Mixture of Formats: Towards LLMs Robustness to Changes in Prompt Formats

Anonymous ACL submission

Abstract

Existing large language models (LLMs) are sensitive to non-semantic changes in prompt formats. In the literature, this problem is commonly referred to as prompt brittleness. Several works study the problem and provide methods for quantifying the performance variation, but there is no simple solution for addressing it. We propose Mixture of Formats (MOF), a simple and efficient technique for addressing prompt brittleness in LLMs by diversifying the styles used in the prompt few-shot examples. Empirical results show that our proposed technique reduces sensitivity in various LLMs and also increases the performance of both the highest and lowest-performing prompt formats.

1 Introduction

Large language models (LLMs) are useful for many applications i.e., content generation, translation, and text analysis. However, they are sensitive to meaning-preserving changes in prompt formats such as adding extra space, replacing two colons with one colon, etc. (Sclar et al., 2024; Lu et al., 2022), as illustrated in Figure 1. This problem is sometimes referred to as prompt brittleness (Zhou et al., 2023). Prompt brittleness makes LLMs unreliable and prevents their adoption in high-risk domains such as healthcare. Existing works have tried to address this problem (Zhou et al., 2023; Wan et al., 2023; Arora et al., 2022), but the proposed techniques are not simple and efficient enough for practical applications. In this work, we propose *Mixture of Formats (MOF)*, a simple and computationally efficient prompting technique created by making sure that each few-shot example in the prompt uses a different style. Additionally, for each example, the model is asked to rewrite the example using different styles as illustrated in Figure 2. MOF was inspired by ideas from computer vision that involve learning from datasets with diverse styles to prevent models from associating styles with the target variable (Arjovsky et al., 2019; Kamath et al., 2021; Yin et al., 2021; Wald et al., 2021; Li et al., 2021). We test the effectiveness of MOF prompting on classification tasks from SuperNaturalInstructions (Wang et al., 2022) by comparing its performance with that of traditional prompts.

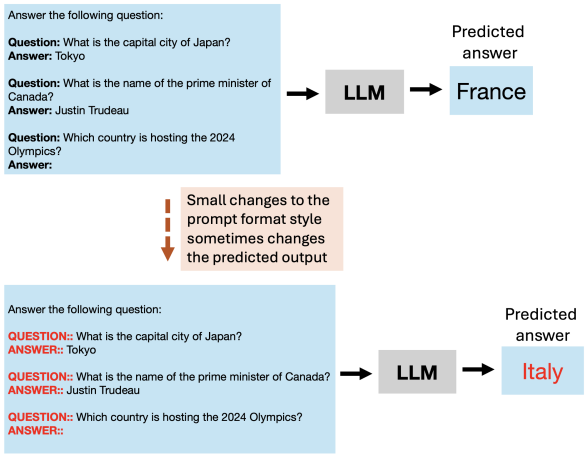

Figure 1: A demonstration of how small changes to the prompt format can sometimes lead to incorrect predictions in LLMs.

2 Related work

Recent works have shown that LLMs are sensitive to changes in prompt formats (Sclar et al., 2024) and to the order of few-shot examples in the prompt (Lu et al., 2022). Sclar et al. (2024) propose a method for efficiently quantifying variations in the performance of LLMs when prompt formats change, by computing the performance difference (spread) between the best-performing format and the worst-performing format. Due to LLMs sensitivity to prompt variations, Polo et al. (2024) propose PromptEval, an efficient method for evaluating LLMs on multiple prompts instead of a single prompt. Methods for finding the best prompt format have also been proposed (Feffer et al., 2024; Sorensen et al., 2022; Yin et al., 2023; Sclar et al.,

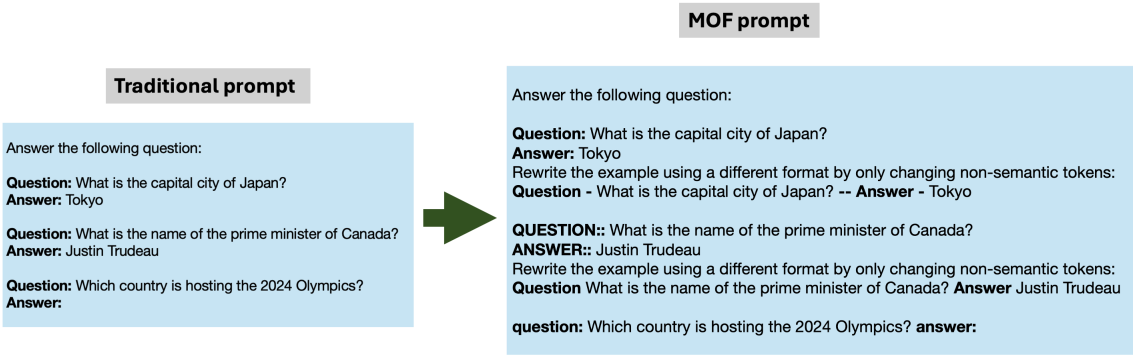

Figure 2: An illustration of how to Convert a traditional prompt into a MOF prompt.

2024; Polo et al., 2024).

There are works that have tried to address the prompt brittleness problem in LLMs. Wan et al. (2023) propose an approach that involves generating zero-shot outputs for given inputs using an LLM and then selecting candidates to use for creating a better prompt based on consistency, diversity, and repetition. Arora et al. (2022) propose a method for addressing prompt brittleness by transforming inputs into a question-answering format that encourages open-ended responses, creating several imperfect prompts, and aggregating the responses from all the prompts using a weak supervision strategy to obtain the final output. Although these methods have shown performance improvements over traditional prompting methods, they are computationally expensive and not simple to adopt in practice. In this work, we attempt to address this problem by proposing a simple and computationally efficient prompting technique.

3 Mixture of Formats

The sensitivity of LLMs to style changes in the prompt format is similar to problems observed in computer vision, where small changes to the image style (eg. color or background) affect the model's ability to make correct predictions (Nagarajan et al., 2020). In computer vision, several techniques involving learning from diverse datasets have been proposed to address the problem (Arjovsky et al., 2019; Kamath et al., 2021; Yin et al., 2021; Wald et al., 2021; Li et al., 2021). The idea is that diverse data points encourage the model to disassociate styles and the target variable. Inspired by these techniques, we propose Mixture of Formats (MOF), a technique that deviates from traditional ways of crafting prompts by using a different style format for every few-shot example in the prompt, as illustrated in Figure 2. We evaluate our proposed technique as described in the next sub-sections.

3.1 Experiments setting

Let X denote input queries in a task, and Y denote the target variable. Given N observations of inputs X and its target Y as data $\mathcal{D} = \{X_n, Y_n\}_{n=1}^{N}$, we automatically build a traditional prompt and a MOF prompt, each with 5 few-shot examples, and use them to perform inference using an LLM via in-context learning (ICL). A traditional prompt is constructed using FormatSpread (Sclar et al., 2024). A MOF prompt is constructed by modifying FormatSpread (Sclar et al., 2024) to accommodate the mixture of formats in the few-shot examples as seen in Figure 2. Using FormatSpread, we create 10 traditional prompt variations and 10 MOF prompt variations. From the 10 prompt variations, for both traditional prompts and MOF prompts, we obtain performance accuracies for each prompt format on classification tasks. The goal is to compare the LLMs' sensitivity to small changes in prompt formats when traditional prompts are used versus when MOF prompts are used. Similar to Sclar et al. (2024), we quantify LLMs' sensitivity by computing the performance spread, the accuracy difference between the best-performing prompt format and worst-performing format.

Datasets We perform experiments on classification tasks from SuperNaturalInstructions (Mishra et al., 2022; Wang et al., 2022). Due to limited computational resources, we perform experiments on 25 classification tasks: `task114`, `task155`, `task158`, `task161`, `task162`, `task163`, `task190`, `task279`, `task280`, `task316`, `task317`, `task319`, `task320`, `task337`, `task607`, `task608`, `task609`, `task904`, `task905`, `task1186`, `task1283`, `task1284`, `task1347`, `task1502`, and

Table 1: `Llama-3-70B-Instruct` results where our proposed MOF prompts perform better than traditional prompts across all metrics. *Spread* is the difference between max accuracy and min accuracy. *Mean Accuracy* is the average accuracy across all the 10 prompt variations.

Task	Traditional Prompts				MOF Prompts			
	Min Accuracy	Max Accuracy	Spread	Mean Accuracy	Min Accuracy	Max Accuracy	Spread	Mean Accuracy
task280	0.811	0.860	0.049	0.837	**0.880**	**0.900**	**0.020**	**0.890**
task317	0.139	0.229	0.090	0.154	**0.712**	**0.795**	**0.083**	**0.770**
task320	0.544	0.862	0.318	0.826	**0.661**	**0.885**	**0.224**	**0.856**
task1347	0.248	0.524	0.276	0.454	**0.464**	**0.535**	**0.071**	**0.502**
task1612	0.624	0.839	0.215	0.800	**0.787**	**0.851**	**0.064**	**0.822**

Table 2: `Llama-2-13b` results where our proposed MOF prompts perform better than traditional prompts across all metrics. *Spread* is the difference between max accuracy and min accuracy. *Mean Accuracy* is the average accuracy across all the 10 prompt variations.

Task	Traditional Prompts				MOF Prompts			
	Min Accuracy	Max Accuracy	Spread	Mean Accuracy	Min Accuracy	Max Accuracy	Spread	Mean Accuracy
task162	0.314	0.427	0.113	0.375	**0.402**	**0.450**	**0.048**	**0.425**
task163	0.389	0.438	0.049	0.414	**0.410**	**0.452**	**0.042**	**0.429**
task280	0.324	0.847	0.523	0.636	**0.808**	**0.864**	**0.056**	**0.842**
task317	0.508	0.626	0.118	0.565	**0.684**	**0.764**	**0.080**	**0.725**
task607	0.503	0.588	0.085	0.548	**0.518**	**0.591**	**0.073**	**0.553**
task1186	0.497	0.530	0.033	0.520	**0.526**	**0.531**	**0.005**	**0.529**
task1284	0.477	0.665	0.188	0.582	**0.530**	**0.684**	**0.154**	**0.593**
task1502	0.260	0.557	0.297	**0.489**	**0.378**	**0.561**	**0.183**	0.485

`task1612`. We use 1000 samples for each task and a batch size of 100.

Baselines and LLMs used FormatSpread (Sclar et al., 2024) is a tool that can create different variations of traditional prompts. We use the traditional prompts created by FormatSpread as our baselines. We then compare the performance of LLMs when we use traditional prompts versus when we use MOF prompts. When the results of MOF prompts improve over traditional prompts by increasing the accuracy of the best prompt, minimizing the spread, and increasing the average accuracy across all the 10 variations of prompts, we consider that to be a success. From the 25 tasks that we perform experiments on, we report three categories of results: results where MOF prompts successfully improved over traditional prompts for each LLM, results where there were partial improvements, and results where MOF prompts failed to improve over traditional prompts. In our experiments we use 4 LLMs: `falcon-11B`, `Llama-2-13b-hf`, `Llama-2-13b-chat-hf`, and `llama-3-70b-instruct`.

3.2 Results

Meta-Llama-3-70B-Instruct results Out of the 25 tasks, our proposed prompting technique, MOF, successfully improves over traditional prompts on 5 tasks as reported in Table 1, partially improves on 12 tasks as reported in Table 9 and fails on 8 tasks as reported in Table 5 in the appendix.

Llama-2-13b results Out of the 25 tasks, our proposed prompting technique, MOF, successfully improves over traditional prompts on 8 tasks as reported in Table 2, partially improves on 11 tasks as reported in Table 10 and fails on 6 tasks as reported in Table 6 in the appendix.

Llama-2-13b-chat results Out of the 25 tasks, our proposed prompting technique, MOF, successfully improves over traditional prompts on 6 tasks as reported in Table 3, partially improves on 13 tasks as reported in Table 11 and fails on 6 tasks as reported in Table 7 in the appendix.

Falcon-11B results Out of the 25 tasks, our proposed prompting technique, MOF, successfully improves over traditional prompts on 4 tasks as reported in Table 4, partially improves on 13 tasks as reported in Table 12 and fails on 8 tasks as reported in Table 8 in the appendix.

4 Conclusion and future work

Addressing the prompt brittleness problem is a difficult task. Existing works have made progress, but proposed methods are complex and computationally inefficient. We propose a simple and efficient prompting technique, MOF, and we explore its effectiveness at addressing prompt brittleness in LLMs. Results are encouraging and we observe significant improvements over the traditional prompt-

Table 3: `Llama-2-13b-chat` results where our proposed MOF prompts perform better than traditional prompts across all metrics. *Spread* is the difference between max accuracy and min accuracy. *Mean Accuracy* is the average accuracy across all the 10 prompt variations.

Task	Traditional Prompts				MOF Prompts			
	Min Accuracy	Max Accuracy	Spread	Mean Accuracy	Min Accuracy	Max Accuracy	Spread	Mean Accuracy
task158	0.495	0.568	0.073	0.537	**0.558**	**0.586**	**0.028**	**0.571**
task161	0.160	0.250	0.090	0.200	**0.225**	**0.303**	**0.078**	**0.278**
task162	0.183	0.310	0.127	0.262	**0.314**	**0.421**	**0.107**	**0.374**
task190	0.458	0.624	0.166	0.533	**0.584**	**0.653**	**0.069**	**0.627**
task1186	0.481	0.526	0.045	0.499	**0.525**	**0.530**	**0.005**	**0.528**
task1284	0.532	0.636	0.104	0.572	**0.560**	**0.639**	**0.079**	**0.598**

Table 4: `Falcon-11B` results where our proposed MOF prompts perform better than traditional prompts across all metrics. *Spread* is the difference between max accuracy and min accuracy. *Mean Accuracy* is the average accuracy across all the 10 prompt variations.

Task	Traditional Prompts				MOF Prompts			
	Min Accuracy	Max Accuracy	Spread	Mean Accuracy	Min Accuracy	Max Accuracy	Spread	Mean Accuracy
task161	0.254	0.457	0.203	**0.390**	**0.285**	**0.464**	**0.179**	0.388
task190	0.402	0.656	0.254	0.579	**0.657**	**0.658**	**0.001**	**0.657**
task316	0.070	0.554	0.484	0.312	**0.257**	**0.563**	**0.306**	**0.476**
task905	0.514	0.680	0.166	0.568	**0.560**	**0.698**	**0.138**	**0.617**

ing strategy on some tasks, as reported in Tables 1, 2, 3, and 4. An interesting direction to explore for future work is performing an extensive analysis to understand why MOF prompting works well on some tasks and fails on others.

5 Limitations

MOF prompting is simple and computationally efficient but fails on some of the tasks. Exploring why this technique fails on those tasks is an interesting direction. We hope this work will inspire more research towards addressing prompt brittleness in LLMs.

References

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2019. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*.

Simran Arora, Avanika Narayan, Mayee F Chen, Laurel Orr, Neel Guha, Kush Bhatia, Ines Chami, and Christopher Re. 2022. Ask me anything: A simple strategy for prompting language models. In *The Eleventh International Conference on Learning Representations*.

Michael Feffer, Ronald Xu, Yuekai Sun, and Mikhail Yurochkin. 2024. Prompt exploration with prompt regression. *arXiv preprint arXiv:2405.11083*.

Pritish Kamath, Akilesh Tangella, Danica Sutherland, and Nathan Srebro. 2021. Does invariant risk minimization capture invariance? In *International Conference on Artificial Intelligence and Statistics*, pages 4069–4077. PMLR.

Shuangning Li, Matteo Sesia, Yaniv Romano, Emmanuel Candès, and Chiara Sabatti. 2021. Searching for consistent associations with a multi-environment knockoff filter. *arXiv preprint arXiv:2106.04118*.

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

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-task generalization via natural language crowdsourcing instructions. In *ACL*.

Vaishnavh Nagarajan, Anders Andreassen, and Behnam Neyshabur. 2020. Understanding the failure modes of out-of-distribution generalization. *arXiv preprint arXiv:2010.15775*.

Felipe Maia Polo, Ronald Xu, Lucas Weber, Mírian Silva, Onkar Bhardwaj, Leshem Choshen, Allysson Flavio Melo de Oliveira, Yuekai Sun, and Mikhail Yurochkin. 2024. Efficient multi-prompt evaluation of llms. *arXiv preprint arXiv:2405.17202*.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations*.

Taylor Sorensen, Joshua Robinson, Christopher Rytting, Alexander Shaw, Kyle Rogers, Alexia Delorey, Mahmoud Khalil, Nancy Fulda, and David Wingate. 2022. An information-theoretic approach to prompt

engineering without ground truth labels. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 819–862, Dublin, Ireland. Association for Computational Linguistics.

Yoav Wald, Amir Feder, Daniel Greenfeld, and Uri Shalit. 2021. On calibration and out-of-domain generalization. *Advances in Neural Information Processing Systems*, 34.

Xingchen Wan, Ruoxi Sun, Hanjun Dai, Sercan Arik, and Tomas Pfister. 2023. Better zero-shot reasoning with self-adaptive prompting. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3493–3514, Toronto, Canada. Association for Computational Linguistics.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, et al. 2022. Super-naturalinstructions:generalization via declarative instructions on 1600+ tasks. In *EMNLP*.

Fan Yin, Jesse Vig, Philippe Laban, Shafiq Joty, Caiming Xiong, and Chien-Sheng Wu. 2023. Did you read the instructions? rethinking the effectiveness of task definitions in instruction learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3063–3079, Toronto, Canada. Association for Computational Linguistics.

Mingzhang Yin, Yixin Wang, and David M Blei. 2021. Optimization-based causal estimation from heterogenous environments. *arXiv preprint arXiv:2109.11990*.

Han Zhou, Xingchen Wan, Lev Proleev, Diana Mincu, Jilin Chen, Katherine Heller, and Subhrajit Roy. 2023. Batch calibration: Rethinking calibration for in-context learning and prompt engineering. *arXiv preprint arXiv:2309.17249*.

A Additional results

Table 5: `Llama-3-70B-Instruct` results where our proposed MOF prompts fail to improve over traditional prompts on all the metrics. *Spread* is the difference between max accuracy and min accuracy. *Mean Accuracy* is the average accuracy across all the 10 prompt variations.

Task	Traditional Prompts				MOF Prompts			
	Min Accuracy	Max Accuracy	Spread	Mean Accuracy	Min Accuracy	Max Accuracy	Spread	Mean Accuracy
task155	0.539	0.567	0.028	0.556	0.372	0.562	0.190	0.512
task162	0.594	0.630	0.036	0.612	0.580	0.622	0.042	0.601
task190	0.644	0.727	0.083	0.662	0.491	0.669	0.178	0.568
task337	0.884	0.893	0.009	0.888	0.876	0.891	0.015	0.885
task607	0.681	0.707	0.026	0.694	0.652	0.685	0.033	0.674
task905	0.742	0.778	0.036	0.763	0.721	0.764	0.043	0.741
task1186	0.575	0.597	0.022	0.585	0.554	0.591	0.037	0.569
task1284	0.721	0.780	0.059	0.753	0.694	0.758	0.064	0.721

Table 6: `Llama-2-13b` results where our proposed MOF prompts fail to improve over traditional prompts on all the metrics. *Spread* is the difference between max accuracy and min accuracy. *Mean Accuracy* is the average accuracy across all the 10 prompt variations.

Task	Traditional Prompts				MOF Prompts			
	Min Accuracy	Max Accuracy	Spread	Mean Accuracy	Min Accuracy	Max Accuracy	Spread	Mean Accuracy
task155	0.337	0.396	0.059	0.370	0.280	0.385	0.105	0.334
task316	0.269	0.475	0.206	0.344	0.122	0.472	0.350	0.268
task320	0.429	0.578	0.149	0.541	0.316	0.546	0.230	0.445
task905	0.490	0.586	0.096	0.526	0.438	0.538	0.100	0.491
task1283	0.545	0.609	0.064	0.567	0.472	0.557	0.085	0.515
task1612	0.519	0.591	0.072	0.564	0.445	0.560	0.115	0.506

Table 7: `Llama-2-13b-chat` results where our proposed MOF prompts fail to improve over traditional prompts on all the metrics. *Spread* is the difference between max accuracy and min accuracy. *Mean Accuracy* is the average accuracy across all the 10 prompt variations.

Task	Traditional Prompts				MOF Prompts			
	Min Accuracy	Max Accuracy	Spread	Mean Accuracy	Min Accuracy	Max Accuracy	Spread	Mean Accuracy
task279	0.388	0.514	0.126	0.444	0.354	0.501	0.147	0.414
task280	0.846	0.861	0.015	0.853	0.817	0.852	0.035	0.841
task608	0.514	0.564	0.050	0.538	0.491	0.549	0.058	0.513
task609	0.593	0.642	0.049	0.625	0.541	0.618	0.077	0.578
task1283	0.490	0.533	0.043	0.519	0.466	0.520	0.054	0.506
task1502	0.583	0.612	0.029	0.597	0.548	0.610	0.062	0.580

Table 8: `Falcon-11B` results where our proposed MOF prompts fail to improve over traditional prompts on all the metrics. *Spread* is the difference between max accuracy and min accuracy. *Mean Accuracy* is the average accuracy across all the 10 prompt variations.

Task	Traditional Prompts				MOF Prompts			
	Min Accuracy	Max Accuracy	Spread	Mean Accuracy	Min Accuracy	Max Accuracy	Spread	Mean Accuracy
task155	0.359	0.432	0.073	0.393	0.333	0.431	0.098	0.379
task279	0.506	0.724	0.218	0.597	0.337	0.700	0.363	0.587
task319	0.586	0.828	0.242	0.722	0.414	0.825	0.411	0.680
task320	0.584	0.861	0.277	0.732	0.113	0.843	0.730	0.562
task609	0.640	0.729	0.089	0.701	0.503	0.692	0.189	0.612
task1284	0.645	0.790	0.145	0.753	0.545	0.778	0.233	0.709
task1347	0.312	0.452	0.140	0.425	0.183	0.389	0.206	0.294
task1612	0.513	0.706	0.193	0.639	0.466	0.699	0.233	0.553

Table 9: `Llama-3-70B-Instruct` results where our proposed MOF prompts perform better than traditional prompts on some of the metrics. *Spread* is the difference between max accuracy and min accuracy. *Mean Accuracy* is the average accuracy across all the 10 prompt variations.

Task	Traditional Prompts				MOF Prompts			
	Min Accuracy	Max Accuracy	Spread	Mean Accuracy	Min Accuracy	Max Accuracy	Spread	Mean Accuracy
task114	0.753	0.802	0.049	0.773	0.705	0.749	0.044	0.731
task158	0.654	0.779	0.125	0.708	0.654	0.796	0.142	0.699
task161	0.472	0.507	0.035	0.496	0.475	0.512	0.037	0.493
task163	0.485	0.513	0.028	0.498	0.475	0.542	0.067	0.502
task279	0.614	0.780	0.166	0.739	0.693	0.762	0.069	0.742
task316	0.534	0.555	0.021	0.548	0.546	0.577	0.031	0.561
task319	0.335	0.886	0.551	0.783	0.358	0.879	0.521	0.775
task608	0.704	0.740	0.036	0.730	0.716	0.769	0.053	0.737
task609	0.669	0.696	0.027	0.684	0.673	0.690	0.017	0.683
task904	0.587	0.684	0.097	0.643	0.576	0.717	0.141	0.626
task1283	0.541	0.564	0.023	0.553	0.535	0.553	0.018	0.544
task1502	0.443	0.666	0.223	0.600	0.479	0.639	0.160	0.594

Table 10: `Llama-2-13b` results where our proposed MOF prompts perform better than traditional prompts on some of the metrics. *Spread* is the difference between max accuracy and min accuracy. *Mean Accuracy* is the average accuracy across all the 10 prompt variations.

Task	Traditional Prompts				MOF Prompts			
	Min Accuracy	Max Accuracy	Spread	Mean Accuracy	Min Accuracy	Max Accuracy	Spread	Mean Accuracy
task114	0.513	0.639	0.126	0.563	0.494	0.684	0.190	0.614
task158	0.521	0.594	0.073	0.564	0.483	0.602	0.119	0.550
task161	0.220	0.282	0.062	0.246	0.299	0.424	0.125	0.371
task190	0.367	0.657	0.290	0.594	0.654	0.657	0.003	0.657
task279	0.060	0.595	0.535	0.460	0.203	0.471	0.268	0.314
task319	0.005	0.481	0.476	0.190	0.035	0.595	0.560	0.357
task337	0.013	0.775	0.762	0.572	0.397	0.593	0.196	0.491
task608	0.472	0.594	0.122	0.530	0.513	0.684	0.171	0.554
task609	0.549	0.650	0.101	0.625	0.541	0.601	0.060	0.567
task904	0.185	0.643	0.458	0.423	0.422	0.567	0.145	0.494
task1347	0.236	0.307	0.071	0.269	0.159	0.216	0.057	0.181

Table 11: `Llama-2-13b-chat` results where our proposed MOF prompts perform better than traditional prompts on some of the metrics. *Spread* is the difference between max accuracy and min accuracy. *Mean Accuracy* is the average accuracy across all the 10 prompt variations.

Task	Traditional Prompts				MOF Prompts			
	Min Accuracy	Max Accuracy	Spread	Mean Accuracy	Min Accuracy	Max Accuracy	Spread	Mean Accuracy
task114	0.599	0.759	0.160	0.655	0.553	0.635	0.082	0.605
task155	0.284	0.332	0.048	0.307	0.217	0.452	0.235	0.381
task163	0.387	0.486	0.099	0.427	0.383	0.457	0.074	0.415
task316	0.479	0.500	0.021	0.484	0.488	0.515	0.027	0.502
task317	0.564	0.600	0.036	0.579	0.719	0.787	0.068	0.750
task319	0.071	0.320	0.249	0.174	0.160	0.286	0.126	0.230
task320	0.521	0.559	0.038	0.543	0.481	0.560	0.079	0.530
task337	0.570	0.797	0.227	0.642	0.553	0.798	0.245	0.657
task607	0.516	0.657	0.141	0.608	0.551	0.616	0.065	0.585
task904	0.495	0.548	0.053	0.516	0.536	0.621	0.085	0.572
task905	0.538	0.636	0.098	0.571	0.527	0.612	0.085	0.581
task1347	0.230	0.275	0.045	0.252	0.180	0.200	0.020	0.189
task1612	0.440	0.513	0.073	0.471	0.375	0.555	0.180	0.491

Table 12: `Falcon-11B` results where our proposed MOF prompts perform better than traditional prompts on some of the metrics. *Spread* is the difference between max accuracy and min accuracy. *Mean Accuracy* is the average accuracy across all the 10 prompt variations.

Task	Traditional Prompts				MOF Prompts			
	Min Accuracy	Max Accuracy	Spread	Mean Accuracy	Min Accuracy	Max Accuracy	Spread	Mean Accuracy
task114	0.511	0.745	0.234	0.632	0.494	0.755	0.261	0.610
task158	0.450	0.536	0.086	0.483	0.520	0.692	0.172	0.567
task162	0.344	0.451	0.107	0.419	0.328	0.415	0.087	0.392
task163	0.443	0.491	0.048	0.464	0.458	0.480	0.022	0.471
task280	0.599	0.862	0.263	0.727	0.765	0.841	0.076	0.803
task317	0.376	0.672	0.296	0.501	0.453	0.795	0.342	0.672
task337	0.625	0.807	0.182	0.731	0.471	0.810	0.339	0.532
task607	0.504	0.705	0.201	0.619	0.499	0.561	0.062	0.507
task608	0.498	0.766	0.268	0.657	0.491	0.688	0.197	0.593
task904	0.233	0.666	0.433	0.420	0.347	0.604	0.257	0.465
task1186	0.469	0.595	0.126	0.556	0.525	0.550	0.025	0.536
task1283	0.528	0.575	0.047	0.546	0.516	0.542	0.026	0.525
task1502	0.001	0.633	0.632	0.305	0.414	0.578	0.164	0.494