Is "A Helpful Assistant" the Best Role for Large Language Models? A Systematic Evaluation of Social Roles in System Prompts

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. But is "a helpful assistant" the best role for LLMs? In this study, we present a systematic evaluation of how social roles in system prompts affect model performance. We create a list of 162 roles covering 6 types of interpersonal relationships and 8 types of occupations. Through extensive analysis of 3 popular LLMs and 2457 questions, we show that adding interpersonal roles in prompts consistently improves the models' performance over a range of questions. Moreover, while we find that using gender-neutral roles and specifying the role as the audience leads to better performances, predicting which role leads to the best performance remains a challenging task, and that frequency, similarity, and perplexity do not fully explain the effect of social roles on model performances. Our results can help inform the design of system prompts for AI systems. Code and data are available at `AnonymizedURL`.

1 Introduction

Social roles define the way people perceive themselves and others (Wolfensberger, 2000) and provide important context for all types of human interactions (Sunstein, 1996). For example, the norms about interpersonal roles affect how people perceive the appropriateness of behaviors (Aune et al., 1994) and communications (Derlega, 1984). Occupational roles (e.g. firemen) are also deeply embedded in our society and define people's identities (Christiansen, 1999).

Building persona- or role-based chatbots has attracted enormous attention from the AI and NLP community due to its potential business and society applications (Pataranutaporn et al., 2021). Recent advances in LLMs also provide huge opportunities to build intelligent agents that are able to behave and talk like certain characters or roles (Wang et al., 2023). However, with all the existing studies on role-playing with LLMs, it is unclear how different types of social roles affect LLMs' objective performance. To address this gap, we conduct a large-scale analysis of 162 roles spanning 6 types of interpersonal relationships and 8 types of occupations. We design prompts and analyze the performance of three popular open-source models including Flan-T5 (Chung et al., 2022), LLaMA2 (Touvron et al., 2023), and OPT-instruct (Iyer et al., 2022). We evaluate the model's performance over 2457 questions sampled from the Massive Multitask Language Understanding (MMLU) dataset (Hendrycks et al., 2021), balanced for categories.

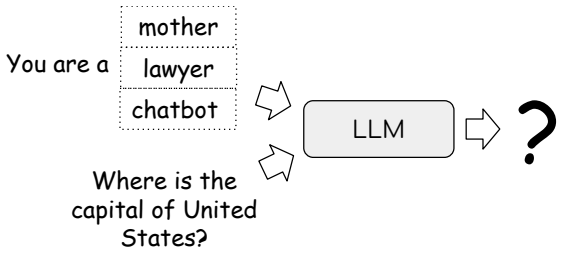


Figure 1: Our overall research question: does adding social roles in prompts affect LLMs' performance?

In this study, we focus on three major research questions: (1) Does adding different types of social roles in prompts affect LLMs' performance? (2) What might explain the effect of different social roles on LLMs? (3) Can we automatically find the best roles for prompting? Through our analysis, we find that the model's performance is sensitive to the types of roles and specific roles. Prompting with social roles generally improves the model's performance in answering questions compared with the control prompt which does not contain any social roles. Furthermore, we observe that interpersonal roles like "friend" and gender-neutral roles

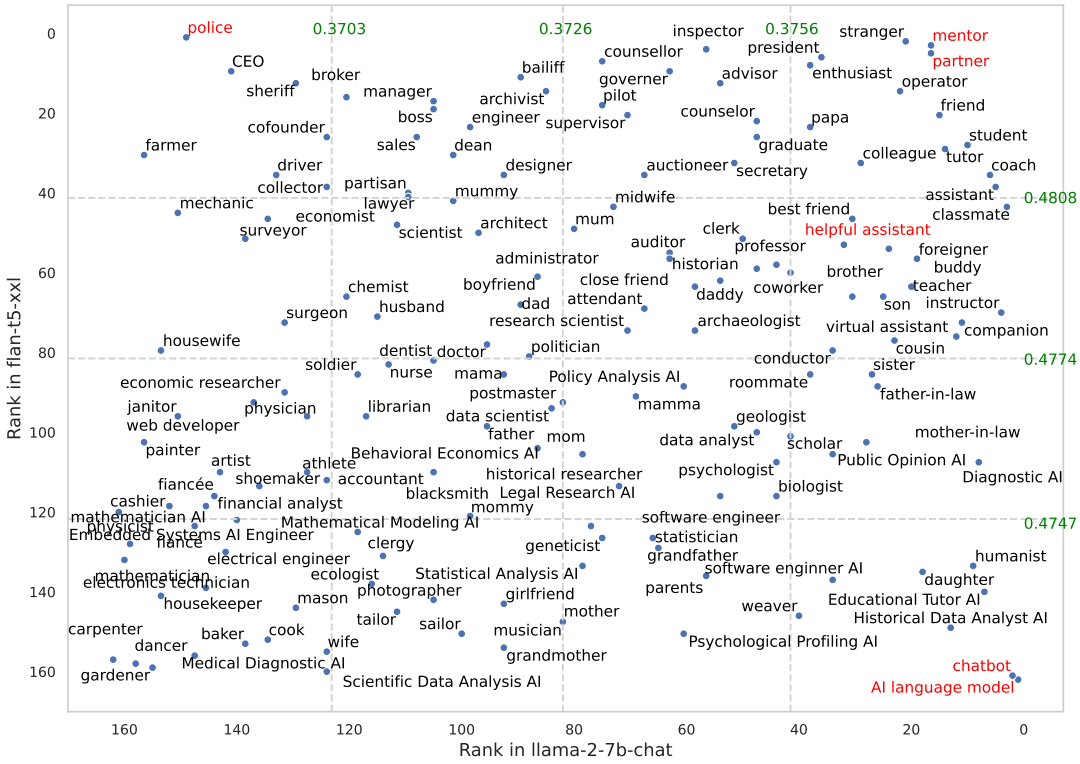


Figure 2: Overall model performance when being prompted with different social roles (e.g. "You are a lawyer.") for FLAN-T5-XXL and LLAMA2-7B-Chat. Tested on 2457 MMLU questions. Best-performing roles are highlighted in red. We also highlight "helpful assistant" as it is commonly used in commercial AI systems like ChatGPT.

are more likely to lead to higher performances. Interpersonal roles consistently lead to higher performance across different models and datasets. However, the effect of occupation-based roles depends on the specific questions and datasets.

What might explain the effect of social roles in prompts? We further analyze the word frequency of the role, the perplexity of the prompt, the similarity between the prompt and questions, and whether the role is aligned with the domain of the question. We find that high-frequency, low-perplexity roles/prompts and prompts that have higher similarities with the questions tend to lead to higher performance but only with a weak correlation. We further explore automatic role-searching strategies which aim at finding the role that leads to the best performance for each question. We find that while predicted roles lead to higher performance compared with several baselines, predicting the best roles remains to be a challenging task.

Our study makes the following three contributions to the community. First, we introduce a new pipeline to systematically evaluate LLMs' performance when being prompted with a wide range of social roles. Second, our experiments reveal insights regarding why prompting with social roles helps with the model performance. Third, our experiments with a wide range of automatic role-searching strategies suggest that optimal roles can meaningfully improve performance, but more work is needed to automatically identify these.

2 Related work

Social Roles Social roles are fundamental in human society and day-to-day interactions (Heiss, 2017; Goffman, 2016). Social roles define the norm of human interactions and affect human behaviors in various contexts (Sunstein, 1996). Two prominent types of social roles 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 social roles 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 social roles holds 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's 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. Despite its wide usage in commercial AI systems, the effect of using social roles 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 social roles in prompts affects LLMs' performances. To answer this question, we design a series of experiments and this section details the dataset, models, prompts, and social roles used in our experiments.

Prompt Type	Example
Control	{question}
Role Prompt	You are a/an {role}
Audience Prompt	You are talking to a/an {role}
Interpersonal Prompt	You are talking to your {role}

Table 1: Types and Examples of prompt templates for social roles used in our experiment. Full prompts for each model are available in the Appendix (Table 6 and Table 7).

3.1 Datasets

MMLU (Hendrycks et al., 2021) is a dataset designed for extensive multitask language understanding and has been widely used as the 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 sample 2457 questions from the MMLU dataset, balanced across 26 subjects. We further map the 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.

3.2 Prompts

Social roles can be incorporated into prompts in various ways. We carefully design three types of prompts: (1) **Role Prompt:** prompts that assign the role to the LLM (i.e. "who you are"). For example, "You are a lawyer". (2) **Audience Prompt:** prompts that specify the audience of the conversation (i.e. "who you are talking to"). For example, "You are talking to a fireman." and (3) **Interpersonal Prompt:** prompts that connotate the relationship between the speaker and listener. For example, "You are talking to your mom." 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 your mom"). 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 Social Roles

For a given question, which role could lead to the best answer compared to others? To answer this question, we carefully create a diverse set of social roles that are actively used in people's daily interactions. We first collect over 300 social roles 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 social roles, 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, 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. Due to the rise of AI systems, we also include a list of AI roles (e.g. "AI language model" and "AI assistant").

3.4 Models

We experiment with three open-source instruction-tuned LLMs whose sizes range from 1.3B to 11B. The models are FLAN-T5-XXL (Chung et al., 2022), LLaMA-2-7b-chat (Touvron et al., 2023), and OPT-iml-max-1.3B (Iyer et al., 2022). We use open-source models as these allow us to easily control the system prompt. For each of the sampled 2457 questions, we create prompts with the 6 types of templates and the 162 social roles.

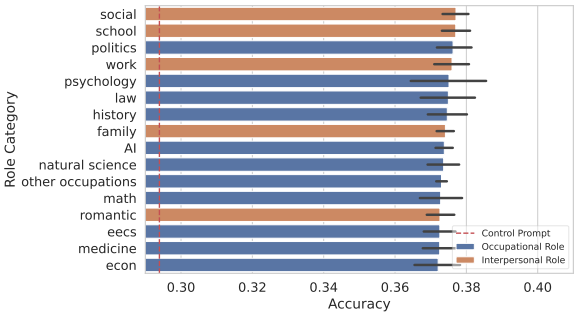


Figure 3: Interpersonal roles (e.g. social, school and work) generally lead to higher performances compared with occupational roles.

4 Which Types of Social Roles Are More Helpful?

Does prompting with social roles affect LLMs' performance in answering questions? Here, we discuss the effect of social roles on LLMs' performance.

Overall Results Figure 3 shows the overall performance of each role category on the 2457 MMLU questions, where scores are averaged across different prompt templates. Specifying a role when prompting can effectively improve the performance of LLMs by at least 20% compared with the control prompt, where no context is given. Such a result suggests that adding a social role in the prompt could benefit LLMs by a large margin. Moreover, we observe that interpersonal roles tend to lead to higher performances except for romantic and family roles.

Within interpersonal roles, we find that non-intimate roles including social, school, and work consistently lead to better performances compared with intimate relationships like romantic (e.g. boyfriend, girlfriend) and family (e.g. mother and daddy), suggesting that the implied social distance of roles may impact answer quality. For occupational roles, we find that politics, psychology, and law-related roles like psychologists, politicians, and lawyers perform better compared with Econ, Medicine, and EECS-related roles like economists and software engineers. Furthermore, we observe that, compared with occupational roles, interpersonal roles tend to perform better, especially for categories like school, social, and work. Our result suggests that models are sensitive to the social roles in prompts. What could be the leading factors behind this finding? We explore three potential factors to explain the disparity among role categories.

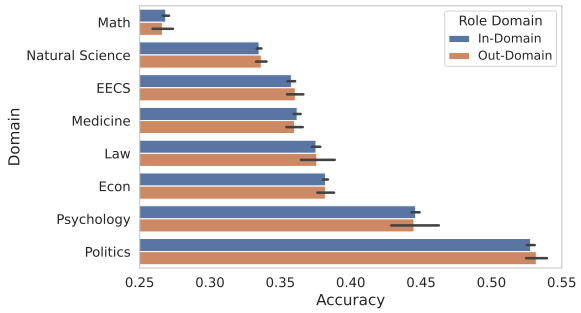


Figure 4: In-domain and out-domain roles do not have significant differences in the final model performance.

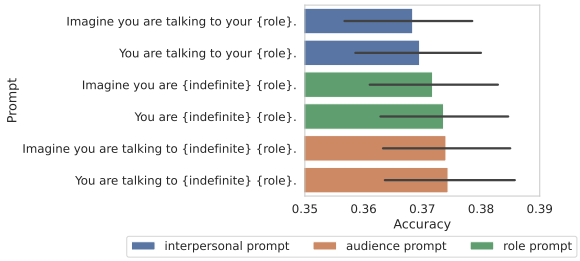


Figure 5: Audience prompts and role prompts lead to better performance than the interpersonal prompt.

Domain Differences The 26 subjects encompassed by the sampled questions are categorized into 8 domains, as illustrated in Table 3. Consequently, occupations associated with these subjects are also mapped to the corresponding domain. This allows each role-question pair to be labeled as "in-domain" or "out-domain".

Figure 4 shows the performance of in-domain and out-domain roles on different datasets. In general, we do not see a significant difference between in-domain and out-domain roles and the effect depends on the specific dataset. For example, in-domain roles perform slightly better than out-domain roles on math questions while out-domain roles perform better on questions related to politics. Our results suggest that the effect of roles is more complicated than the alignment of the domain.

Gender Differences Gender roles are one of the most prominent and widely studied social roles in the literature of sociology (Blackstone, 2003; Acker, 1992) and society as they are embedded in various types of social roles 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 analyze 50 interpersonal roles and categorize them as male, female, or neutral, resulting in 11 male roles, 15 female roles, and 24 neutral roles. Table 5 in the Appendix shows the full mapping of gender and roles. As a stricter comparison, we select three relationship categories with different gender roles. For example, "significant other" can be referred to as "husband", "wife", or "partner". Such a setting allows us to control the effect of relationship types and reveal the nuanced effects of gendered roles.

As shown in Figure 12, gender-neutral words tend to perform better than other gendered roles and male roles perform better than female roles. The p-value associated with "neutral" is 0.0045, with "female" as the reference group in a mixed linear model. This suggests a significant difference in the performance of gender-neutral roles compared to female roles, whereas the difference between female and male roles is not significant. A similar pattern is observed even when we control for the specific role, suggesting that LLMs might contain implicit bias of gender roles. Furthermore, as detailed in Appendix Figure 13, the influence of gender is consistent across these domains, with male roles exhibiting equal or superior performance compared to female roles.

Perspective Taking vs. Role Playing For each question and role, we design three types of prompts to guide LLMs in generating answers. These include the audience prompt, role prompt, and interpersonal prompt, which specify "who you are talking to", "who you are", and "what the relationship between us is", respectively. Does the way we incorporate social roles affect model performances?

As shown in Figure 5, a consistent pattern emerges: specifying the audience yields the highest performance, followed by role prompts. Furthermore, interpersonal prompts exhibit the lowest performance, and prompts beginning with "imagine" consistently generate poorer answers compared to their counterparts. Such a result suggests that when prompting LLMs, it is more effective to specify the targeted audience rather than positioning the model as someone within your social network. This implies that instructing the model to assume the role of "a doctor" is preferable to "your doctor". Furthermore, directly specifying the role or audience is a more effective strategy than prompting the model to "Imagine" the role, suggesting that LLMs are sensitive to nuances in the way that the role is specified.

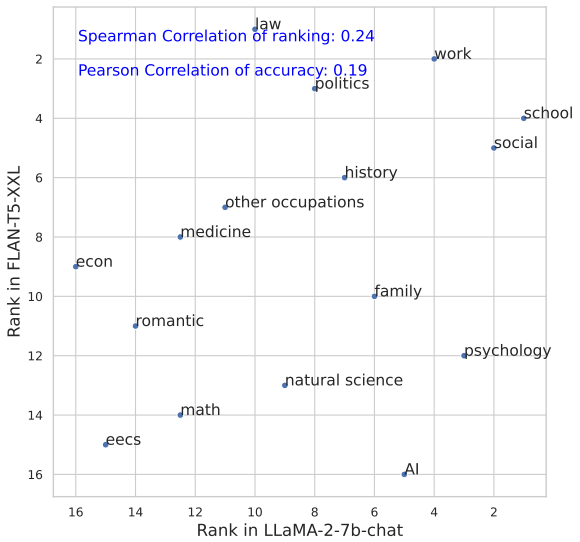


Figure 6: Interpersonal roles (e.g. work, social, and school) lead to good performance for both FLAN-T5 and LLAMA2, while the effect of occupational roles is more model-dependent.

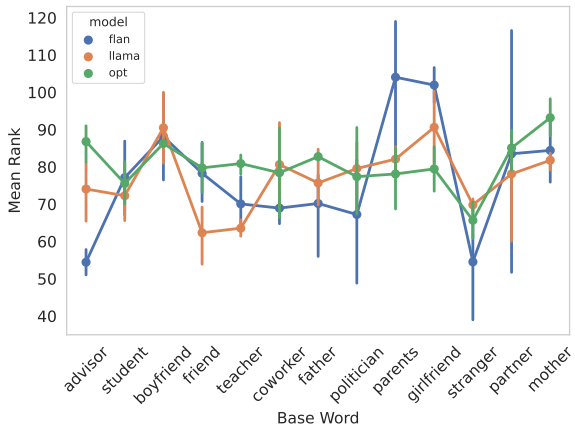


Figure 7: Average performance ranking of roles with multiple hyponyms across models. The ranking of roles with hyponyms is generally consistent across models, especially for LLAMA-2 and OPT.

Robustness Check In previous sections, we show aggregated results from 162 roles, six prompts, and three popular open-source instruction-tuned models. Do different models have different behaviors when prompted with different social roles? Here we conduct a robustness check for models as well as hyponyms of roles.

We focus on FLAN-T5-XXL and LLaMA2-7b-chat, representing an encoder-decoder model and a causal model, respectively. Additionally, comparing models of similar size ensures a more equitable comparison. Figure 6 suggests a low to moderate positive correlation (Pearson's $r = 0.35$ and Spearman's $r = 0.32$) between the rankings of role categories in terms of their mean accuracy across FLAN-T5-XXL and LLaMA2-7b-chat. Interpersonal roles like "work", "school", and "social" consistently perform better on both models while occupational roles can be more model-dependent. For example, AI is the worst role category for FLAN-T5 while it is ranked as the 5th for LLAMA-2.

Hyponyms refer to words that hold similar meanings to the root words (similar to synonyms). Are the social role effects caused by specific words or does the pattern persist across different hyponyms? We analyze the performance of 13 roles and their hyponyms across three models. As shown in Figure 7, the effect of different hyponyms is generally consistent across different models, suggesting that the performance change is generally caused by the types of role instead of the specific framing of it. Furthermore, we observe that LLAMA-2 and OPT tend to have similar performances across different roles, while FLAN-T5 diverges for certain roles like advisor and parents. This is potentially because OPT and LLAMA2 have similar architecture and training datasets.

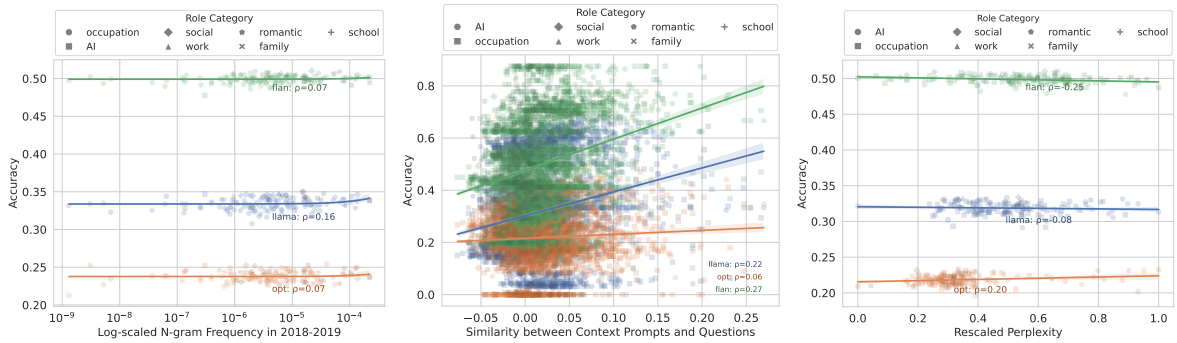
5 Potential Mechanism

Adding different social roles to the context prompt has a substantial impact on LLMs' performance. But where does this disparity originate? In this section, we test whether specific characteristics of the prompt and social roles 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 Social Roles 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 [1].

Figure 8a 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 0.16 for LLaMA2, the highest among the three, suggesting that word frequency

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



(a) Word frequency (b) Prompt-question similarity (c) Prompt perplexity

Figure 8: (a) Social roles' 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.

does not fully explain the effect of social roles 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 8b, we observe a positive correlation between similarity and accuracy at the role level. Specifically, the correlation for FLAN-T5-XXL and LLaMA2-7b-chat are both higher than 0.2, whereas the correlation for OPT-instruct is significantly lower, suggesting that the effect of similarity might depend on specific models.

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 LLaMA and OPT, perplexity is computed for an entire prompt, consisting of a context prompt followed by a question with options. For FLAN-T5, we use a pair of context prompts and the questions as the input. We further rescaled the calculated perplexity scores to a range of 0 to 1 to allow easier comparisons across models.

As shown in Figure 8c, the mean accuracy is negatively correlated with the rescaled perplexity at the role level on FLAN-T5 and LLaMA2. Such a result suggests that prompts with higher logical coherence and inherent reasonability are more likely to result in more accurate responses. However, we observe a slightly positive correlation between perplexity and accuracy on OPT, suggesting that the effect of perplexity might be different for smaller-size models.

6 Finding the Best Roles for Prompting

In previous sections, we demonstrate that prompting with social roles helps the models to perform better on various types of questions and the effect depends on specific datasets and models. A natural question is: instead of manually choosing roles, 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 FLAN-T5 and LLaMA2.

6.1 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

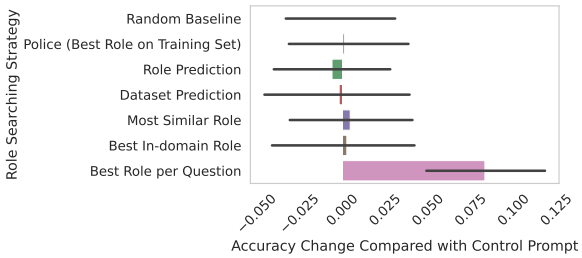


Figure 9: Performance change (compared with the control prompt) of different role searching strategies for FLAN-T5.

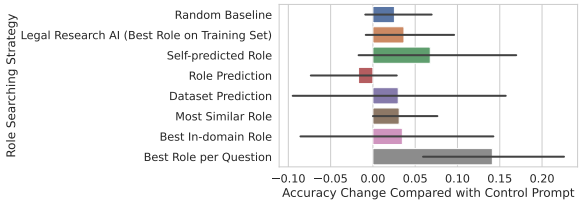


Figure 10: Performance change (compared with the control prompt) of different role searching strategies for LLAMA2.

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 2,457 questions are divided into a 7:1:2 ratio 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 social roles. The prediction target is the 162 roles and the classifier achieved 0.33 accuracy for FLAN-T5 and 0.2 for LLaMA. For reference, the accuracies of a random guess on FLAN-T5 and LLaMA are 1.11% and 0.48% respectively. **Self-Pick:** Can LLMs automatically pick the role for answering questions? To test this, we prompt the model to choose the role that it believes will yield the most accurate answer and then ask it to generate the answer based on the self-select role. We experiment with a series of prompt templates which could lead to valid answers for both roles and questions[2].

6.2 Results

Figure 9 and Figure 10 show the performance comparisons using different role-searching strategies on two models relative to the control group. The best role per question can be considered as the theoretical upper limit for the role predictor, where the model can accurately pick the best role for each question. Similarly, the best in-domain role serves as the theoretical upper limit for the dataset predictor, with the assumption that in-domain roles outperform out-of-domain roles. The similar performance observed between the most similar role and the best role per question reflects the previous findings that emphasize the role of similarity in explaining model performance. We perform McNemar's test to compare the results of "best role" and "control", using a threshold of 0.5 for the predicted accuracy of "best role". The resulting p-values are 2.9e-12 for LLaMA and 0.029 for FLAN-T5, indicating a statistically significant difference in model performance between the "best role" and "control" conditions. On the other hand, the poor performance of the predicted role given by the role classifier and the inconsistency in the model's ability to select the best role suggest that automatically determining appropriate roles is challenging.

[2]This method is only implemented on LLaMA. We experimented with several different prompts on FLAN-T5 but even the best prompt only resulted in 63% valid answers that include both the chosen role and answer.

7 Conclusion

Incorporating social roles 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 social roles in 26 categories to explore how prompting with social roles affects model performances. Through our analysis, we show that adding social roles consistently improves LLMs' performance over a wide range of types of questions. While we observe that interpersonal roles and gender-neutral roles 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, and could not be easily explained by word frequency, similarity, and perplexity. 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.

8 Limitations

Our study has the following limitations: First, we only studied three 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 social roles, we were not able to experiment with all the social roles beyond the 162 ones in our current experiment. We will release the full list of our social roles to support future research in this direction.

9 Ethical Considerations

Our study has the following ethical implications. First, to ensure the robustness of our results, we experimented with 162 roles, 6 prompt templates, and 3 models over 2457 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 social roles. 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.

Krystyna Strzyzewski Aune, R Kelly Aune, and David B Buller. 1994. The experience, expression, and perceived appropriateness of emotions across levels of relationship development. *The Journal of social psychology*, 134(2):141–150.

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.

Charles H Christiansen. 1999. Defining lives: Occupation as identity: An essay on competence, coherence, and the creation of meaning. *The American Journal of Occupational Therapy*, 53(6):547–558.

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.

Valerian J Derlega. 1984. Self-disclosure and intimate relationships. *Communication, intimacy, and close relationships*, pages 1–9.

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.

Srinivasan Iyer, Xi Victoria Lin, Ramakanth Pasunuru, Todor Mihaylov, Dániel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, et al. 2022. Opt-iml: Scaling language model instruction meta learning through the lens of generalization.

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 gender impact is listed in Table 5. The six 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-2	https://ai.meta.com/llama/license/
OPT	https://huggingface.co/facebook/opt-1.3b/blob/main/LICENSE.md

Table 2: List of licenses

Computational infrastructure and budget The GPU hours required for running experiments on Flan-T5-XXL and LLaMA-2 are each around 100 hours on 8 NVIDIA RTX A6000. For OPT-iml-max-1.3B, it took less than 24 hours on 8 NVIDIA RTX A6000.

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 Model Comparisons

Different LLMs perform exhibit varying performance across datasets, as shown in Figure 11. This variation suggests differences in model capacities and task complexities.

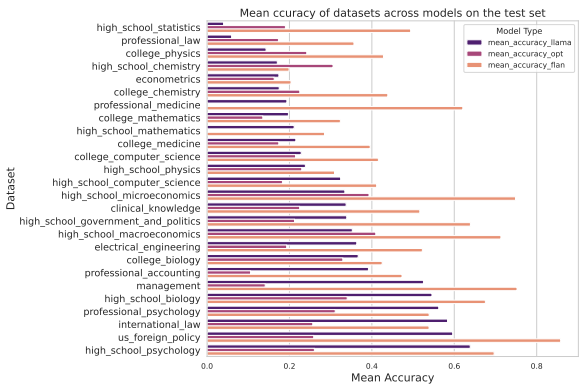


Figure 11: Mean accuracy across datasets on each model on the test set

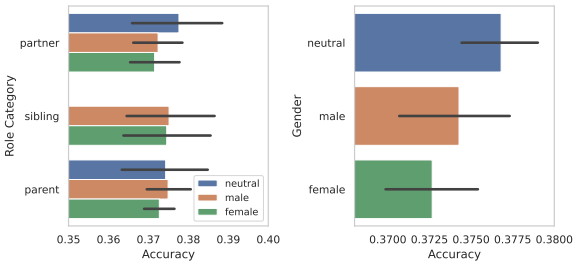


Figure 12: Gender-neutral and male roles lead to higher model performances in both settings.

C Gender Impact Analysis

The performances of gender-neutral and gender-specific roles across domains are shown in Figure 13. The gender impact within the same role type is shown in Figure 12.

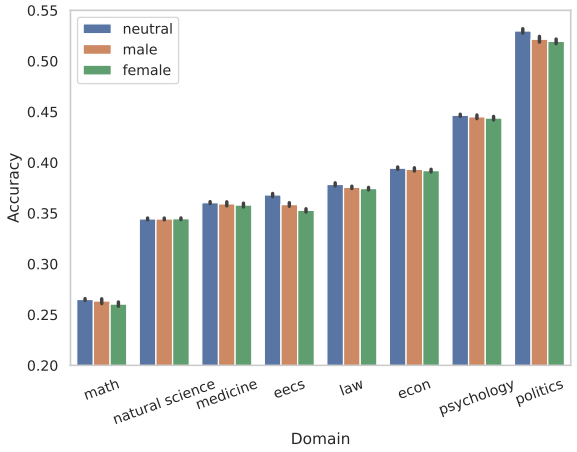


Figure 13: Gender-neutral and male roles lead to higher performances compared with female roles.

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,parents, 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

Gender	Roles
Male	brother, father, daddy, dad, papa, father-in-law, grandfather, husband, son, boyfriend, fiancé
Female	sister, mother, mummy, mamma, mommy, mom, mum, mama, daughter, mother-in-law, grandmother, wife, girlfriend, fiancée, housewife
Neutral	professor, supervisor, instructor, student, coach, tutor, dean, graduate, partner, classmate, companion, buddy, roommate, coworker, boss, colleague, mentor, friend, stranger, foreigner, best friend, close friend, parents, cousin

Table 5: List of roles categorized by gender

Prompt Type	Example
Audience Prompt	You are talking to a/an {role}. Imagine you are talking to a/an {role}.
Role Prompt	You are a/an {role}. Imagine you are a/an {role}.
Interpersonal Prompt	You are talking to your {role}. Imagine you are talking to your {role}.

Table 6: Context prompts

Model Type	Prompt Template
FLAN-T5	{context_prompt} {question} {options} Please select the correct answer number:
LLaMa2/OPT	Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request. Instruction:\n{{You will be presented with a role-playing context followed by a multiple-choice question. {context_prompt} Select only the option number that corresponds to the correct answer for the following question.}}\n\n Input:\n{{{{{question}}}}} Provide the number of the correct option without explaining your reasoning.}} \n\n Response:

Table 7: Context Prompts for each model

Model Type	Prompt Template
FLAN-T5	{question} Please select the correct answer number:
LLaMa2/OPT	Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.\n\n Instruction:\n{{You will be presented with a multiple-choice question. Select only the option number that corresponds to the correct answer for the following question.}}\n \n Input:\n{{{{{question}}}}} Provide the number of the correct option without explaining your reasoning.}} \n \n Response:

Table 8: Control Prompts for each model