Steering Conversations via Logit Bias in LLMs

Anonymous ACL submission

Abstract

Recent advancements in open-domain conversational agents highlight challenges in creating dynamic chatbot personalities. This paper explores Logit Bias as a novel mechanism for customizing LLM outputs, enabling seamless personality shifts without relying on static, dataset-constrained training. Unlike fine-tuning or prompt tuning, this method personalizes interactions without additional training, offering a flexible and efficient alternative. Through extensive experiments, we show that this approach effectively modifies model behavior while maintaining overall performance, influencing conversational quality and linguistic properties. This scalable solution allows for dynamically adaptable language models, meeting user expectations across diverse applications without requiring fine-tuning.

1 Introduction

Large language models (LLMs) have emerged in recent times as powerful tools capable of generating human-like language and engaging in sophisticated dialogues. These models have found applications across a wide range of domains, including customer service, content creation, and educational technology. The challenge of tailoring LLM traits to align with specific conversational contexts or user preferences is yet to be adequately addressed however. Traditional approaches often rely on crafting precise prompts with respect to desired traits or, alternatively, involve fine-tuning extensive parameters on new datasets – a process that is both time-consuming and increasingly constrained by data availability.

Considering the cost of retraining in term of time and resources, approaches to automatic dialogue that can be adapted without extensive retraining are needed. We subsequently propose a new method of steering of LLM conversation style that enables conversation control over nuanced traits, including reasoning, writing, information extraction, etc., that does not require retraining, prompt engineering, or data-specific fine-tuning. Our approach works by manipulating predicted logit scores from pre-trained models prior to the application of softmax, allowing Logit Bias to manipulate model outputs by altering the probability distribution of possible responses.

This paper investigates the application of Logit Bias as a strategic mechanism for steering LLM conversational traits. We conduct a series of experiments to test the degree to which it is possible to effectively steer model traits with this approach. Experiment results indicate that the approach is capable of steering conversational traits on demand without any retraining. By leveraging the adaptability of models that do not rely on specific training datasets, we propose a scalable and efficient path to deploying language models that meet the diverse expectations of users and applications. Through this work, we aim to contribute to efficient methods of LLM customization and generalization, presenting Logit Bias as a promising tool for enhancing model adaptability without the traditional burdens of data and resource-intensive retraining processes.

2 Background

The customization of language models has progressed significantly, with traditional fine-tuning approaches such as those by (Radford and Narasimhan, 2018), (Devlin et al., 2019), and (Brown et al., 2020) tailoring models to specific domains. However, these methods are computationally expensive and impractical for real-time adaptability. Parameter-efficient tuning methods like LoRA (Hu et al., 2021) have improved efficiency by reducing resource requirements while maintaining performance (Hayou et al., 2024; Liu et al., 2024). Despite these benefits, LoRA still necessitates training and lacks real-time adaptability. In

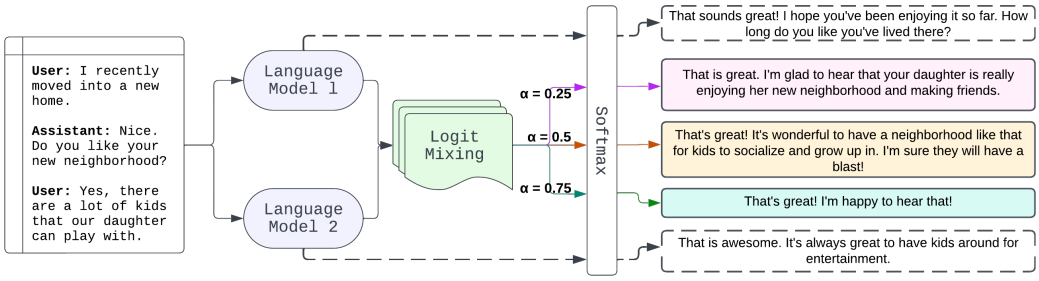

Figure 1: An example from Empathetic Dataset tested on Llama3.2B-instruct-3B model with input and potential outputs using proposed architecture.

parallel, Mixture of Experts (MoE) models (Artetxe et al., 2022) have emerged, dynamically selecting specialized experts for different inputs, enhancing efficiency and scalability. Hybrid methods integrating MoE with LoRA, such as (Li et al., 2024; Wang et al., 2024; Qing et al., 2024), have demonstrated further improvements. However, MoE models require multiple trained sub-models, leading to increased resource demands and complexity. One may also argue that adjusting temperature can result in varied responses but it lacks precise control and is difficult to align to the predefined requirements.

Prompt engineering has gained popularity as an alternative strategy (Lester et al., 2021; Liu et al., 2022; Tu et al., 2022; Fan et al., 2023), using structured prompts to guide model outputs. This approach offers flexibility and reduces computational overhead compared to fine-tuning but often results in inconsistent outputs and necessitates careful prompt design (Macedo et al., 2024; Ronanki et al., 2024; Ye et al., 2024; Fagadau et al., 2024). Furthermore, prompt effectiveness varies across models, making deployment consistency challenging. Efforts such as (Yang et al., 2024) have attempted automated prompt optimization, yet difficulties in ensuring stability across tasks and architectures remain unresolved.

Recent advancements in Logit Bias techniques have introduced novel ways to modify LLM outputs without altering model parameters. The Offset Unlearning framework (Huang et al., 2024) applies logit offsets to remove undesired knowledge, while Dynamic Logits Fusion (DLF) (Fan et al., 2024) combines logits from smaller fine-tuned models to enhance performance without additional training. Although these methods improve adaptability, they rely on complex operations such as KL divergence at each decoding step, increasing computational overhead and limiting scalability. Additionally, existing works focus primarily on performance improvement rather than qualitative aspects like conversational personality. This study proposes a simpler mechanism for adjusting logit probabilities, allowing for controlled personality shifts in LLMs without requiring extensive retraining or intricate prompt engineering.

3 Method

Our approach takes advantage of the availability of vast numbers of readily available pretrained models fine-tuned for specific use cases, and the autoregressive nature of LLMs, as they generate output by predicting one token at a time, with each prediction conditioned on all previously generated tokens. This sequential generation make LLMs particularly suitable for fine-grained control over generation through Logit Bias manipulation.

Almost all modern LLMs are designed to generate logit scores along with their final output, as logit scores represent the unnormalised scores for each token in the vocabulary. Due to the autoregressive nature of LLMs, these scores are generated at the token level to maintain coherence of the output sentence during generation. This property allows us to steer the model's prediction using Logit Bias. Logit Bias allows control over whether the model is more or less likely to output a specific word and can be used, for example, to ban a particular word from the vocabulary, such as expletives. In contrast, our approach employs Logit Bias manipulation in order to steer outputs towards specific conversational traits such as making the dialogue agent more knowledgeable or empathetic. This allows generation of new hybrid models that show traits from distinct datasets without fine-tuning.

For a given set of n base fine-tuned models specialized for different traits, we propose using Logit Bias mixing at the token level to generate novel models that pose the weighted traits of all the base fine-tuned models. We achieve this by passing prompts through the n base models and extract the distribution of logits from each model just before feeding them into the softmax function. We can then apply arithmetical operation on these distributions to get a unified distribution.

For the scope of this work, we choose weighted average of the logits as the arithmetic operation for logit combination. As with weighted averaging, we can assign varying importance to individual models and their specialization, thereby steering the overall output towards the desired blend of traits. We achieve this with supplying the weights for each trait to define their influence in the final response. The following function is used to merge the distribution:

$$x = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \tag{1}$$

where x is the new set of unified logits; x_i are logits from the model fined-tuned on trait i (e.g. empathy, knowledge, sarcasm, etc.), and w_i is the weight assigned to trait i.

For two base models, for example, $n = 2$, equation 1 can be rewritten as:

$$x = \alpha \times x_1 + (1 - \alpha) \times x_2 \tag{2}$$

where α is a tunable parameter controlling the influence of both traits in the final response.

As the generation is autoregressive, the Logit Bias computation is carried out at the token level. In case of two base models, this suggests that even after generating two different preferred tokens, both models get fed the same token at each generation step after Logit manipulation. This allows re-aligning of the base models and influences them to stick to the current context at the token level, enabling fine-grained control over the conversational trajectory of the models; resulting in stylistically varied conversations. Allowing the user to steer the traits of model while varying α to create innumerably varied models using just two (or more) base models with distinct conversation traits.

4 Experiments

We utilized two variants of Llama 3.2-instruct[1][2] with 1 billion and 3 billion parameters as base models to enhance scalability and performance. To establish distinct conversational traits, we focused on two personality attributes: knowledge grounding and empathetic responses, following (Smith et al., 2020). We fine-tuned the base models separately on the Wizard of Wikipedia (WoW) dataset (Dinan et al., 2019) and the Empathetic Dialogues (ED) dataset (Rashkin et al., 2019) comprising of 19,533 and 18,446 prompts, respectively, after applying the Llama chat-template to each conversation. We used 8-bit quantized LoRA fine-tuning (Hu et al., 2021) for the same.

To explore the impact of Logit Bias mixing, we created three hybrid models by merging the two fine-tuned models with fixed α values of 0.25, 0.50, and 0.75, corresponding to models 25E-75W, 50E-50W, and 75E-25W. Additionally, we retained the base models 100E-0W ($\alpha = 1$) and 0E-100W ($\alpha = 0$). Here, α represents the influence of the ED-trained model, inversely affecting the contribution of the WoW-trained model. The goal was to investigate how Logit Bias influences model behavior, balancing strengths from both models while introducing response variety.

Each of the five variants was evaluated on 1,000 randomly selected test samples from EMP and WoW datasets. We employed an automatic evaluation framework to analyze the performance gradient across different α values, including BLEU (Papineni et al., 2002), Word Error Rate (WER), Average Response Length, BERT similarity score (Zhang et al., 2019), and lexical diversity. Additionally, we applied the *LLM as Judge* approach using MT Bench (Zheng et al., 2023) to qualitatively assess conversational attributes like reasoning, writing, and roleplay. This evaluation provides insight into how Logit Bias can be used to steer language generation patterns and model adaptability. We provide the anonymized code to help future research.[3]

5 Results

Tables 1 and 2 present the performance of the (steered) hybrid models based on Llama3.2-instruct

[1] huggingface.co/meta-llama/Llama-3.2-1B-Instruct
[2] huggingface.co/meta-llama/Llama-3.2-3B-Instruct
[3] https://anonymous.4open.science/r/logit_bias_steering-6ADB

variants (1B and 3B parameters) tested on ED and WoW datasets. While automatic metrics are not perfect, trends across WER, ARL, BLEU, and BERT score confirm that Logit Bias influences model performance in the desired way. WER decreases as α increases for both datasets, while ARL is highest for 0E-100W due to WoW's prompts explanatory nature and decreases significantly with ED's influence. BLEU improves for ED with increasing α as expected, while merged models outperform base models on WoW, suggesting Logit Bias aids dataset alignment or at least does not hurt the baseline performance. Stable BERT scores also indicate performance retention despite diversified responses. Similar trends are observed for Lexical diversity presented in Table 3.

	Model	WER	ARL	BLEU	BERTScore
ED	0E-100W	7.4081	66.9165	0.3623	0.8463
	25E-75W	4.5601	41.7769	0.6443	0.8537
	50E-50W	2.9972	27.1934	0.9731	0.8591
	75E-25W	1.9709	17.6826	1.4557	0.8631
	100E-0W	1.4082	11.5314	1.8738	0.8634
WoW	0E-100W	3.9522	66.4440	0.85097	0.8510
	25E-75W	2.8478	47.4070	0.8546	0.8546
	50E-50W	2.0224	33.3090	0.8575	0.8576
	75E-25W	1.4282	21.6260	0.8581	0.8581
	100E-0W	1.1391	14.1120	0.8567	0.8568

Table 1: Llama 1B Performance for ED and WoW

	Model	WER	ARL	BLEU	BERTScore
ED	0E-100W	3.0905	27.46	0.7439	0.7961
	25E-75W	2.5526	22.527	1.1348	0.8409
	50E-50W	1.9121	17.078	1.5614	0.8622
	75E-25W	1.5488	13.623	1.7771	0.8649
	100E-0W	1.3639	11.134	1.8450	0.8639
WoW	0E-100W	1.6800	21.91	1.81738	0.6158
	25E-75W	1.5229	21.61	2.2302	2.2302
	50E-50W	1.2690	18.485	2.3831	2.3831
	75E-25W	1.2427	17.448	2.02066	2.0206
	100E-0W	1.1455	15.318	1.9307	1.9307

Table 2: Llama 3B Performance for ED and WoW

For qualitative evaluation, we use MT-Bench. Table 4 for Llama-1B shows a large gap between the scores of both base models, likely due to the performance drop by fine-tuning on ED, however, increasing WoW influence boosts the performance of as we decrease α, mitigating the drop in performance, demonstrating Logit Bias effectiveness. In Llama-3B experiments, the merged models outperform base models, highlighting the benefits of controlled logit blending in enhancing conversational ability when equally performing base models are provided. Figure 3 in Appendix A shows that hybrid models maintain overall performance while excelling in specific conversational traits, making them more suitable for targeted applications. Additional details for 1B and 3B models are in Tables 8, 7, and Figure 2.

Both automatic and qualitative evaluations confirm a gradual shift in response properties as α changes, making it a key hyperparameter for controlled output variation without extra fine-tuning. This work demonstrates the potential of Logit Bias for adapting generative model outputs to different use cases while maintaining overall performance, even improving in some cases. Future research will explore dynamic weight optimization at the response level, enabling real-time trait adjustments for improved replies.

	Llama 1B		Llama 3B	
Model	ED	WoW	ED	WoW
0E-100W	0.0652	0.1913	0.1786	0.2735
25E-75W	0.0707	0.2048	0.1675	0.2609
50E-50W	0.0727	0.2167	0.1708	0.2507
75E-25W	0.0782	0.2245	0.1763	0.2428
100E-0W	0.0842	0.2429	0.1694	0.2497

Table 3: Lexical scores for Llama 1B and Llama 3B model

Model	Llama 1B	Llama 3B
0E-100W	6.0815	5.5032
25E-75W	5.7563	6.0751
50E-50W	4.1310	6.2806
75E-25W	3.8875	6.1075
100E-0W	2.7340	5.3516

Table 4: Ave. MT-Bench scores for Llama 1B and 3B model

6 Conclusion

In this paper, we demonstrate the effectiveness of Logit Bias in modulating LLM outputs without additional fine-tuning. By adjusting weight distribution between two base models, we observed a controlled shift in conversational properties, establishing Logit Bias as a potential methodology for tuning model personalities. Automatic metrics such as WER, BLEU, BERT score, and lexical diversity showed predictable trends, confirming the influence of Logit Bias. Qualitative evaluation via MT-Bench further validated hybrid models, revealing enhanced conversational abilities, often surpassing base models by balancing conversation traits. While this work focuses on fixed-weight blending, future research can explore dynamic weight adjustments for real-time adaptation.

Limitations

Despite its advantages, our approach has several limitations similar to other logit bias based methods. Firstly, the inference process requires running infernce on multiple base models simultaneously, which increases computational overhead. However, this can be seen as a trade-off with the cost and inefficiency of training new models for every static persona. Secondly, as aligning models using Logit Bias fairly recent development, it lacks established baselines regarding the shifts in conversational traits of models. This possed a challenge of finding a well established baseline to evaluate our apparoach. Lastly, while our technique is theoretically applicable to merging logits from models with different architectures, its effectiveness in such scenarios remains unverified. We plan to explore cross-architecture logit blending to assess its viability and limitations in future work.

References

Mikel Artetxe, Shruti Bhosale, Naman Goyal, Todor Mihaylov, Myle Ott, Sam Shleifer, Xi Victoria Lin, Jingfei Du, Srinivasan Iyer, Ramakanth Pasunuru, Giri Anantharaman, Xian Li, Shuohui Chen, Halil Akin, Mandeep Baines, Louis Martin, Xing Zhou, Punit Singh Koura, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Mona Diab, Zornitsa Kozareva, and Ves Stoyanov. 2022. Efficient large scale language modeling with mixtures of experts. *Preprint*, arXiv:2112.10684.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. 33:1877–1901.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. *Preprint*, arXiv:1811.01241.

Ionut Daniel Fagadau, Leonardo Mariani, Daniela Micucci, and Oliviero Riganelli. 2024. Analyzing prompt influence on automated method generation: An empirical study with copilot. In *Proceedings of the 32nd IEEE/ACM International Conference on Program Comprehension*, ICPC '24, page 24–34, New York, NY, USA. Association for Computing Machinery.

Caoyun Fan, Jidong Tian, Yitian Li, Wenqing Chen, Hao He, and Yaohui Jin. 2023. Chain-of-thought tuning: Masked language models can also think step by step in natural language understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14774–14785, Singapore. Association for Computational Linguistics.

Chenghao Fan, Zhenyi Lu, Wei Wei, Jie Tian, Xiaoye Qu, Dangyang Chen, and Yu Cheng. 2024. On giant's shoulders: Effortless weak to strong by dynamic logits fusion. *Preprint*, arXiv:2406.15480.

Soufiane Hayou, Nikhil Ghosh, and Bin Yu. 2024. Lora+: Efficient low rank adaptation of large models. *Preprint*, arXiv:2402.12354.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

James Y. Huang, Wenxuan Zhou, Fei Wang, Fred Morstatter, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2024. Offset unlearning for large language models. *Preprint*, arXiv:2404.11045.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Dengchun Li, Yingzi Ma, Naizheng Wang, Zhengmao Ye, Zhiyuan Cheng, Yinghao Tang, Yan Zhang, Lei Duan, Jie Zuo, Cal Yang, and Mingjie Tang. 2024. Mixlora: Enhancing large language models fine-tuning with lora-based mixture of experts. *Preprint*, arXiv:2404.15159.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2022. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 61–68, Dublin, Ireland. Association for Computational Linguistics.

Zequan Liu, Jiawen Lyn, Wei Zhu, Xing Tian, and Yvette Graham. 2024. ALoRA: Allocating low-rank

adaptation for fine-tuning large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 622–641, Mexico City, Mexico. Association for Computational Linguistics.

Marcos Macedo, Yuan Tian, Filipe Cogo, and Bram Adams. 2024. Exploring the impact of the output format on the evaluation of large language models for code translation. In *Proceedings of the 2024 IEEE/ACM First International Conference on AI Foundation Models and Software Engineering*, FORGE '24, page 57–68, New York, NY, USA. Association for Computing Machinery.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Peijun Qing, Chongyang Gao, Yefan Zhou, Xingjian Diao, Yaoqing Yang, and Soroush Vosoughi. 2024. AlphaLoRA: Assigning LoRA experts based on layer training quality. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20511–20523, Miami, Florida, USA. Association for Computational Linguistics.

Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pre-training.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: a new benchmark and dataset. *Preprint*, arXiv:1811.00207.

Krishna Ronanki, Beatriz Cabrero-Daniel, and Christian Berger. 2024. Prompt smells: An omen for undesirable generative ai outputs. In *Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering - Software Engineering for AI*, CAIN '24, page 286–287, New York, NY, USA. Association for Computing Machinery.

Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. 2020. Can you put it all together: Evaluating conversational agents' ability to blend skills. *Preprint*, arXiv:2004.08449.

Lifu Tu, Caiming Xiong, and Yingbo Zhou. 2022. Prompt-tuning can be much better than fine-tuning on cross-lingual understanding with multilingual language models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5478–5485, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

P. Wang, M. Wang, Z. Ma, X. Yang, S. Feng, and Y. Zhang. 2024. Language models as continuous self-evolving data engineers.

Sheng Yang, Yurong Wu, Yan Gao, Zineng Zhou, Bin Benjamin Zhu, Xiaodi Sun, Jian-Guang Lou, Zhiming Ding, Anbang Hu, Yuan Fang, Yunsong Li, Junyan Chen, and Linjun Yang. 2024. AMPO: Automatic multi-branched prompt optimization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20267–20279, Miami, Florida, USA. Association for Computational Linguistics.

Qinyuan Ye, Mohamed Ahmed, Reid Pryzant, and Fereshte Khani. 2024. Prompt engineering a prompt engineer. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 355–385, Bangkok, Thailand. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

A Appendix

A.1 Additional Results

Model	Turn 1	Turn 2	Ave. Score
0E-100W	6.0815	5.2927	5.7098
25E-75W	5.7563	4.7500	5.2531
50E-50W	4.1310	3.1943	3.6782
75E-25W	3.8875	3.4250	3.6563
100E-0W	2.7340	2.2941	2.5251

Table 5: MT-Bench scores For llama 1B model

Model	Turn 1	Turn 2	Ave. Score
0E-100W	5.4187	5.5945	5.5032
25E-75W	6.1812	5.9589	6.0751
50E-50W	6.4062	6.1466	6.2806
75E-25W	6.3875	5.8205	6.1075
100E-0W	5.5812	5.1066	5.3516

Table 6: MT-Bench scores For llama 3B model

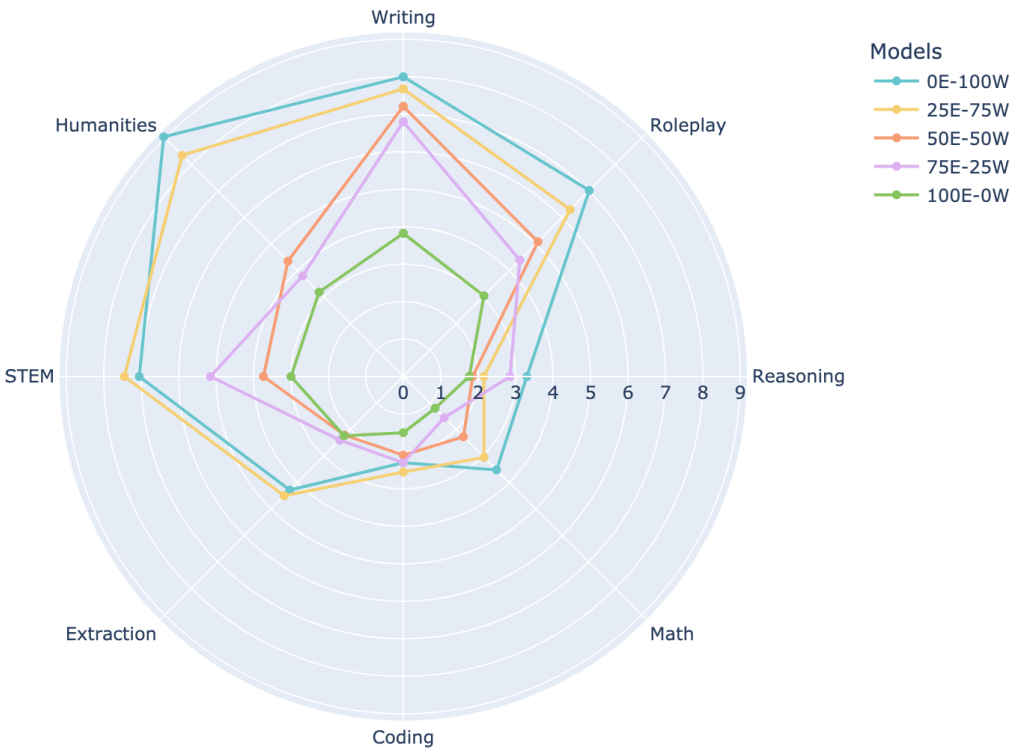

Figure 2: Scores of Llama3.2 1B based model on different conversational traits from LLM-Bench.

Model	Writing	Roleplay	Reasoning	Math	Coding	Extraction	STEM	Humanities
orig.	7.5000	6.3810	2.3810	4.2727	3.5000	4.2857	6.3571	8.7917
0E-100W	8.0000	7.0227	3.3000	3.5238	2.3000	4.2857	7.0500	9.0476
25E-75W	7.6750	6.3000	2.1500	3.0500	2.5500	4.5000	7.4500	8.3500
50E-50W	7.2100	5.0851	1.8542	2.2727	2.0976	2.2143	3.7333	4.3556
75E-25W	6.8000	4.4000	2.8500	1.5500	2.3000	2.4000	5.1500	3.8000
100E-0W	3.8214	3.0455	1.7600	1.2000	1.5000	2.2381	3.0000	3.1818

Table 7: Comparison of Llama 1B based models evaluated with respect to a range of on conversational properties, where orig. = Llama3.2_1b

Model	Writing	Roleplay	Reasoning	Math	Coding	Extraction	STEM	Humanities
orig.	8.8947	8.1111	4.1053	5.3000	6.2500	8.2632	7.6316	8.4722
0E-100W	7.4000	7.4750	4.2000	3.1053	3.1111	5.1579	6.2000	7.1667
25E-75W	7.2000	7.0500	4.8947	3.8333	4.2222	6.3000	6.8421	7.9211
50E-50W	8.1500	7.1500	4.0000	4.3158	4.1875	6.8500	6.8500	8.2250
75E-25W	6.6500	6.5250	4.9474	3.8500	5.7368	6.3000	7.4000	7.3750
100E-0W	6.3500	5.5000	3.4444	3.8421	3.5556	7.3000	5.4750	6.9000

Table 8: Comparison of Llama 3B based models evaluated with respect to a range of on conversational properties, where orig. = Llama3.2_3b

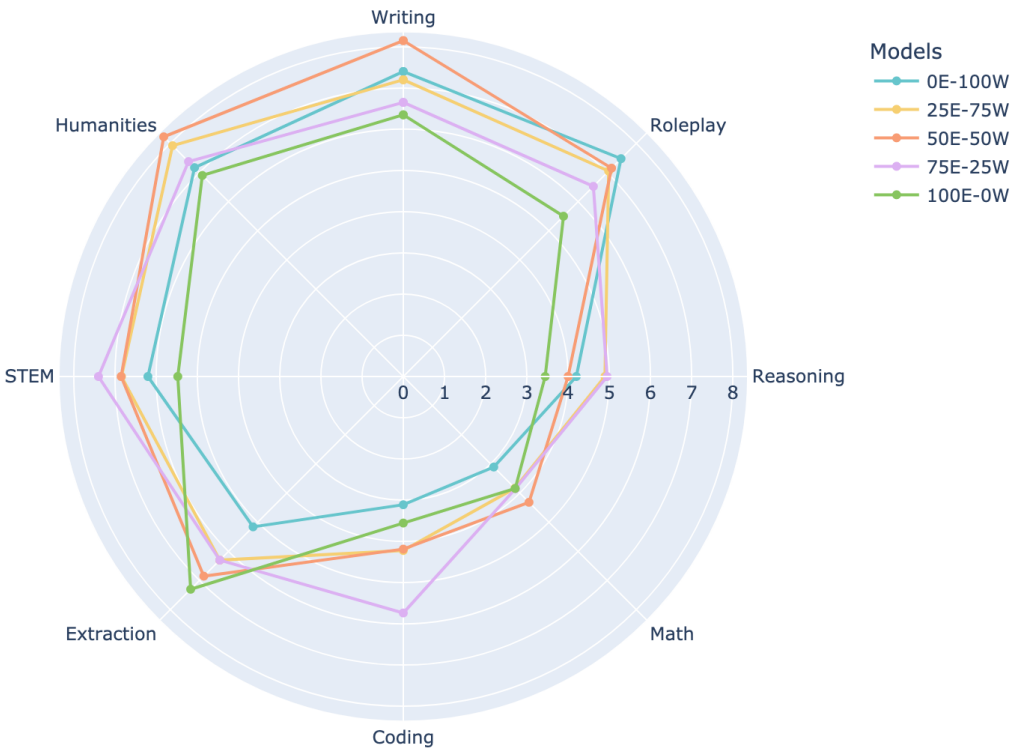

Figure 3: Scores of Llama3.2 3B based model on different conversational traits from MT-Bench.