A Lightweight Multi Aspect Controlled Text Generation Solution For Large Language Models

Chenyang Zhang[*], **Jiayi Lin**[*], **Haibo Tong, Bingxuan Hou,**
Dongyu Zhang, Jialin Li, Junli Wang [†]
Key Laboratory of Embedded System and Service Computing (Tongji University),
Ministry of Education, Shanghai 201804, China.
National (Province-Ministry Joint) Collaborative Innovation Center
for Financial Network Security, Tongji University, Shanghai 201804, China.
{inkzhangcy,2331908,2151130,2052643,yidu,2233032,junliwang}@tongji.edu.cn

Abstract

Multi-Aspect Controllable Text Generation (MCTG) introduces fine-grained multiple constraints in natural language generation, i.e. control attributes in topics, sentiments, and detoxification. MCTG demonstrates application prospects for trustworthy generation of Large Language Models (LLMs) but is limited by generalization issues. Existing work exploits additional structures and strategies for solutions, requiring LLMs' modifications. To activate LLMs' MCTG ability, we propose a lightweight MCTG pipeline based on data augmentation and instruction tuning. We analyze aspect bias and correlations in traditional datasets and address these concerns with augmented control attributes and sentences. Augmented datasets are feasible for instruction tuning. We conduct experiments for various LLMs backbone and parameter sizes, demonstrating general effectiveness on MCTG performance.

1 Introduction

Multi-Aspect Controlled Text Generation (Gu et al., 2022) is an emerging natural language generation task. MCTG alleviates multiple constraints (e.g. detoxification requirements) in language generation and contributes to a secure, faithful, and trustworthy generation. Existing methods (Gu et al., 2022; Liu et al., 2024b; Ding et al., 2023; Kumar et al., 2021) mainly focus on additional structures or decoding procedures, limiting extrapolation to LLMs. Due to enormous parameters and complex inference processes, refactoring LLMs with existing methods is unavailable in terms of cost and performance.

Instruction tuning (IT) on target datasets is a general solution for various LLM tasks, e.g. Role Playing(Chen et al., 2023b; Shao et al., 2023), Mathematical(Li et al., 2024). However, MCTG suffers from the absence of high-quality IT datasets. Existing work (Dathathri et al., 2020; Qian et al., 2022) relies on combinations of single-aspect datasets for supervised learning, which fails to achieve the ideal performance due to issues like aspects bias and correlations (Gu et al., 2022; Liu et al., 2024b).

From the perspective of datasets, we propose a lightweight MCTG solution for LLMs. We analyze concerns in existing MCTG datasets and address them with an LLM-based data augmentation pipeline. First, we delve into control attributes and sentences in existing datasets and analyze potential concerns for aspect bias and correlations. Then, we construct a data augmentation pipeline to produce augmented datasets. We provide mechanisms to ensure the effectiveness and quality of augmentation. The data format is conveniently consistent with IT datasets. Consequently, data augmentation is beneficial to common LLMs without specific structures. We validate the effectiveness on various scales and the backbone of our LLMs. The result shows that the augmented dataset contributes to the performance of MCTG, especially in aspect de-biasing and overall accuracy among 3 aspects.

2 Task Formulation

For MCTG tasks, controls may contain various n aspects $A = \{A_1, \ldots, A_n\}$. The i-th aspect contains $|A_t|$ exclusive attributes $\{a_i^1, \ldots, a_i^{|A_t|}\}$(Liu et al., 2024b). MCTG requires a control combination, which selects one attribute from each aspect. The combination is a vector of attribute indices $\boldsymbol{c} = [c_1, \ldots, c_n]$, where $c_i \in \{1, \ldots, |A_i|\}$ stands for attribute index of i-th aspect. With the input of control combinations $\boldsymbol{c}$ and generation prompt m, generation of language model should follow multiple control aspects $(a_1^{c_1}, \ldots, a_n^{c_n})$.

Existing MCTG tasks are trained on a set of single-aspect datasets. For i-th aspect, training set

[*]Equally Contribution.

[†]Corresponding author. This work was supported by the National Key Research and Development Program of China under Grant 2023YFB3002201.

Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP 2025), pages 542–551

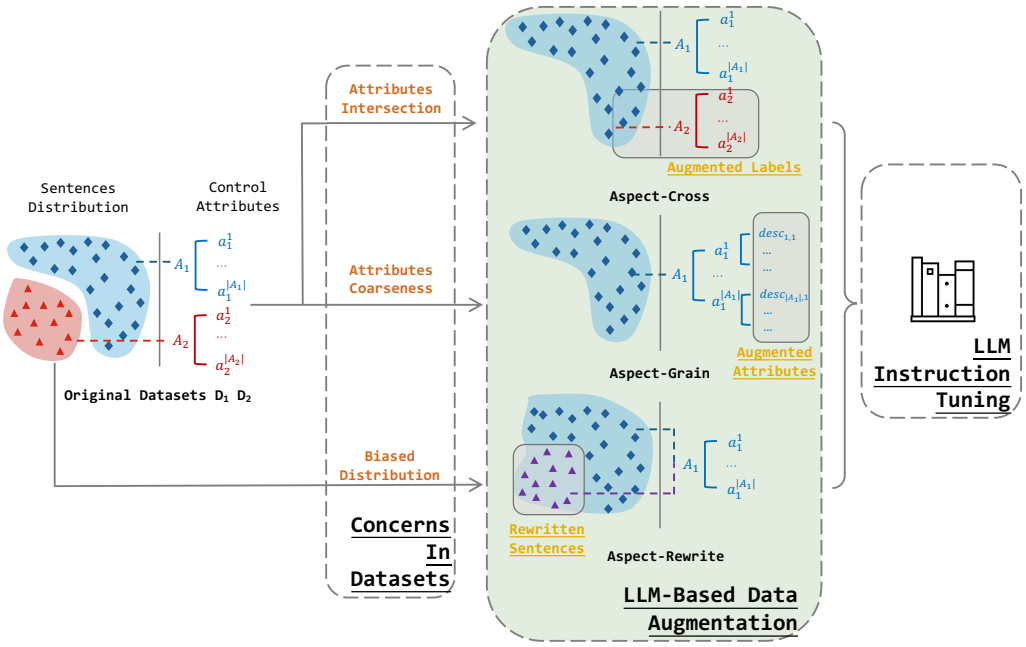

Figure 1: An overview of our lightweight MCTG solution.

$\mathcal{D}_i$ is composed of sentences x with its corresponding attribute label y in aspect A_i, notated in Eq. 1.

$$\mathcal{D}_i = \{(x, y)|x \sim (a_i^y), 1 \leq y \leq |A_i|\} \quad (1)$$

3 Methodology

As shown in Fig. 1, we first analyze 3 representative concerns in existing MCTG datasets, then propose an LLM-based data augmentation pipeline correspondingly, and finally transform augmented datasets for instruction tuning of LLMs.

3.1 Concerns In Existing MCTG Dataset

Concerns in Control Attributes Attributes from different aspects may share some common concepts, notated as **attributes intersection**. For example, IMDB (Maas et al., 2011) demonstrates positive and negative attributes in sentiment. Unfortunately, negative attributes include toxic attributes like sarcasm for the detoxification aspect.

Secondly, control attributes $a_i^t \in A_i$ are predefined, which is not specific and accurate, notated as **attributes coarseness**. Taking AGNews (Zhang et al., 2015) as an instance, it provides control aspects of *topic* only in four choices: *Sci/Tech*, *Sports*, *World* and *Business*. *World* consists of various subtopics, and sentences inside the training set struggle to cover all of the world news, which integrates the bias. General and ambiguous control attributes obstruct the further application of LLMs.

Concerns in Sentences Distributions Selections of sentences x in the training set are not uniform, with **biased distribution**. The distribution of x is biased during dataset construction. For example, IMDB datasets extract sentences from online movie reviews. However, corresponding control attributes may have instances other than movie reviews, limiting the generalization of models.

3.2 LLM-Based Data Augmentation Pipeline

We propose a data augmentation pipeline, addressing aforementioned concerns in MCTG datasets [1].

3.2.1 Aspect-Cross Augmentation

To address attribute intersection, we exploit LLMs to assign label $\tilde{y}$ in other aspects, as Eq. 2 shows.

$$\text{cross}(\mathcal{D}_i) = \{(x, \tilde{y})|x \sim (a_j^{\tilde{y}}), 1 \leq \tilde{y} \leq |A_j|, j \neq i\} \quad (2)$$

Contrasting In-Context Learning Design Although LLMs exhibit the ability for zero-shot natural language processing, direct prompting is not trustworthy. To avoid bias in labeling, we randomly sample examples for every target aspect in each prompt, known as in-context learning (ICL) examples (Brown et al., 2020).

Reject Options To enhance labeling confidence, we allow LLM to reject (e.g. output "None") for

[1] In practice, we prompt GPT-3.5-Turbo-0125 for augmentation, more details are provided in Appendix. B.

formidable scenarios. We will neglect all rejected options to drop unreasonable augmentation.

Consistency Validation Considering the randomness of LLMs, we repeat each prompt 3 times and only keep consistent responses.

3.2.2 Aspect-Grained Augmentation

The development of LLM provides an opportunity to address control coarseness. We extract unrestricted control attributes for input sentences, extrapolating the label space. For $\mathcal{D}_i$, we regenerate detailed attribute $desc(x, a_i^y)$ for sentence x with original attribute a_i^y. This process is demonstrated in Eq. 3. Taking the sentiment aspect as an instance, aspect-grained augmentation provides a detailed sentiment like *disappointed* instead of *negative*.

$$\text{grained}(\mathcal{D}_i) = \{(x, desc(x, a_i^y)) | x \sim desc(x, a_i^y)\} \tag{3}$$

In practical prompting, we provide sentences and their original control attributes. LLMs are instructed to output detailed descriptions based on original attributes with similar rejected options.

3.2.3 Aspect-Rewrite Augmentation

For concerns in sentence distribution, we rewrite sentences outside current aspect $\tilde{x} \notin \mathcal{D}_i$ with control attribute in A_i, as notated in Eq. 4. The rewritten sentences extrapolate an imbalanced distribution in the original dataset.

$$\begin{aligned}\text{rewrite}(\mathcal{D}_i) = \{(\tilde{x}, y) \,|\, \tilde{x} \sim (a_i^y),\\ 1 \le y \le |A_i|, \tilde{x} \notin \mathcal{D}_i\}\end{aligned} \tag{4}$$

In practice, we select sentences in other aspects and rewrite them with current aspect controls, contrastive ICL examples, and rejected options.

We eliminate instances that deviate from statistical norms (e.g. very short sentences). Additionally, we filter unsuccessful rewriting. In practice, LLMs may copy the input or output abnormal responses. We compare semantic similarity [2] before and after rewriting, then eliminate top 50% and bottom 10% of similar instances.

3.3 Instruction Tuning Dataset Construction

Augmented datasets share a common format with original datasets, and we transform them into IT datasets for training. An instance of an IT dataset consists of instruction I and response R. LLMs should output R with the input of I.

For an instance $(x, y) \in \mathcal{D}_i$, we provide simple task descriptions, target control attribute a_i^y, and generation prefix [3]. We simply use controlled sentence x as R. An instance is in Appendix. B.4.

4 Experiments

4.1 Datasets

Following Gu et al. (2022), we select IMDB (Maas et al., 2011), AGNews (Zhang et al., 2015) and Toxic Comment [4] for sentiment, topic and detoxification aspects as basic datasets. Then we conduct the aforementioned data augmentation. We provide two categories for training. **Vanilla** datasets include all basic datasets. **Augmented** datasets contain vanilla datasets and their corresponding augmented version. We integrate universal IT datasets to keep an identical volume of two categories, statistics are in Appendix. C.1.

4.2 Model Training

We select Qwen-2.5-3B (Yang et al., 2024) as LLM backbone in main experiments, and Qwen-2.5-0.5B, Llama-3.2-3B (Dubey et al., 2024) for supplementary experiments. Hyperparameters and more details are in Appendix. C.2.

4.3 Evaluation

Following Gu et al. (2022); Pascual et al. (2021), we provide control combinations and prefixes for model generation. We calculate the ratio of controlled sentences by classifiers in Gu et al. (2022) as **accuracy**, and the ratio of generations fits all 3 control aspects as **total accuracy**. We additionally repeat each generation 10 times and set the temperature to 0.2 for LLMs to weaken randomness.

4.4 Experiment Results

As shown in Table 1, augmented datasets enhance the performance of MCTG, especially in total combinations and certain aspects. Augmented datasets enhance the total accuracy significantly(20%). Vanilla datasets have a bias on sentiment aspects, and neglect the learning of the other two aspects due to unprocessed aspect correlations and bias. Augmented datasets successfully address these concerns and re-balance three aspects in the generation. Therefore, the total and each aspect's accuracy are enhanced. As for the ablation study, aspect rewrite

[2] We use bge-large-en-v1.5 as semantic embedder and calculate the cosine similarity between two sentences.

[3] Following Gu et al. (2022); Dathathri et al. (2020), we provide certain prefix in training and evaluation.

[4] https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/.

Model	Dataset	Total Accuracy	Sentiment	Topic	Detoxification
Qwen-2.5-3B	Vanilla	22.14	98.86	41.89	51.35
	Augmented(Ours)	47.57	77.75	71.11	82.75
	w/o Cross.	44.03	77.32	61.46	85.39
	w/o Grained.	35.25	84.36	59.89	71.18
	w/o Rewrite.	29.67	93.27	55.61	59.68

Table 1: MCTG performance of Qwen-2.5-3B trained on various datasets combinations.

Model	Dataset	Total Accuracy	Sentiment	Topic	Detoxification
Qwen-2.5-0.5B	Vanilla	18.92	95.93	32.04	24.43
	Augmented(Ours)	34.89	86.21	39.57	49.25
Llama-3.2-3B	Vanilla	25.72	94.11	39.34	59.46
	Augmented(Ours)	44.46	80.46	75.79	69.81

Table 2: MCTG performance on various LLM backbones and sizes

	Augmented(Ours)	Vanilla
MI(A_1,A_2,A_3)	0.280	0.508
MI(A_1,A_2)	0.042	0.173
MI(A_1,A_3)	0.231	0.331
MI(A_2,A_3)	0.016	0.074

Table 3: MI of three aspects for Qwen-2.5-3B. A_1,A_2,A_3 stand for sentiment, topic and detoxification.

	ARC-c	gsm8k	IFEval-P	IFEval-I
Vanilla	28.81	72.48	37.71	50.84
Augmented (Ours)	30.85	74.07	39.74	52.16

Table 4: Accuracy of general LLM benchmarks for models trained on Qwen-2.5-3B. IFEval-P and IFEval-I stand for accuracy of prompt level and instruction level.

is the most influential one for performance, which indicates LLMs are more sensitive to sentence features during instruction tuning. All augmentation methods are demonstrated beneficial to MCTG performance in ablation study. In Appendix. D, we conduct a case study on model generations.

5 Discussion

Aspect Correlations To demonstrate aspect correlations learned by LLMs, we record predicted attribute distribution and their mutual information (MI) (Shannon, 1948; Kreer, 1957). We calculate the MI of all three aspects and each two of them, results are shown in Table 3. Control attributes are combined orthogonally in instructions, so ideal MI items should be 0. Augmented datasets weaken correlations among aspects, but the two datasets still share an identical impact trend for all correlations.

General LLM Capabilities Assessment We experiment with models on general LLMs benchmarks for Qwen-2.5-3B trained on Vanilla and Augmented datasets. Investigated benchmarks consist of **ARC-c** (Commonsense Machine Reading Comprehension), **gsm8k** (Mathematical problems) and **IFEval** (Instruction Following). Results are shown in Table 4, after integrating augmented datasets, LLMs have a slight performance enhancement since augmentation corrects some bias brought by original MCTG datasets and improves the performance of instruction tuning. The result indicates that LLMs do not lose general abilities after integrating augmented datasets.

Experiments on Various Model Backbone We conduct similar experiments on more model backbones, including Llama-3.2-3B and Qwen-2.5-0.5B, results are shown in Table 2. Augmented datasets show the effectiveness of enhancing MCTG performance identically, with a similar aspect of performance balancing phenomena.

6 Conclusion

In this work, we construct a lightweight MCTG solution for LLMs. We analyze concerns in original MCTG datasets and provide an LLM-based data augmentation pipeline for better MCTG instruction-tuning, including generating cross labels, fine-grained label descriptions and rewriting heterogeneous sentences for target aspects. In experiments, training LLM with augmented data exhibits enhanced and balanced performances among aspects.

7 Limitations

In this work, we propose a lightweight solution to activate MCTG ability for LLMs. Our work still leaves some limitations for future discussion as follows:

(1) The data augmentation pipeline relies on advanced LLMs like GPT3.5, which is a compromising option for complex data synthetic tasks. We leave the self-conditioned manner of data augmentation for future work.

(2) The quality control of augmentation relies on a strict and simple filter policy, we expect more explainable filter strategies to enhance data productivity.

(3) Our work focuses on instruction tuning of LLMs for MCTG but leaves other post-training manners for future discussions.

8 Ethical Considerations And Broader Impact Discussion

In this work, the trained model includes a toxic aspect, which may result in the generation of toxic content during evaluation. However, the inclusion of the toxic aspect is solely to evaluate the model's capabilities. We assure we will not require the model to generate toxic content in real-world applications.

For broader impact, our work provides a lightweight solution for fine-grained controlled generation of LLMs without model structure refactoring. From the perspective of instruction tuning datasets, our work may contribute to trustworthy generation for various domain of LLM application.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher 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. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Derek Chen, Celine Lee, Yunan Lu, Domenic Rosati, and Zhou Yu. 2023a. Mixture of soft prompts for controllable data generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14815–14833, Singapore. Association for Computational Linguistics.

Nuo Chen, Yan Wang, Haiyun Jiang, Deng Cai, Yuhan Li, Ziyang Chen, Longyue Wang, and Jia Li. 2023b. Large language models meet harry potter: A dataset for aligning dialogue agents with characters. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8506–8520, Singapore. Association for Computational Linguistics.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and play language models: A simple approach to controlled text generation. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Hanxing Ding, Liang Pang, Zihao Wei, Huawei Shen, Xueqi Cheng, and Tat-Seng Chua. 2023. MacLaSa: Multi-aspect controllable text generation via efficient sampling from compact latent space. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4424–4436, Singapore. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Yuxuan Gu, Xiaocheng Feng, Sicheng Ma, Lingyuan Zhang, Heng Gong, and Bing Qin. 2022. A distributional lens for multi-aspect controllable text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1023–1043, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xuancheng Huang, Zijun Liu, Peng Li, Tao Li, Maosong Sun, and Yang Liu. 2023. An extensible plug-and-play method for multi-aspect controllable text generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15233–15256, Toronto, Canada. Association for Computational Linguistics.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Damjan Kalajdzievski. 2023. A rank stabilization scaling factor for fine-tuning with lora. *CoRR*, abs/2312.03732.

J. Kreer. 1957. A question of terminology. *IRE Transactions on Information Theory*, 3(3):208–208.

Sachin Kumar, Eric Malmi, Aliaksei Severyn, and Yulia Tsvetkov. 2021. Controlled text generation as continuous optimization with multiple constraints. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA. Curran Associates Inc.

Haolong Li, Yu Ma, Yinqi Zhang, Chen Ye, and Jie Chen. 2024. Exploring mathematical extrapolation of large language models with synthetic data. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 936–946, Bangkok, Thailand. Association for Computational Linguistics.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597.

Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and Junxian He. 2024a. What makes good data for alignment? A comprehensive study of automatic data selection in instruction tuning. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Yi Liu, Xiangyu Liu, Xiangrong Zhu, and Wei Hu. 2024b. Multi-aspect controllable text generation with disentangled counterfactual augmentation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9231–9253, Bangkok, Thailand. Association for Computational Linguistics.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Damian Pascual, Beni Egressy, Clara Meister, Ryan Cotterell, and Roger Wattenhofer. 2021. A plug-and-play method for controlled text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3973–3997, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jing Qian, Li Dong, Yelong Shen, Furu Wei, and Weizhu Chen. 2022. Controllable natural language generation with contrastive prefixes. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2912–2924, Dublin, Ireland. Association for Computational Linguistics.

Nazneen Rajani, Lewis Tunstall, Edward Beeching, Nathan Lambert, Alexander M. Rush, and Thomas Wolf. 2023. No robots. `https://huggingface.co/datasets/HuggingFaceH4/no_robots`.

C. E. Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-LLM: A trainable agent for role-playing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13153–13187, Singapore. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Kexin Yang, Dayiheng Liu, Wenqiang Lei, Baosong Yang, Mingfeng Xue, Boxing Chen, and Jun Xie. 2023. Tailor: A soft-prompt-based approach to attribute-based controlled text generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 410–427, Toronto, Canada. Association for Computational Linguistics.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 649–657.

Yusen Zhang, Yang Liu, Ziyi Yang, Yuwei Fang, Yulong Chen, Dragomir Radev, Chenguang Zhu, Michael

Zeng, and Rui Zhang. 2023. MACSum: Controllable summarization with mixed attributes. *Transactions of the Association for Computational Linguistics*, 11:787–803.

A Related Work

Large Language Models Large language models (LLMs), such as LLaMA (Touvron et al., 2023; Dubey et al., 2024) and GPT-4 (Achiam et al., 2023), refer to a series of Transformer-based models undergoing extensive pretraining with massive corpora. By scaling up the data volume and model capacity, LLMs demonstrate remarkable emergent capabilities, such as In-Context Learning (ICL) (Brown, 2020) and Chain-of-Thought (CoT) prompting (Wei et al., 2022), enable them to comprehend human instructions and handle complex tasks with minimal or even no supervision. Despite their exceptional performance, LLMs still produce nonsensical or incongruent information in practical applications (e.g. "hallucination"(Ji et al., 2023)). In this paper, our method leverages the knowledge and generative capabilities of LLMs.

Multi-aspect Controlled Text Generation From the perspective of parameter fusion, Huang et al. (2023) have improved MCTG in prefix tuning(Li and Liang, 2021) by adjusting the positions where prefixes are added, thereby reducing the mutual influence of multiple prefixes. Tailor (Yang et al., 2023) adjusts the multi-attribute prompt mask and re-indexes the position sequence to bridge the gap between the training phase (where each task uses a single-attribute prompt) and the testing phase (where two prompts are connected).

On the other hand, Gu et al. (2022) approaches this issue from the perspective of distribution within semantic space. After obtaining the intersection of attribute distributions, the language model's distribution is biased toward this region. However, the intersection of different attribute distributions may not overlap. To address this, MacLaSa (Ding et al., 2023) estimates a compact latent space to improve control ability and text quality, mitigating interference between different aspects. Liu et al. (2024b) propose MAGIC, which uses counterfactual feature vectors in the latent space to disentangle attributes, alleviating the imbalance in attribute correlation during training.

Regarding the scarcity of training data for MCTG, Zhang et al. (2023) propose MACSUM, a human-annotated dataset containing summaries with mixed control attributes. Chen et al. (2023a) use a strategy of mixing soft prompts to help large models generate training data that aligns with multi-aspect control attributes.

B Data Augmentation Details

B.1 Data Augmentation Prompts

Aspect-Cross Augmentation Fig. 2 shows the prompt of Aspect-Cross Augmentation. Aspects descriptions are colored green; attributes descriptions are colored red; ICL examples of target attributes are colored purple; target sentences for label are colored blue. Bold fonts are written in markdown format like ***Example***.

Aspect-Grained Augmentation Fig. 3 shows the prompt of Aspect-Grained Augmentation. Aspects descriptions are colored green; attributes descriptions are colored red; target sentences for grained augmentation are colored blue.

Aspect-Rewrite Augmentation Fig. 4 shows the prompt of Aspect-Rewrite Augmentation. Aspects descriptions are colored green; attributes descriptions are colored red; ICL examples for rewriting are colored purple; sentences need to be rewritten are colored blue.

B.2 Augmentation Settings

We conduct aspect-cross augmentation for each two of control aspects and aspect-grained augmentation for all of the basic datasets. For aspect-rewrite augmentation, we select each aspect and rewrite sentences of the other aspects for current aspect control [5].

B.3 Rejection Rate Details

Aspect	Target Datasets	Rejection Rate
Sentiment	AGNews	10.4%
	Toxic Comment	9.2%
Topic	IMDB	69.4%
	Toxic Comment	71.2%
Detoxification	AGNews	<0.1%
	IMDB	<0.1%

Table 5: Rejection Rate Details

In Table 5, we report details of rejection that occurred in various aspects and datasets. The aspect "detoxification" has a lower rejection rate,

[5]Detoxification is skipped in rewriting since GPT-3.5 is aligned not to generate harmful expressions.

Aspect Cross Prompts:

Now you should judge sentiment of given sentences.

Here is some examples of "Positive" sentences.
In the year of 1990, the world of Disney TV cartoons was certainly at it's prime. Shows like Chip n Dale Rescue Rangers, DuckTales and Gummi Bears was already popular, and now Disney made another great cartoon……

Here is some examples of "Negative" sentences.
I love watching Jerry as much as the rest of the world, but this poor excuse for a soft-core porno flick is needlessly offensive, lacks anything resembling wit……

Here is the sentence you need judge.
Jose Guillen and Jeff DaVanon homer off Esteban Loaiza, who failed to make it out of the fourth inning Saturday……

Output Format:
You should only output a word, "Positive" stands for positive sentiment, and "Negative" for negative sentiment. If you can't judge, just output "None".
Notice that you should output "Positive" or "Negative" **in best effort**.

Figure 2: The prompt of Aspect-Cross Augmentation

Aspect Grained Prompts:

Now you need to summarize the sentiment in the following sentence with a single word:
Please notice that you should use accurate word to describe. DO NOT use coarse-grained words like "negative".

The following sentence is:
So you think a talking parrot is not your cup of tea huh? …… Don't miss it! It is available on home video.

You only need to output **a single word** to indicate the sentiment of this sentence in best effort.
If given questions are not available to answer, output "None" directly.

Figure 3: The prompt of Aspect-Grained Augmentation

Aspect Rewrite Prompts:

Now you need to rewrite the following sentence into the requirements: Topic: Business.
To help you understand the requirements Topic: Business, here are some instances satisfying the requirement:

Families who are approved will …… through the Angel Tree program. Those applying need to ……

When Aloft Group Inc. chief executive Matt Bowen first saw ……

Here is the original sentence you need to rewrite:
BASEketball is indeed a really funny movie. David Zucker manages to ……

Please notice that:
1. Except for the requirements Topic: Business, you should keep other sentence meaning SAME WITH original sentence in best effort.
2. You should always output a shorter sentence than original one.
3. Only output the rewritten sentence, DO NOT contain other information.

Figure 4: The prompt of Aspect-Rewrite Augmentation

since labels in detoxification are in the range of {toxic, non-toxic}, LLMs can assign one of these to target sentences conveniently. For cross-labeling of the aspect "Sentiment", LLMs have a moderate rejection rate. Rejection occurs when the sentence has an unspecific sentiment tendency. For cross-labeling of the aspect "Topic", the rejection occurs most frequently. Sentences in IMDB and Toxic Comment may have topics other than {Sci/Tech, Sports, World, Business}. We provide "Others." as a rejection word, and find LLMs output them when sentences are not in provided topics.

B.4 Details Of Instruction Tuning Dataset Construction

Fig. 5 shows the final instruction and response pair of an IT dataset instance. Aspects descriptions are colored green; attributes descriptions are colored red; prefixes for generation are colored pink.

C Instruction Tuning Details

C.1 Datasets Statistics

In our instruction tuning process, we conduct three categories of datasets as follows:

Data Augmentation Augmented datasets including aspect-cross augmentation (notated as **Cross.**), aspect-grained augmentation (notated as **Grained.**), and aspect-rewrite augmentation (notated as **Rewrite.**).

Universal Instruction Tuning Datasets (notated as **Univ.**) We exploit a mixture of Deita-10k-v0 [6] (Liu et al., 2024a), Airobos3.2 [7], Capybara [8], no-robots (Rajani et al., 2023) [9] for universal IT datasets. They are all popular instruction-tuning datasets in the community, whose instructions cover a wide range of universal tasks for LLMs.

Original CTG Datasets (notated as **Original**) We exploit the original version of IMDB (Maas et al., 2011), AGNews (Zhang et al., 2015), and Jigsaw Toxic Comment, transforming them into IT format like Sec. 3.3.

We conduct random sampling on these datasets, to keep the dataset volume identical, as demonstrated in Table 6.

[6] https://huggingface.co/datasets/hkust-nlp/deita-10k-v0

[7] https://huggingface.co/datasets/HuggingFaceH4/airoboros-3.2

[8] https://huggingface.co/datasets/LDJnr/Capybara

[9] https://huggingface.co/datasets/HuggingFaceH4/no_robots

C.2 Hyperparameter Settings

Hyperparameter settings for instruction tuning and generation are shown in Table 7. Training loss is only calculated for response tokens. We train models on 3 NVIDIA V100 GPUs for 6 hours in each experiment.

D Case Study

Warning: This section may contain offensive and toxic sentences. Fig. 6 presents a detailed example, where the model is required to generate text with a negative sentiment, a sports title, and without toxic expressions. The sentence generated by Qwen-2.5-3B trained on vanilla datasets meets the sentiment requirement but fails to align with the topic and toxic criteria, and these sentences are relatively verbose. In contrast, the sentence generated by Qwen-2.5-3B trained on augmented datasets meets all requirements and is more concise and elegant. This indicates that the data augmentation method enables the model to generate sentences that better adhere to multiple aspects.

Datasets	
Augmented	28.5k Univ. + 9k Original + 3k Cross. + 3k Grained. + 1.5k Rewrite.
w/o Cross.	31.5k Univ. + 9k Original + 3k Grained. + 1.5k Rewrite.
w/o Grained.	31.5k Univ. + 9k Original + 3k Cross. + 1.5k Rewrite.
w/o Rewrite.	30k Univ. + 9k Original + 3k Cross. + 3k Grained.
Vanilla	36k Univ. + 9k Original

Table 6: Training dataset statistics.

Instructions

Generate a sentence in the following requirements:
Topics: Figure skating.
Your generation must start with "Olympic silver medalist"

Responses

Olympic silver medalist and 2002 world champion Irina Slutskaya presented an almost flawless program to take the lead after the short program at the Cup of Russia on Friday.

Figure 5: An instance of instruction datasets for MCTG.

Inputs:
Generate a sentence in the following requirements: Sentiment: Negative, Topics: Sports, Without Toxic Expression. Your generation must start with "To conclude".

Vanilla MCTG:	**Augmented CTG:**
*To conclude, this is one of the worst movies I have ever seen. The acting was **piece of a shit,** the story was boring and the music was awful. I can't believe this movie got made. I am sure that it was made by some people who don't know what they are doing. I would never recommend this movie to anyone.*	*To conclude, **Sports** are not my **cup of tea**.*
Sentiment: √	**Sentiment:** √
Topics: ×	**Topics:** √
Toxic: ×	**Toxic:** √

Figure 6: A simple case study. Key sentence components demonstrating control attributes are in **bold**.

Hyperparameter	**Value**
Learning Rate	5e-5
Learning Rate Scheduler	Cosine
Warmup Steps	20
Training Batch Size	144
Max Input Length	3072
Max Generated Length	128
Precision of Tensor	Float32
Vocabulary Size	151642
Random Seed	1996
Epochs	2
Optimizer	Adam
LoRA Rank	32
LoRA α	32
LoRA Dropout	0.1
Rank-Stabilized LoRA (Kalajdzievski, 2023)	Enabled
Chat Template	ChatML

Table 7: Hyperparameter Settings