Routing LoRA Adapters for Controllable Text Generation

Anonymous ACL submission

Abstract

Large language models are routinely used as writing assistants for scientific text, where outputs must be both topically accurate and stylistically appropriate. We propose a compositional LoRA framework that routes style adapters to the first n transformer layers and topic adapters to the remaining layers, composing them at inference time without retraining. We release a multi-style scientific writing benchmark with five stylistic directions and an evaluation protocol combining a listwise ranker and a pointwise scorer, calibrated against human pairwise judgments. On Llama-2-7B and Llama-3-8B, our method outperforms prompt-only control, single-adapter baselines, and naive merging, ranking first or second among methods that do not require a style-specific downstream dataset. Style adapters are trained once and swapped against new topic adapters at inference time, giving a modular alternative to dataset-specific fine-tuning.

1 Introduction

Modern large language models generate fluent text across many domains (Minaee et al., 2024; Annepaka and Pakray, 2025), but in applications such as scientific writing assistants, topical accuracy alone is not enough. The output also has to match a target tone, level of formality, and communicative style. Fine-grained control over content and style at the same time remains hard (Shen et al., 2017; Zhang et al., 2023; Minaee et al., 2024): prompt-based control often fails to enforce both stylistic and topical constraints when several attributes are combined (Alhafni et al., 2024; Yang et al., 2025), and full fine-tuning is expensive and prone to catastrophic forgetting (Luo et al., 2023; Li et al., 2024).

Parameter-Efficient Fine-Tuning (PEFT), in particular Low-Rank Adaptation (LoRA), adapts large language models without updating the base parameters (Hu et al., 2022; Mao et al., 2024). LoRA is naturally modular: separate adapters can be trained for attributes such as style, sentiment, or domain, and selected at inference time (Mao et al., 2024). Composing independently trained adapters, however, remains hard (Roy et al., 2025; Adsul et al., 2026). Naive merging or simultaneous activation degrades fluency and causes interference between the encoded attributes (Yadav et al., 2023; Buehler and Buehler, 2024).

Previous work has explored adapter fusion and multi-adapter composition without modifying the underlying pretrained model or the adapters themselves (Pfeiffer et al., 2021; Karimi Mahabadi et al., 2021). Recent LoRA-based methods further improve training-free composition (Yang et al., 2024; Roy et al., 2025; Zheng et al., 2025; Peng et al., 2025; Zhang and Xiong, 2025), but still struggle to adapt to multiple control dimensions simultaneously. K-LoRA (Ouyang et al., 2025) is closest to our setting, as it performs a training-free fusion of LoRA subject- and style adapters. However, it was developed for diffusion models, whereas in autoregressive language models adapter updates directly affect next-token distributions across layers and timesteps, making such fusion more prone to interference and semantic drift. In Section 3.4 we show that K-LoRA underperforms on topic–style composition tasks.

In this work we introduce a compositional LoRA framework for controllable text generation that separates style and topic control across complementary layer groups. Style and topic adapters are trained once and combined at inference time, without reconstructing datasets or running extra fine-tuning for each style–topic configuration. We evaluate the approach on scientific text generation, where both terminological accuracy and stylistic consistency matter. Our contributions are: (1) a compositional framework for combining LoRA adapters in language models; (2) style and topic LoRA modules for academic text synthesis; (3) a

layer-partitioned composition strategy that reuses the same adapters across configurations without rebuilding the dataset; (4) an evaluation protocol that ranks generated texts by stylistic and topical relevance; and (5) a reusable dataset of scientific texts written in multiple styles.

2 Method

We do not propose a new fine-tuning recipe. We propose a way to combine LoRA adapters such that style and topic can be swapped at inference time without retraining. Each adapter can be trained with any standard objective (SFT, DPO (Rafailov et al., 2023), etc.), independently of the composition step. In this work we use SFT for the style adapters, the simplest option that does not require preference pairs for every style and topic combination.

Construction of datasets. We build a stylization dataset by sampling short, topic-neutral fragments from a pretraining corpus and prompting an LLM to rewrite each fragment in a target style. We use five styles: *sad/melancholic*, *cheerful/lighthearted*, *thoughtful/reflective*, *promotional/persuasive*, and *aggressive*. Sourcing from topic-neutral text keeps the style adapters topic-agnostic and transferable to arbitrary downstream domains. The full pipeline is available at `https://anonymous.4open.science/r/style-lora/`.

Selection of layers to target style and topic. In transformer models, initial layers tend to encode general grammar and syntax, while deeper layers encode higher-level semantics and task-relevant features (Langedijk et al., 2024; He et al., 2024; Alshomary et al., 2025; Nadipalli, 2025; Veprikov et al., 2025). We exploit this split: the first n layers carry the style adapters, and the remaining layers ($i > n$) carry the topic adapters. The two adapter groups therefore act on disjoint parameter blocks and do not interfere.

Training style and topic LoRA adapters. Style adapters (B_s, A_s) on the first n layers and topic adapters (B_t, A_t) on the remaining layers are trained independently with causal LM, the first on a stylistic dataset and the second on a downstream dataset. A small n already suffices (we use n=7; see Appendix C).

Combining style and topic adapters. At inference time the weights are $W_0^{(i)} + (\alpha_s/r_s)B_s^{(i)}A_s^{(i)}$ for $i \leq n$ and $W_0^{(i)} + (\alpha_t/r_t)B_t^{(i)}A_t^{(i)}$ for $i > n$. Style can therefore be replaced by swapping the LoRA adapters on the first n layers without touching the topic adapters.

Style	# examples	avg. stylized words
aggressive	2,400	60
cheerful/lighthearted	2,390	76
thoughtful/reflective	2,400	77
promotional/persuasive	2,381	81
sad/melancholic	2,400	73

Table 1: Per-style statistics of the released stylized dataset. Source passages have a near-uniform average length of 58 words. Each split mixes Wikipedia, CNN/DailyMail, and arXiv-astrophysics, with an extra 200 astrophysics-domain examples per style.

3 Experiments

3.1 Stylization Dataset Generation Pipeline

We construct a training dataset for stylized text generation using an automated pipeline based on large language model rewriting.

Data sources. The base corpus consists of short text fragments (2–3 sentences, 120–500 characters) extracted from three sources: general-domain texts (Wikipedia (Wikimedia Foundation, 2026), CNN/DailyMail (Hermann et al., 2024)) and abstracts from arXiv (Mishra, 2025), which are standard for LLM pretraining (Gao et al., 2020). The released dataset spans five styles with roughly 2,400 stylized examples each. Table 1 reports per-style sizes and the average stylization expansion in word count. A full breakdown by subset and source is in Table 5 in Appendix.

Stylization model. We use Qwen2.5-32B-Instruct (Team, 2024) (goes under Apache License Version 2.0) as the rewriting model. For each source text and target style, the model is instructed to generate a stylized version that preserves the semantic content while adopting the specified tone (the prompts are listed in D.1).

3.2 Baselines

We compare against several baselines. The simplest does not involve any fine-tuning: the model is prompted with an instruction that specifies both the target style and the target topic at the start of the input. We call this approach *Instructed style and topic*. Its quality depends strongly on the exact prompt (see Appendix C). We further consider adapter-based variants in which LoRA modules are applied to all layers to specialize the model for either style or topic, while the complementary

	aggressive							promotional/persuasive							cheerful/lighthearted							reflective/thoughtful							sad/melancholic						
	OURS	Ideal score	K-LoRA	Half sum	Instructed style and topic	Instructed style	Instructed topic	OURS	Ideal score	K-LoRA	Half sum	Instructed style and topic	Instructed style	Instructed topic	OURS	Ideal score	K-LoRA	Half sum	Instructed style and topic	Instructed style	Instructed topic	OURS	Ideal score	K-LoRA	Half sum	Instructed style and topic	Instructed style	Instructed topic	OURS	Ideal score	K-LoRA	Half sum	Instructed style and topic	Instructed style	Instructed topic
Norm. rank	**0.67**	0.93	0.49	0.32	0.28	0.21	0.60	**0.66**	0.85	0.41	0.38	0.28	0.59	0.33	**0.63**	0.88	0.36	0.36	0.26	0.38	**0.63**	0.54	0.83	0.44	0.49	0.36	**0.58**	0.24	**0.70**	0.92	0.49	0.33	0.29	0.24	0.54
Top-2 rate	**0.42**	0.90	0.23	0.01	0.03	0.00	0.41	**0.41**	0.77	0.17	0.07	0.06	0.36	0.16	0.35	0.84	0.10	0.07	0.06	0.06	**0.52**	**0.32**	0.77	0.23	0.17	0.06	**0.32**	0.13	**0.51**	0.91	0.20	0.01	0.04	0.01	0.30
Top-3 rate	**0.75**	0.99	0.32	0.17	0.13	0.03	0.61	**0.71**	0.86	0.30	0.20	0.10	0.57	0.26	**0.71**	0.91	0.28	0.13	0.13	0.26	0.58	0.51	0.81	0.32	0.42	0.25	**0.54**	0.16	**0.74**	0.99	0.32	0.17	0.19	0.04	0.55
Topic Hit Rate	**0.90**	0.93	0.25	0.36	0.36	0.74	0.67	0.64	0.77	0.29	0.27	0.29	**0.69**	0.40	**0.74**	0.88	0.20	0.41	0.32	0.70	0.38	**0.71**	0.84	0.32	0.38	0.41	**0.71**	0.22	**0.90**	0.93	0.25	0.36	0.36	0.74	0.67
Style Quality	0.69	0.80	0.63	0.36	0.37	0.36	**0.71**	0.48	0.73	**0.59**	0.50	0.33	0.42	**0.59**	0.44	0.75	0.60	0.44	0.31	0.29	**0.72**	0.49	0.68	**0.53**	0.44	0.41	0.45	0.49	**0.72**	0.80	0.63	0.36	0.37	0.36	0.71
Topic Quality	**0.53**	0.60	0.16	0.21	0.15	0.45	0.35	0.34	0.42	0.16	0.17	0.13	**0.40**	0.17	0.40	0.45	0.12	0.23	0.14	**0.42**	0.14	0.42	0.54	0.17	0.22	0.19	**0.46**	0.11	**0.55**	0.60	0.16	0.21	0.15	0.45	0.35
Coherence Quality	**0.63**	0.75	0.34	0.19	0.19	0.35	0.57	0.38	0.60	0.32	0.25	0.22	**0.40**	0.35	**0.38**	0.63	0.23	0.23	0.20	0.28	0.31	0.44	0.66	0.24	0.23	0.25	**0.46**	0.20	**0.65**	0.75	0.34	0.19	0.19	0.35	0.58
Aggregate Score	**0.60**	0.69	0.17	0.17	0.16	0.34	0.44	0.34	0.50	0.15	0.15	0.12	**0.35**	0.22	**0.37**	0.58	0.12	0.19	0.12	0.29	0.23	**0.39**	0.58	0.20	0.19	0.19	0.38	0.13	**0.60**	0.69	0.17	0.17	0.16	0.34	0.44

Table 2: Metrics across styles for Llama-2-7B (higher is better). Best result per row within each style is bolded and second-best is underlined. We exclude *Ideal score*, as it uses a dedicated downstream dataset and consistently outperforms all other methods.

attribute is specified via instruction at inference time. When the model is fine-tuned for style and prompted for topic, we denote the method as *Instructed topic*; and, when it is fine-tuned for topic and prompted for style, we refer to it as *Instructed style*. We also consider combining LoRA adapters using their half sum with $A = \frac{1}{2}(A_s + A_t)$ and $B = \frac{1}{2}(B_s + B_t)$ for all layers; we refer to this approach as *Half sum*.

We implement a variant of the K-LoRA approach (Ouyang et al., 2025), in which the style or topic adapters B_x and A_x (with $x \in \{s, t\}$) are selected for each module across all layers based on a scoring criterion $S_x = \sum \left(\text{Top-}K(|\Delta W_x|)\right)$, i.e., the sum of the K largest absolute values in ΔW_x. If $S_t \geq S_s$, the topic adapters are applied; otherwise, the style adapters are selected.

Lastly, *Ideal score* uses the pipeline of Section 2 on a domain-specific stylistic dataset where the target style and the target topic are jointly represented, and trains LoRA adapters on it. It serves as an approximate upper bound for fully supervised style–domain adaptation, not as a directly comparable modular method, since our setting trains style and topic adapters separately and composes them at inference time.

3.3 Evaluation Pipeline

To anchor the evaluation we run a human study. We ask 5 assessors to pick which of two outputs better matches the target style, comparing our method against each baseline and against the *Ideal Score* (Fig. 1). Our method outperforms the baselines and stays close to the *Ideal Score*. We then run the same pairwise questions through Qwen-2.5-32B-Instruct (Team, 2024) (5 trials per pair). Its scores agree with the human ones at $R^2 = 0.766$, with human ratings as reference, therefore we rely on the LLM judge for the rest of the evaluation.

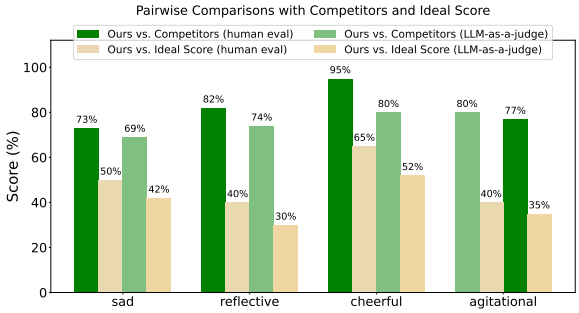

Figure 1: Human and LLM-based pairwise evaluations (our methods vs. competitors and our method vs. Ideal Score). Human eval results are averaged over 5 annotators who measured which approach provides better stylistic conformity, while LLM (Qwen-2.5-32B-Instruct) answered the same question 5 times too.

The automated pipeline reports two families of metrics with Qwen2.5-32B-Instruct as the judge: a *listwise ranker* and a *pointwise scorer*. The ranker gives **Norm. Rank** ($\frac{1}{N}\sum_{i=1}^{N}\frac{K-\rho_i}{K-1} \in [0,1]$) and **Top M rate** (fraction of texts ranked among the top M). The scorer gives **Topic Hit Rate** (whether the text addresses the target topic), **Style Quality** (how strongly the target style is expressed), **Topic Quality** (substantive engagement with the topic), and **Coherence Quality** (naturalness of style and topic integration). The *Aggregate Score* is the mean of the three quality dimensions, gated by topic relevance: $r \cdot \frac{q_{\text{style}}+q_{\text{topic}}+q_{\text{coh}}}{3}$. We release the full evaluation protocol, including the listwise and pointwise prompts, the judge calibration rubric, the ag-

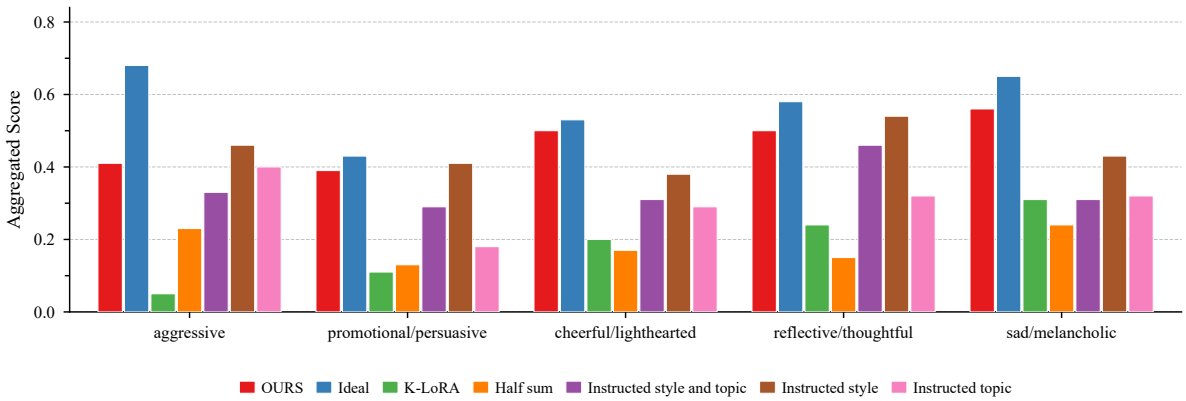

Figure 2: Comparison of aggregated scores across methods and setups.

gregation script, and the human study templates, at `https://anonymous.4open.science/r/style-lora/` and Appendix D.2.

3.4 Experiment Results

We use Llama-2-7B (Touvron et al., 2023) and Llama-3-8B (Grattafiori et al., 2024) (both under Community License) as backbone models. The downstream data set is built from short arXiv abstract fragments (Mishra, 2025) with astrophysics as the target topic, which gives a controlled testbed for the composition of the style and topic adapter under a fixed backbone. It took approximately 100 GPU-hours on a single NVIDIA A100 GPU.

As shown in Table 2, the *Ideal score* is the highest, as expected, since it is trained on a dedicated style–domain dataset. Among the remaining methods, our approach ranks first or second in most configurations and leads the listwise metrics for four styles, indicating that the judge consistently prefers our outputs over competing baselines.

Prompt-only control (*Instructed style and topic*) is the weakest setup: it gets the lowest Normalized Rank and Aggregate Score in nearly every style, which shows that natural-language instructions alone are not enough for joint stylistic and topical control. The adapter-based variants *Instructed style* and *Instructed topic* are uneven. *Instructed topic* reaches competitive Style Quality but loses topic relevance, while *Instructed style* preserves topic but underperforms on style. Naive parameter averaging (*Half sum*) and K-LoRA also fall short, often producing incoherent outputs with low Topic Hit Rate and Coherence Quality. Partitioning style and topic adapters across disjoint layer groups therefore gives a more effective fusion than prompting or naive merging.

Where the gains come from. The Aggregate Score $r \cdot \frac{q_{\text{style}}+q_{\text{topic}}+q_{\text{coh}}}{3}$ is gated by the Topic Hit Rate r: a candidate that fails on topic relevance scores zero regardless of stylistic quality. Our method keeps the topic hit rate around 0.8 for all five styles in Table 2 (0.9 in three of them), near the *Ideal score*, while baselines that put all LoRA capacity to a single attribute drop notably on the other. The layer-partitioned design protects topic coverage with a dedicated adapter group while leaving the first n layers free for style, which prompt-based and weight-merging baselines cannot reproduce.

In Table 2 the listwise and pointwise metrics track each other across methods. For Llama-3-8B (Grattafiori et al., 2024) we therefore report only the *Aggregate Score* (Fig. 2), the cleanest single proxy for end-to-end quality. Prompt-only control (*Instructed style*) does well on *agitational* and *reflective*: the former is hard to combine with astrophysics, where weak stylization can still score by staying on topic, while the latter is close to scientific writing itself, leaving little stylistic differentiation and a narrow gap to the *Ideal Score*.

Conclusion

We propose a compositional framework for controllable text generation that routes style and topic adapters to disjoint layer groups. On scientific text generation it outperforms prompt-based control, single-adapter baselines, and naive merging, ranking first or second among all methods that do not require a style-specific downstream dataset. The fully supervised *Ideal Score* baseline scores higher, but must be retrained from a rewritten dataset for every new style and topic pair, while our composition reuses the same adapters at inference time.

Limitations

This work is limited to scientific writing and a small set of styles and domains. Although this setting is appropriate for studying terminological precision and stylistic control, the results may not generalize to other genres, such as customer support, news, legal, or medical text, where stylistic norms, safety constraints, and topic–style interactions differ. Our style adapters are also trained only with supervised fine-tuning; future work could explore preference-based refinement, such as DPO (Rafailov et al., 2023), to improve stylistic fidelity and coherence while preserving adapter modularity.

Controllable style generation also introduces risks. Persuasive or high-confidence styles, such as *aggressive* or *promotional / persuasive*, can make incorrect, biased, or unsupported content appear more authoritative. Because our method enables modular reuse and recombination of style and topic adapters, unsafe stylistic behavior could transfer across domains if deployed without safeguards. In addition, our evaluation relies on an LLM judge, which may inherit model biases and miss subtle factual errors. We therefore recommend domain-specific validation, grounding, or citation checks, safety filters, calibration, and human oversight for high-impact applications.

References

Riya Adsul, Balachandra Devarangadi Sunil, Isha Nalawade, and Sudharshan Govindan. 2026. Task-aware lora adapter composition via similarity retrieval in vector databases. *arXiv preprint arXiv:2602.21222*.

Bashar Alhafni, Vivek Kulkarni, Dhruv Kumar, and Vipul Raheja. 2024. Personalized text generation with fine-grained linguistic control. In *Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)*, pages 88–101. Association for Computational Linguistics.

Milad Alshomary, Nikhil Reddy Varimalla, Vishal Anand, Smaranda Muresan, and Kathleen McKeown. 2025. Layered insights: Generalizable analysis of authorial style by leveraging all transformer layers. *arXiv preprint arXiv:2503.00958*.

Yadagiri Annepaka and Partha Pakray. 2025. Large language models: A survey of their development, capabilities, and applications. *Knowledge and Information Systems*.

Eric L. Buehler and Markus J. Buehler. 2024. X-LoRA: Mixture of low-rank adapter experts, a flexible framework for large language models with applications in protein mechanics and molecular design. *APL Machine Learning*, 2(2).

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 *International Conference on Learning Representations*.

Jessica Ficler and Yoav Goldberg. 2017. Controlling linguistic style aspects in neural language generation. In *Proceedings of the Workshop on Stylistic Variation*.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The Pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Linyang He, Peili Chen, Ercong Nie, Yuanning Li, and Jonathan R. Brennan. 2024. Decoding probing: Revealing internal linguistic structures in neural language models using minimal pairs. *arXiv preprint arXiv:2403.17299*.

Karl Moritz Hermann, Tomáš Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2024. Cnn/dailymail dataset. Dataset available via TIB LDM service.

Amir Hertz, Andrey Voynov, Shlomi Fruchter, and Daniel Cohen-Or. 2023. Style aligned image generation via shared attention. *arXiv preprint arXiv:2312.02133*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, pages 2790–2799.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric Xing. 2017. Toward controlled generation of text. In *International Conference on Machine Learning (ICML)*.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2023. Editing models with task arithmetic. In *International Conference on Learning Representations*.

Rabeeh Karimi Mahabadi, James Henderson, and Sebastian Ruder. 2021. Compacter: Efficient low-rank hypercomplex adapter layers. *Advances in neural information processing systems*, 34:1022–1035.

Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. 2021. GeDi: Generative discriminator guided sequence generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*.

Anna Langedijk, Hosein Mohebbi, Gabriele Sarti, Willem Zuidema, and Jaap Jumelet. 2024. Decoderlens: Layerwise interpretation of encoder-decoder transformers. In *Findings of ACL*.

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.

Hongyu Li, Liang Ding, Meng Fang, and Dacheng Tao. 2024. Revisiting catastrophic forgetting in large language model tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4297–4308. Association for Computational Linguistics.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. 2021. DExperts: Decoding-time controlled text generation with experts and anti-experts. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pages 6691–6706.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. 2022a. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. In *Advances in Neural Information Processing Systems*.

Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B. Tenenbaum. 2022b. Compositional visual generation with composable diffusion models. In *European Conference on Computer Vision (ECCV)*.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. 2023. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *arXiv preprint arXiv:2308.08747*.

Yuren Mao, Yuhang Ge, Yijiang Fan, Wenyi Xu, Yu Mi, and Zhonghao Hu. 2024. A survey on lora of large language models. *arXiv preprint arXiv:2407.11046*.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large language models: A survey. *arXiv preprint arXiv:2402.06196*.

Sumit Mishra. 2025. arxiv scientific research paper dataset. Mendeley Data. Includes titles, abstracts, categories, and metadata from arXiv.

Suneel Nadipalli. 2025. Layer-wise evolution of representations in fine-tuned transformers: Insights from sparse AutoEncoders. *arXiv preprint arXiv:2502.16722*.

Ziheng Ouyang, Zhen Li, and Qibin Hou. 2025. K-lora: Unlocking training-free fusion of any subject and style loras. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. Accepted at CVPR 2025.

Yuqi Peng, Lingtao Zheng, Yufeng Yang, Yi Huang, Mingfu Yan, Jianzhuang Liu, and Shifeng Chen. 2025. Tara: Token-aware lora for composable personalization in diffusion models. *arXiv preprint arXiv:2508.08812*.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. Adapterfusion: Non-destructive task composition for transfer learning. In *EACL*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Aniket Roy, Maitreya Suin, Ketul Shah, and Rama Chellappa. 2025. Multlfg: Training-free multi-lora composition using frequency-domain guidance. *arXiv preprint*.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Kihyuk Sohn, Nataniel Ruiz, Kimin Lee, Daniel Castro Chin, Irina Blok, Huiwen Chang, Jarred Barber, Lu Jiang, Glenn Entis, Yuanzhen Li, Yuan Hao, Irfan Essa, Michael Rubinstein, and Dilip Krishnan. 2023. StyleDrop: Text-to-image generation in any style. In *Advances in Neural Information Processing Systems*.

Qwen Team. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Andrey Veprikov, Vladimir Solodkin, Alexander Zyl, Andrey Savchenko, and Aleksandr Beznosikov. 2025. Weightlora: Keep only necessary adapters. *arXiv preprint arXiv:2506.02724*.

Wikimedia Foundation. 2026. Wikipedia. https://dumps.wikimedia.org/. Accessed: 2026-02-12. Wikipedia database dump used as a pretraining corpus.

Prateek Yadav, Derek Tam, and 1 others. 2023. Ties-merging: Resolving interference when merging models. *arXiv preprint arXiv:2306.01708*.

Kevin Yang and Dan Klein. 2021. FUDGE: Controlled text generation with future discriminators. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3511–3535.

Yang Yang, Wen Wang, Liang Peng, Chaotian Song, Yao Chen, Hengjia Li, Xiaolong Yang, Qinglin Lu, Deng Cai, Boxi Wu, and Wei Liu. 2024. Loracomposer: Leveraging low-rank adaptation for multi-concept customization in training-free diffusion models. *arXiv preprint*.

Zhe Yang, Yi Huang, Yaqin Chen, Xiaoting Wu, Junlan Feng, and Chao Deng. 2025. Palette of language models: A solver for controlled text generation. *arXiv preprint arXiv:2503.11182*.

Elad Ben Zaken, Shauli Ravfogel, and Yoav Goldberg. 2022. BitFit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 1–9.

Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2023. A survey of controllable text generation using transformer-based pre-trained language models. *ACM Computing Surveys*, 56(3):1–37.

Jia-Chen Zhang and Yu-Jie Xiong. 2025. Subject or style: Adaptive and training-free mixture of loras. *arXiv preprint*.

Guoqing Zhao, Qi Zhang, Shaopeng Zhai, Dazhong Shen, Tianyi Zhang, Yu Qiao, and Tong Xu. 2025. I-lora: Iterative merging of routing-tuned low-rank adapters for multi-task learning. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE.

Peng Zheng, Ye Wang, Rui Ma, and Zuxuan Wu. 2025. FreeLoRA: Enabling training-free LoRA fusion for autoregressive multi-subject personalization. *arXiv preprint arXiv:2507.01792*.

A Related Work

Controllable text generation. Controllable text generation steers outputs toward target attributes such as topic, sentiment, formality, persona, or style. Early neural work relied on explicit control variables and disentangled representations (Ficler and Goldberg, 2017; Hu et al., 2017; Shen et al., 2017). With pretrained LLMs, control is typically done through prompting, decoding-time guidance, or lightweight adaptation. Plug-and-Play LMs perturb hidden activations using attribute classifiers (Dathathri et al., 2020), and GeDi (Krause et al., 2021), FUDGE (Yang and Klein, 2021), and DExperts (Liu et al., 2021) steer decoding through discriminators or expert mixtures. These methods are flexible at inference time but depend on auxiliary models and add decoding cost. Recent work targets fine-grained linguistic control and instruction-based controllability in LLMs (Alhafni et al., 2024; Zhang et al., 2023; Yang et al., 2025). We instead use modular adapter-based control: style and topic are represented by separately trained LoRA modules and composed at inference time.

Parameter-efficient adaptation of language models. Parameter-efficient fine-tuning (PEFT) reduces the cost of adapting LLMs by training a small subset of additional or selected parameters. Adapter modules insert compact trainable layers into frozen Transformers (Houlsby et al., 2019). Prefix tuning and prompt tuning instead optimize continuous task-specific prompts (Li and Liang, 2021; Lester et al., 2021), and sparse or activation-based methods such as BitFit and IA^3 update only small subsets of existing parameters (Zaken et al., 2022; Liu et al., 2022a). LoRA injects low-rank updates into weight matrices while keeping the backbone frozen (Hu et al., 2022; Mao et al., 2024), which lends itself naturally to modular control: adapters trained for different attributes can be swapped at inference time. Independent LoRA adapters, however, do not compose cleanly out of the box, especially across different control dimensions such as style and topic.

Adapter composition and model merging. Several methods combine task-specific modules or fine-tuned models. AdapterFusion composes previously trained adapters without modifying the backbone (Pfeiffer et al., 2021), while Compacter improves adapter parameter efficiency through low-

rank hypercomplex layers (Karimi Mahabadi et al., 2021). Model merging combines independently fine-tuned models in weight space, but naive averaging suffers from parameter interference. Task arithmetic and TIES-Merging show that model updates can encode reusable task vectors, and also that interference and sign conflicts can degrade merged models (Ilharco et al., 2023; Yadav et al., 2023). The same risk applies to LoRA composition: averaging or simultaneously activating multiple adapters can destroy attribute-specific behavior or cause semantic drift. Our method avoids full weight-space merging by routing style and topic LoRA updates to disjoint layer groups.

Composable LoRA and multi-adapter routing. Recent LoRA methods aim at training-free or lightweight composition through adaptive weighting, frequency-domain modulation, dynamic routing, token-aware activation, sparsity-aware selection, and retrieval-based adapter lookup (Yang et al., 2024; Roy et al., 2025; Zheng et al., 2025; Peng et al., 2025; Zhang and Xiong, 2025; Adsul et al., 2026). X-LoRA treats adapters as experts and learns a mixture over them (Buehler and Buehler, 2024), and iterative routing-tuned merging has been used for multi-task adapter combination (Zhao et al., 2025). Much of this work is motivated by multi-concept or multi-style personalization in diffusion models, where LoRA adapters correspond to visual subjects or artistic styles. K-LoRA is the closest to our setting because it performs training-free subject and style LoRA fusion (Ouyang et al., 2025). K-LoRA is designed for diffusion models, however, while in autoregressive LLMs adapter updates reshape next-token distributions across layers and decoding steps, which makes direct selection or merging more vulnerable to topic loss, style conflicts, and incoherence. We instead study LoRA composition specifically for language generation and propose a layer-partitioned routing that separates stylistic and topical control.

Style–content composition. Style–content composition has been studied in image generation, where methods combine visual subjects, styles, or concepts in diffusion models (Liu et al., 2022b; Sohn et al., 2023; Hertz et al., 2023). The general lesson is that separating content from style is useful, but the mechanisms do not transfer to text directly: in text, style and semantics are tightly coupled at the token level, since changes in lexical choice, rhetorical structure, or tone can also affect factual content. Earlier text style-transfer work focused on predefined attributes such as sentiment or formality, often with RNNs, VAEs, or non-parallel objectives (Ficler and Goldberg, 2017; Hu et al., 2017; Shen et al., 2017). We instead target transformer-based LLMs, scientific text generation, and modular recombination of independently trained style and topic adapters, which allows a style adapter to be reused across topics without rebuilding a rewritten dataset for every new style–domain pair.

Component	Description
Source texts	Short passages of 2–3 sentences, approximately 120–500 characters long.
Domains	General-domain texts from Wikipedia and CNN/DailyMail, plus scientific abstracts from arXiv.
Output format	A rewritten version of each source text in a specified target style, preserving the original meaning.
Main use	Training style-specific LoRA adapters and evaluating style–topic controllability.

Table 3: Overview of the stylization dataset construction.

B Dataset summary

The released dataset supports controllable text generation with separate control over writing style and topical/domain content. It is built around a rewriting pipeline: short source passages are sampled from general and scientific corpora, then rewritten by an instruction-tuned LLM into a target style while preserving the original semantic content. This design yields parallel pairs of neutral and stylized texts, making the dataset suitable for training style-specific LoRA adapters and evaluating style–topic composition.

The dataset contains five target writing styles: *sad/melancholic*, *cheerful/lighthearted*, *reflective/thoughtful*, *agitational/persuasive*, and *aggressive*. These styles were chosen to cover a broad range of affective and rhetorical transformations: negative affect, positive affect, introspective tone, persuasive framing, and high-intensity argumentative language. As a result, the dataset tests whether a model can learn not only sentiment polarity, but also more complex stylistic signals such as rhetorical pressure, emotional framing, and reflective phrasing.

The dataset construction process can be summa-

Style	Expected stylistic signal
Sad	Negative emotional framing, subdued tone, loss-oriented or regretful phrasing.
Cheerful	Positive emotional framing, optimistic wording, lighter and more accessible tone.
Reflective	Introspective phrasing, careful reasoning, moderate and contemplative tone.
Agitational	Benefit-oriented language, emphasis, calls to attention, persuasive framing.
Aggressive	Forceful wording, confrontational tone, high-intensity rhetorical pressure.

Table 4: Target styles and their intended stylistic properties.

rized as follows:

1. Sample a short source passage from Wikipedia, CNN/DailyMail, or arXiv.
2. Select one target style from the predefined style set.
3. Prompt Qwen2.5-32B-Instruct to rewrite the passage in the target style.
4. Preserve the original semantic content while changing stylistic realization.
5. Store the neutral–stylized pair for adapter training and evaluation.

First, topic-neutral or broadly applicable source passages are collected. Second, each passage is paired with a target style instruction. Third, the rewriting model generates a stylized version that keeps the original content, but changes the tone, word choice, and rhetorical presentation. Finally, the resulting examples are used for the causal language modeling of style adapters.

A key motivation behind the dataset is modularity. Since the style data is constructed from broadly applicable source texts, the learned style adapters are intended to remain topic-agnostic and reusable across downstream domains. This differs from directly training on a domain-specific stylized corpus, which can produce stronger in-domain results but requires rebuilding the rewritten dataset for every new topic–style combination.

The repository also supports the construction of a domain-specific stylistic dataset used for the *Super baseline*. This dataset follows the same rewriting pipeline, but the source texts come from the target downstream domain. In the paper, this baseline serves as an approximate upper bound: it is expected to align strongly with both the target topic and target style, but it sacrifices the main advantage of the proposed method, namely the ability to swap style and topic adapters independently.

Dataset statistics. The dataset contains two main groups of stylized data. The first group contains 2,200 source passages: 1,000 from Wikipedia, 1,000 from CNN/DailyMail, and 200 from astrophysics texts. These examples are rewritten into five styles: *aggressive*, *cheerful/lighthearted*, *promotional/persuasive*, *sad/melancholic*, and *thoughtful/reflective*. Most style files contain 2,200 examples, while *cheerful/lighthearted* contains 2,190 valid examples and *promotional/persuasive* contains 2,181 valid examples, with one malformed JSONL line. The second group is built from 200 astrophysics source passages and contains exactly 200 rewritten examples for each of the five styles. In total, the dataset contains 11,971 valid rewritten examples: 10,971 in the general/mixed-domain stylized subset and 1,000 in the extended astrophysics subset.

Overall, the dataset provides parallel neutral–stylized examples for studying controllable generation under multiple stylistic conditions. Its main contribution is not merely a set of rewritten texts, but a reusable framework for separating style learning from topic learning: style adapters can be trained once on the stylization dataset, while topic adapters can be trained separately on downstream domain data and composed at inference time.

C Ablation Study

Number of layers. To justify $n = 7$ we measure how the choice of n changes the model's predictions under different style adapters. For each n, we apply style adapters to the first n layers and train one LoRA set per style for four target styles: *sad*, *cheerful*, *reflective*, and *promotional*. We then take a subset of neutral texts that we previously rewrote in each of these styles and measure the average token-level variance of the model's predictions. For a given neutral text, we run the model with each of the four style adapters in turn and get four log-probability vectors per token. We compute the variance across these four vectors at each token position, then average over tokens to get a scalar per text, and finally average over the subset.

The averaged variance quantifies how strongly different style adapters alter the model's token-level predictions. A low value means the adapters have little effect, a higher value means stronger stylistic differentiation. As n grows, the averaged

Subset	Style	# valid examples	Avg. source words	Avg. stylized words
`base`	Aggressive	2,200	54.7	57.8
`base`	Cheerful	2,190	54.8	73.1
`base`	Agitational	2,181	54.9	79.7
`base`	Sad	2,200	54.7	70.7
`base`	Reflective	2,200	54.7	75.3
`extended`	Aggressive	200	85.1	85.9
`extended`	Cheerful	200	85.1	104.7
`extended`	Agitational	200	85.1	101.4
`extended`	Sad	200	85.1	102.1
`extended`	Reflective	200	85.1	100.5
Total	–	11,971	–	–

Table 5: Statistics of the stylized JSONL files in `generate-dataset/data`. The general `styled` subset is derived from mixed-domain source passages, while `styled_astroph_extended` contains rewritten astrophysics-domain passages.

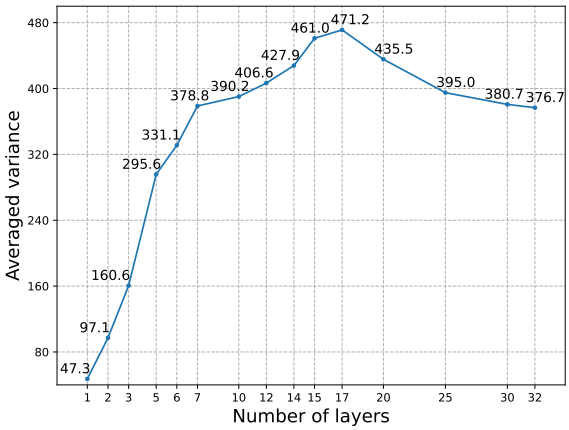

Figure 3: Averaged variance of Llama-3-8B predictions at token level; $n = 7$ has strong impact, while being relatively small.

variance grows with it: adapters applied to a larger portion of the network can encode more stylistic capacity (Fig. 3).

For the *aggressive* style, some inputs produce weak or absent stylization. We attribute this to the smaller and differently sourced training set for that style. To recover, we mix neutral and stylized texts during training and prepend a special style token to distinguish them. At inference time we prepend the same token to trigger stylized generation.

The *aggressive* style is excluded from this analysis because its data comes from a different set of neutral texts and is substantially smaller. For the other four styles the datasets share the same neutral source texts rewritten in different ways, which removes the confounding effect of content memorization and isolates the stylistic contribution of the adapters.

Prompt tuning. To reduce prompt-sensitivity concerns, we manually tested several variants of the *Instructed style and topic* prompt on a held-out development subset and selected the prompt that most consistently preserved both the requested style and topic. The final prompt explicitly includes the target topic, the target style, and an instruction to maintain fluency and coherence. We report results using this fixed prompt across all styles and inputs. Although this does not constitute exhaustive prompt optimization, it provides a consistent and reasonably tuned instruction-only baseline.

D Experiment Details

LoRA adapters are injected into all self-attention and MLP modules. The LM head stays frozen for both style and topic adapters. We tried applying LoRA only to self-attention layers and found that for the style adapters, this is not enough: adding LoRA to the MLP blocks gives more stable training and better generation. We attribute this to the role of MLP layers in semantic and lexical transformation, which is what style adaptation operates on, while self-attention mostly captures contextual co-occurrence patterns.

Both style and topic adapters are trained until the training loss reaches a plateau. We use the same training procedure, the same LoRA hyperparameters (learning rate$=$ 0.0002, rank$=$ 16, scaling coefficient$=$ 32, dropout$=$ 0.05) and also 4-bit quantization for both adapter types, although this uniform setup is not required. The topic adapter is applied to all layers after the n-th one. Finer layer allocation is left for future work.

D.1 Stylization prompts

We use Qwen2.5-32B-Instruct (Team, 2024) as the model for rewriting the texts according to the required style. For each source text and target style, the model is instructed to generate a stylized version that preserves the semantic content while adopting the specified tone. Generation is controlled by a structured rewriting prompt that explicitly specifies the desired style while constraining semantic fidelity.

System prompt:

```
You are a rewriting model.
Always return only the
rewritten text.  Answer
strictly in English.
Do not add greetings,
explanations, or comments.
```

User prompt:

```
Rewrite the following text
in the style:  {style}.
Make it distinctively
{style}, but keep the
original meaning.

Text:  {text}

Output only the rewritten
text in English.
```

The resulting dataset contains paired examples $(x, s, \tilde{x})$ suitable for supervised LoRA fine-tuning of a smaller model for controllable text generation. The full prompts are available at https://anonymous.4open.science/r/style-lora/README.md.

D.2 Evaluation details

All baselines generate a candidate for the same input and target style. We compare candidates with the listwise ranker and pointwise scorer from the main text, both using Qwen2.5-32B-Instruct (Team, 2024) as the judge. Code is at https://anonymous.4open.science/r/style-lora/README.md.

List-wise Ranker. For each evaluation texts, all candidates are presented to the judge model, which produces a strict ordering from best to worst. The ranking criteria prioritize texts matching both attributes, then either one and finally neither. Given a set of K methods evaluated on N rows, each method receives a rank $\rho_i \in \{1, \ldots, K\}$ per row (1 = best). We report metrics:

1. **Norm. Rank**: $\frac{1}{N}\sum_{i=1}^{N}\frac{K-\rho_i}{K-1} \in [0,1]$, where 1 means that the method was always ranked first.
2. **Top-M Rate**: fraction of texts where method was ranked top M.

Point-wise Scorer. Each generated text is independently evaluated based on four criteria:

1. **Topic Hit Rate** ($r \in \{0,1\}$). Binary indicator of whether the text genuinely addresses the specified topic. This acts as a hard gate: texts that fail receive a final score of zero.
2. **Style Quality** ($q_{\text{style}} \in [0,1]$). Degree to which the text exhibits target stylistic properties (tone, word choice, rhetorical devices).
3. **Topic Quality** ($q_{\text{topic}} \in [0,1]$). Level of substantive engagement with domain-specific content, from superficial to expert-level mentions.
4. **Coherence Quality** ($q_{\text{coh}} \in [0,1]$). How naturally the style integrates with the topic content, from forced/disconnected to organic fusion.

The judge assigns a score for each criterion based on a prompt that includes detailed rubrics for each scoring range (e.g. 0.0 – -0.2 indicates absence, 0.7 – -0.9 indicates strong presence) to ensure consistent calibration. We also compute a metric called *Aggregate Score*, which is the arithmetic mean of the three quality dimensions, gated by topic relevance: $r \cdot \frac{q_{\text{style}}+q_{\text{topic}}+q_{\text{coh}}}{3}$. When reporting results on a set of texts N, we independently report the mean of each criterion (*Topic Hit Rate*, *Style Quality*, *Topic Quality*, and *Coherence Quality*) as well as the mean *Aggregate Score*.

The judge model receives the following prompt for each text:

System: `You are a precise text evaluator. Output ONLY valid JSON, no explanations.`

User:

```
Evaluate this text on multiple
criteria.
Topic required:  "{topic}"
Style required:  "{style}"

TEXT TO EVALUATE: """{text}"""

EVALUATION CRITERIA:
1.  topic_relevant (true/false):
Is this text ACTUALLY ABOUT
"{topic}"?
- true = text genuinely
discusses {topic} concepts/ideas
- false = text is about a
DIFFERENT subject
2.  style_match (0.0-1.0):  How
well does the text exhibit the
"{style}" style?
3.  topic_depth (0.0-1.0):  How
deeply does the text engage with
```

```
"{topic}"?
4. topic_style_coherence
(0.0-1.0): How naturally does
the style integrate with the
topic?

OUTPUT FORMAT - respond with
ONLY this JSON:
{"topic_relevant": true/false,
"style_match": 0.XX,
"topic_depth": 0.XX,
"topic_style_coherence": 0.XX}
```

The prompt includes detailed rubrics for each score range (e.g., $0.0 - 0.2$ indicates absence, $0.7 - 0.9$ indicates strong presence) to ensure consistent calibration.

Human Evaluation. We used Amazon Mechanical Turk crowdsourcing marketplace, asking 5 assessors to evaluate pairwise comparisons of stylized texts generated by our approach and competitors (we chose the top-2 competitors for each style and asked annotators to find the most stylistically aligned text out of three proposed texts). We estimated the expected completion time of each task during pilot runs and set compensation to correspond to approximately 10 per hour, above the U.S. federal minimum wage. Each assessor received the following task:

```
Task: Compare two texts and
decide which one best matches
the requested writing style.

You will be shown:

1. A target writing style.
2. Two candidate texts, Text A
   and Text B.

Your goal is to choose the text
that better aligns with the
target style. Focus only on
how well the writing style
matches the requested style.
Do not choose based on which
text is more factually detailed,
longer, more informative, or
more interesting unless that
directly affects the requested
style.

Target style: [STYLE]

Text A: [TEXT A]

Text B: [TEXT B]

Question: Which text better
matches the target style?

Options:

• Text A
• Text B
• Tie / both match the style
  equally well

Guidelines:

• Choose "Text A" if Text A
  expresses the requested style
  more clearly and naturally
  than Text B.
• Choose "Text B" if Text B
  expresses the requested style
  more clearly and naturally
  than Text A.
• Choose "Tie" if both texts
  match the style about equally
  well.

Style descriptions:

• Sad / melancholic: the
  text should sound sorrowful,
  gloomy, regretful, emotionally
  heavy, or marked by loss or
  disappointment.
• Cheerful / lighthearted: the
  text should sound positive,
  bright, optimistic, playful, or
  emotionally uplifting.
• Thoughtful / reflective:
  the text should sound
  contemplative, introspective,
  careful, balanced, or
  intellectually reflective.
• Agitational / persuasive: the
  text should sound convincing,
  benefit-oriented, enthusiastic,
  advertising-like, or designed
  to persuade the reader.
• Aggressive / confrontational:
  the text should sound forceful,
  intense, harsh, combative, or
  strongly assertive.

Important: Do not judge whether
you personally like the text.
Do not judge whether the text
is factually correct unless
factual errors make the style
unnatural. Do not prefer a text
only because it is longer. Do
not prefer a text only because
it has more technical details.
Judge only which text best
follows the requested writing
style.
```

In general, we conducted 20 comparisons per style and competitor (including the Ideal Score), resulting in 80 comparisons in total.