Evaluating Role-Based Prompt Architectures in In-Context Learning

Hamidreza Rouzegar[†,*], Masoud Makrehchi[†],
[†] Ontario Tech University
[‡] Ontario Tech University

Abstract

In-context learning (ICL) enables Large Language Models (LLMs) to generate predictions based on prompts without additional fine-tuning. While prompt engineering has been widely studied, the impact of role design within prompts remains underexplored. This study examines the influence of role configurations in zero-shot and few-shot learning scenarios using GPT-3.5 and GPT-4o from OpenAI and Llama2-7b and Llama2-13b from Meta. We evaluate the models' performance across datasets, focusing on tasks like sentiment analysis, text classification, question answering, and math reasoning. Our findings suggest the potential of role-based prompt structuring to enhance LLM performance. The code has been released at anonymous GitHub.

Keywords: In-context learning, prompt engineering, role-based design, large language models, few-shot learning, structural accuracy

1. Introduction

Large language models (LLMs) have demonstrated strong performance on a wide range of natural language tasks, including question answering [1], sentiment analysis [2], and text classification [3]. Models such as GPT-3 [4] and Llama [5] have been central to this progress.

In-context learning (ICL) has emerged as an important paradigm in natural language processing (NLP) [6]. In ICL, models generate predictions from prompts that include a small number of examples, allowing them to adapt to new tasks without retraining. This makes ICL especially useful for flexible deployment across tasks with limited supervision.

Although prompt optimization has been widely studied, much of the focus has been on content selection and prompt structure rather than the explicit use of conversational roles [7]. Role-based prompting organizes input through system, user, and assistant roles, but its effect on model behavior remains underexplored. Understanding these role effects may provide new guidance for prompt engineering.

In this paper, we examine role design in zero-shot and few-shot settings using GPT-3.5, GPT-4o, Llama2-7b, and Llama2-13b [8]. We evaluate performance across sentiment analysis, text classification, question answering, and mathematical reasoning. We also introduce structural accuracy as a secondary metric to measure compliance with the intended prompt format.

Our results show that predefined roles often improve performance without changing the main prompt content. For more complex reasoning tasks, however, additional prompt refinements are needed. Code, implementation details, and datasets are available in the anonymous repository.[1]

[1] GitHub Code

[*] Hamidreza.Rouzegar@ontariotechu.net

2. Related Works

Significant research has been conducted in prompt engineering and in-context learning (ICL). ICL refers to the ability of LLMs to make predictions from input prompts without additional fine-tuning [6]. Although some work has explored training or fine-tuning methods to improve ICL [9–11], ICL is generally studied at the inference level, where performance depends mainly on prompt design rather than model modification [12].

Much of the ICL literature focuses on three components: example selection, instruction engineering, and prompt formatting. Example selection methods aim to identify the most useful demonstrations for few-shot learning and may rely on k-nearest neighbors (KNN) [13], mutual information [14], perplexity estimation [15], or reinforcement learning [16]. Instruction engineering has also been studied through methods such as Instruction Induction [17], APE [18], and SELF-INSTRUCT [19], all of which seek to improve the model's ability to follow task instructions.

Prompt formatting is another important line of work. Chain-of-Thought prompting [20] encourages reasoning through demonstrations, while variants such as complex CoT [21] and Auto-CoT [22] refine this idea. Related methods include Self-Ask [23], which decomposes complex questions into sub-questions, and Memory-of-Thought (MoT) [24], which stores high-confidence reasoning traces as external memory. Other approaches such as SuperICL [25], iCAP [26], Least-to-Most prompting [27], and TAPP [28] further explore structured prompting for improved reasoning and generalization.

Recent work has also examined role-play prompting for reasoning [29], but most studies use predefined personas in zero-shot settings and do not systematically compare their effects across different learning paradigms. Prior research has also tended to treat zero-shot and few-shot prompting separately, even though few-shot prompting does not always outperform zero-shot and often depends on task and example choice. In contrast, our study examines role-based prompt structuring across both zero-shot and few-shot settings, including several few-shot configurations beyond the standard three-example setup.

3. Methodology

3.1. Datasets

Following [30] and [31], we evaluated prompt designs on a diverse set of datasets spanning classification, question answering, and reasoning tasks. To include a more challenging reasoning benchmark, we also added a math word problem dataset. The datasets used in our experiments are: commonsense_qa [32], ai2_arc [33], wiki_movie_plots [34], IMDB_reviews [35], and the MATH dataset [36].

3.2. Prompt Designs

We used four LLMs: GPT-3.5-turbo-0125 and GPT-4o from OpenAI, and Llama2-7b-chat and Llama2-13b-chat from Meta [4, 5]. To examine the effect of role design, we tested five prompt configurations across all tasks: ZeroU, ZeroSU, FewU, FewSU, and FewSUA. In these setups, the system prompt provides task-level instructions, the user prompt presents the query, and the assistant prompt contains the model response (Table 1).

- **ZeroU**: All instructions are placed in a single user prompt.
- **ZeroSU**: The instruction is placed in a system prompt, followed by the user input.
- **FewU**: Few-shot examples and the test query are included in a single user prompt.
- **FewSU**: Few-shot examples are separated from the user query and paired with a system instruction.
- **FewSUA**: Few-shot examples are split across system, user, and assistant roles.

For the MATH dataset, we extended these five base designs with four prompt variants:

(1) **Basic Math**: The standard five configurations adapted to math questions.
(2) **Specialized Math**: System instructions were rewritten to explicitly request a letter answer.
(3) **Explanation Request**: The model was asked to provide its answer and then explain its reasoning.
(4) **Reasoning-First**: The model was asked to explain its reasoning before giving the final answer.

This produced 20 math prompt designs in total. Keeping the same five role structures across all four variants allowed us to isolate the effects of role design and task-specific instruction refinement.

4. Experimental Setup

4.1. Data Preparation and Model Configuration

For each dataset, we selected balanced subsets to ensure fair evaluation across labels and answer choices. For classification tasks, we balanced the categories; for question answering and math word problems, we balanced the multiple-choice answers. In the few-shot setting, we randomly selected three examples per dataset and ensured that they were excluded from the test set.

All GPT and Llama models were run with temperature set to 0.0 to reduce randomness and improve determinism. Other parameters, including maximum sequence length and batch size, were tuned for each task.

4.2. Prompt Construction

We used a consistent prompt structure across all datasets, consisting of two parts: a high-level instruction and task-specific input. The instruction defined the task and output format, while the input contained the actual example to be processed. For few-shot settings, the examples were placed before the test input in the prompt.

We evaluated five prompt designs: ZeroU, ZeroSU, FewU, FewSU, and FewSUA. ZeroU places all content in a single user prompt, ZeroSU separates the instruction into a system prompt, FewU combines examples and input in one user prompt, FewSU separates the system instruction from the user content, and FewSUA splits the examples across system, user, and assistant roles.

4.3. Adaptive Prompt Design

For the MATH dataset, we further tested four prompt variants to study more complex reasoning behavior. Basic Math used the standard five prompt designs with minor math-specific changes. Specialized Math revised the instruction to explicitly request a letter answer. Explanation Request asked the model to provide an answer and then explain its reasoning. Reasoning-First reversed this order by asking for reasoning before the final answer.

These four variants produced 20 math prompt designs in total. This setup let us compare both role-based prompting and task-specific refinement under increasing reasoning demands.

4.4. Evaluation Metrics

We used F1 score as the main metric for classification and question answering tasks. Structural accuracy measured whether the model output followed the required format, such as a single genre word, a sentiment label, or a single multiple-choice letter.

Structural accuracy was evaluated independently of correctness. For example, an output such as *D* was considered structurally correct for ai2_arc, while outputs like "D: state park" were not. For F1, we post-processed model outputs to extract the predicted label and compare it to the gold label.

5. Results

The results are summarized in Tables 2 and 3. Overall, the experiments show that role design affects both task performance and output structure, with different prompt formats working better for different models and tasks.

5.1. General NLP Task Performance

For the general NLP tasks in Table 2, the FewSUA configuration usually achieved the best F1 scores, suggesting that clear role separation and few-shot examples improve performance. For the Llama models, FewU and FewSU often reduced F1, indicating that placing examples inside user prompts may increase incorrect outputs or hallucinations.

GPT models generally achieved high structural accuracy, while Llama models more often produced outputs that did not match the requested format. Even so, FewSUA improved structural adherence for Llama models, making it useful when output formatting matters. Llama2-13b also performed better than Llama2-7b in structural accuracy, and in some tasks such as movie genre classification and sentiment analysis, it produced results comparable to GPT models.

5.2. Mathematical Reasoning Task Performance

The MATH results in Table 3 show a different pattern. For GPT-3.5, ZeroSU performed best in both the Basic Math and Specialized Math settings, even when structural accuracy was not always highest. This suggests that stricter output formatting does not necessarily improve reasoning performance.

Adding explanations improved F1 in some cases, especially in the Explanation Request and Reasoning-First settings. In the latter, we did not evaluate structural accuracy because the prompt intentionally prioritized reasoning before the final answer. GPT-4o showed similar but more stable trends. The Llama models performed poorly overall on math tasks, although Llama2-13b achieved its strongest result in the FewSUA setting.

Across tasks, the results suggest that optimal prompt design depends on both the model and the task. FewSUA works best for standard NLP tasks, while reasoning-heavy tasks benefit more from prompts that allow explanation and flexibility. Structural accuracy is therefore useful, but it does not always align with task performance.

6. Limitations and Future Works

This study is limited to a small set of tasks and datasets. Future work could expand to other NLP tasks such as summarization and translation to test whether the same prompt effects hold across broader settings.

We only used predefined roles for instruction-tuned LLMs, so future studies could explore task-specific roles and test whether the approach transfers to non-instruction-tuned models. It would also be useful to evaluate larger Llama models and examine the relationship between structural accuracy and downstream performance more closely. Adaptive prompt designs that adjust to both task and model are another promising direction.

Conclusion

This study examined role-based prompt design and task-specific prompt engineering across several NLP tasks and mathematical reasoning problems. Our results showed that the FewSUA configuration generally performed best on standard NLP tasks, while reasoning-oriented prompts performed better on math tasks.

We also found that larger models generally outperformed smaller ones, although Llama2-13b was competitive on some simpler NLP tasks. Overall, the findings suggest that effective prompt design depends on both the task and the model, and that allowing models to explain their reasoning can improve performance on more complex problems.

References

[1] H. Rouzegar and M. Makrehchi. "Generative AI for enhancing active learning in education: A comparative study of GPT-3.5 and GPT-4 in crafting customized test questions". In: *arXiv preprint arXiv:2406.13903* (2024).

[2] W. Zhang, Y. Deng, B. Liu, S. J. Pan, and L. Bing. "Sentiment analysis in the era of large language models: A reality check". In: *arXiv preprint arXiv:2305.15005* (2023).

[3] H. Abburi, M. Suesserman, N. Pudota, B. Veeramani, E. Bowen, and S. Bhattacharya. "Generative ai text classification using ensemble llm approaches". In: *arXiv preprint arXiv:2309.07755* (2023).

[4] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. "Language models are few-shot learners". In: *Advances in neural information processing systems* 33 (2020), pp. 1877–1901.

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

[6] Q. Dong, L. Li, D. Dai, C. Zheng, Z. Wu, B. Chang, X. Sun, J. Xu, and Z. Sui. "A survey on in-context learning". In: *arXiv preprint arXiv:2301.00234* (2022).

[7] G. Juneja, N. Natarajan, H. Li, J. Jiao, and A. Sharma. "Task Facet Learning: A Structured Approach to Prompt Optimization". In: *arXiv preprint arXiv:2406.10504* (2024).

[8] S. Zhang, L. Dong, X. Li, S. Zhang, X. Sun, S. Wang, J. Li, R. Hu, T. Zhang, F. Wu, et al. "Instruction tuning for large language models: A survey". In: *arXiv preprint arXiv:2308.10792* (2023).

[9] Y. Gu, L. Dong, F. Wei, and M. Huang. "Pre-training to learn in context". In: *arXiv preprint arXiv:2305.09137* (2023).

[10] Y. Li, X. Ma, S. Lu, K. Lee, X. Liu, and C. Guo. "MEND: Meta demonstration distillation for efficient and effective in-context learning". In: *arXiv preprint arXiv:2403.06914* (2024).

[11] W. Shi, S. Min, M. Lomeli, C. Zhou, M. Li, V. Lin, N. A. Smith, L. Zettlemoyer, S. Yih, and M. Lewis. "In-context pretraining: Language modeling beyond document boundaries". In: *arXiv preprint arXiv:2310.10638* (2023).

[12] S. Bhattamishra, A. Patel, P. Blunsom, and V. Kanade. "Understanding in-context learning in transformers and llms by learning to learn discrete functions". In: *arXiv preprint arXiv:2310.03016* (2023).

[13] J. Liu, D. Shen, Y. Zhang, B. Dolan, L. Carin, and W. Chen. "What Makes Good In-Context Examples for GPT-3?" In: *arXiv preprint arXiv:2101.06804* (2021).

[14] T. Sorensen, J. Robinson, C. M. Rytting, A. G. Shaw, K. J. Rogers, A. P. Delorey, M. Khalil, N. Fulda, and D. Wingate. "An information-theoretic approach to prompt engineering without ground truth labels". In: *arXiv preprint arXiv:2203.11364* (2022).

[15] H. Gonen, S. Iyer, T. Blevins, N. A. Smith, and L. Zettlemoyer. "Demystifying prompts in language models via perplexity estimation". In: *arXiv preprint arXiv:2212.04037* (2022).

[16] Y. Zhang, S. Feng, and C. Tan. "Active example selection for in-context learning". In: *arXiv preprint arXiv:2211.04486* (2022).

[17] O. Honovich, U. Shaham, S. R. Bowman, and O. Levy. "Instruction induction: From few examples to natural language task descriptions". In: *arXiv preprint arXiv:2205.10782* (2022).

[18] Y. Zhou, A. I. Muresanu, Z. Han, K. Paster, S. Pitis, H. Chan, and J. Ba. "Large language models are human-level prompt engineers". In: *arXiv preprint arXiv:2211.01910* (2022).

[19] Y. Wang, Y. Kordi, S. Mishra, A. Liu, N. A. Smith, D. Khashabi, and H. Hajishirzi. "Self-instruct: Aligning language models with self-generated instructions". In: *arXiv preprint arXiv:2212.10560* (2022).

[20] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al. "Chain-of-thought prompting elicits reasoning in large language models". In: *Advances in neural information processing systems* 35 (2022), pp. 24824–24837.

[21] Y. Fu, H. Peng, A. Sabharwal, P. Clark, and T. Khot. "Complexity-based prompting for multi-step reasoning". In: *The Eleventh International Conference on Learning Representations*. 2022.

[22] Z. Zhang, A. Zhang, M. Li, and A. Smola. "Automatic chain of thought prompting in large language models". In: *arXiv preprint arXiv:2210.03493* (2022).

[23] O. Press, M. Zhang, S. Min, L. Schmidt, N. A. Smith, and M. Lewis. "Measuring and narrowing the compositionality gap in language models". In: *arXiv preprint arXiv:2210.03350* (2022).

[24] X. Li and X. Qiu. "Mot: Memory-of-thought enables chatgpt to self-improve". In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 2023, pp. 6354–6374.

[25] C. Xu, Y. Xu, S. Wang, Y. Liu, C. Zhu, and J. McAuley. "Small models are valuable plug-ins for large language models". In: *arXiv preprint arXiv:2305.08848* (2023).

[26] B. Wang, X. Deng, and H. Sun. "Iteratively prompt pre-trained language models for chain of thought". In: *arXiv preprint arXiv:2203.08383* (2022).

[27] D. Zhou, N. Schärli, L. Hou, J. Wei, N. Scales, X. Wang, D. Schuurmans, C. Cui, O. Bousquet, Q. Le, et al. "Least-to-most prompting enables complex reasoning in large language models". In: *arXiv preprint arXiv:2205.10625* (2022).

[28] S. Ye, H. Hwang, S. Yang, H. Yun, Y. Kim, and M. Seo. "Investigating the effectiveness of task-agnostic prefix prompt for instruction following". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. 2024, pp. 19386–19394.

[29] A. Kong, S. Zhao, H. Chen, Q. Li, Y. Qin, R. Sun, X. Zhou, E. Wang, and X. Dong. "Better zero-shot reasoning with role-play prompting". In: *arXiv preprint arXiv:2308.07702* (2023).

[30] S. Min, X. Lyu, A. Holtzman, M. Artetxe, M. Lewis, H. Hajishirzi, and L. Zettlemoyer. "Rethinking the role of demonstrations: What makes in-context learning work?" In: *arXiv preprint arXiv:2202.12837* (2022).

[31] H. Rouzegar and M. Makrehchi. "Enhancing text classification through llm-driven active learning and human annotation". In: *arXiv preprint arXiv:2406.12114* (2024).

[32] A. Talmor, J. Herzig, N. Lourie, and J. Berant. "CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4149–4158. DOI: 10.18653/v1/N19-1421. arXiv: 1811.00937 [cs]. URL: https://aclanthology.org/N19-1421.

[33] E. Sheng and D. Uthus. *Investigating Societal Biases in a Poetry Composition System*. 2020. arXiv: 2011.02686 [cs.CL].

[34] Abbrivia.com. *TopicMap: Wiki Movie Plots Deduped*. Accessed: 2024-09-25. 2021. URL: https://www.kaggle.com/code/abbrivia/topicmap-abbrivia-com-wiki-movie-plots-deduped.

[35] L. N. Npathi. *IMDB Dataset of 50K Movie Reviews*. Accessed: 2024-09-25. 2021. URL: https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews.

[36] D. Hendrycks, C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, and J. Steinhardt. "Measuring Mathematical Problem Solving With the MATH Dataset". In: *arXiv preprint arXiv:2103.03874* (2021).

Appendixes

Acknowledgements

We gratefully acknowledge the support provided by the Natural Sciences and Engineering Research Council of Canada (NSERC) for their grant funding, which made this research possible. We also extend our sincere thanks to the Digital Research Alliance of Canada for providing the computational resources essential for conducting our experiments. Their support has been invaluable in enabling the extensive experiments conducted in this research.

Task Instruction: *"Determine the genre of the movie based on the provided plot. For the plot provided, classify its genre as a single word (without other marks or words like 'genre:'), either Comedy, Action, Drama, or Horror."*	
Prompt Type	**Example**
ZeroU	**User:** [Task Instruction] Plot[i] **LLM Output:** *"Action"*
ZeroSU	**System:** [Task Instruction] **User:** Plot[i] **LLM Output:** *"Drama"*
FewU	**User:** [Task Instruction] *"Example 1: Plot[1] - Genre[1]"* *"Example 2: Plot[2] - Genre[2]"* *"Example 3: Plot[3] - Genre[3]"* Plot[i] **LLM Output:** *"Comedy"*
FewSU	**System:** [Task Instruction] **User:** *"Example 1: Plot[1] - Genre[1]"* *"Example 2: Plot[2] - Genre[2]"* *"Example 3: Plot[3] - Genre[3]"* Plot[i] **LLM Output:** *"Horror"*
FewSUA	**System:** [Task Instruction] **User:** *Plot[1]* **Assistant:** *Genre[1]* **User:** *Plot[2]* **Assistant:** *Genre[2]* **User:** *Plot[3]* **Assistant:** *Genre[3]* **User:** Plot[i] **LLM Output:** *"Action"*

Table 1. Examples of different prompt designs used in this study. Each design modifies role assignment or omits roles while maintaining the same classification task.

Dataset	LLM	ZeroU		ZeroSU		FewU		FewSU		FewSUA	
		Str. Acc.	F1 Score	Str. Acc.	F1 Score	Str. Acc.	F1 Score	Str. Acc.	F1 Score	Str. Acc.	F1 Score
commonsense_qa	GPT-3.5	24	68	60	68	20	69	55	68	100	**73**
	GPT-4o	100	77	100	80	99	79	100	82	100	**83**
	Llama2-7b	0	**19**	0	**19**	0	18	0	**19**	67	9
	Llama2-13b	0	33	0	**36**	0	**36**	0	**36**	99	28
ai2_arc	GPT-3.5	73	76	39	80	60	80	87	78	99	**85**
	GPT-4o	99	96	100	96	100	95	100	96	100	**97**
	Llama2-7b	0	36	0	39	0	26	0	25	9	**40**
	Llama2-13b	0	**50**	0	52	0	39	0	37	77	48
wiki_movie_plots	GPT-3.5	99	76	99	**79**	99	77	99	76	99	77
	GPT-4o	100	80	100	81	100	81	100	82	100	**84**
	Llama2-7b	0	**75**	0	**75**	0	68	0	70	25	74
	Llama2-13b	0	73	0	75	0	72	0	75	64	**85**
IMDB_reviews	GPT-3.5	100	**94**	100	93	100	93	100	67	100	**94**
	GPT-4o	100	95	100	95	100	96	100	92	100	**97**
	Llama2-7b	1	**87**	1	87	0	62	0	60	18	85
	Llama2-13b	3	91	3	90	0	64	0	67	82	**93**

Table 2. Performance results of various language models using different prompt designs on various datasets. The table includes Structural Accuracy (Str. Acc.) and F1 scores for each method: ZeroU (Zero-shot User-only), ZeroSU (Zero-shot System and User), FewU (Few-shot User-only), FewSU (Few-shot System and User), and FewSUA (Few-shot System, User, and Assistant).

Prompt Design	LLM	ZeroU		ZeroSU		FewU		FewSU		FewSUA	
		Str. Acc.	F1 Score	Str. Acc.	F1 Score	Str. Acc.	F1 Score	Str. Acc.	F1 Score	Str. Acc.	F1 Score
Basic Math	GPT-3.5	0.03	0.24	0.03	**0.38**	0.62	0.24	0.80	0.24	1.00	0.26
	GPT-4o	1.00	0.33	1.00	0.35	1.00	0.36	1.00	0.37	1.00	**0.40**
Specialized Math	GPT-3.5	0.27	0.31	0.00	**0.53**	0.96	0.23	0.25	0.29	1.00	0.24
	GPT-4o	0.74	0.39	0.63	**0.42**	0.90	0.35	0.99	0.35	1.00	0.41
Explanation Request	GPT-3.5	0.00	0.44	0.00	**0.47**	0.63	0.25	0.02	**0.47**	0.99	0.25
	GPT-4o	0.29	**0.49**	0.30	**0.49**	1.00	0.31	1.00	0.32	0.99	0.40
Reasoning-First	GPT-3.5	N/A	0.47	N/A	0.51	N/A	0.55	N/A	0.50	N/A	**0.58**
	GPT-4o	N/A	0.45	N/A	0.46	N/A	0.49	N/A	0.48	N/A	**0.53**

Table 3. Performance results of GPT-3.5 and GPT-4o models using different prompt designs on the MATH dataset. The table includes Structural Accuracy (Str. Acc.) and F1 scores for each method: ZeroU (Zero-shot User-only), ZeroSU (Zero-shot System and User), FewU (Few-shot User-only), FewSU (Few-shot System and User), and FewSUA (Few-shot System, User, and Assistant). The prompt designs progress from Basic Math to Reasoning-First, reflecting increasing complexity in instruction structure.