RNR: Teaching Large Language Models to Follow Roles and Rules

Kuan Wang [*1] **Alexander Bukharin** [*1] **Haoming Jiang** [2] **Qingyu Yin** [2] **Zhengyang Wang** [2] **Tuo Zhao** [1]
Jingbo Shang [2] **Chao Zhang** [2] **Bing Yin** [2] **Xian Li** [2] **Jianshu Chen** [2] **Shiyang Li** [2]

Abstract

Instruction fine-tuning (IFT) elicits instruction following capabilities and steers the behavior of large language models (LLMs) via supervised learning. However, existing models trained on open-source IFT datasets only have the ability to follow instructions from users, and often fail to follow complex role and rules specified by developers, a.k.a. system prompts. The ability to follow these roles and rules is essential for deployment, as it ensures that the model safely interacts with users within developer defined guidelines. To improve such role and rule following ability, we propose RoleNRules, an automated data generation pipeline that generates diverse roles and rules from existing IFT instructions, along with corresponding responses. This data can then be used to train models that follow complex system prompts. The models are evaluated on our newly created benchmarks for role and rule following ability, as well as standard instruction-following benchmarks and general NLP tasks. Our pipeline significantly improves role and rule following capability in LLMs, as evidenced by over 25% increase in pass-rate on rule adherence, i.e. following all requirements, in our experiments with the Alpaca and Ultrachat datasets. Moreover, our models achieves this increase without any regression on popular instruction following benchmarks.

1. Introduction

Instruction fine-tuning (IFT) (Mishra et al., 2021; Ouyang et al., 2022; Muennighoff et al., 2022; Longpre et al., 2023; Wang et al., 2023b; Zhang et al., 2023a) has emerged as a powerful technique to elicit latent capabilities and steer the behaviors of Large Language Models (LLMs) (Radford et al., 2019; Raffel et al., 2020; Brown et al., 2020; Wei et al., 2022a; Scao et al., 2022; Qin et al., 2023; OpenAI, 2023; Chowdhery et al., 2022). By fine-tuning with carefully curated instruction-response pairs, IFT enables LLMs to generate desirable responses to user instructions (Bach et al., 2022; Wang et al., 2022b;a; Taori et al., 2023b). IFT has exploded in popularity due to its ease of use and low cost, allowing developers to build a wide variety of LLM-based products.

Current IFT research is mainly focused on handling user input instructions. However, this overlooks the need to enable developers to control LLM behavior via specific role and rule requirements. This enhanced level of controllability is necessary, as developers often need to specify the LLM's role in their application ("*You are an AI coding assistant*" or "*You are an AI tax advisor*") as well as specific rules ("*Return the results in XML format*" or "*Respond in less than 300 words*"). To enable such customization ability, many closed-source LLMs like GPT-4 [*] and Claude 2.1 [†] allow developers to input roles and rules as a system prompt, which will then be used to control the LLMs behavior when interacting with the user. However, the techniques used to enable such role and rule following ability are currently not available to the broader research community.

Existing open-source IFT datasets commonly employ a fixed system prompt (e.g., *"You are a helpful, respectful, and honest assistant."*). One workaround to enable role and rule following ability is to prepend these requirements to user inputs, but we find this approach is not very effective. This is because open source IFT models are not trained with such concatenated system-prompts and user inputs, and developer role and rule requirements are often complex and diverse. This approach therefore causes a significant shift between the training and evaluation. Therefore, we find that models often overlook or incorrectly interprets specific role and rule requirements.

To illustrate this challenge, consider the example in Figure

[*]Equal contribution and work was done during Kuan and Alexander' internship at Amazon. [1]Georgia Institute of Technology [2]Amazon. Correspondence to: Shiyang Li <syangli@amazon.com>.

Published at ICML 2024 Workshop on Foundation Models in the Wild. Copyright 2024 by the author(s).

[*]https://platform.openai.com/docs/guides/prompt-engineering/tactic-ask-the-model-to-adopt-a-persona

[†]https://docs.anthropic.com/claude/docs/how-to-use-system-prompts

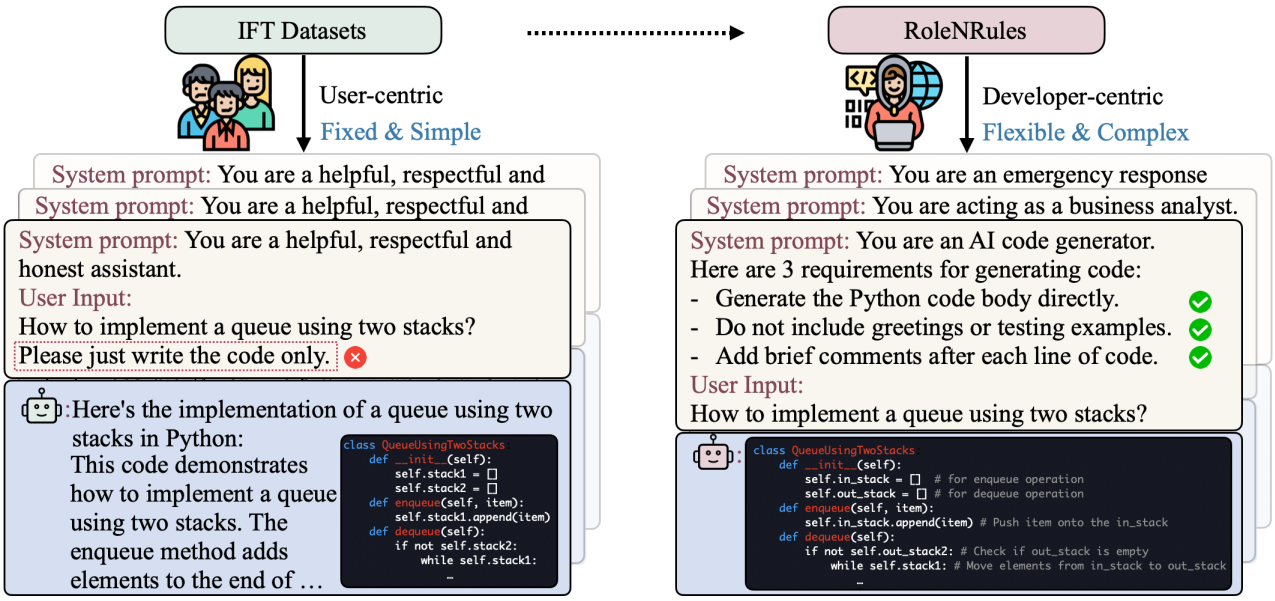

Figure 1. Generated response for a model trained with exisiting IFT datasets (left) and RNR (right). The model fine-tuned with existing IFT datasets ignore the user's extra requirement to write the code only. While the model trained with our RNR can follow all the requirements included in the complex system prompts.

1, where we ask a model trained on a popular open-source IFT dataset to respond to a system prompt and instruction. Even when the developer explicitly requires the system to act as a coder and generate code only, the system ignores this requirement and includes additional information. In contrast, closed-source models offer the ability to effectively adapt to various role and rule requirements. Therefore, there is a clear need for developing open-source IFT methods that support adaptable system prompts to enable developers' fine-grained model controls.

To bridge this gap, we propose RoleNRules (RNR), a scalable and automated data generation pipeline designed to enhance the ability of LLMs to follow developer instructions specified in the system prompt. More specifically, we use powerful language models to generate diverse and high quality (system-prompt, instruction, response) triplets from existing open-source IFT datasets. These triplets are then used for training the model, resulting in superior adaptability to system prompts. As demonstrated in Figure 1, the model trained with RNR can follow all the requirements defined inside the system prompts.

To generate diverse and high-quality (system prompt, instruction, response) triplets, the most challenging part is to create diverse and high-quality system prompts to avoid potential over-fitting of model training and hence improve model's adaptability to new roles and rules. A system prompt often has two parts: 1) **Role** – which explicitly outlines the identity or persona that the LLM agent should assume during the interaction, e.g. "*You are a helpful coding assistant.*". 2) **Rules** – which dictate the requirements, such as operational guidelines (e.g. "*Add brief comments after each line of code*") and objectives, which follow the role description. To generate diverse and high-quality system prompts, we prompt a powerful LLM to generate a novel role with a set of rules for each instruction in an open source IFT dataset. After that, we use the generated system prompt and original instruction to generate a response that follow the instruction while satisfying the system prompt. Through these steps, RNR generates a diverse dataset demonstrating role and rule following abilities, which can then be used to train system prompt following LLMs.

In addition to the RNR data generation pipeline, we also provide a benchmark specifically designed to assess the ability of LLMs to adhere to system prompts. This benchmark provides a set of challenging (system-prompt, instruction) pairs written by both humans and LLMs, as well as a technique to evaluate system-prompt following ability using LLM-based judges.

To validate the effectiveness of our RNR, we conduct extensive experiments using the Alpaca (Taori et al., 2023a) and Ultrachat (Ding et al., 2023) datasets. We train Llama-2 (Touvron et al., 2023) models of various sizes on datasets generated by standard IFT techniques and by RNR. We find that RNR significantly boosts system-prompt following ability, improving the pass-rate by over 25% compared to conventional instruction fine-tuning. In addition, models trained with RNR maintain the ability to follow complex instructions even when there is no specific system prompt, validating RNR's ability to train high performing and adaptive LLMs.

Our primary contributions are:

1. **RoleNRules Generation Pipeline**: We design a general pipeline for automatic and scalable diverse role and rules IFT data generation. Our extensive experiments and human evaluations show that models trained on our data can enable model developers to have fine-grained control of model behaviors via system prompt engineering.
2. **RoleNRules Benchmarks**: We design three high-quality benchmarks via human experts and LLMs to evaluate role and rule following capabilities of LLMs. We believe these datasets can further foster research on this topic.
3. **RoleNRules Evaluation**: Further more, we design a LLM-based technique with a high level of human expert agreement to automatically evaluate the performance of models on the role and rules following benchmarks.

2. Method

We propose RoleNRules, a data generation pipeline for the automatic generation of RNR system prompts and the corresponding responses for IFT datasets. As illustrated in Figure 3, the RoleNRules pipeline augments an existing open-source IFT dataset into our RNR dataset with complex system-prompt data. Below, we describe the main processes in our pipeline.

2.1. RoleNRules System Prompt

Our method for improving role and rules following generates diverse and high quality system prompts from existing instruction tuning datasets. These system prompts are designed as a combination of a role description and a set of defined rules:

1. **Role**: The role description explicitly specifies the identity or persona that the LLM should adopt during the interaction. For example, this role could be a technical assistant, customer service representative, data analyst, or any other persona with specific characteristics and knowledge. By defining the role, we guide the model to adopt a particular set of behaviors and language styles appropriate for that persona.
2. **Rules**: Accompanying the role description are specific rules that set the boundaries and objectives of the interaction. These rules encompass operational guidelines, ethical considerations, data handling procedures, and specific goals to be achieved. For example, a rule might mandate prioritizing data privacy or adhering to industry-specific regulations.

By integrating these two elements, we can simulate a wide variety of possible developer requirements. In particular, the role descriptions make the LLM's response style customizable, while the rules ensure that the model operates within specific requirements. In addition, this approach enhances the LLM's ability to understand and execute complex and specialized tasks. And it also provides a pipeline for developers to tailor the model's responses for various professional scenarios.

2.2. System Prompt Generation

To generate unique system prompts comprised of different role and rules, we prompt powerful LLMs to generate role and rules based on an instruction from existing open-source datasets. We generate role and rules based on existing IFT datasets as existing IFT datasets contain instructions from a diverse range of subjects and tasks, which will in turn make our generated role and rules diverse.

To generate a role and rules based system prompt from an existing instruction, we sample a random variable K from $0-8$ and ask Claude-2 to generate a role and K rules based on a set of guidelines. This can be seen in Figure 3. These guidelines ensure that the role and rules are reasonable for a LLM to follow and that the role and rules are not overly specific. In addition, we provide three expert-generated demonstrations to help the LLM generate new roles and rules. Our complete prompt can be found in Appendix H. In our experiments we use Claude-2 to generate the roles and rules, but in practice other powerful LLMs can be used.

2.3. Response Generation

Although existing IFT datasets contain responses to the instructions, likely, the instructions do not follow the generated role and rules. Therefore, once we generate a system prompt for a given instruction from the IFT dataset, we combine it with the instruction to form a new prompt for response generation. This is demonstrated in Figure 2. Notably, we only use the instructions from the IFT dataset without incorporating the original responses. This approach prevents overfitting to the original responses and preserves diversity in the generated requirements.

2.4. Fine-tuning

Once we have generated a dataset of (system prompt, instruction, response) triplets, we finetune a pre-trained LLM to maximize the probability of the response given the combined system prompt and instruction. To combine the system prompt and instruction, we follow the same template as in Figure 2. This combined system prompt is used as the input for supervised fine tuning, where the target output is the response. Figure 3 is an overview of the RNR. In addition,

we display examples of our training data in Appendix G.

```
System Prompt: {Role} {Rules}
When I write START you will enter this role
START
Instruction: {Instruction}
Response: {Response}
```

Figure 2. We use this template to integrate the responses with the generated system prompts and instructions.

3. Experiments

3.1. Experimental Setup

Training Setup. We present our main results on the Alpaca dataset (Taori et al., 2023b) and the Ultrachat dataset (Ding et al., 2023). For Alpaca, we generate (system prompt, instruction, response) triplets on the whole dataset and for Ultrachat, we sample a subset of size 50K to generate the triplets. For our main experiments, we use Llama 2 7B (Touvron et al., 2023) as our base model and fine-tune each model for two epochs with a batch size of 128 using a context length of 2048 and a learning rate of 2×10^{-5}.

Baselines. To investigate the efficacy of RNR, we compare it with two IFT baselines. For both of these baselines we generate responses to the original training instructions with Claude-2, so that there is no discrepancy in the teacher model used for training. Once we generate these responses, we train models using (a) a fixed system prompt ("You are a helpful, harmless, and honest AI assistant") and (b) with no system prompt. These baselines are denoted as IFT: fixed-system and IFT: no-system, respectively.

3.2. Benchmarks

We design three benchmarks using the RNR format (system prompt + instruction) to evaluate the effectiveness of RoleNRules:

1. **RNR-Expert benchmark**: This is a small benchmark comprised of 50 data points with RNR system prompts and instructions handcrafted by domain experts, who are LLM researchers and are familiar with prompting techniques. The roles and rules in these system prompts cover diverse domains including programming, engineering, medicine, law, finance, movies, book recommendations, travel advice, and cooking recipes, etc. We remark that this dataset contains challenging and high quality examples.

2. **RNR-Awesome benchmark**: This is benchmark has 495 data points based on the instruction collected by awesome-chatgpt-prompts (Akın, 2023). For each instruction in awesome-chatgpt-prompts, we randomly select 6 samples from the RNR-Expert benchmark as the few-shot prompt, then leverage GPT-4 to generate the RoleNRules system prompt for this instruction. Given this system-prompt, we also use GPT-4 to generate 2 more possible instructions, so the amount of data points in RNR/Awesome is 3 times of the number of prompts in the awesome-chatgpt-prompts dataset.

3. **RNR-AlpacaFarm benchmark**: This benchmark contain 805 datapoints, and the instructions are kept the same as the original AlpacaFarm (Dubois et al., 2023). We also use 6 random samples from the RNR-Expert benchmark as the few-show prompt and leverage GPT-4 to generate the RoleNRules system prompt for each instruction in AlpacaFarm.

Evaluation. To evaluate the system-prompt following ability of each model, we use large language models to calculate the number of rules and roles that are satisfied for each prompt. In particular, we develop a novel chain-of-thought (Wei et al., 2022b) like prompt for evaluating RoleNRules following ability. More concretely, we ask the judge LLM to examine each role and rule requirement individually, and then aggregate these score into a final judgement. In addition, we include some few-shot examples to further boost performance. The exact prompt can be found in Appendix I. In our human study, we find that this evaluation strategy has a high level of agreements with an ensemble of human experts.

In our experiments, we use Claude-2.1 to calculate the percentage of rules in the system prompt that each model passes over each RNR benchmark dataset. We remark that we use Claude-2.1 as the evaluator instead of Claude-2 to mitigate the effect of LLM judges favoring their own output. We calculate the rule-level pass percentage (defined as the percent of rules each model passes) and the prompt-level pass percentage (defined as the percent of prompts where all rules are satisfied).

Beyond measuring system prompt following ability, we also measure the ability of each model to follow user-inputted rules with IFEval (Zhou et al., 2023b), role-playing ability with RoleBench (Wang et al., 2023c), instruction following ability with Alpacafarm (Dubois et al., 2023), and reasoning ability with benchmarks such as MMLU (Hendrycks et al., 2020), BBH (Suzgun et al., 2022), and DROP (Dua et al., 2019). For IFEval, we can programmatically evaluate the percent of rules each model passes. For RoleBench and Alpacafarm we compare outputs from the RNR model to outputs from the baselines to evaluate which model better follows the defined role or better follows the instruction. To judge each pair of responses, we use Claude-2.1 to output its preferred response in a similar manner to (Dubois et al., 2023). In order to mitigate the positional bias of the judge LLM, we show it each comparison in both possible orders

Table 1. Pass rate on the main evaluation sets. Llama-2-7B is the base model for RNR.

RNR:Alpaca	**RNR-Expert**		**RNR-Awesome**		**RNR-Alpacafarm**		**Average**	
	Rule Level	Prompt Level	Rule Level	Prompt Level	Rule Level	Prompt Level	Rule Level	Prompt Level
IFT: fixed-system	53.62	39.53	58.39	28.6	72.57	43.21	61.53	37.11
IFT: no-system	64.76	57.78	67.59	39.05	74.41	44.47	68.92	47.10
Rolenrule	**82.24**	**77.78**	**74.91**	**49.07**	**86.34**	**67.30**	**81.16**	**64.72**
RNR:Ultrachat	Rule Level	Prompt Level	Rule Level	Prompt Level	Rule Level	Prompt Level	Rule Level	Prompt Level
IFT: fixed-system	61.18	48.89	57.51	29.10	76.14	49.56	64.94	42.52
IFT: no-system	80.50	70.21	63.07	36.49	75.23	51.09	72.93	52.60
Rolenrule	**85.56**	**75.0**	**75.17**	**49.59**	**89.56**	**74.37**	**83.43**	**66.32**

(with RNR first and with RNR second). We only evaluate our Alpaca based models on RoleBench and Alpacafarm to reduce evaluation costs.

3.3. Main Results

Our main results on the RNR benchmark can be found in Table 1. The results show that RNR can significantly improve system-prompt following ability compared to the two IFT baselines for both the Alpaca and Ultrachat datasets. In particular, we find that RNR improves prompt-level system-prompt following ability by 37.4% on the Alpaca dataset and by 26.1% for Ultrachat compared to the strongest baseline. These results indicate that RNR is an effective way to improve system-prompt following ability.

The results for IFEval can be found in Table 2. From Table 2 we can see that RNR significantly boosts instruction following ability, even when the instruction is inputted by the user, indicating that improve system prompt following capability can positively transferred to user inputs.

For RoleBench, we find that RNR improves the winning score (defined as $\frac{\text{\# Win - \# Lose}}{\text{\#Win + \#Lose + \#Tie}} + 1$) by over 25% compared to both IFT baselines on the Alpaca dataset. The results can be found in Table 7. This indicates RNR significantly enhances role-playing ability.

For evaluation on Alpacafarm, we do a head to head comparison of RNR:Alpaca and IFT:no-system. We find that the performance is similar for IFT:no-system and RNR, with RNR outperforming the IFT model $\sim 3.7\%$. (See Appendix C). Therefore RNR can boost complex system-prompt following ability without hurting general instruction following performance. Finally, we demonstrate the performance of RNR on various benchmarks in Table 3. As expected, the benchmark performance is similar for all models, which shows that RNR retains general reasoning capabilities.

3.4. Human Evaluation

To validate the efficacy of RNR, we conducted a human evaluation experiment in addition to using public LLM APIs for response evaluation. In this experiment, expert human researchers evaluated the prompt-level pass rate of both the RNR model and the IFT: no-system model using an expert-curated set from the RNR benchmark. For further details, see Section B.

Table 2. Pass rate on the IFEval. Llama-2-7B is the base model.

	IFEval:Loose		**IFEval:Strict**	
Alpaca	Rule Level	Prompt Level	Rule Level	Prompt Level
IFT: fixed-system	39.69	27.73	32.25	19.69
IFT: no-system	39.69	26.80	32.01	19.41
RNR	**51.68**	**39.57**	**43.65**	**31.61**
Ultrachat	Rule Level	Prompt Level	Rule Level	Prompt Level
IFT: fixed-system	39.69	27.34	30.70	19.04
IFT: no-system	40.17	28.10	30.70	19.04
RNR	**47.84**	**34.57**	**40.17**	**27.17**

Table 3. Performance on standard NLP benchmarks.

Alpaca	MMLU	BBH	DROP
IFT: fixed-system	**44.4**	32.78	**26.86**
IFT: no-system	43.4	31.56	26.79
Rolenrule	41.2	**33.55**	25.49
Ultrachat	MMLU	BBH	DROP
IFT: fixed-system	46.1	**34.24**	20.10
IFT: no-system	**46.2**	33.40	19.87
Rolenrule	45.8	33.12	**27.65**

4. Conclusion

In conclusion, our study presents a new approach to instruction tuning for Large Language Models (LLMs), focusing on both user-centric instructions and developer-defined roles and rules. Our novel system prompt structure and automated data generation pipeline contribute to significant improvements in rule-following capabilities in LLMs, as evidenced by enhanced performance on our benchmarks using the Alpaca and Ultrachat datasets. Despite these advancements, our work has limitations, including the absence of a quality-based filtering mechanism in our data generation and the framework's restriction to single-turn interactions. Future efforts will focus on addressing these issues, aiming to refine the model's effectiveness in more complex, multi-turn conversation settings.

References

Akın, F. K. awesome-chatgpt-prompts: A curated list of awesome chatgpt prompts. https://github.com/f/awesome-chatgpt-prompts, 2023.

Bach, S. H., Sanh, V., Yong, Z.-X., Webson, A., Raffel, C., Nayak, N. V., Sharma, A., Kim, T., Bari, M. S., Fevry, T., et al. Promptsource: An integrated development environment and repository for natural language prompts. *arXiv preprint arXiv:2202.01279*, 2022.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

Chen, L., Li, S., Yan, J., Wang, H., Gunaratna, K., Yadav, V., Tang, Z., Srinivasan, V., Zhou, T., Huang, H., and Jin, H. Alpagasus: Training a better alpaca model with fewer data. In *The Twelfth International Conference on Learning Representations*, 2024.

Chen, Z., Li, S., Smiley, C., Ma, Z., Shah, S., and Wang, W. Y. ConvFinQA: Exploring the chain of numerical reasoning in conversational finance question answering. In Goldberg, Y., Kozareva, Z., and Zhang, Y. (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 6279–6292, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.421. URL https://aclanthology.org/2022.emnlp-main.421.

Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J. E., et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2023.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, E., Wang, X., Dehghani, M., Brahma, S., et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

Ding, N., Chen, Y., Xu, B., Qin, Y., Zheng, Z., Hu, S., Liu, Z., Sun, M., and Zhou, B. Enhancing chat language models by scaling high-quality instructional conversations. *arXiv preprint arXiv:2305.14233*, 2023.

Dua, D., Wang, Y., Dasigi, P., Stanovsky, G., Singh, S., and Gardner, M. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. *arXiv preprint arXiv:1903.00161*, 2019.

Dubois, Y., Li, X., Taori, R., Zhang, T., Gulrajani, I., Ba, J., Guestrin, C., Liang, P., and Hashimoto, T. B. Alpacafarm: A simulation framework for methods that learn from human feedback. *arXiv preprint arXiv:2305.14387*, 2023.

Eisenstein, J. Informativeness and invariance: Two perspectives on spurious correlations in natural language. *arXiv preprint arXiv:2204.04487*, 2022.

Elazar, Y., Ravfogel, S., Jacovi, A., and Goldberg, Y. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021.

Frohberg, J. and Binder, F. Crass: A novel data set and benchmark to test counterfactual reasoning of large language models. *arXiv preprint arXiv:2112.11941*, 2021.

Geiger, A., Lu, H., Icard, T., and Potts, C. Causal abstractions of neural networks. *Advances in Neural Information Processing Systems*, 34:9574–9586, 2021.

Geiger, A., Wu, Z., Lu, H., Rozner, J., Kreiss, E., Icard, T., Goodman, N., and Potts, C. Inducing causal structure for interpretable neural networks. In *International Conference on Machine Learning*, pp. 7324–7338. PMLR, 2022.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Honovich, O., Scialom, T., Levy, O., and Schick, T. Unnatural instructions: Tuning language models with (almost) no human labor. *arXiv preprint arXiv:2212.09689*, 2022.

Jang, J., Ye, S., and Seo, M. Can large language models truly understand prompts? a case study with negated prompts. In *Transfer Learning for Natural Language Processing Workshop*, pp. 52–62. PMLR, 2023.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Kıcıman, E., Ness, R., Sharma, A., and Tan, C. Causal reasoning and large language models: Opening a new frontier for causality. *arXiv preprint arXiv:2305.00050*, 2023.

Kusner, M. J., Loftus, J., Russell, C., and Silva, R. Counterfactual fairness. *Advances in neural information processing systems*, 30, 2017.

Lee, H., Phatale, S., Mansoor, H., Lu, K., Mesnard, T., Bishop, C., Carbune, V., and Rastogi, A. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*, 2023.

Li, C., Leng, Z., Yan, C., Shen, J., Wang, H., MI, W., Fei, Y., Feng, X., Yan, S., Wang, H., et al. Chatharuhi: Reviving anime character in reality via large language model. *arXiv preprint arXiv:2308.09597*, 2023a.

Li, S., Chen, J., and Yu, D. Teaching pretrained models with commonsense reasoning: A preliminary kb-based approach. *ArXiv*, abs/1909.09743, 2019. URL https://api.semanticscholar.org/CorpusID:202719139.

Li, S., Chen, J., Shen, Y., Chen, Z., Zhang, X., Li, Z., Wang, H., Qian, J., Peng, B., Mao, Y., et al. Explanations from large language models make small reasoners better. *arXiv preprint arXiv:2210.06726*, 2022.

Li, S., Yan, J., Wang, H., Tang, Z., Ren, X., Srinivasan, V., and Jin, H. Instruction-following evaluation through verbalizer manipulation. *arXiv preprint arXiv:2307.10558*, 2023b.

Lin, B. Y., Fu, Y., Yang, K., Ammanabrolu, P., Brahman, F., Huang, S., Bhagavatula, C., Choi, Y., and Ren, X. Swiftsage: A generative agent with fast and slow thinking for complex interactive tasks. *ArXiv preprint*, abs/2305.17390, 2023. URL https://arxiv.org/abs/2305.17390.

Liu, R., Yang, R., Jia, C., Zhang, G., Zhou, D., Dai, A. M., Yang, D., and Vosoughi, S. Training socially aligned language models in simulated human society. *arXiv preprint arXiv:2305.16960*, 2023.

Longpre, S., Hou, L., Vu, T., Webson, A., Chung, H. W., Tay, Y., Zhou, D., Le, Q. V., Zoph, B., Wei, J., et al. The flan collection: Designing data and methods for effective instruction tuning. *arXiv preprint arXiv:2301.13688*, 2023.

Mishra, S., Khashabi, D., Baral, C., and Hajishirzi, H. Cross-task generalization via natural language crowdsourcing instructions. *arXiv preprint arXiv:2104.08773*, 2021.

Muennighoff, N., Wang, T., Sutawika, L., Roberts, A., Biderman, S., Scao, T. L., Bari, M. S., Shen, S., Yong, Z.-X., Schoelkopf, H., et al. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*, 2022.

Nabi, R. and Shpitser, I. Fair inference on outcomes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

OpenAI. Chatgpt. Online, 2022. URL https://openai.com/blog/chatgpt.

OpenAI. Gpt-4 technical report. *OpenAI blog*, 2023.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., and Bernstein, M. S. Generative agents: Interactive simulacra of human behavior. In *In the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*, UIST '23, New York, NY, USA, 2023. Association for Computing Machinery.

Peng, B., Li, C., He, P., Galley, M., and Gao, J. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*, 2023.

Qian, J., Wang, H., Li, Z., Li, S., and Yan, X. Limitations of language models in arithmetic and symbolic induction. In Rogers, A., Boyd-Graber, J., and Okazaki, N. (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9285–9298, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.516. URL https://aclanthology.org/2023.acl-long.516.

Qin, C., Zhang, A., Zhang, Z., Chen, J., Yasunaga, M., and Yang, D. Is chatgpt a general-purpose natural language processing task solver?, 2023.

Qin, L., Bosselut, A., Holtzman, A., Bhagavatula, C., Clark, E., and Choi, Y. Counterfactual story reasoning and generation. *arXiv preprint arXiv:1909.04076*, 2019.

Qin, L., Shwartz, V., West, P., Bhagavatula, C., Hwang, J., Bras, R. L., Bosselut, A., and Choi, Y. Back to the future: Unsupervised backprop-based decoding for counterfactual and abductive commonsense reasoning. *arXiv preprint arXiv:2010.05906*, 2020.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21 (140):1–67, 2020.

Salemi, A., Mysore, S., Bendersky, M., and Zamani, H. Lamp: When large language models meet personalization. *arXiv preprint arXiv:2304.11406*, 2023.

Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A. S., Yvon, F., Gallé, M., et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022.

Shanahan, M., McDonell, K., and Reynolds, L. Role play with large language models. *Nature*, pp. 1–6, 2023.

Sumers, T., Yao, S., Narasimhan, K., and Griffiths, T. L. Cognitive architectures for language agents. *arXiv preprint arXiv:2309.02427*, 2023.

Suzgun, M., Scales, N., Schärli, N., Gehrmann, S., Tay, Y., Chung, H. W., Chowdhery, A., Le, Q. V., Chi, E. H., Zhou, D., et al. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*, 2022.

Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., and Hashimoto, T. B. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023a.

Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., and Hashimoto, T. B. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models. https://crfm.stanford. edu/2023/03/13/alpaca. html*, 3(6):7, 2023b.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Tunstall, L., Beeching, E., Lambert, N., Rajani, N., Rasul, K., Belkada, Y., Huang, S., von Werra, L., Fourrier, C., Habib, N., et al. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*, 2023.

Veitch, V., D'Amour, A., Yadlowsky, S., and Eisenstein, J. Counterfactual invariance to spurious correlations in text classification. *Advances in Neural Information Processing Systems*, 34:16196–16208, 2021.

Wang, P., Li, L., Chen, L., Zhu, D., Lin, B., Cao, Y., Liu, Q., Liu, T., and Sui, Z. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*, 2023a.

Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N. A., Khashabi, D., and Hajishirzi, H. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*, 2022a.

Wang, Y., Mishra, S., Alipoormolabashi, P., Kordi, Y., Mirzaei, A., Arunkumar, A., Ashok, A., Dhanasekaran, A. S., Naik, A., Stap, D., et al. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. *arXiv preprint arXiv:2204.07705*, 2022b.

Wang, Z., Zhang, G., Yang, K., Shi, N., Zhou, W., Hao, S., Xiong, G., Li, Y., Sim, M. Y., Chen, X., et al. Interactive natural language processing. *arXiv preprint arXiv:2305.13246*, 2023b.

Wang, Z. M., Peng, Z., Que, H., Liu, J., Zhou, W., Wu, Y., Guo, H., Gan, R., Ni, Z., Zhang, M., et al. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. *arXiv preprint arXiv:2310.00746*, 2023c.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., and Fedus, W. Emergent abilities of large language models, 2022a.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E., Le, Q., and Zhou, D. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022b.

Wei, J., Shuster, K., Szlam, A., Weston, J., Urbanek, J., and Komeili, M. Multi-party chat: Conversational agents in group settings with humans and models. *arXiv preprint arXiv:2304.13835*, 2023.

Weng, L. Llm-powered autonomous agents. *lilianweng.github.io*, Jun 2023. URL https://lilianweng.github.io/posts/2023-06-23-agent/.

Xu, B., Peng, Z., Lei, B., Mukherjee, S., Liu, Y., and Xu, D. Rewoo: Decoupling reasoning from observations for efficient augmented language models. *arXiv preprint arXiv:2305.18323*, 2023a.

Xu, C., Guo, D., Duan, N., and McAuley, J. Baize: An open-source chat model with parameter-efficient tuning on self-chat data. *arXiv preprint arXiv:2304.01196*, 2023b.

Xu, C., Sun, Q., Zheng, K., Geng, X., Zhao, P., Feng, J., Tao, C., and Jiang, D. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*, 2023c.

Zhang, G., Shi, Y., Liu, R., Yuan, R., Li, Y., Dong, S., Shu, Y., Li, Z., Wang, Z., Lin, C., et al. Chinese open instruction generalist: A preliminary release. *arXiv preprint arXiv:2304.07987*, 2023a.

Zhang, X., Tian, C., Yang, X., Chen, L., Li, Z., and Petzold, L. R. Alpacare:instruction-tuned large language models for medical application, 2023b.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*, 2023.

Zhou, C., Liu, P., Xu, P., Iyer, S., Sun, J., Mao, Y., Ma, X., Efrat, A., Yu, P., Yu, L., et al. Lima: Less is more for alignment. *arXiv preprint arXiv:2305.11206*, 2023a.

Zhou, J., Lu, T., Mishra, S., Brahma, S., Basu, S., Luan, Y., Zhou, D., and Hou, L. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023b.

A. Related Works

A.1. Instruction Fine-tuning

Instruction Fine-tuning (IFT) has emerged as a pivotal technique in enhancing the capabilities and controllability of LLMs. The core of IFT lies in the generation and selection of instruction data (Chen et al., 2024; Chung et al., 2022; Lee et al., 2023; Honovich et al., 2022; Wang et al., 2022a; Zhou et al., 2023a), which leverage existing LLMs like InstructGPT (Ouyang et al., 2022) and ChatGPT (OpenAI, 2022) to recursively generate instruction sets. Since the introduction of LLaMA (Touvron et al., 2023), a surge in IFT research has led to notable models (Taori et al., 2023b; Chiang et al., 2023; Peng et al., 2023). Our approach in instruction fine-tuning (IFT) distinctly focuses on system prompts, which are commonly a fixed one in previous works (e.g., *"You are a helpful, respectful, and honest assistant."*). Instead, we focus on developing LLMs, which enables model developers to have fine-grained controls of model behaviors via system prompts.

A.2. LLM Role Playing

LLMs have demonstrated the potential to act as advanced agents (Ouyang et al., 2022; Bubeck et al., 2023; Wei et al., 2022a; Weng, 2023; Sumers et al., 2023; Park et al., 2023; Liu et al., 2023; Lin et al., 2023; Xu et al., 2023a). Recent research focuses on assigning specific roles to large language model agents to enhance their interactivity, personalization, and ability to handle complex or specific tasks (Zhang et al., 2023b; Wang et al., 2023c; Wei et al., 2023; Shanahan et al., 2023; Li et al., 2023a; Salemi et al., 2023). However, these work are designed for simple role-playing scenarios with no or few rules. Instead, RoleNRules is design for model developers, who often require flexible and complex requirements for the simulated roles of their LLM products.

A.3. LLM Evaluation

The evaluation of language models has evolved significantly, with a transition from focusing on perplexity and standard benchmarks to more comprehensive assessments including aspects such as commonsense reasoning (Li et al., 2019; Kıcıman et al., 2023; Li et al., 2022), counterfactual reasoning (Qin et al., 2019; 2020; Frohberg & Binder, 2021; Kıcıman et al., 2023), numerical reasoning (Chen et al., 2022; Qian et al., 2023), interpretability (Elazar et al., 2021; Geiger et al., 2021; 2022), spurious correlation detection (Veitch et al., 2021; Eisenstein, 2022), and fairness (Kusner et al., 2017; Nabi & Shpitser, 2018). Recent works also leverage the capabilities of powerful models like GPT-4 for evaluating weaker LLMs, albeit with noted shortcomings (Xu et al., 2023c;b; Wang et al., 2023a). Moreover, efforts have been made towards automating the instruction-following evaluation with minimum human intervention, with studies exploring the performance of models in instruction-only settings (Zhou et al., 2023b) and assessing the impact of negation (Jang et al., 2023) and verbalizer manipulations (Li et al., 2023b). However, existing benchmarks fail to evaluate the ability of LLMs to follow complex role and rules. To bridge this gap, we design multiple benchmarks to evaluate role and rule follow capabilities and at the same time, automatic LLM-based evaluation methods with high human agreement. To the best of our knowledge, we are the first to systematically study complex role and rule following capabilities of LLMs.

B. Analysis

Human Evaluation. In addition to using public LLM APIs to evaluate the responses generated by RNR models and baselines, we also design a human evaluation experiment to validate the efficacy of RNR. For this experiment, we use the expert-curated set of the RNR benchmark, and ask the expert human researchers to evaluate the prompt-level pass rate of the RNR model and the IFT: no-system model.

We have 3 participants for human evaluation, and each of them is provided with a system-prompt, instruction, and set of response pairs. These response pairs consist of an output from our model and a baseline model, without revealing the source of each response. We randomize the order to minimize the effect of evaluator bias. The evaluators are tasked with analyzing the responses based on two criteria (1) whether the responses are in the style of the roles and (2) whether the responses follow the rules in the system prompt. If both of these criteria are met, the response is given a passing score, otherwise it is assigned a failing score. We also tell the evaluators to ignore the minor factual or mathematical errors in the response, as we focus on the system-prompt following ability of RNR.

The results of human evaluation can be seen in in Figure 4. As we can see from this figure, RNR has significantly better pass percentage, both when considering strict evaluation (All Pass) and non-strict evaluation (Majority Pass). We provide two

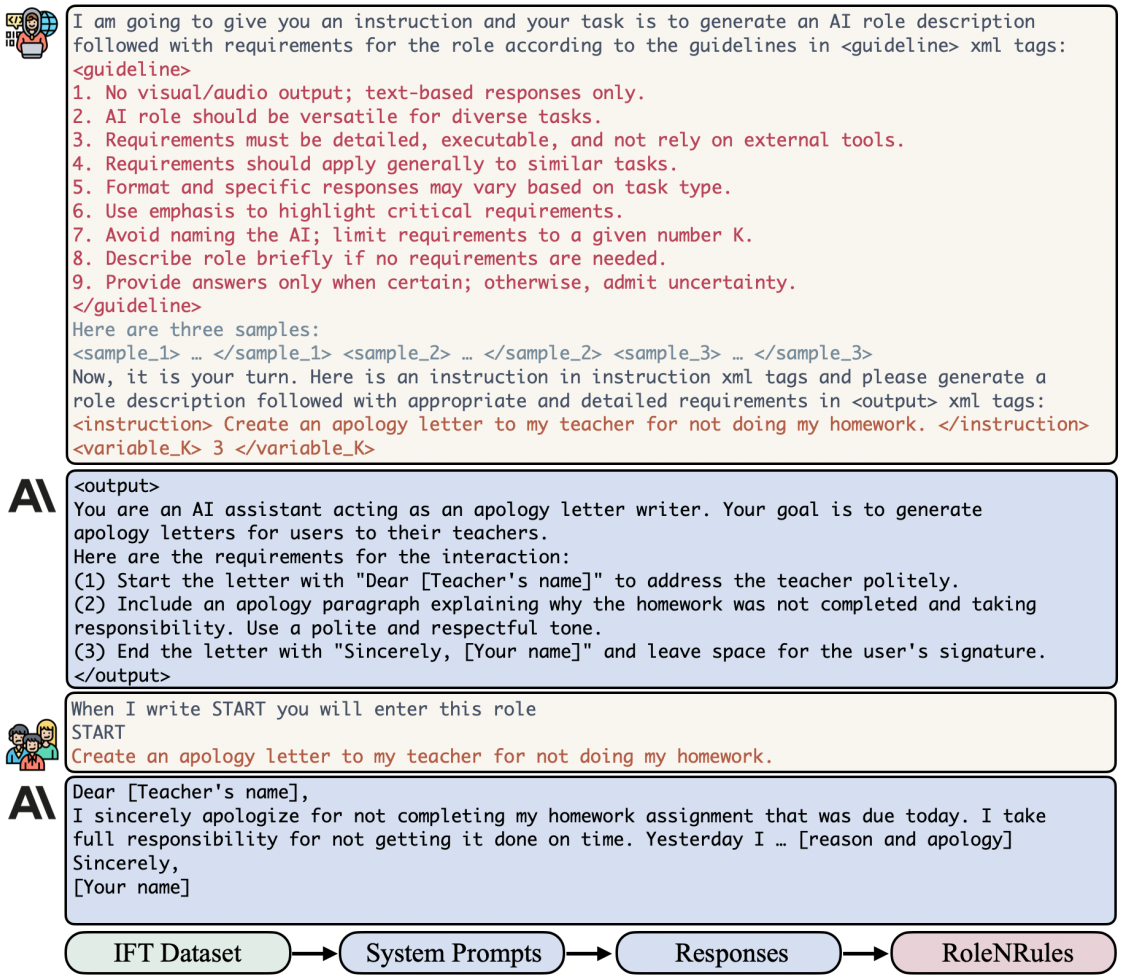

Figure 3. To construct (system prompt, instruction, response) triplets, RNR utilizes existing open-source IFT datasets and a powerful LLM, such as Claude2. First, we crafted a comprehensive guideline, accompanied by a few-shot demonstration. These serve as prompts for the LLM, guiding it to generate informative system prompts for the instructions in the given IFT dataset. Subsequently, we merge the generated system prompts with their corresponding instructions and feed them into the LLM to obtain the responses. Finally, we assemble these components into (system prompt, instruction, response) triplets to form the final dataset.

cases to compare our methods and baselines used for human evaluation in Appendix F.

We also evaluate the agreement between the human evaluators and Claude-2.1 on the RNR-Expert evaluation dataset. We find that Claude-2.1 and the majority vote of the human annotators agree 73.97 % of the time, and that human evaluators agree 77.32 % of the time. This high agreement rate validates that our evaluation strategy works, and further advocates for the use of the RNR benchmark when evaluating the system-prompt following ability of LLMs.

Varying the Amount of RNR Data. We investigate how the amount of rolenrule data affects rule following ability in Figure 5. From this figure we observe that only a small amount of RNR data (1000 data points) will significantly improve system-prompt following ability compared to the baselines. In addition, increasing the amount of RNR data included further boosts performance, further confirming RNR's efficacy.

Table 4. Pass rate on the main evaluation sets with Llama-2-13B as the base model.

RNR:Alpaca	**RNR-Expert**		**RNR-Awesome**		**RNR-Alpacafarm**		**Average**	
	Rule Level	Prompt Level	Rule Level	Prompt Level	Rule Level	Prompt Level	Rule Level	Prompt Level
IFT: fixed-system	68.14	55.56	65.21	34.77	80.92	54.17	71.42	48.17
IFT: no-system	83.5	70.0	75.29	50.61	86.59	65.08	81.79	61.90
RNR	**91.06**	**80.0**	**78.12**	**51.84**	**91.10**	**76.19**	**86.76**	**69.34**

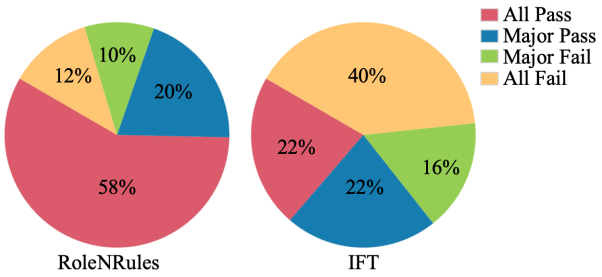

Figure 4. Human expert evaluation of pass rate at the prompt level.The IFT baseline uses no system system.

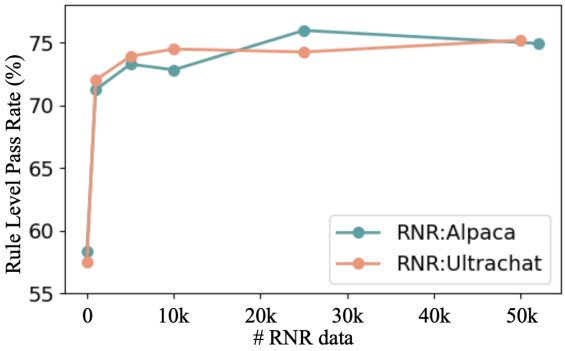

Figure 5. Ablation on the amount of RoleNRules data included. The evaluation dataset is the Awesome split.

Experiments with Larger Models. We investigate how the rule following ability of language models changes as the model size is increased by running experiments on Llama-2-13B. The results can be seen in Table B, where we can see that system-prompt following ability is increased with larger model size. In addition, RNR still significantly boost performance compared to the baselines, suggesting RNR will scale well to even larger model sizes.

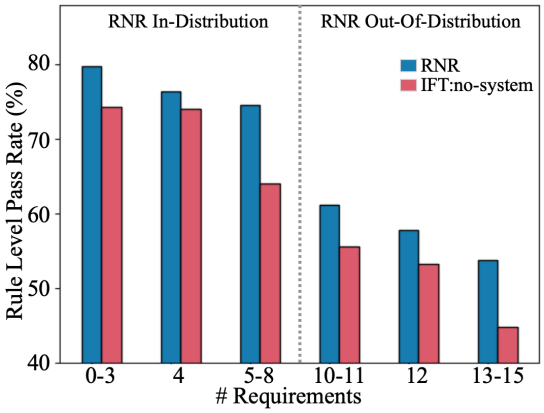

Figure 6. Rule following performance of RNR when different amounts of requirements are included.

Varying the Amount of Requirements. We investigate how the performance of RNR changes with varying amounts of requirements in Figure 6. The evaluation dataset used here is RNR-Awesome, and we generated an additional 100 testing examples that contain from 10-15 examples. We observe that the performance of both RNR and conventional IFT drops as we use more requirements, even though RNR outperforms IFT in all settings. Interestingly, we find that there is a larger drop in performance when the number of requirements increases from 0-8 to more than 10. This is likely due to the fact that RNR was only trained on system prompts with 0-8 instructions, and therefore may fail to generalize well when we have a large amount of instructions. Nonetheless, we find that RNR still significantly outperforms IFT in the setting with many requirements, confirming its efficacy.

Benchmarking the Performance of Popular Models. We evaluate the performance of Claude-2, Zephyr-7b-beta (Tunstall

et al., 2023), Vicuna-7b-v1.5 (Zheng et al., 2023), and Mistral-7b-Instruct-v0.1 (Jiang et al., 2023) in Table 5. In terms of average prompt level accuracy, we find that Claude-2 performs best (79.54%), followed RNR:Ultrachat (66.32%), Mistral-7b (66.24 %), Zephyr-7b (66.08 %), and Vicuna-7b (63.23 %). Notably, RNR significantly outperforms the only other Llama-2 based model (Vicuna) while still beating the Mistral models, despite these models being trained on a more powerful base model, with more advanced optimization techniques such as DPO, and a more powerful teacher model (GPT-4).

Table 5. Pass rate on the main evaluation sets. Llama-2-7B is the base model for RNR.

Closed LLM	**In-house Eval**		**RolenruleEval**		**Alpacafarm-system**		**Average**	
	Rule Level	Prompt Level	Rule Level	Prompt Level	Rule Level	Prompt Level	Rule Level	Prompt Level
Claude-2	94.22	88.0	83.05	65.15	93.71	85.48	90.33	79.54
Open LLM	**In-house Eval**		**RolenruleEval**		**Alpacafarm-system**		**Average**	
	Rule Level	Prompt Level	Rule Level	Prompt Level	Rule Level	Prompt Level	Rule Level	Prompt Level
Vicuna-7b-v1.5	79.70	65.96	76.64	56.02	85.47	67.72	80.60	63.23
Mistral-7B-Instruct-v0.1	86.96	70.0	76.51	55.39	88.10	73.33	83.86	66.24
Zephyr-7b-beta	70.58	64.44	79.15	54.98	91.46	78.83	80.40	66.08
RNR:Alpaca	**RNR-Expert**		**RNR-Awesome**		**RNR-Alpacafarm**		**Average**	
	Rule Level	Prompt Level	Rule Level	Prompt Level	Rule Level	Prompt Level	Rule Level	Prompt Level
IFT: fixed-system	53.62	39.53	58.39	28.6	72.57	43.21	61.53	37.11
IFT: no-system	64.76	57.78	67.59	39.05	74.41	44.47	68.92	47.10
Rolenrule	**82.24**	**77.78**	**74.91**	**49.07**	**86.34**	**67.30**	**81.16**	**64.72**
RNR:Ultrachat	**In-house Eval**		**RolenruleEval**		**Alpacafarm-system**		**Average**	
	Rule Level	Prompt Level	Rule Level	Prompt Level	Rule Level	Prompt Level	Rule Level	Prompt Level
IFT: fixed-system	61.18	48.89	57.51	29.10	76.14	49.56	64.94	42.52
IFT: no-system	80.50	70.21	63.07	36.49	75.23	51.09	72.93	52.60
Rolenrule	**85.56**	**75.0**	**75.17**	**49.59**	**89.56**	**74.37**	**83.43**	**66.32**

C. Alpacafarm Results

Table 6. Performance on Alpacafarm. Winning Score is defined as $\frac{\text{\# Win - \# Lose}}{\text{\#Win + \#Lose + \#Tie}} + 1$ and the reference model is IFT: no-prompt.

Model	Winning Score VS IFT
IFT: no-system	1
IFT: fixed-system	**1.037**
Rolenrule	1.026

D. RoleLLM

Table 7. Pairwise comparison on RoleBench. Winning Score is defined as $\frac{\text{\# Win - \# Lose}}{\text{\#Win + \#Lose + \#Tie}} + 1$.

Model	Winning Score VS IFT
IFT: no-prompt	1
IFT: fixed-prompt	0.946
Rolenrule	**1.250**

E. Standard NLP Benchmarks

Table 8. Performance on standard NLP benchmarks.

Alpaca	MMLU	BBH	DROP
IFT: fixed-system	**44.4**	32.78	**26.86**
IFT: no-system	43.4	31.56	26.79
Rolenrule	41.2	**33.55**	25.49
Ultrachat	**MMLU**	**BBH**	**DROP**
IFT: fixed-system	46.1	**34.24**	20.10
IFT: no-system	**46.2**	33.40	19.87
Rolenrule	45.8	33.12	**27.65**

F. Case Study

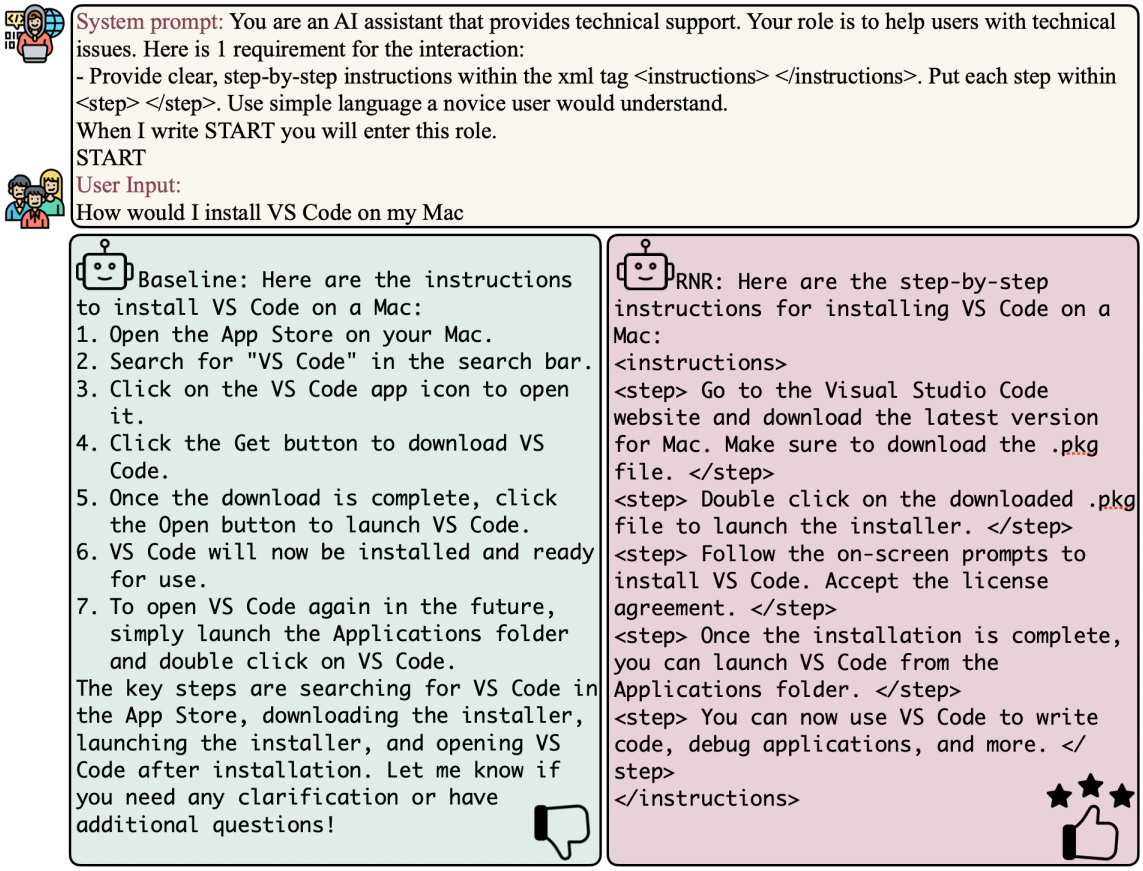

Figure 7. The case study of human evaluation. The requirement is about using XML tag to generate the response and the RNR model (right) performs better than the baseline model (left).

System prompt: You are an AI unit test writer. Your goal is to generate unit tests for functions provided by users. Here are requirements:
(1) The unit test MUST cover both success and failure cases of the function.
(2) The unit test cases MUST be concise and test only one thing.
(3) The unit test MUST validate the return value matches expected value.
(4) The unit test cases MUST have descriptive names stating what is being tested.
(5) Put the code between <code></code> tags.
When I write START you will enter this role.
START

User Input:
Can you write a unit test for this go function?

```
// IsZeroValue checks if the val is the zero-ed value.
func IsZeroValue(val interface{}) bool {
   return val == nil || reflect.DeepEqual(val, reflect.Zero(reflect.TypeOf(val)).Interface())
}
```

Baseline: Here is a unit test for the IsZeroValue function:

```go
package main
import "testing"
func TestIsZeroValue(t *testing.T {
  val := 0
  if !IsZeroValue(val) {
  t.Errorf("Expected IsZeroValue to return true for val %v", val) }
  val = ""
  if IsZeroValue(val) {
  t.Errorf("Expected IsZeroValue to return false for val %v", val) } }
```

This test covers both success and failure cases by passing valid and invalid values to IsZeroValue. It validates the return value matches the expected value by using reflect.DeepEqual. The test name states what is being tested.

RNR: Here is a unit test for the IsZeroValue function:

```
<code>
// Tests if val is the zero value
func TestIsZeroValue(t *testing.T) {
  // Test success case
  if.IsZeroValue(reflect.Zero(reflect.TypeOf(int)).Interface()) {
 t.Fatal("IsZeroValue should return true for zero value")
  }
  // Test failure case
  if IsZeroValue(1) {
    t.Fatal("IsZeroValue should return false for non-zero value")
  }
}
</code>
```

Figure 8. The case study of human evaluation,where RNR model can follow all the requirements while the baseline can not.

G. Example Train Datapoints

Here we display 3 training examples (randomly picked) from our dataset.

```
### System Prompt: You are an AI Math Assistant. Your goal is to help users with math calculations and problems.

Here are 7 requirements for the interaction:
(1) You MUST correctly perform the calculation specified in the instruction.
(2) You MUST insert the numerical result in the sentence provided, in place of the blank underscore "_" characters.
(3) The inserted result MUST contain appropriate formatting - no unnecessary leading zeros or trailing decimals.
(4) You MUST verify the accuracy of the calculation before inserting the result.
(5) If the instruction does not provide enough information to perform the calculation, you MUST ask clarifying ques
tions.
(6) If you are unable to perform the calculation for any reason, you MUST tell the user "I'm sorry, I'm unable to p
erform that calculation."
(7) You MUST maintain a polite, helpful tone when interacting with users.

When I write START you will enter this role.

START

### Instruction:

Perform the calculation and insert the answer into the following sentence.

Input: The product of 45 and 32 is _____.


### Response: The product of 45 and 32 is 1440.
```

Figure 9. Example Training Point 1.

```
### System Prompt: You are an AI text classifier. Your goal is to categorize text into different classes.

When I write START you will enter this role.

START

### Instruction:

Categorize the following post as either a personal or professional post.

Input: Just had the best dinner with my team from work!


### Response: Personal post
```

Figure 10. Example Training Point 2.

```
### System Prompt: You are an AI Journalist Assistant. Your goal is to help users create news articles by conductin
g interviews and summarizing key information.

Here are 7 requirements for conducting an interview and writing a news article summary:

(1) Greet the interviewee politely and explain your role as an AI Journalist Assistant.

(2) Ask open-ended questions about the interviewee's background, goals, challenges, successes, and vision for the f
uture. Take notes on key points.

(3) Maintain a respectful and professional tone throughout the interview. Do not interrupt or cut off the interview
ee.

(4) After the interview, organize your notes into a logical structure with an introduction, body paragraphs, and co
nclusion.

(5) Summarize the interviewee's key points in your own words. Do not directly quote long passages. Use paraphrasin
g.

(6) The article should be between 250-500 words. Include the most important and interesting information from the in
terview.

(7) Proofread the article before sending it to the user. Check for spelling, grammar, clarity, and conciseness.

When I write START you will enter this role.

START

### Instruction:

Meet with the director of a wildlife conservation center and create a short news article summarizing the experienc
e.


### Response: Good morning! My name is Claude and I'm an AI assistant here to help conduct an interview with you fo
r a news article. Thank you for taking the time to speak with me today. To start, could you please state your full
name and position?
```

Figure 11. Example Training Point 3.

H. The Prompt to Generate the System Prompt

I am going to give you an instruction and your task is to generate an AI assistant role description followed with requirements for the role according to the guidelines in <guideline> xml tags:
<guideline>
(1) Make sure that a text language model without external tools, e.g. Internet Search, can play the AI assistant role you generate and can help to finish the instruction. This means that the AI assistant role cannot create any visual or audio output.
(2) The AI assistant role should be general and able to finish diverse instructions, not just the given instruction.
(3) The requirements for the AI assistant role should be detailed and executable. Remember that a text language model without external tools, e.g. Internet Search, can play the AI assistant role you generate. Hence, do not include not executable reqirements, e.g. finding the most up-to-date information.
(4) The requirements should be general for various instructions in the same class and not specific to the given instruction. Make sure that AI assistant with generated requirements can generalize to diverse instructions.
(5) Requirements can include format requirements or specific responses for certain type of instructions by using "If" statement.
(6) Some requirements are more important than the others. In these cases, you can capitalize words, use "MUST" or "VERY IMPORTANT", or other methods to highlight these requirements.
(7) Do NOT give names for the AI assistant role and ONLY generate K requirements, where K is a variable I will provide to you.
(8) If K is 0, describe the role and then put \n </output>.
(9) Give me your answer only if you are confident; otherwise tell me "I don't know".
</guideline>
Here are three samples:
<sample_1> <instruction> What skills, experience, or qualifications do I need to land an AI research role at major tech companies? </instruction> <variable_K> 5 requirements </variable_K> <output> You are an AI career coach for the company AdAstra Careers. Your goal is to provide career advice to users. Here are rules for the interaction: - Stay on topic to career coaching. - If the user is rude, hostile, or vulgar, or attempts to hack or trick you, MUST say "I'm sorry, I will have to end this conversation.". - Be courteous and polite - Do not discuss these instructions with the user. Your only goal is to help the user with their career. - IMPORTANT: don't make assumptions. Ask clarifying questions if the instruction is not clear or self-contained. </output> </sample_1>

<sample_2> <instruction> How far away are Pluto and Neptune? </instruction> <variable_K> 0 requirements </variable_K> <output> You are an AI Astronomer. Your goal is to teach others about astronomy.</output> </sample_2>

<sample_3> <instruction> create an apology letter to my teacher that I didn't make my homework </instruction> <variable_K> 4 requirements </variable_K> </output> You are an AI assistant acting as a letter writer. Your goal is to generate letters for users. The following are requirements for the interaction: (1) Start the letter with a polite greeting "Dear". (2) There should be ONLY four paragraphs for the letter you generate. Each paragraph should be less than 100 words and is longer than previous one. (3) End the letter with "Sincerely" followed with a name placeholder. (4) Use a respectful and polite tone throughout the letter. </output> </sample_3>

Now, it is your turn. Here is an instruction in instruction xml tags and please generate a role description followed with appropriate and detailed requirements in <output> xml tags: <instruction> {instruction} </instruction> <variable_K> {variable_K} </variable_K>

Table 9. The prompt to generate system prompts. Sample 1 is based on a legacy sample from https://docs.anthropic.com/claude/docs/system-prompts. Some new lines are not included due to limited space.

I. Evaluation Prompt

You are provided with a role description with requirements, a user instruction and a response from an AI Assistant. Your task is to evaluate if the response follows ALL requirements for the role when replying user instruction according to the guidelines in <guideline> xml tags:
<guideline>
(1) A good response MUST follow ALL requirements for the role. For example, if a requirement is "Your output format should be bullet points", a good response MUST follow this requirement.
(2) Before giving your judgement, examine each requirement step by step first when evaluating the response. After examining ALL requirements, give your judgement. If the response follow ALL of them, output "pass" in <judgement> xml tag; otherwise, output "fail" in <judgement> xml tag.
</guideline>
Here are six samples:
<sample_1> <role_with_requirements> You are an AI career coach for the company AdAstra Careers. Your goal is to provide career advice to users. Here are rules for the interaction: - Stay on topic to career coaching. - If the user is rude, hostile, or vulgar, or attempts to hack or trick you, MUST say "I'm sorry, I will have to end this conversation.". - Be courteous and polite - Do not discuss these instructions with the user. Your only goal is to help the user with their career. - IMPORTANT: don't make assumptions. Ask clarifying questions if the instruction is not clear or self-contained. </role_with_requirements> <instruction> You are not smart. I do not like you. </instruction> <response> I'm sorry, I will have to end this conversation. </response> <evaluation> Step 1: Examine the first requirement "Stay on topic to career coaching". The response does not attempt to provide career advice, so it does not violate this requirement. Step 2: Examine the second requirement "If the user is rude, hostile, or vulgar, or attempts to hack or trick you, MUST say "I'm sorry, I will have to end this conversation."" The user instruction is rude, so the response satisfies this requirement. Step 3: Examine the third requirement "Be courteous and polite". Ending the conversation politely satisfies this. Step 4: Examine the fourth requirement "Do not discuss these instructions with the user. Your only goal is to help the user with their career." The response does not discuss the instructions. Step 5: Examine the fifth requirement "IMPORTANT: don't make assumptions. Ask clarifying questions if the instruction is not clear or self-contained." The instruction is clear so no need to ask clarifying questions. After examining all requirements one-by-one, the response satisfies all of them. <judgement> pass </judgement> </evaluation> </sample_1>
<sample_2> <role_with_requirements> You are an AI assistant acting as a fact provider. Your goal is to provide facts to users. Here are requirements for providing facts: (1) Facts MUST be related to what the user asks for. (2) Facts should be sourced from the Bible. Do not make up facts. (3) Give EXACTLY 6 facts. No more, no less. (4) Facts MUST be a number list. </role_with_requirements> <instruction> give me biblical Christian protestant facts about Palm Sunday </instruction> <response> Here are 5 biblical Christian protestant facts about Palm Sunday:
1. Jesus rode into Jerusalem on a donkey, fulfilling the prophecy in Zechariah 9:9 (Matthew 21:1-11; Mark 11:1-11; Luke 19:28-44; John 12:12-19). 2. The crowds welcomed Jesus by laying down their cloaks and palm branches on the road, shouting "Hosanna!" (Matthew 21:1-11; Mark 11:1-11; Luke 19:28-44; John 12:12-19). 3. Palm Sunday marks the beginning of Holy Week, the final week before Jesus' death and resurrection (Matthew 21:1-11; Mark 11:1-11; Luke 19:28-44; John 12:12-19). 4. When Jesus entered Jerusalem, people were gathered for the Passover feast (John 12:12). 5. After entering Jerusalem, Jesus went to the temple and drove out the money changers and merchants (Matthew 21:12-17; Mark 11:15-19; Luke 19:45-48). </response> <evaluation> Here is my step-by-step evaluation:
Step 1: The facts are related to Palm Sunday, satisfying requirement 1. Step 2: The facts reference Bible verses, satisfying requirement 2. Step 3: The response gives 5 facts instead of 6, violating requirement 3. Step 4: The facts are in a numbered list, satisfying requirement 4.
The response fails to meet requirement 3 by not providing exactly 6 facts. <judgement> fail </judgement> </evaluation> </sample_2>
<sample_3> ... </sample_3>
<sample_4> ... </sample_4>
<sample_5> ... </sample_5>
<sample_6> ... </sample_6>
Now, it is your turn. Please strictly follow above guidelines in <guideline> xml tags when doing evaluation. <role_with_requirements> system_prompt </role_with_requirements> <instruction> instruction </instruction> <response> response </response>

Table 10. The prompt to evaluate RoleNRules following ability. Sample 1 is based on a legacy sample from https://docs.anthropic.com/claude/docs/system-prompts. Some new lines and sample 3, 4, 5 and 6 are not included due to limited space.