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

Anonymous ACL submission

Abstract

Large language models (LLMs) show remarkable abilities with instruction tuning. However, they fail to achieve ideal tasks when lacking high-quality instruction tuning data on target tasks. Multi-Aspect Controllable Text Generation (MCTG) is a representative task for this dilemma, where aspect datasets are usually biased and correlated. Existing work exploits additional model structures and strategies for solutions, limiting adaptability to LLMs. To activate MCTG ability of LLMs, we propose a lightweight MCTG pipeline based on data augmentation. We analyze bias and correlations in traditional datasets, and address these concerns with augmented control attributes and sentences. Augmented datasets are feasible for instruction tuning. In our experiments, LLMs perform better in MCTG after data augmentation, with a 20% accuracy rise and less aspect correlations.

1 Introduction

Large language models (LLMs) exhibit ideal abilities in various natural language processing tasks (Brown et al., 2020; Kojima et al., 2022; Qin et al., 2023; Wei et al., 2022a; Ganguli et al., 2022). LLMs rely on ideal training datasets for task performance enhancement, especially for instruction tuning (IT) (Bai et al., 2022; Touvron et al., 2023; Chung et al., 2024) dataset.

However, LLMs struggle on certain downstream tasks since the absence of high-quality IT datasets. MCTG task suffers from this dilemma. 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). Recent work addresses corresponding issues through designed models structures (Carlsson et al., 2022; Liu et al., 2024b; Gu et al., 2022; Yang et al., 2023b). Unfortunately, LLMs have enormous model parameters and complex generation process, which is costly to adapt to existing approaches.

In this work, we propose a lightweight MCTG solution for LLMs from the perspective of instruction tuning datasets. We analyze concerns in existing MCTG datasets and address them in a LLM-based data augmentation pipeline. For control attributes in existing datasets, different aspect datasets may possess intersection parts. Attributes are provided in limited label spaces, inaccurate labels fail to recall certain knowledge of models. For sentences in existing datasets, they exhibit bias which is introduced in dataset construction. To address aspect bias and correlation, we conduct data augmentation by prompting advanced LLMs. For control attributes, we provide labels in other aspects for existing sentences and obtain fine-grained accurate attribute descriptions. For sentences, we rewrite heterogeneous sentences for corresponding control attributes. We provide mechanisms to ensure effectiveness, diversity and quality of augmentation. Form of augmented datasets is consistent with original datasets, which can be conveniently transformed into IT datasets. Consequently, data augmentation is beneficial to common LLMs without specific structures.

We validate the effectiveness of our experiments for LLMs up to 3B scale. The result shows that the augmented dataset contributes to performance of MCTG, especially exhibiting a more balanced performance for various aspects. We additionally test mutual information of 3 aspects in generation, result shows that data augmentation diminishes correlations among aspects.

2 Task Formulation

Control Aspects And Attributes For MCTG tasks, controls may contain various n aspects $A = \{A_1, \ldots, A_n\}$. The i-th aspect contains $|A_t|$ exclu-

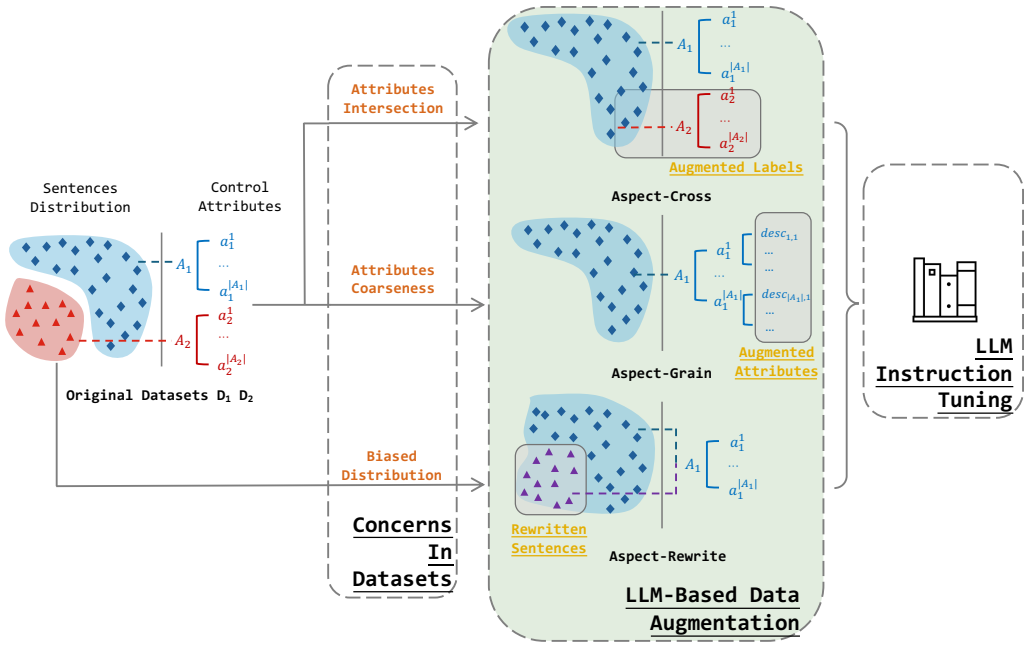

Figure 1: An overview of our lightweight MCTG solution.

sive 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 can be notated as 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.

Generation Task Formulation With the input of control combinations $\boldsymbol{c}$ and generation prompt m, generation of language model LM should follow multiple control aspects, notated in Eq. 1.

$$LM(m|\boldsymbol{c}) \sim (a_1^{c_1}, \ldots, a_n^{c_n}) \quad (1)$$

Dataset Application in MCTG Existing MCTG tasks are trained on a set of single aspect datasets. As for i-th aspect, training set $\mathcal{D}_i$ is composed of sentences x with its corresponding attribute label y in aspect A_i, notated in Eq. 2.

$$\mathcal{D}_i = \{(x, y)|x \sim (a_i^y), 1 \leq y \leq |A_i|\} \quad (2)$$

3 Methodology

As shown in Fig. 1, we first analyze 3 representative concerns in existing MCTG datasets. Then we propose an LLM-based data augmentation pipeline to address the 3 issues correspondingly. Finally, augmentation data is transformed into format of IT data, 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 attributes positive and negative in sentiment aspect. Unfortunately, negative includes toxic attributes like sarcasm for 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 training set struggle to cover all of the world news, which integrates the bias. General and ambiguous control attributes obstruct further application on LLMs.

Concerns in Sentences Distributions Selections of sentences x in training set are not uniform, with **biased distribution**. x limits feature distribution due to dataset construction process. For example, IMDB datasets provide sentences with negative and positive sentiments through crawling movie reviews. But corresponding control attributes may have instances other than movie reviews, limiting generalization of models.

3.2 LLM-Based Data Augmentation Pipeline

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

[1] In practice, we prompt GPT-3.5-Turbo-0125 for data augmentation, detailed prompts can be found in Appendix. B.

3.2.1 Aspect-Cross Augmentation

To address attribute intersection, we exploit LLMs to assign label $\tilde{y}$ in other aspects. We prompt an advanced LLM for dataset generation. Augmented dataset is described in Eq. 3.

$$\text{cross}(\mathcal{D}_i) = \{(x, \tilde{y}) | x \sim (a_j^{\tilde{y}}), \\ 1 \leq \tilde{y} \leq |A_j|, j \neq i\} \quad (3)$$

Contrasting In-Context Learning Demonstrations Though LLMs exhibit ability for zero-shot natural language processing, direct prompting is always 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 [2] for formidable scenarios. We will neglect all rejected options since some cross aspect labeling is not reasonable.

Consistency Validation Considering randomness of LLMs, we repeat each prompt for 3 times and collect all answers. After normalization of case and format, we 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. 4. Taking sentiment aspect as an instance, aspect-grained augmentation provides a detailed sentiment like *disappointed* instead of *negative*.

$$\text{grained}(\mathcal{D}_i) = \{(x, desc(x, a_i^y)) | x \sim desc(x, a_i^y)\} \quad (4)$$

In practical prompting, we provide sentences and original control attributes. LLMs are instructed to output detailed descriptions of given attributes but with 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. 5. The rewritten sentences extrapolate imbalanced distribution in original dataset.

$$\text{rewrite}(\mathcal{D}_i) = \{(\tilde{x}, y) \,|\, \tilde{x} \sim (a_i^y), \\ 1 \leq y \leq |A_i|, \tilde{x} \notin \mathcal{D}_i\} \quad (5)$$

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

Quality control We eliminate instances that evidently deviate from statistical norms (i.e. very short sentences). Additionally, we filter unsuccessful rewriting due to the task difficulty. In practice, LLMs may copy the input or output abnormal responses. We compare semantic similarity [3] before and after rewriting, then eliminate top 50% and bottom 10% of similar instances.

3.3 Instruction Tuning Dataset Construction

Augmented datasets share common format with original datasets, and we transform them into IT dataset for training. An instance of IT dataset consist 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 [4] in I. We simply use controlled sentence x as R. An instance is in Appendix. C.

4 Experiments

4.1 Datasets Selection

Basic Datasets Following Gu et al. (2022), we select IMDB (Maas et al., 2011), AGNews (Zhang et al., 2015) and Jigsaw Toxic Comment [5] for sentiment, topic and detoxification aspects.

Augmented Datasets We conduct aspect-cross augmentation for each two of basic datasets, and aspect-grained augmentation for all of basic datasets. For aspect-rewrite augmentation, we select each aspect and rewrite sentences of the other two aspects for current aspect control [6].

[2] For example, LLM can output *None* in response.

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

[4] In Gu et al. (2022); Dathathri et al. (2020), MCTG should start with certain generation prompts. Consequently, downstream evaluation experiments contain corresponding prefix. We also provide this requirement in instruction, like *starting with* $p(x)$ (standing for first three words of response).

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

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

Baselines	Total Accuracy↑(%)	Sentiment↑(%)	Topic↑(%)	Detoxification↑(%)
Augmented MCTG	**47.57**	77.75	71.11	82.75
w/o Cross	44.03	77.32	61.46	85.39
w/o Grain	35.25	84.36	59.89	71.18
w/o Rewrite	29.67	93.27	55.61	59.68
Vanilla MCTG	22.14	98.86	41.89	51.35

Table 1: Overall result on MCTG, best total accuracy is bold. Accuracy indicates ratio of controlled sentences evaluated by classifiers in Gu et al. (2022) **Total accuracy** indicates ratio of generations fit all 3 control aspects.

Datasets Mixture Distributions of IT datasets influence LLM performance (Lu et al., 2024). We integrate universal IT datasets with MCTG datasets, to avoid overfitting and instruction-ability degradation. For baselines, **Augmented MCTG** consists of universal IT data, vanilla MCTG dataset, and all augmented MCTG data. While **Vanilla MCTG** replaces augmented data with incremental universal IT datasets. Details are in Appendix. D.

4.2 Model Training

We apply Qwen-2.5-3B (Yang et al., 2024) [7] as LLM backbone, and conduct LoRA (Hu et al., 2022) during tuning. Details are in Appendix. E.

4.3 Evaluation

We experiment with the same control combinations, prefix and evaluation models to Gu et al. (2022); Pascual et al. (2021). We additionally repeat each generation 10 times and set temperature to 0.2 for LLMs to weaken randomness.

4.4 Experiment Results

MCTG Performance As shown in Table. 1, augmented MCTG datasets enhance the performance of MCTG, especially in total combinations and certain aspects. Augmented MCTG datasets enhance the total accuracy significantly(20%). Original 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, total and each aspect accuracy are enhanced. As for ablation study, aspect rewrite is the most influential one for performance, which indicates LLMs are more sensitive to sentence features during instruction tuning. Sentences are trained as response so that more uniform and diverse responses are beneficial for LLMs in MCTG. In Appendix. F, we conduct a case study on for model generations.

[7] https://huggingface.co/Qwen/Qwen2.5-3B

	Augmented MCTG	Vanilla MCTG
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 generations. A_1,A_2,A_3 stand for sentiment, topic and detoxification aspects.

Aspect Correlations To demonstrate aspect correlations learned by LLMs, we record predicted attributes distribution and their mutual information (MI) (Shannon, 1948; Kreer, 1957). We calculate MI of all three aspects and each two of them, results are shown in Table. 2. Control attributes are combined orthogonally in instructions, so ideal MI items should be 0. Augmented MCTG weakens correlations among aspects, but two baselines still share an identical trend of correlations, necessitating further processing with correlations.

5 Conclusion

In this work, we construct a lightweight MCTG solution for LLMs. Starting from a perspective of datasets, we analyze concerns in traditional MCTG datasets including attributes intersection, attributes coarseness, and biased distribution. Then we provide a LLM-based data augmentation pipeline for better instruction tuning datasets, including generating cross labels, generating fine-grained label descriptions and rewriting heterogeneous sentences for target aspects. In experiments, training LLM with augmented data exhibits enhanced and balanced performances on overall aspects. The result indicates our solution is effective for LLMs. Result of MI shows that augmented dataset weakens correlations between certain aspects.

6 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 (Chan et al., 2024; Yang et al., 2023a). But a self-conditioned augmentation pipeline is more feasible for lightweight solutions, where data augmenter LLMs and trained LLMs remain the same like self-distill (Dubey et al., 2024; Xu et al., 2023).

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

(3) Our work focuses on instruction tuning of LLMs for MCTG, but leaves other post-training processes like RLHF (Ouyang et al., 2022) and DPO (Rafailov et al., 2023) for future discussions.

7 Ethical Considerations

In this work, the trained MCTG 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 for the purpose of evaluating the model's capabilities. We assure that we will not require the model to generate toxic content in real-world applications.

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*.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *Preprint*, arXiv:2204.05862.

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.

Fredrik Carlsson, Joey Öhman, Fangyu Liu, Severine Verlinden, Joakim Nivre, and Magnus Sahlgren. 2022. Fine-grained controllable text generation using non-residual prompting. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6837–6857, Dublin, Ireland. Association for Computational Linguistics.

Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2024. Scaling synthetic data creation with 1,000,000,000 personas. *ArXiv*, abs/2406.20094.

Derek Chen, Celine Lee, Yunan Lu, Domenic Rosati, and Zhou Yu. 2023. 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.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

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*.

Deep Ganguli, Danny Hernandez, Liane Lovitt, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova Dassarma, Dawn Drain, Nelson Elhage, Sheer El Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Scott Johnston, Andy Jones, Nicholas Joseph, Jackson Kernian, Shauna Kravec, Ben Mann, Neel Nanda, Kamal Ndousse, Catherine Olsson, Daniela Amodei, Tom Brown, Jared Kaplan, Sam McCandlish, Christopher Olah, Dario Amodei, and Jack Clark. 2022. Predictability and surprise in large generative models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 1747–1764, New York, NY, USA. Association for Computing Machinery.

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.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

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.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

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

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.

Keming Lu, Hongyi Yuan, Zheng Yuan, Runji Lin, Junyang Lin, Chuanqi Tan, Chang Zhou, and Jingren Zhou. 2024. #instag: Instruction tagging for analyzing supervised fine-tuning of large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

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.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

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.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is

ChatGPT a general-purpose natural language processing task solver? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1339–1384, Singapore. Association for Computational Linguistics.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *ArXiv*, abs/2305.18290.

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.

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, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022a. Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.

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

Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. 2023. Baize: An open-source chat model with parameter-efficient tuning on self-chat data. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6268–6278, Singapore. Association for Computational Linguistics.

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*.

Dongjie Yang, Ruifeng Yuan, Yuantao Fan, Yifei Yang, Zili Wang, Shusen Wang, and Hai Zhao. 2023a. RefGPT: Dialogue generation of GPT, by GPT, and for GPT. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2511–2535, Singapore. Association for Computational Linguistics.

Kexin Yang, Dayiheng Liu, Wenqiang Lei, Baosong Yang, Mingfeng Xue, Boxing Chen, and Jun Xie. 2023b. 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., 2022b), 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 MACTG in prefix tuning(Li and Liang, 2021) by adjusting the positions

Baselines	Datasets
Augmented MCTG	28.5k **Univ.** + 9k **Vanilla** + 3k **Cross.** + 3k **Grained.** + 1.5k **Rewrite.**
w/o *Cross.*	31.5k **Univ.** + 9k **Vanilla** + 3k **Grained.** + 1.5k **Rewrite.**
w/o *Grained.*	31.5k **Univ.** + 9k **Vanilla** + 3k **Cross.** + 1.5k **Rewrite.**
w/o *Rewrite.*	30k **Univ.** + 9k **Vanilla** + 3k **Cross.** + 3k **Grained.**
Vanilla MCTG	36k **Univ.** + 19k **Vanilla**

Table 4: Training dataset statistics of all baselines in experiments.

where prefixes are added, thereby reducing the mutual influence of multiple prefixes. Tailor (Yang et al., 2023b) adjust the multi-attribute prompt mask and re-index 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. (2023) use a strategy of mixing soft prompts to help large models generate training data that aligns with multi-aspect control attributes.

B 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.

C 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.

D Datasets Statistics

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

Augmented Datasets 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 [8] (Liu et al., 2024a), Airobos3.2 [9], Capybara [10], no-robots (Rajani et al., 2023) [11] for universal IT datasets. They are all popular instruction-tuning datasets in community, whose instructions cover a wide range of universal tasks for LLMs.

Vanilla CTG Datasets (notated as **Vanilla**) We exploit 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.

[8] https://huggingface.co/datasets/hkust-nlp/deita-10k-v0
[9] https://huggingface.co/datasets/HuggingFaceH4/airoboros-3.2
[10] https://huggingface.co/datasets/LDJnr/Capybara
[11] https://huggingface.co/datasets/HuggingFaceH4/no_robots

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 5: Hyperparameter Settings

We conduct random sample on these datasets, to keep dataset volume of each baseline identical. Final statistics of all baselines are demonstrated in Table. 4.

E Hyperparameter Settings

Hyperparameter settings for instruction tuning and generation are shown in Table. 5. Training loss is only calculated for response tokens. We train models on 3 NVIDIA V100 GPUs for 6 hours in each experiment.

F Case Study

Warning: This sections may contains offensive and toxic sentences. Fig. 6 presents a detailed example, where model is required to generating text with a negative sentiment, a title of sports and without toxic expressions. The sentences generated by Vanilla MCTG meet the sentiment requirement but fail to align with the topic and toxic criteria, and these sentences are relatively verbose. In contrast, the sentences generated by Augmented CTG meet all requirements and are more concise and elegant. This indicates that the Augmented CTG method enables the model to generate sentences that better adhere to multiple aspects.

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

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: *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.* **Sentiment:** √ **Topics:** × **Toxic:** ×	**Augmented CTG:** *To conclude, **Sports** are not my **cup of tea**.* **Sentiment:** √ **Topics:** √ **Toxic:** √

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