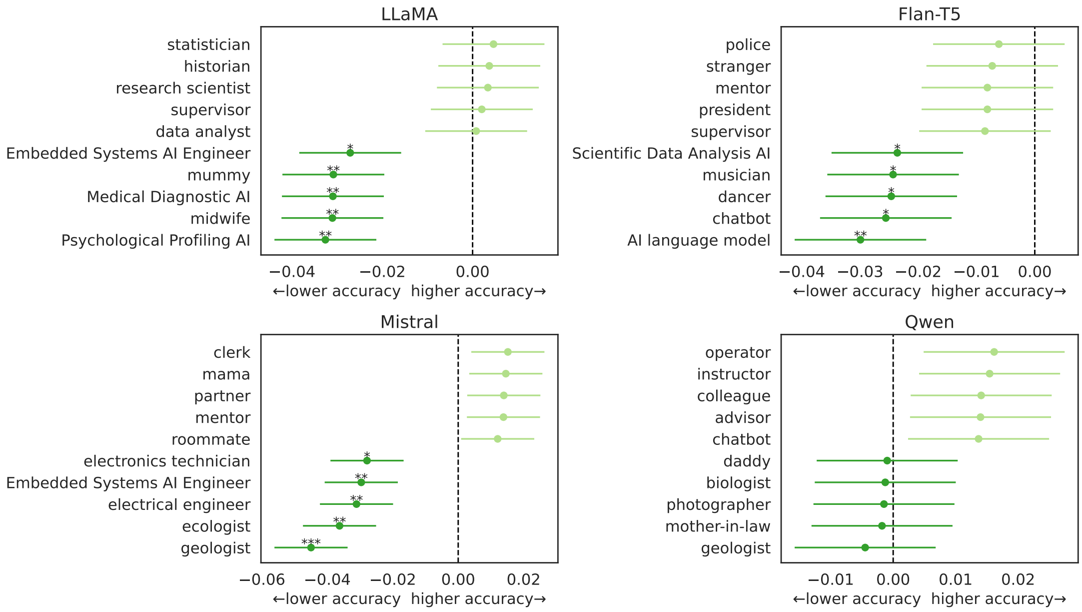


Figure 2: The first and last 5 coefficients ranked by scale of the regression model accuracy∼role for each model.

friend, romantic, work, and school. For important roles that do not fit into the above categories (e.g. stranger), we add them into the category of "social". We further augment the role list by adding hyponyms from WordNet (Miller, 1995) to selected roles as a robustness check. For example, for the word "mother", we also include "mama", "mamma", "mom" and "mommy".

Occupational Roles We compile our set of occupations from Garg et al. (2018). Additionally, we manually add occupations that are relevant to the subjects of the sampled MMLU questions. For example, we add "software engineer" under the category of EECS. Furthermore, given the wide adoption of AI systems in our society, we also include a list of AI roles (e.g. "AI language model" and "AI assistant").

3.4 Models

We experiment with four popular open-source instruction-tuned LLMs whose sizes range from 7B to 11B: FLAN-T5-XXL (Chung et al., 2022), LLaMA-3-8b-Instruct (AI@Meta, 2024), Mistral-7B-Instruct-v0.2 (Jiang et al., 2023) and Qwen-7B-Chat (Bai et al., 2023). All of the four models are fine-tuned to follow instructions, and three of them (except Flan-t5) allow a chat template that contains both a system prompt and a user prompt. We choose open-source models ranging from 7B to 11B majorly because of the following reasons: (1) 7 to 11B open-source models have shown promising performances on a wide range of tasks, especially LLaMA-3 and Qwen. Smaller-size models may not have enough role-playing or instruction-following capabilities; (2) Our experiment requires running inference tasks over 2410 questions with 4 prompt templates and 162 personas, making it computationally and financially expensive to query API-based or bigger models. (3) experimenting with open-source models allows other researchers to easily replicate our experiment results.

4 Does Prompting with Personas Improve LLMs' Performance?

To assess whether adding personas helps improve model performance for answering factual questions, for each model, we fit linear regressions for each model that use the added persona to predict the inference accuracy. The control setting, where no role is added to the system prompts, is used as the reference category. Figure 2 shows the first 5 and last 5 coefficients ranked by scale for each model. The coefficients of all roles are detailed in Section B in the Appendix. We observe no significant differences between the best-performing personas and the control setting. On the contrary, certain personas may actually lead to *lower* performance (e.g., ecologist for Mistral). Furthermore, as shown in Figure 3, most of the personas have no statistically significant effect on the model's prediction accuracy compared with the control setting, and such a pattern is consistent across all four models.

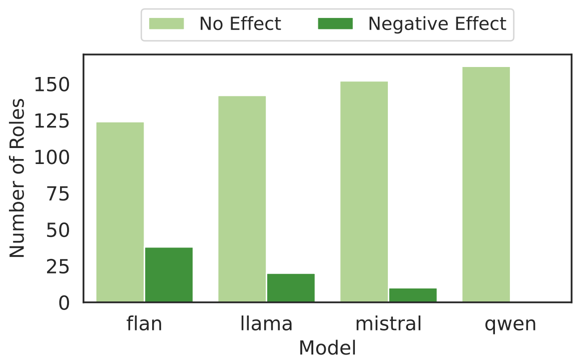


Figure 3: Most of personas have no or negative impact on model performance.

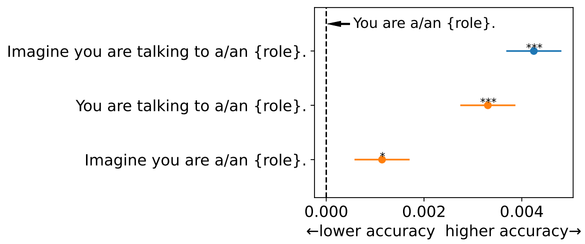


Figure 4: Audience-specific prompts are significantly better than speaker-specific prompts with small effect sizes.

Our results suggest that there might not exist a single persona that can consistently help to improve LLMs' performance across diverse questions.

Does the framing of the prompt affect the model's performance? To answer this question, we run a mixed-effects model on the relationship between accuracy and prompt type, controlling for each model as a random effect. Figure 4 shows the regression coefficients for each prompt template. We observe that audience-specific prompts perform better than speaker-specific prompts, and the difference is statistically significant. However, we must note that the effect size is relatively small, suggesting that different framings of the prompt have limited impacts on model performance.

5 Are Certain Personas Better Than Others?

While adding a persona might not be better than the control setting where no role is added, in practice, LLM service providers or users may still need to define the role of the system for various reasons (e.g., security and language styles). Therefore, it is still worth discussing whether different types of personas could lead to different model performances.

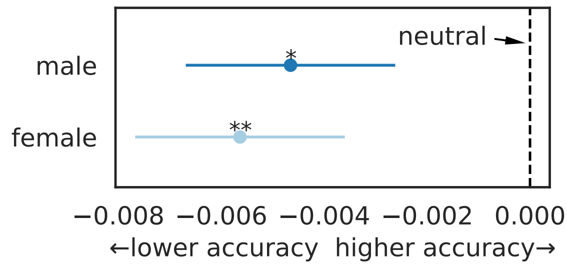


Figure 5: Gender-neutral roles lead to better performances than gendered roles.

Gender Gender roles are one of the most prominent and widely studied personas in the literature of sociology (Blackstone, 2003; Acker, 1992) and society as they are embedded in various types of personas like father and wife. Do LLMs exhibit a tendency whereby a "father" role is more likely to yield accurate responses compared to a "mother" role? To quantify the impact of gender, we assess interpersonal roles and occupational roles separately, by analyzing the explicit and implicit gender impact respectively.

For interpersonal roles, we analyze 16 aligned roles and categorize them as male, female, or neutral, resulting in 7 male roles, 7 female roles, and 2 neutral roles. Table 5 in the Appendix shows the mapping of gender and roles. Such a setting allows us to control the effects of role types and reveal the nuanced effects of gender. We employ a mixed-effects model to analyze the relationship between accuracy and gender, with "accuracy" as the dependent variable, "gender" as an independent categorical variable of values "male", "female" and "neutral", and we include a random effect for each model to account for potential variability across different models. As shown in Figure 5, gender-neutral roles perform significantly better than gendered roles, and male roles perform slightly less worse than female roles with a small effect size.

For occupational roles, we use the percentages of workers belonging to each gender in 65 occupational roles, extracted from historical US census data (Garg et al., 2018). We fit a similar mixed-effects model with the percentage of male workers as the independent variable, and include random intercepts for each model. The p-value associated with "Male", the percentage of male workers for each occupation, is 0.247, indicating that the gender percentage is not a significant predictor of model performance. The results of the two mixed-effects models for gender impact collectively lead

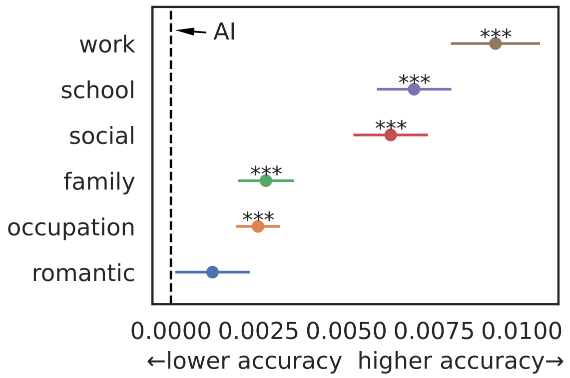


Figure 6: Work- and School-related Roles lead to better performances than other types of roles across models.

to the conclusion that the gender nature of personas has a very limited impact on the models' performance in terms of accuracy.

Role Category The 162 roles are categorized into 7 groups: work, school, social, family, romantic, occupation, and AI. These role categories differentiate the roles based on the social relationships and settings they typically involve. The mixed-effects model shows that the role category is an insignificant predictor of accuracy across models.

Domain Alignment While we observe no significant differences between most of the personas and the control setting, it is possible that certain roles might still lead to better answers for specific questions. For example, many prompt engineering guidebooks suggest adding roles that are aligned with the current conversation context [1]. Do domain-aligned personas really lead to better model performances? To test this question, we label each role-question pair with "in-domain" and "out-domain" based on its category. For example, if the persona is "software engineer" and the question is in Computer Science, we consider it as an in-domain pair.

To assess the effect of domain alignment, we fit another mixed-effects model using the binary in-domain indicator as the sole predictor and include a random effect for each model. The coefficient for "in-domain" is 0.005 ($p < 0.01$), suggesting that in-domain roles generally lead to better performances than out-domain roles. For example, lawyers are more likely to give accurate answers to law-related questions than doctors.

[1] https://llama.meta.com/docs/how-to-guides/prompting/

6 Why Certain Personas Lead to Higher Accuracies?

Why do certain personas lead to better performances than others? Despite the complexity across personas, we assess several potential mechanisms. In this section, we propose a method to calculate persona embedding that enables an overall performance comparison. Furthermore, we test whether specific characteristics of the prompt and personas might be driving the behavior: the n-gram frequency of role words, the perplexity of the context prompts, and the similarity between context prompts and questions.

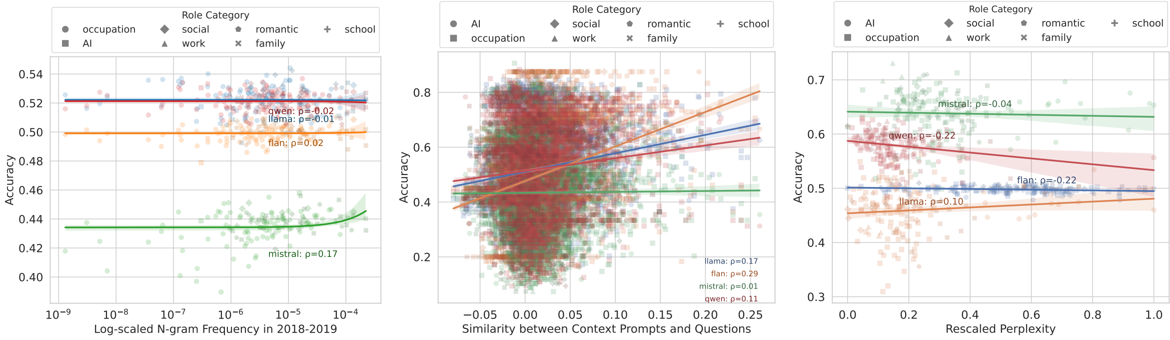
Word Frequency of Personas Model performance could be explained by familiarity with the role word itself in training. Therefore, for each role, we obtain its n-gram frequency for the period between 2018 and 2019 (the most recent data available) from the Google Ngram Viewer [2]. The value of "n" depends on the specific role. For example, for the role "mom", n = 1, and for the role "software engineer", n = 2.

Figure 7a illustrates the aggregated relationship between accuracy and role word frequency for each model, where each point represents a role and is characterized by its role category. Roles' n-gram frequency is weakly correlated to their accuracy, as evidenced by the Pearson correlation coefficients at the role level being 0.17 for Mistral, the highest among the three, suggesting that word frequency does not fully explain the effect of personas on model performances.

Prompt-Question Similarity Are context prompts that closely resemble the questions more likely to generate accurate answers? To answer this question, we utilize MiniLM (Wang et al., 2020) from Sentence-BERT package (Reimers and Gurevych, 2019) to encode a set of context prompts and full questions with options, and then compute the cosine similarity between the two vectors as a measure of distance between the question and prompt.

As shown in Figure 7b, we observe a weakly correlation between similarity and accuracy at the role level. Specifically, the highest correlation is 0.29 on FLAN-T5-XXL, whereas the correlation for Mistral-7B-Instruct-v0.2 is 0.01, suggesting that the effect of similarity might depend on specific models.

[2] https://books.google.com/ngrams/



(a) Word frequency

(b) Prompt-question similarity

(c) Prompt perplexity

Figure 7: (a) personas' word frequency is weakly correlated with model performances. (b) prompt-question similarity shows weak to moderate correlations with the models' performance. (c) The perplexity of the prompt has a negative and weak correlation with the models' performance.

Prompt Perplexity Perplexity quantifies the overall probability of a piece of text for a given language model. It serves as an indicator of the model's uncertainty, with lower perplexity reflecting higher prediction accuracy. We use each model's tokenizer and architecture to compute model-specific perplexities. For FLAN-T5, we use a pair of context prompts and the questions as the input. For the other three models, perplexity is computed for an entire prompt, consisting of a context prompt followed by a question with options. We further rescale the calculated perplexity scores to a range of 0 to 1 to allow easier comparisons across models. As shown in Figure 7c, the mean accuracy is negatively correlated with the rescaled perplexity at the role level on FLAN-T5, Qwen and Mistral, whereas the correlation is positive on LLaMA. These results suggest that logical coherence and inherent reasonability of prompts do not necessarily result in more accurate responses. The impact of perplexity is model-dependent as well.

Overall Regression Analysis To perform a comprehensive analysis of all the attributes of roles mentioned previously, we fit a mixed-effects model using three independent variables: the role's n-gram frequency, prompt-question perplexity, and prompt-question similarity. Random intercepts are included for each model. The model results lead to the conclusion that higher frequency, lower perplexity, and higher similarity will lead to better performance in general. Furthermore, all of these three predictors are significant at the 0.01 level, and the VIF scores are all below 5, indicating no colinearity. Table 9 in the Appendix details the coefficients and p-value for each predictor.

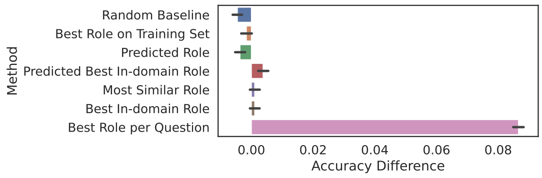
7 Finding the Best Roles for Prompting

In previous sections, we demonstrate that there might not exist a single persona that consistently improves the performance of diverse sets of questions. However, we also observe that personas might help in cases where their domains are aligned with the questions or when they have higher similarities. A natural question arises: instead of manually choosing roles for all questions, could we automatically find the best roles for prompting in various settings? We experiment with a list of search strategies to find the best role using data obtained from each of the four models.

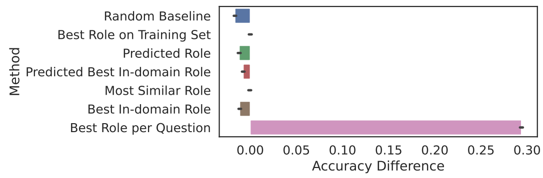
7.1 Methods

We experiment with the following baselines in selecting the best roles for prompting. **Random:** Randomly select a role from the predefined role list for each question. **In-domain best role:** Automatically select the best in-domain role in the training set. **Best role:** Automatically select the best role in the training data. **Best role per question:** Automatically select the best role per question in the test data, this is the performance upper bound.

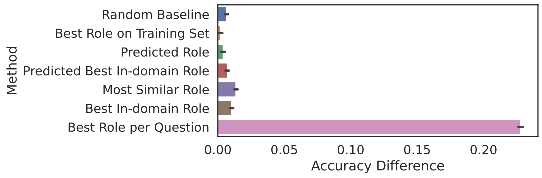
We further design the following methods to automatically select the best roles. **Similarity-based Method:** Select the role that has the highest similarity to the question. **Dataset Classifier:** aims at finding the correct domain for each question. We first fine-tune a `roberta-base` model to predict the domain of the question. We concatenate the entire question with its options as the input and the output is the domain of the question. We further select the best in-domain role from the training set. The 2,410 questions are divided into a 7:1:2 ratio



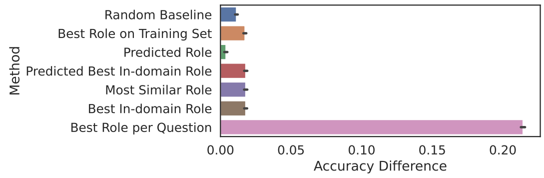
(a) FLAN-T5



(b) LLAMA3



(c) Mistral



(d) Qwen

Figure 8: Performance change for each model (compared with the control prompt) across different role-selection strategies reveals that the best-performing role per question is often idiosyncratic and different strategies for selecting the appropriate role offer limited (if any) improvement over picking a random role.

for training, validation, and the test set, respectively. The overall accuracy of the domain classifier is 78.1% on the test set. For reference, the accuracies of a random guess and choosing the most frequent class are 5.2% and 6.9% respectively. **Role Classifier:** aims at predicting the best role for each question. We fine-tune a `roberta-base` model and use it as a multi-label classifier for personas. The prediction target is the 162 roles, and the classifier achieved an accuracy of 0.34 for FLAN-T5, 0.37 for LLaMA, 0.39 for Mistral, and 0.30 for Qwen.

7.2 Results

Figure 8 shows performance comparisons using different role-searching strategies on four models relative to the control group (i.e., prompting with no role). The best role per question can be considered as the theoretical upper limit for the role predictor, where the model accurately picks the best role for each question. We find that when automatically selecting the best role, the aggregated result can lead to significantly better overall performance. This suggests that for each specific question, there exist certain roles that can lead to better prediction accuracy. However, all the automatic role-searching strategies are far away from this theoretical upper bound. On the contrary, while the most similar role and the best in-domain role generally perform better than the random baseline, most of the role-searching strategies are barely better than randomly selecting a persona for each question. This result suggests that while choosing in-domain or more similar personas could help to improve the prediction accuracy by a small margin, the effect of personas on model performance is largely random.

8 Conclusion

Incorporating personas in prompts has been an important approach for the design of system prompts as well as role-playing with LLMs. In this study, we present a systematic analysis of 162 personas in 26 categories to explore how prompting with personas affects model performances. Through our analysis, we show that adding person does not necessarily improves LLMs' performance over a wide range of types of questions. While we observe that roles with higher frequency in web corpus, prompts with lower perplexity and prompt-question pair with higher similarity potentially lead to better performances, predicting the role that leads to the best performance remains challenging and the best role depends on a specific question, dataset, and model. Our studies can help inform the future design of system prompts and role-playing strategies with LLMs. All data, results, and experiment code are available at *http://anon*, which we hope will encourage testing of future models.

9 Limitations

Our study has the following limitations: First, we only studied four open-source LLMs and didn't include closed-source models like GPT3.5 and GPT4. This is due to the computational cost of running such a large experiment. We will release the script to run the experiment and we welcome other researchers to explore how role-playing affects LLM performances on other models. Second, while we aimed to be comprehensive when selecting the personas, we were not able to experiment with all the personas beyond the 162 ones in our current experiment. We will release the full list of our personas to support future research in this direction. Third, given the computational costs of our experiments, we only used MMLU as our testbed, overlooking other factual question datasets and open-ended questions. While we believe that our current analysis provides important findings regarding how personas affect the models' performances, we acknowledge this limitation and plan to extend our analysis to more settings.

10 Ethical Considerations

Our study has the following ethical implications. First, to ensure the robustness of our results, we experimented with 162 roles, 4 prompt templates, and 4 models over 2410 MMLU questions. Running such an experiment is computationally expensive and is likely to result in a substantial release of carbon dioxide. Second, some of our analyses may reinforce existing stereotypes regarding personas. For example, our results suggest that male roles lead to better performances than female roles, which might inadvertently reinforce traditional gender stereotypes. However, our results also show that gender-neutral roles lead to higher performances than gendered roles, suggesting that developers should consider using gender-neutral roles when creating system prompts. On the other hand, our results also reveal potential model biases originating from implicit societal stereotypes regarding gender roles. We call for future research in this direction to study de-biasing technologies when training or aligning LLMs.

References

Joan Acker. 1992. From sex roles to gendered institutions. *Contemporary sociology*, 21(5):565–569.

AI@Meta. 2024. Llama 3 model card.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Ellen Berscheid. 1994. Interpersonal relationships. *Annual review of psychology*, 45(1):79–129.

Amy M Blackstone. 2003. Gender roles and society.

Thomas Brante. 1988. Sociological approaches to the professions. *Acta sociologica*, 31(2):119–142.

Rue Bucher and Anselm Strauss. 1961. Professions in process. *American journal of sociology*, 66(4):325–334.

Minje Choi, Ceren Budak, Daniel M Romero, and David Jurgens. 2021. More than meets the tie: Examining the role of interpersonal relationships in social networks. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 105–116.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Orianna Demasi, Yu Li, and Zhou Yu. 2020. A multi-persona chatbot for hotline counselor training. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3623–3636.

Kerstin Denecke, Sayan Vaaheesan, and Aaganya Arulnathan. 2020. A mental health chatbot for regulating emotions (sermo)-concept and usability test. *IEEE Transactions on Emerging Topics in Computing*, 9(3):1170–1182.

Jan Deriu, Alvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. 2021. Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review*, 54:755–810.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Erving Goffman. 2016. The presentation of self in everyday life. In *Social Theory Re-Wired*, pages 482–493. Routledge.

Jerold Heiss. 2017. Social roles. In *Social psychology*, pages 94–130. Routledge.

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

Thomas F Heston and Charya Khun. 2023. Prompt engineering in medical education. *International Medical Education*, 2(3):198–205.

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*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, and Xin Zhou. 2023. Better zero-shot reasoning with role-play prompting. *arXiv preprint arXiv:2308.07702*.

Zhaojiang Lin, Zihan Liu, Genta Indra Winata, Samuel Cahyawijaya, Andrea Madotto, Yejin Bang, Etsuko Ishii, and Pascale Fung. 2020. Xpersona: Evaluating multilingual personalized chatbot. *arXiv preprint arXiv:2003.07568*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2021. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *arXiv preprint arXiv:2104.08786*.

Philip Massey, Patrick Xia, David Bamman, and Noah A Smith. 2015. Annotating character relationships in literary texts. *arXiv preprint arXiv:1512.00728*.

Bertalan Meskó. 2023. Prompt engineering as an important emerging skill for medical professionals: tutorial. *Journal of Medical Internet Research*, 25:e50638.

Risto Miikkulainen. 2007. Creating intelligent agents in games. In *Frontiers of engineering: Reports on leading-edge engineering from the 2006 symposium*, page 15. National Academies Press.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Jonas Oppenlaender. 2022. Prompt engineering for text-based generative art. *arXiv preprint arXiv:2204.13988*.

Pat Pataranutaporn, Valdemar Danry, Joanne Leong, Parinya Punpongsanon, Dan Novy, Pattie Maes, and Misha Sra. 2021. Ai-generated characters for supporting personalized learning and well-being. *Nature Machine Intelligence*, 3(12):1013–1022.

Hongjin Qian, Xiaohe Li, Hanxun Zhong, Yu Guo, Yueyuan Ma, Yutao Zhu, Zhanliang Liu, Zhicheng Dou, and Ji-Rong Wen. 2021. Pchatbot: a large-scale dataset for personalized chatbot. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 2470–2477.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. Role play with large language models. *Nature*, pages 1–6.

Cass R Sunstein. 1996. Social norms and social roles. *Colum. L. Rev.*, 96:903.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, 33:5776–5788.

Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Man Zhang, et al. 2023. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. *arXiv preprint arXiv:2310.00746*.

Wolf Wolfensberger. 2000. A brief overview of social role valorization. *Mental retardation*, 38(2):105–123.

A Experiment Settings

Dataset and Models The dataset and models used in this study along with their licenses are listed in Table 2. All of them are open-source and our use is consistent with their intended purpose. The mapping between sampled subsets of MMLU and their domains are illustrated in Table 3.

Roles and Prompts The full list of roles is shown in Table 4 and the roles used for explicit gender impact is listed in Table 5. The 4 prompt templates are listed in Table 6 and the deailed context prompts and control prompts are shown in Table 7 and Table 8.

Model/Dataset	License
MMLU	MIT
Flan-T5	Apache-2.0
LLaMA-3	llama3
Mistral-7B	Apache-2.0
Qwen	tongyi-qianwen-license-agreement

Table 2: List of licenses

Computational infrastructure and budget The GPU hours required for running experiments on Flan-T5-XXL are around 100 hours on 8 NVIDIA RTX A6000. For LLaMA-3, Mistral and Qwen, it took around 24 hours for each using 2 NVIDIA RTX A6000 with the "vllm" package.

Classification Parameters We train the classifiers using roberta-base. The parameters are set as follows: learning rate=1e-5, epochs=50 and weight_decay=0.01.

Used Packages We primarily utilize the "transformer" and "torch" packages for model inference. For data analysis and visualization, we rely on the "pandas" and "seaborn" packages. To calculate similarity between prompts and questions, we employ "sentence_transformers" to obtain sentence embeddings, and we use "lmppl" to acquire perplexity scores.

B Regression Results

Persona Impact Figure 9, Figure 10, Figure 11, and Figure 12 show the coefficients of "role" in the linear relationship between accuracy and role type for each model.

Overall Regression Table 9 lists the coefficients and p-values for the mixed-effects model on the impact of frequency, similarity and perplexity on prediction accuracy, controlling for each model as a random effect.

C Persona Embeddings

To quantify the performance differences of various personas, we build embeddings for each persona and analyze the similarity across these embeddings. For each persona, we first calculate the average accuracy of each question, resulting in a vector of length 2410. Then, we use Uniform Manifold Approximation and Projection (UMAP) for dimension reduction to map these embeddings to two dimensions. The persona embeddings calculated from each model are illustrated in Figure 13, Figure 14, Figure 15, and Figure 16. The distributions of pairwise cosine similarity for each model are shown in Figure 17. The skewed distributions in models LLaMA, Mistral, and Qwen towards the right around value 1 demonstrate the high similarity across roles, whereas the embeddings are relatively more divergent in Flan-T5.

D Model Consistency

The correlation between personas' mean accuracy over 2410 questions and 4 prompts across 4 models are illustrated in Figure 18.

Domain	Datasets
Law	professional_law, international_law
Medicine	clinical_knowledge, college_medicine, professional_medicine
EECS	electrical_engineering, college_computer_science, high_school_computer_science
Math	high_school_statistics, college_mathematics, high_school_mathematics
Politics	us_foreign_policy, high_school_government_and_politics
Psychology	professional_psychology, high_school_psychology
Natural Science	college_physics, college_biology, high_school_physics, high_school_chemistry, college_chemistry, high_school_biology
Econ	management, professional_accounting, econometrics, high_school_macroeconomics, high_school_microeconomics

Table 3: Domain Dictionary

Category	Roles
family	sister, son, father-in-law, mother-in-law, brother,parent, father,mother, daddy, dad, papa, mummy, mamma, mommy, mom, mum, mama, daughter, cousin, grandfather, grandmother
romantic	partner, husband, wife, boyfriend, housewife, girlfriend, fiancée, fiancé
school	professor, instructor, student, coach, tutor, dean, graduate, classmate
work	supervisor, coworker, boss, colleague, mentor
social	companion, buddy, roommate, friend, stranger, foreigner, best friend, close friend
AI	chatbot, assistant, virtual assistant , AI language model, mathematician AI, software enginner AI, Educational Tutor AI, Medical Diagnostic AI, helpful assistant, Behavioral Economics AI, Historical Data Analyst AI, Legal Research AI, Mathematical Modeling AI, Statistical Analysis AI, Diagnostic AI, Policy Analysis AI, Public Opinion AI, Psychological Profiling AI, Scientific Data Analysis AI, Embedded Systems AI Engineer
econ	economic researcher, economist, financial analyst
eecs	electronics technician, data scientist, electrical engineer, software engineer, web developer
history	historian, archivist, historical researcher, archaeologist
law	bailiff, lawyer
math	data analyst, mathematician, statistician
medicine	nurse, doctor, physician, dentist, surgeon
natural science	geneticist, biologist, physicist, teacher, chemist, ecologist
other occupations	painter, auctioneer, musician, scientist, driver, accountant, geologist, janitor, architect, mason, baker, administrator, research scientist, weaver, postmaster, cook, clerk, broker, dancer, surveyor, clergy, secretary, soldier, housekeeper, collector, carpenter, cashier, conductor, mechanic, engineer, photographer, manager, farmer, tailor, shoemaker, sales, librarian, blacksmith, artist, pilot, inspector, police, gardener, attendant, athlete, operator, sailor, designer, midwife, president, humanist, auditor, scholar, CEO, advisor, counsellor, counselor, cofounder
politics	politician, sheriff, governer, enthusiast, partisan
psychology	psychologist

Table 4: Role Dictionary

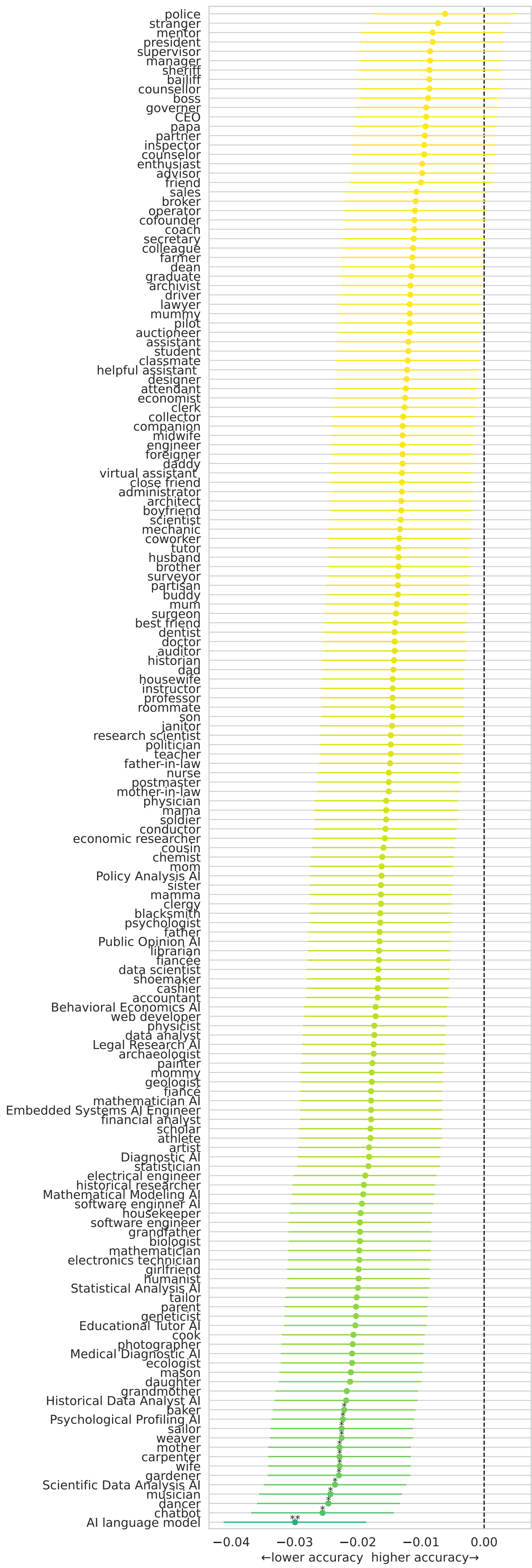


Figure 9: Coefficients of the regression model on the relationship between accuracy and role with random intercepts for Flan-T5-XXL

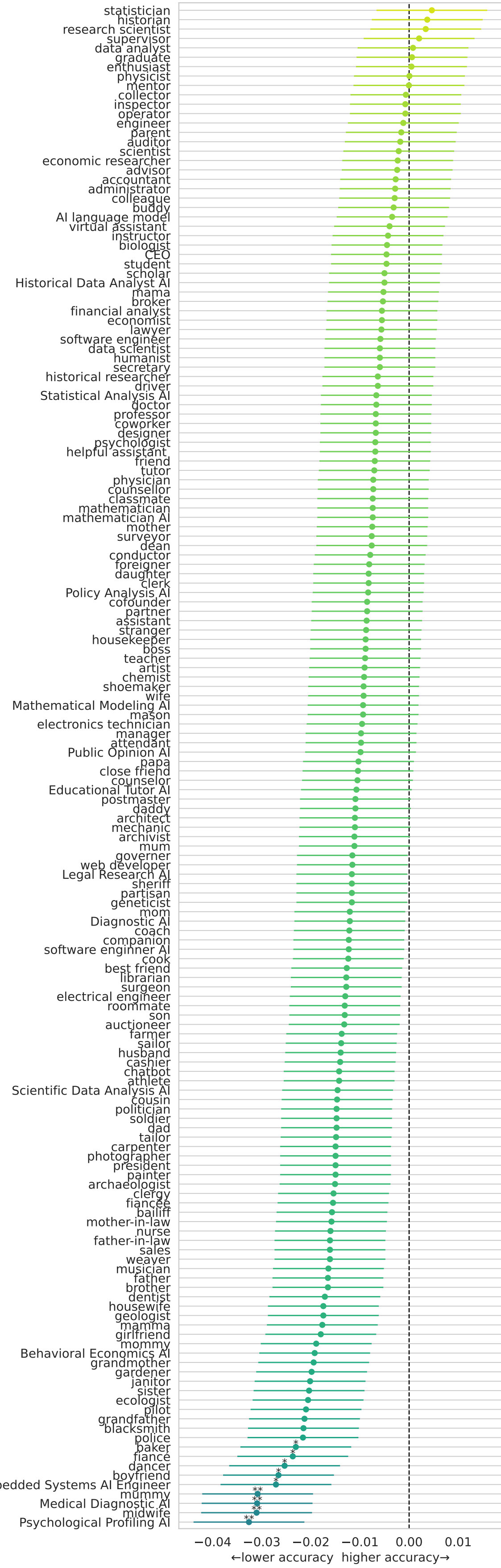


Figure 10: Coefficients of the regression model on the relationship between accuracy and role with random intercepts for LLaMA-3

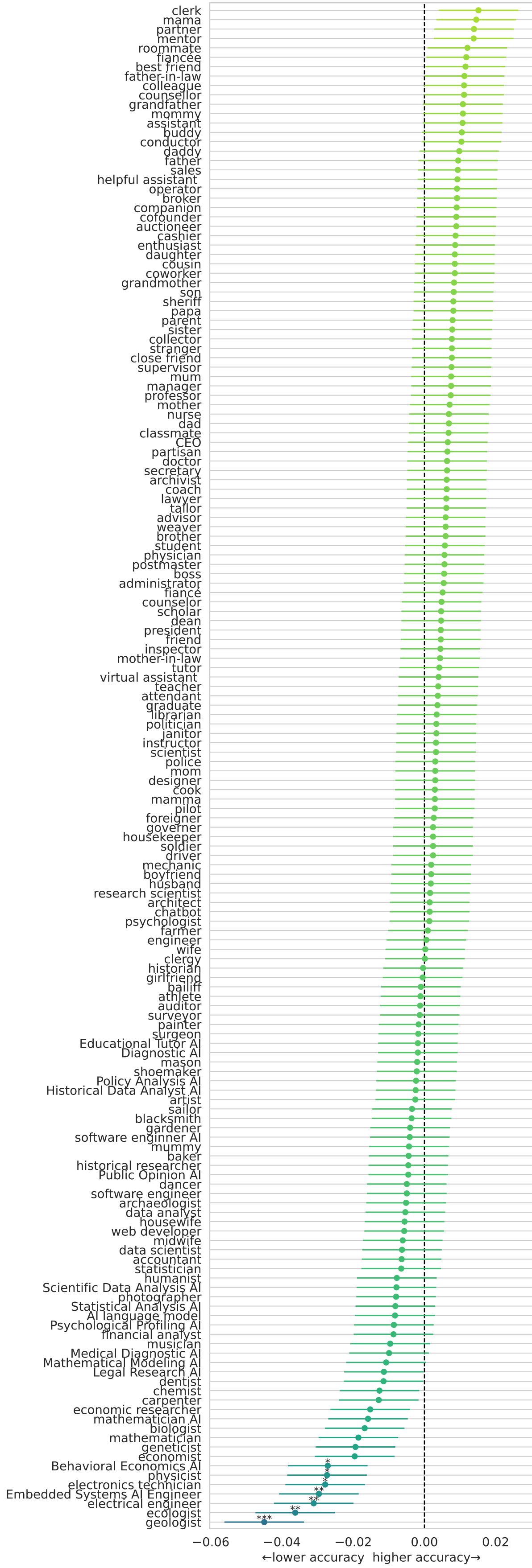


Figure 11: Coefficients of the regression model on the relationship between accuracy and role with random intercepts for Mistral

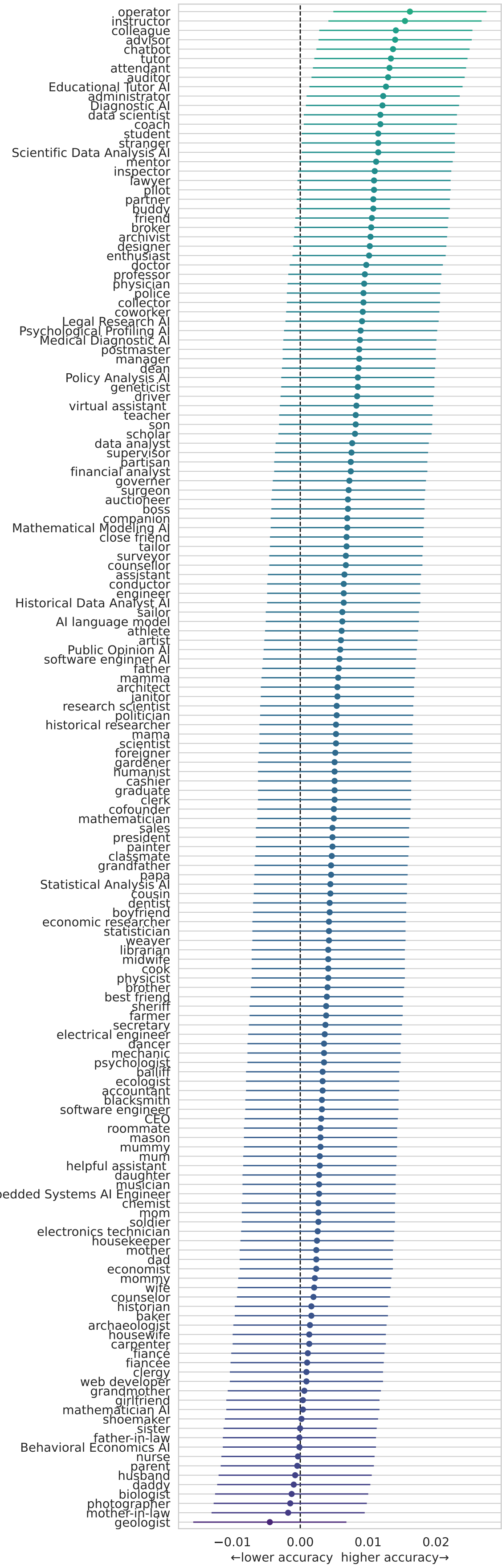


Figure 12: Coefficients of the regression model on the relationship between accuracy and role with random intercepts for Qwen

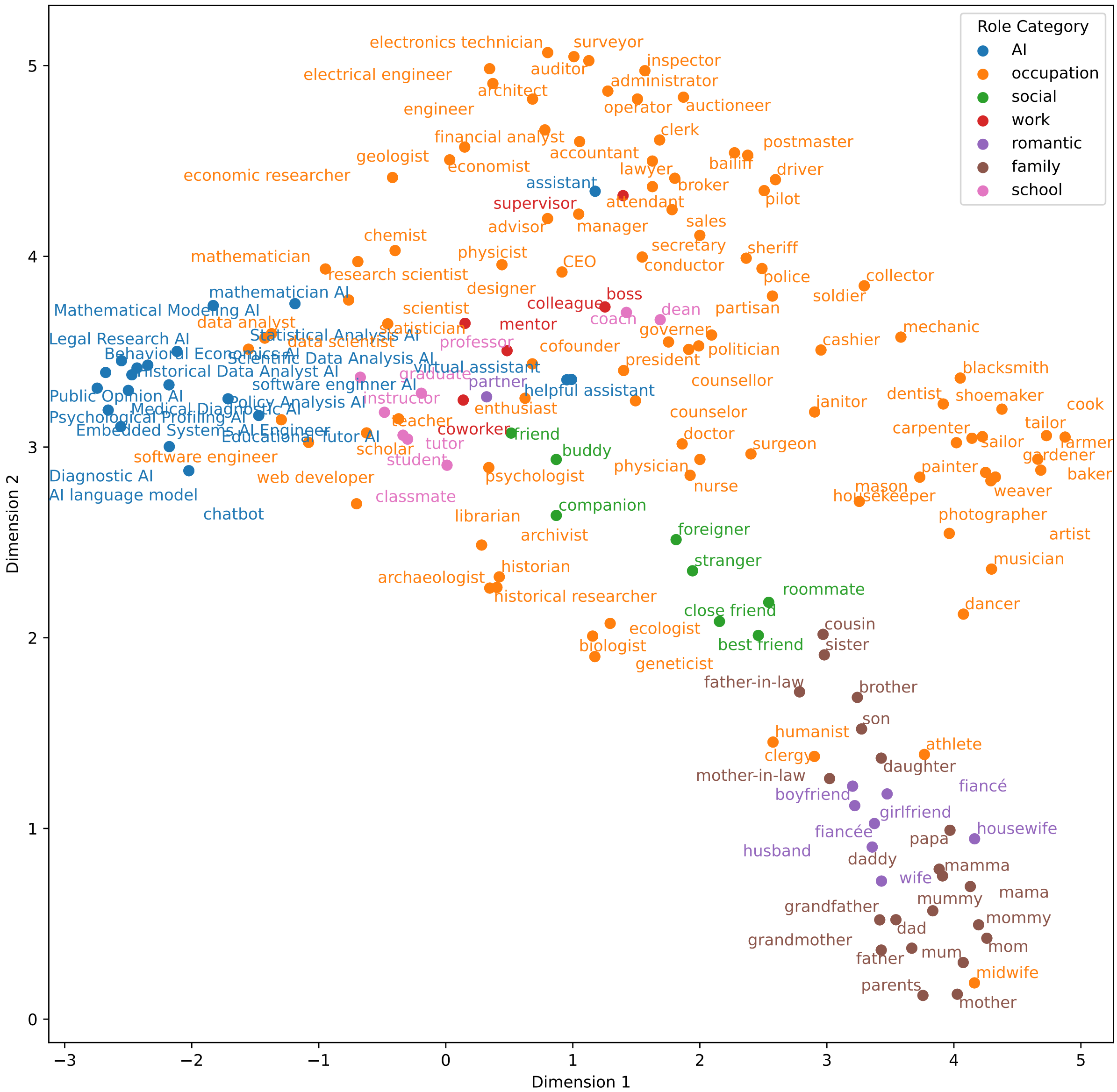


Figure 13: Role embeddings calculated by UMAP for Flan

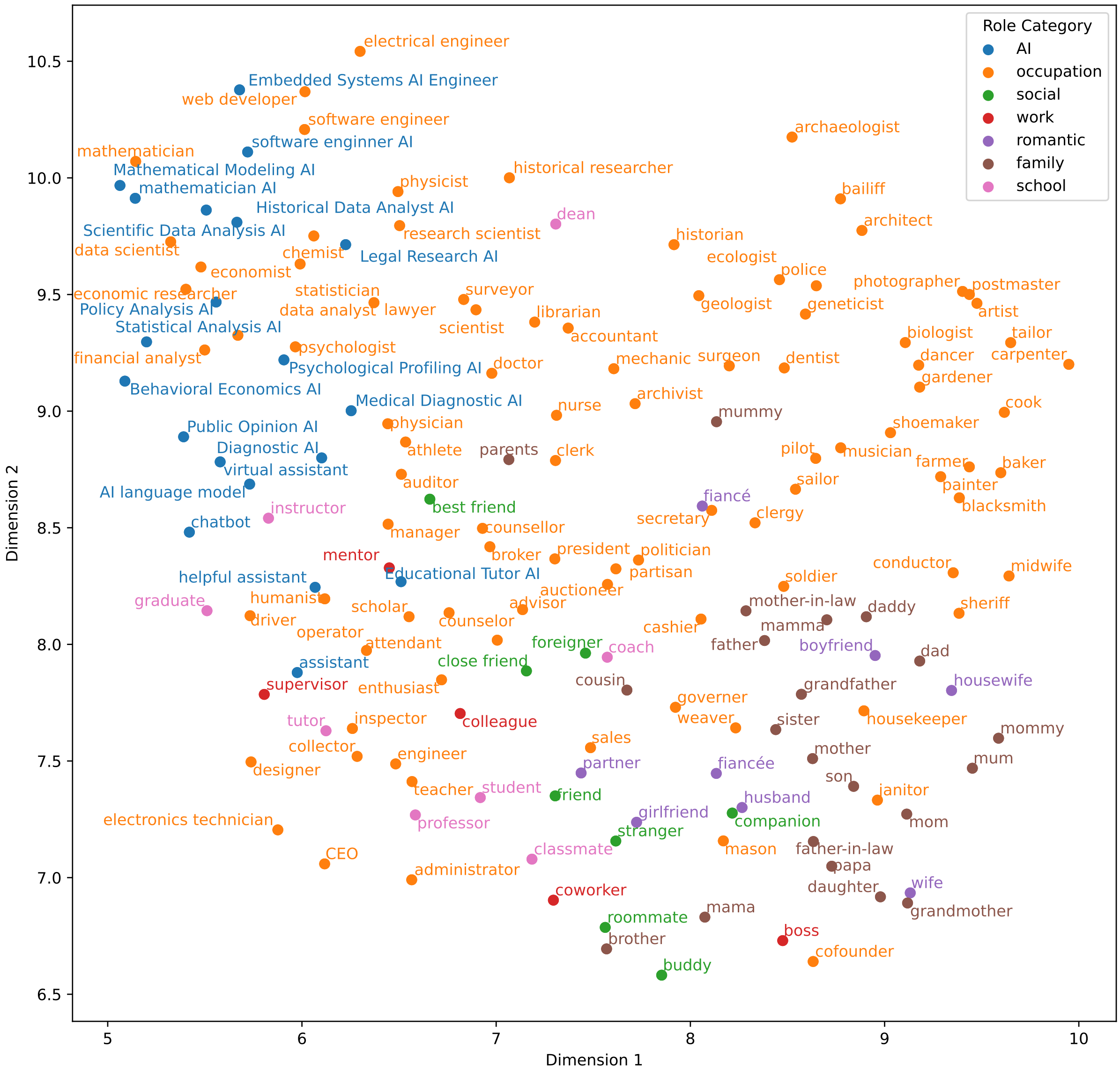


Figure 14: Role embeddings calculated by UMAP for Llama

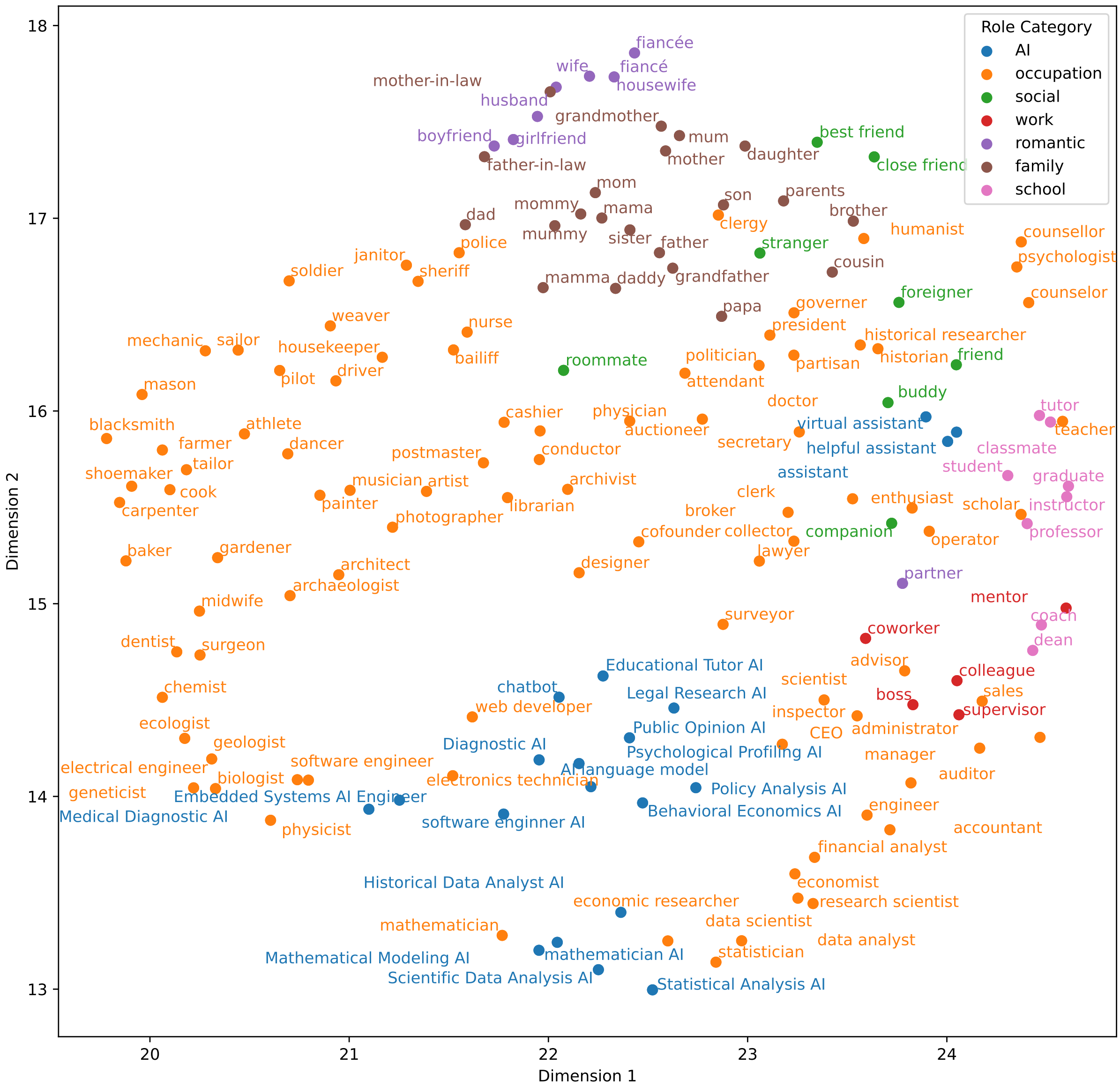


Figure 15: Role embeddings calculated by UMAP for Mistral

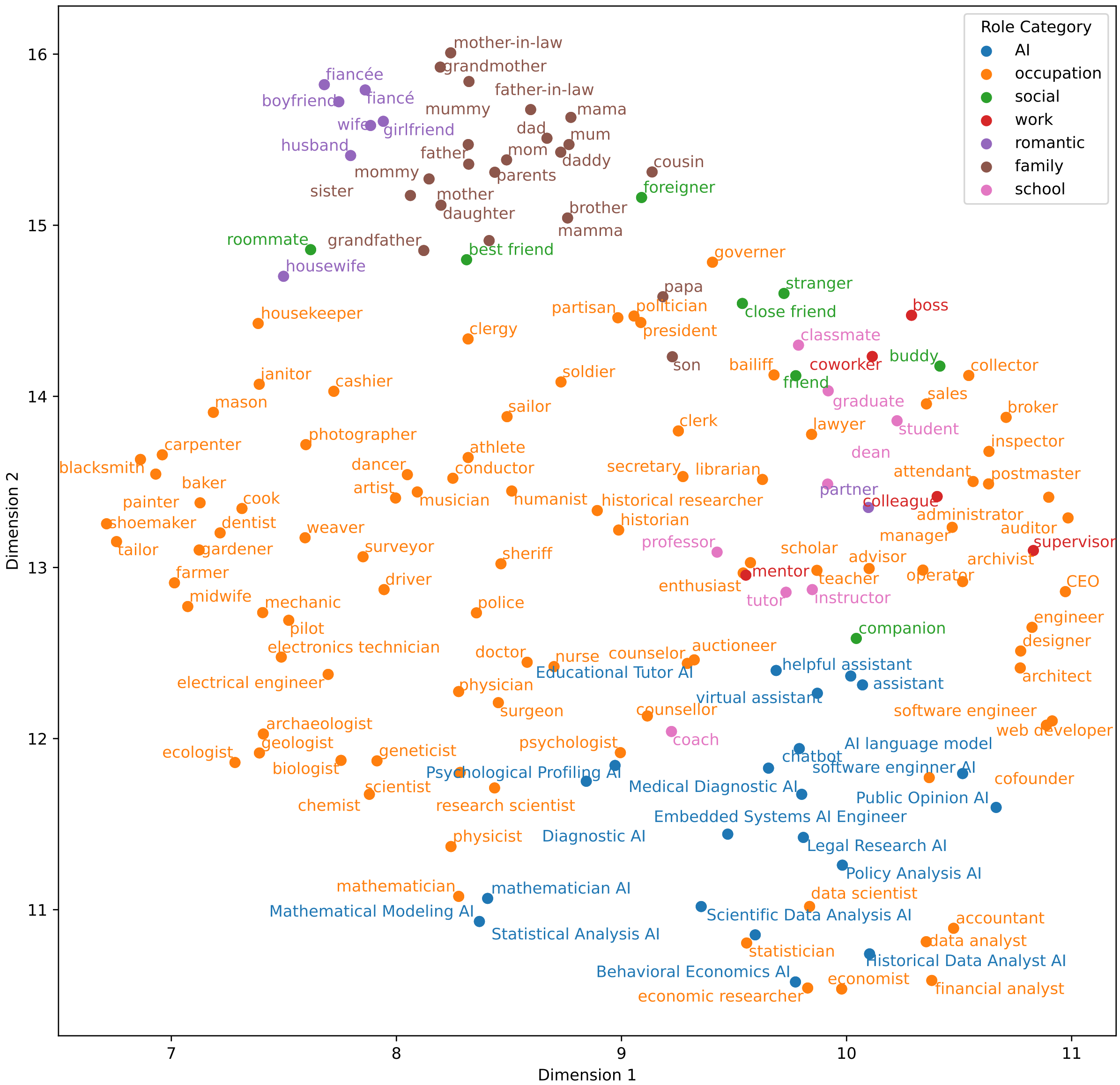
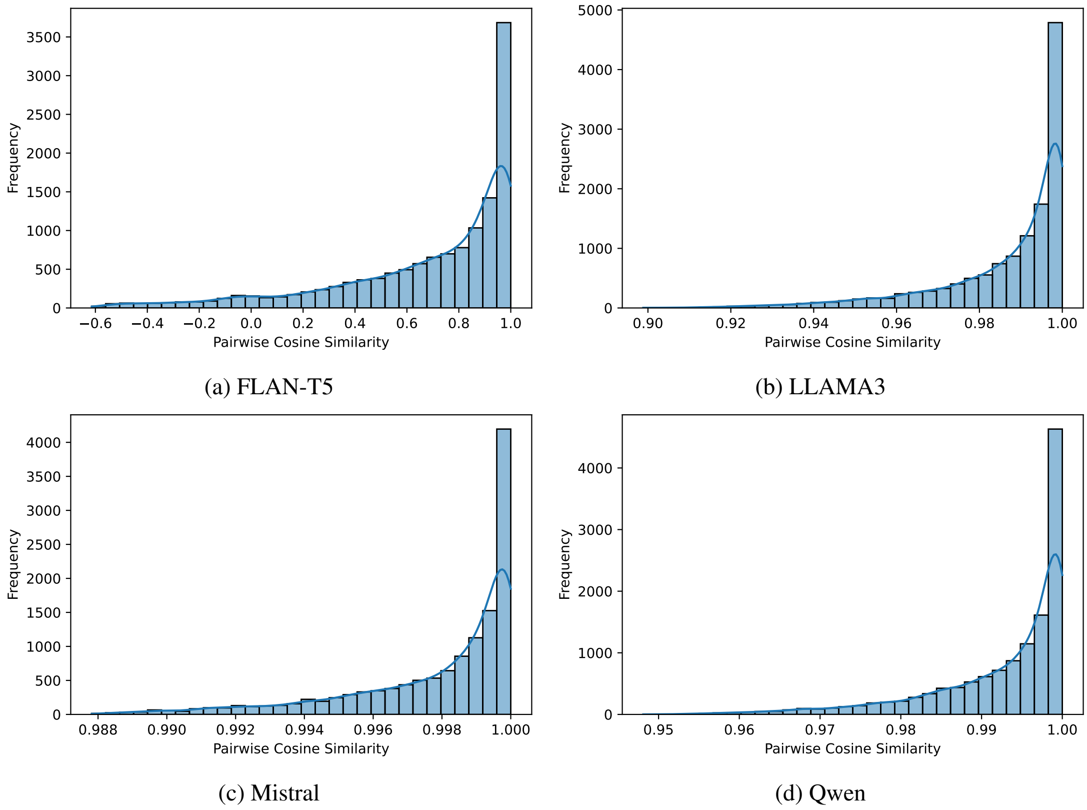


Figure 16: Role embeddings calculated by UMAP for Qwen



(a) FLAN-T5

(b) LLAMA3

(c) Mistral

(d) Qwen

Figure 17: Cosine similarity distribution of role mebeddings for each model.

Gender	Roles
Male	father, daddy, dad, papa, father-in-law, grandfather, husband, son, boyfriend, fiancé
Female	mother, mommy, mom, mamma, mother-in-law, grandmother, wife, daughter, girlfriend, fiancée
Neutral	partner, parent

Table 5: List of aligned roles categorized by gender

Prompt Type	Prompt
Audience-Specific	You are talking to a/an {role}. Imagine you are talking to a/an {role}.
Speaker-Specific	You are a/an {role}. Imagine you are a/an {role}.

Table 6: Context prompts

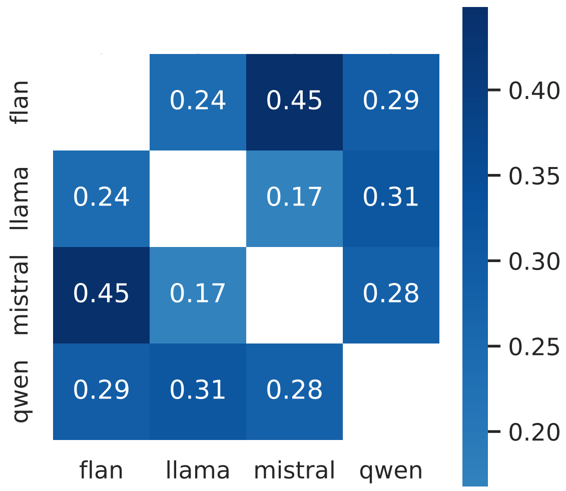


Figure 18: Heatmap of the correlation between personas' mean accuracy across models.

Model Type	Prompt Template
FLAN-T5	{context_prompt} {question} Please select the correct answer number:
LLaMa3, Mistral, Qwen	{“role”: “system”, “content”: {context_prompt}}, {“role”: “user”, “content”: The following is a multiple choice question (with answers). Reply with only the option number. {question}}

Table 7: Context Prompts for each model

Model Type	Prompt Template
FLAN-T5	{question} Please select the correct answer number:
LLaMa3, Mistral, Qwen	{“role”: “user”, “content”: The following is a multiple choice question (with answers). Reply with only the option number. {question}}

Table 8: Control Prompts for each model

Term	Coefficient	p-value
Frequency	106.714	3.81e-02
Perplexity	-0.000281	4.71e-04
Similarity	0.321	4.36e-38

Table 9: Coefficients of the mixed-effects model on the relationship between accuracy and all the role attributes