Evaluating Style-Personalized Text Generation: Challenges and Directions

Anonymous ACL submission

Abstract

While prior research has built tools and benchmarks towards style personalized text generation, there has been limited exploration of evaluation in low-resource author style personalized text generation space. Through this work, we question the effectiveness of the widely adopted evaluation metrics like BLEU and ROUGE, and explore other evaluation paradigms such as style embeddings and LLM-as-judge to holistically evaluate the style personalized text generation task. We evaluate these metrics and their ensembles using our style discrimination benchmark, that spans eight writing tasks, and evaluates across three settings, *domain discrimination*, *authorship attribution*, and *LLM personalized vs non-personalized discrimination*. We provide conclusive evidence to adopt ensemble of diverse evaluation metrics to effectively evaluate style personalized text generation.

1 Introduction

Recent empirical studies have highlighted a sharp increase in adoption of large language models (LLMs) for writing assistance across various facets of society. LLM-assisted writing has permeated various professional areas like journalism (Diakopoulos, 2019), legal services (Magesh et al., 2024), healthcare (Baker et al., 2024; Bongurala et al., 2024; Rengers et al., 2024), marketing (Kumar et al., 2024b), academic writing (Khalifa and Albadawy, 2024; Nguyen et al., 2024) *etc*. Adoption of AI-based writing assistants have also gained popularity at personal scale, with use cases varying from emails (Li et al., 2025), resume and cover letters (Zinjad et al., 2024), social media posts (Long et al., 2023; Jain et al., 2023), blogs (Kaisen et al., 2024) *etc*. This growing reliance on LLMs brings forth the need for style personalized text generation (Mysore et al., 2025). Lack of appropriate techniques to generate text that reflects a user's writing style can do more harm than good, affecting their interpersonal relationships and professional branding (Kadoma et al., 2025).

Recent years have seen increased efforts to utilize the style transferring capabilities of large language models to develop style personalized text generation frameworks, like a personalized prompt rewriter to generate desired style output (Li et al., 2024a), linguistic feature controlled multi-attribute style personalized text generation system (Alhafni et al., 2024), and generation calibrated retriever for personalized writing assistance (Mysore et al., 2023). Prior work has also developed generation benchmarks like `LaMP` (Salemi et al., 2023) and `LongLaMP` (Kumar et al., 2024a) to to evaluate the style personalization capabilities of text generation models. Yet evaluating the style personalization capabilities of LLMs remains a primarily underexplored problem, with prior efforts majorly adopting rudimentary ngram-overlap based evaluation metrics like `BLEU` (Papineni et al., 2002), `ROUGE` (Lin, 2004) and `METEOR` (Banerjee and Lavie, 2005) against a reference text to measure the quality of style personalization in the generated output. In this work, we explore ngram evaluation metrics along with other evaluation paradigms like style-aware embeddings (Wegmann et al., 2022; Patel et al., 2024) and LLM-as-judge (Gu et al., 2024; Li et al., 2024b) to develop guidelines on evaluation of low-resource style personalized text generation. The major contributions of our work are as follows -

1. We develop an evaluation benchmark spanning eight diverse writing tasks across three different evaluation settings to rigorously evaluate the style discrimination task.
2. To the best of our knowledge, we are the first work to extensively evaluate low-resource style personalized text generation using ngram, style embeddings and LLM-as-judge metrics and their ensembles.

Table 1: Statistics for our extended style personalization evaluation benchmark across the three evaluation settings - *domain discrimination (DD)*, *authorship attribution (AA)*, and *LLM personalized vs non-personalized (LLM)*, with corresponding columns describing the average number of words in the T_{ref}, T_+, and T_- respectively.

Dataset Name	#authors	*Domain Discrimination*	*Authorship Attribution*	*LLM*
Amazon Food Reviews (McAuley and Leskovec, 2013)	100	157/267/388	161/267/263	267/272/300
ArXiV (arXiv.org submitters, 2024)	100	144/224/447	144/224/229	224/406/505
Blogs (Schler et al., 2006)	95	173/329/373	164/329/349	329/369/396
Enron (Klimt and Yang, 2004)	100	123/328/520	119/328/334	328/193/219
Lyrics (Edenbd, 2020)	95	210/278/495	211/278/288	278/308/315
Reddit (Patel et al., 2022)	56	87/38/410	93/38/38	38/222/261
Reuters (Lewis, 1987)	49	363/506/461	366/506/507	506/416/456
Short Stories (Carney and Robertson, 2019)	41	903/1289/267	911/1289/1364	1289/543/515

2 Problem Statement

Given a reference text T_{ref}, and two candidate texts T_+ and T_-, we define the style discrimination task, where the objective is to identify which of the two candidates is stylistically closer to the reference. Specifically, assuming the existence of an underlying style distribution ($T_x \sim S_x$), such that $S_{ref} \approx S_+$ and $S_{ref} \not\approx S_-$. The task is to learn a discriminator $f : (T_{ref}, T_+, T_-) \mapsto \{+, -\}$ that predicts which of $\{T_+, T_-\}$ is stylistically closer to T_{ref}.

3 Experimental Settings

3.1 Dataset

We extensively evaluate the metrics over eight writing domains across three evaluation settings. We provide the overall statistics of the benchmark and source datasets in Table 1. We explain the three evaluation settings below, and discuss the dataset construction details in Appendix A.

Based on how we obtained the candidates $\{T_+.T_-\}$, we propose three evaluation settings. For a reference text $T_{ref} \in D_{d_1}^{a_1}$, where D_d^a denotes set of text written by author a in domain d, we obtain T_+ and T_- as follows -

Domain discrimination (*DD*): sample $T_+ \in D_{d_1}^{a_1}$ such that $T_+ \neq T_{ref}$, and $T_- \in D_{d_2}^{a_2}$.

Authorship attribution (*AA*): sample $T_+ \in D_{d_1}^{a_1}$ such that $T_+ \neq T_{ref}$, and $T_- \in D_{d_1}^{a_2}$.

LLM personalized vs non-personalized (*LLM*): given a query reconstruction function $f_{query} : \mathbb{T} \rightarrow \mathbb{Q}$, that outputs a query ($q \in \mathbb{Q}$) corresponding to an input text ($T \in \mathbb{T}$), we obtain the original user query $q_{ref} = f_{query}(T_{ref})$. Using a generation model $\mathbf{M}$, we obtain the candidates $T_+ = \mathbf{M}(q_{ref}|T'_{ref})$ and $T_- = \mathbf{M}(q_{ref})$, where $T'_{ref} \in D_{d_1}^{a_1}$ such that $T'_{ref} \neq T_{ref}$ (respective prompts in Figures 9 and 4).

3.2 Evaluation Metrics

We explore three different paradigms of evaluation metrics in this study -

Ngram overlap-based evaluation metrics. We explore three widely adopted ngram evaluation metrics, namely `BLEU` (Papineni et al., 2002), `ROUGE` (Lin, 2004), and `METEOR` (Banerjee and Lavie, 2005). Since these evaluation metrics work on a text pair, we obtain the binary label for similarity measure S as follows:

$$label = \begin{cases} 1 & \text{if} \quad S(T_{ref}, T_+) > S(T_{ref}, T_-) \\ 0 & otherwise \end{cases} \tag{1}$$

Style Embedding-based evaluation metrics. We explore two different style embedding models, `Wegmann`[1] (Wegmann et al., 2022), and `StyleDistance` (Patel et al., 2024). Since these embedding models map a piece of text into the latent vector space, we use cosine similarity as the similarity measure S to measure similarity between T_{ref} and $T_X, X \in \{+, -\}$, and compute the label using Eq. 1.

LLM-as-judge evaluation metrics. We use `Ministral-3B`, `Llama-3.1-8B` (Grattafiori et al., 2024), `Mistral-24B`, `Qwen3-32B` (Yang et al., 2025), `DeepSeek-V3`[2] (Liu et al., 2024) as our open-source models, and we also evaluate OpenAI's closed source models `o4-mini` and `gpt-4.1` (Achiam et al., 2023), spanning several model families across various parameter size (see Figure 6 for the evaluation prompt).

3.3 Ensemble of Metrics

Motivated by the effectiveness of ensemble methods achieving superior performance (Dong et al., 2020; Mienye and Sun, 2022), we explore several

[1]For simplicity, we refer to this baseline as "Wegmann" after the first author in Wegmann et al. (2022).

[2]We use `Mistral-24B` and `DeepSeek-V3` to refer to `Mistral-Small-24B-Instruct-2501` and `DeepSeek-V3-0324` models respectively.

Table 2: Accuracy of evaluation metrics for *domain discrimination* ***(DD)***, *authorship attribution* ***(AA)***, and *LLM personalized vs non-personalized* ***(LLM)***.

Evaluation Metric	***DD***	***AA***	***LLM***	Mean
Random 100	0.502	0.498	0.498	0.499
BLEU	0.869	0.700	0.631	0.733
ROUGE1	0.858	0.695	0.613	0.722
ROUGE2	0.803	0.643	0.634	0.693
ROUGEL	0.854	0.673	0.631	0.719
METEOR	0.783	0.654	0.547	0.661
Wegmann	0.910	0.670	0.558	0.713
StyleDistance	0.926	0.665	0.575	0.722
Ministral-3B	0.285	0.272	0.417	0.324
Llama-3.1-8B	0.157	0.263	0.418	0.279
Mistral-24B	0.478	0.453	0.470	0.467
Qwen3-32B	0.689	0.608	0.525	0.607
DeepSeek-V3	0.693	0.679	0.601	0.658
o4-mini	0.917	0.750	0.597	0.755
gpt-4.1	**0.961**	**0.807**	**0.678**	**0.815**

ensembles based on the availability of evaluation metrics as follows -

- ρ^X_{ngram}: ensembling over ngram evaluation metrics.
- $\rho^X_{\neg LLM}$: ensembling over all non-LLM evaluation metrics.
- ρ^X_{LLM}: ensembling over LLM-as-judge evaluation metrics.
- ρ^X_{OS}: ensembling over all open-source evaluation metrics.
- ρ^X_{all}: ensembling over all evaluation metrics detailed in Section 3.2.

where X denotes the ensembling strategy. We explore two ensembling strategies - majority voting (MV), and performance-weighted voting (PWV).

4 Results and Discussion

We present the results of evaluation metrics on our proposed benchmark in Table 2. We observe certain trends in the performance across different evaluation settings and different metric paradigms that we discuss below.

Performance across different evaluation settings. We found the performance to be highest for *domain discrimination (DD)* evaluation setting, then *authorship attribution (AA)* and then *LLM personalized vs non-personalized (LLM)* across almost all evaluation metrics. As *domain discrimination* requires the metric to discriminate across different writing styles, it was the simplest style discrimination setting, with four metrics achieving over 0.9 accuracy. Comparatively, the *authorship attribution* evaluation setting requires a metric to discriminate within the same writing domain, and was a harder evaluation task, with the highest accuracy of 0.807 (dropping 16% from *DD*). Whereas, the task of classifying personalized and non-personalized generated text (*LLM*) was the hardest to distinguish, with the best performance of 0.678 accuracy (dropping 28.6% from *DD*). We posit two key reasons for this low performance: a) both T_+ and T_- text are generated from the same generation model (**M**) for the same query (q_{ref}), with the main difference being the presence of reference style text for the generation of T_+. As the generation models have their own stylistic preferences when generating text (Reinhart et al., 2025), both T_+ and T_- might could have overlapping stylistic features making it harder to distinguish, and b) unlike the *DD* and *AA* evaluation settings that have different content across the input triplet $\{T_{ref}, T_+, T_-\}$, the input in *LLM* setting share the same writing task, making it a harder evaluation task for style evaluation metrics. From Figure 1 notice that there's a significant increase in BertScore (Zhang et al., 2019) semantic similarity in *LLM* evaluation setting, with the highest similarity between T_- and T_+.

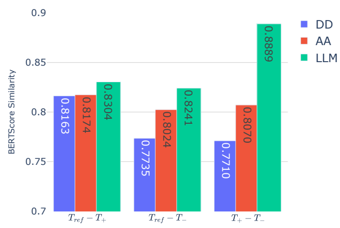

Figure 1: Average pairwise BertScore (Zhang et al., 2019) similarity across T_{ref}, T_+, and T_-.

Performance across different evaluation metrics. We found the ngram evaluation metrics to be surprisingly competent in this low-resource style discrimination evaluation task, contradicting prior work deeming them incapable of style transfer evaluation (Pang and Gimpel, 2018). We believe this to be because Pang and Gimpel (2018) explored text style transfer at a sentence level, where lack of sufficient reference ngrams makes it difficult to capture nuanced stylistic features for evaluation. Nonetheless, we find the ngram metrics to fall behind other complex neural evaluation metrics. gpt-4.1 achieves the highest overall evaluation score of 0.815 accuracy, compared to BLEU at 0.733. We observe that not all language models perform well on the task, with a clear trend of larger and closed-source models performing better than smaller models. Three smallest mod-

Table 3: Accuracy of best performing ensemble across the three evaluation settings.

Metric	**Ensemble Composition**	*DD*	*AA*	*LLM*	Mean
Random 100	-	0.502	0.498	0.498	0.499
BLEU	-	0.869	0.700	0.631	0.733
StyleDistance	-	0.926	0.665	0.575	0.722
gpt-4.1	-	**0.961**	**0.807**	**0.678**	**0.815**
ρ^{MV}_{ngram}	BLEU ROUGE2 ROUGEL	0.866	0.704	0.654	0.742
ρ^{PWV}_{ngram}	BLEU ROUGE1 ROUGE2 ROUGEL	0.862	0.714	0.654	0.743
$\rho^{MV}_{\neg LLM}$	BLEU ROUGE2 ROUGEL Wegmann StyleDistance	0.914	0.725	0.682	0.774
$\rho^{PWV}_{\neg LLM}$	BLEU ROUGE2 Wegmann StyleDistance	0.948	0.722	0.670	0.780
ρ^{MV}_{LLM}	Qwen3-32B o4-mini gpt-4.1	0.937	0.781	0.640	0.786
ρ^{PWV}_{LLM}	Qwen3-32B DeepSeek-V3 o4-mini gpt4.1	0.945	0.772	0.662	0.793
ρ^{MV}_{OS}	BLEU rougeL Wegmann StyleDistance DeepSeek-V3	0.925	0.750	0.684	0.786
ρ^{PWV}_{OS}	BLEU ROUGEL Wegmann StyleDistance Qwen3-32B DeepSeek-V3	0.937	0.750	0.682	0.790
ρ^{MV}_{all}	BLEU ROUGEL Wegmann StyleDistance gpt-4.1	0.962	0.786	**0.697**	0.815
ρ^{PWV}_{all}	BLEU ROUGE1 StyleDistance gpt-4.1	**0.967**	**0.802**	0.693	**0.821**

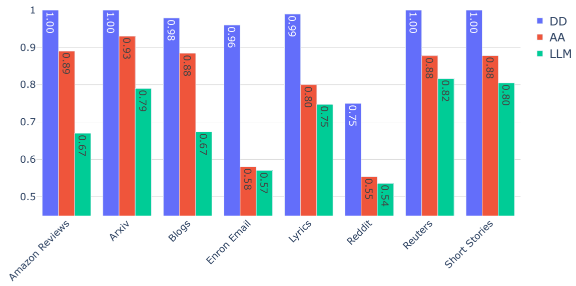

Figure 2: Accuracy of ρ^{PWV}_{all} across all domains.

els we evaluate yield below random performance, and this happens because our prompt restricts responses beyond T_+ and T_-, with other acceptable response being 'Both' and 'None' (see Fig. 5).

Performance of ensemble of metrics In Table 3, we present the results for best performing ensemble combination over different metrics. Using an ensemble of evaluation metrics is generally able to achieve better performance than its best constituent; *e.g.*, $\rho^{MV}_{\neg LLM}$ achieves 5.6% better accuracy than its best performing constituent BLEU. We also obeserve that using performance-weighted voting (*PWV*) ensemble always slightly outperforms it's corresponding majority vote counterpart. It is interesting to observe that the best performing ensembles have an fairly uniform distribution of metrics, emphasizing the significance of multiple evaluation paradigms to yield robust performance. This becomes further evident when we observe the pairwise disagreement of the evaluation metrics, where we see high disagreement across different evaluation paradigms (see Fig. 3).

Quantitative analysis of ρ^{PWV}_{all}. We compute agreement between ρ^{PWV}_{all} and its constituents to explore the effective contribution of each metric (BLEU: 0.68, ROUGE1: 0.68, StyleDistance: 0.66, gpt-4.1: 0.78). We find that all metrics except gpt-4.1 contribute equally, which can be credited to the PWV ensembling strategy.

Investigating domain-wise accuracy (Figure 2), we observe that *Reddit* yields the poorest performance, even for *domain discrimination*, due to it's short reference text (Table 1). This performance further degrades to 0.54 for *LLM* evaluation setting (barely better than Random 100). We see similar trends for *Enron* emails, which also contains small reference text, with significant content overlap across authors as well (due to corporate emails belonging to the same organization). Besides these two domains, for the *LLM* evaluation setting, we notice a clear divide in performance between the standard writing domains (*arxiv*, *reuters*, *short stories*), and the informal domains (*amazon food reviews*, *blogs*), with about 0.1 gap in accuracy. Our manual investigation revealed that the generated responses don't capture certain stylistic traits of informal writing like ellipsis ('...'), non-standard capitalization, niche slangs *etc.*

5 Conclusion

Our experiments reveal two key takeaways for evaluation in future research. **First**, under low-resource evaluation, metrics are most effective at distinguishing texts across domains (*DD*), moderately effective at within-domain authorship attribution (*AA*), and least effective at detecting differences between personalized and non-personalized LLM-generated text (*LLM*). **Second**, the best performance across all three style discrimination tasks is achieved from an ensemble of diverse evaluation paradigms, such as ngram matching, style embedding and LLM-as-judge. This approach yields up to 12% improvement over the best-performing ngram evaluation metric used in prior work.

Limitations

While we develop an extensive evaluation benchmark to extensively evaluate the style personalized text generation task, there are three main limitations of our work. **First**, we limit our research to the binary classification task of discriminating two responses given a reference text, and leave the exploration of fine-grained explainable evaluation systems to future work. **Second**, we only explore two ensembling approaches, majority voting, and weighted voting, leaving exploration of other ensembling techniques like likelihood-based voting, threshold-tuned voting and meta ensemble learners to future work. **Third**, the binary nature of our task formulation doesn't take into account the first-person perspective of style personalized text generation in our evaluation benchmark. We acknowledge the challenge of developing a diverse first-person evaluation benchmark to cater to individual user's specific preferences, and adopt a set of simpler style discrimination tasks to evaluate existing and proposed evaluation metrics.

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*.

Bashar Alhafni, Vivek Kulkarni, Dhruv Kumar, and Vipul Raheja. 2024. Personalized text generation with fine-grained linguistic control. *arXiv preprint arXiv:2402.04914*.

arXiv.org submitters. 2024. arxiv dataset.

Hayden P Baker, Emma Dwyer, Senthooran Kalidoss, Kelly Hynes, Jennifer Wolf, and Jason A Strelzow. 2024. Chatgpt's ability to assist with clinical documentation: a randomized controlled trial. *JAAOS-Journal of the American Academy of Orthopaedic Surgeons*, 32(3):123–129.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Archana Reddy Bongurala, Dhaval Save, Ankit Virmani, and Rahul Kashyap. 2024. Transforming health care with artificial intelligence: redefining medical documentation. *Mayo Clinic Proceedings: Digital Health*, 2(3):342–347.

James Carney and Cole Robertson. 2019. 4000 stories with sentiment analysis dataset. brunel university london.

Nicholas Diakopoulos. 2019. *Automating the news: How algorithms are rewriting the media*. Harvard University Press.

Xibin Dong, Zhiwen Yu, Wenming Cao, Yifan Shi, and Qianli Ma. 2020. A survey on ensemble learning. *Frontiers of Computer Science*, 14(2):241–258.

Edenbd. 2020. 150k lyrics labeled with spotify valence.

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

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Vibhu Jain, Yash Goel, M Uma, et al. 2023. Ai powered transformative post generator for linkedin using llm and explicit filter. In *2023 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES)*, pages 1–7. IEEE.

Kowe Kadoma, Danaë Metaxa, and Mor Naaman. 2025. Generative ai and perceptual harms: Who's suspected of using llms? In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–17.

Joshua Kaisen, Meng Li, and Shijie Lu. 2024. Ai and productivity: The impact of chatgpt's release on blogging. *Available at SSRN 4858507*.

Mohamed Khalifa and Mona Albadawy. 2024. Using artificial intelligence in academic writing and research: An essential productivity tool. *Computer Methods and Programs in Biomedicine Update*, page 100145.

Bryan Klimt and Yiming Yang. 2004. Introducing the enron corpus. In *CEAS*, volume 4, page 1.

Ishita Kumar, Snigdha Viswanathan, Sushrita Yerra, Alireza Salemi, Ryan A Rossi, Franck Dernoncourt, Hanieh Deilamsalehy, Xiang Chen, Ruiyi Zhang, Shubham Agarwal, et al. 2024a. Longlamp: A benchmark for personalized long-form text generation. *arXiv preprint arXiv:2407.11016*.

V Kumar, Abdul R Ashraf, and Waqar Nadeem. 2024b. Ai-powered marketing: What, where, and how? *International Journal of Information Management*, 77:102783.

David Lewis. 1987. Reuters-21578 Text Categorization Collection. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C52G6M.

Cheng Li, Mingyang Zhang, Qiaozhu Mei, Weize Kong, and Michael Bendersky. 2024a. Learning to rewrite prompts for personalized text generation. In *Proceedings of the ACM on Web Conference 2024*, pages 3367–3378.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024b. Llms-as-judges: a comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*.

Weijiang Li, Yinmeng Lai, Sandeep Soni, and Koustuv Saha. 2025. Emails by llms: A comparison of language in ai-generated and human-written emails. In *Proceedings of the 17th ACM Web Science Conference 2025*, pages 391–403.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Tao Long, Dorothy Zhang, Grace Li, Batool Taraif, Samia Menon, Kynnedy Simone Smith, Sitong Wang, Katy Ilonka Gero, and Lydia B Chilton. 2023. Tweetorial hooks: generative ai tools to motivate science on social media. *arXiv preprint arXiv:2305.12265*.

Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D Manning, and Daniel E Ho. 2024. Hallucination-free? assessing the reliability of leading ai legal research tools. *Journal of Empirical Legal Studies*.

Julian John McAuley and Jure Leskovec. 2013. From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews. In *Proceedings of the 22nd international conference on World Wide Web*, pages 897–908.

Ibomoiye Domor Mienye and Yanxia Sun. 2022. A survey of ensemble learning: Concepts, algorithms, applications, and prospects. *Ieee Access*, 10:99129–99149.

Sheshera Mysore, Debarati Das, Hancheng Cao, and Bahareh Sarrafzadeh. 2025. Prototypical human-ai collaboration behaviors from llm-assisted writing in the wild. *arXiv preprint arXiv:2505.16023*.

Sheshera Mysore, Zhuoran Lu, Mengting Wan, Longqi Yang, Steve Menezes, Tina Baghaee, Emmanuel Barajas Gonzalez, Jennifer Neville, and Tara Safavi. 2023. Pearl: Personalizing large language model writing assistants with generation-calibrated retrievers. *arXiv preprint arXiv:2311.09180*.

Andy Nguyen, Yvonne Hong, Belle Dang, and Xiaoshan Huang. 2024. Human-ai collaboration patterns in ai-assisted academic writing. *Studies in Higher Education*, 49(5):847–864.

Richard Yuanzhe Pang and Kevin Gimpel. 2018. Unsupervised evaluation metrics and learning criteria for non-parallel textual transfer. *arXiv preprint arXiv:1810.11878*.

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

Ajay Patel, Nicholas Andrews, and Chris Callison-Burch. 2022. Low-resource authorship style transfer: Can non-famous authors be imitated? *arXiv preprint arXiv:2212.08986*.

Ajay Patel, Jiacheng Zhu, Justin Qiu, Zachary Horvitz, Marianna Apidianaki, Kathleen McKeown, and Chris Callison-Burch. 2024. Styledistance: Stronger content-independent style embeddings with synthetic parallel examples. *arXiv preprint arXiv:2410.12757*.

Alex Reinhart, Ben Markey, Michael Laudenbach, Kachatad Pantusen, Ronald Yurko, Gordon Weinberg, and David West Brown. 2025. Do llms write like humans? variation in grammatical and rhetorical styles. *Proceedings of the National Academy of Sciences*, 122(8):e2422455122.

Timothy A Rengers, Cornelius A Thiels, and Hojjat Salehinejad. 2024. Academic surgery in the era of large language models: a review. *JAMA surgery*, 159(4):445–450.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2023. Lamp: When large language models meet personalization. *arXiv preprint arXiv:2304.11406*.

Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W Pennebaker. 2006. Effects of age and gender on blogging. In *AAAI spring symposium: Computational approaches to analyzing weblogs*, volume 6, pages 199–205.

Anna Wegmann, Marijn Schraagen, and Dong Nguyen. 2022. Same author or just same topic? towards content-independent style representations. *arXiv preprint arXiv:2204.04907*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

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

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024. Wildchat: 1m chatGPT interaction logs in the wild. In *The Twelfth International Conference on Learning Representations*.

Saurabh Bhausaheb Zinjad, Amrita Bhattacharjee, Amey Bhilegaonkar, and Huan Liu. 2024. Resumeflow: An llm-facilitated pipeline for personalized resume generation and refinement. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2781–2785.

A Dataset Curation (Contd.)

Source dataset selection. To rigorously explore the evaluation metrics, we selected eight NLP datasets spanning diverse writing tasks and domains. We selected datasets that - i) contain multiple writing instances from same author, ii) reflect a real-world writing task, and iii) that are publicly available and licensed for reuse in research publications[3]. We selected the datasets spanning diverse writing tasks like creative writing (*short stories*, *lyrics*), formal writing (*reuters news*, *enron emails*, *arxiv scientific abstracts*), and casual writing (*blogs*, *amazon food reviews*, *reddit microblogs*).

Domain-level downsampling. In order to develop a diverse evaluation corpus to test the capabilities of style-personalized document generation, we sampled instances from each dataset based on the number of tokens written by each author. We used OpenAI's tiktoken[4] library to determine the number of tokens in each article. For each author a in domain d, we develop a bag of articles ($\mathcal{D}_d^a$), by randomly sampling the articles written by the author with an upper limit of 500 tokens and randomly selected a target article with 200 to 700 tokens. Reddit microblog and short stories datasets had different token lengths due to their non-standard text lengths. For the reddit dataset, because all the written microblogs were too short, we set the minimum token length for target article to zero; whereas for the short stories dataset, because the written stories were too long, we increased the limit of style grounding articles to 1500 tokens and set the upper bound for target article to 2000 tokens. After the sampling process, we obtained about 50 instances for reddit, reuters and short stories datasets, while 100 instances for the remaining five datasets. However, we removed a few examples from the dataset where the personalized text generation model ($\mathbf{M}$) didn't yield appropriate response.

User Query Reconstruction and Personalized Text Generation. To develop the $< T_{ref}, T_+, T_- >$ instance triplet for the *llm* evaluation setting, we first develop a query reconstruction prompt using a large language model[5]. In order to replicate human-written queries, we made use of the queries from WildChat dataset as the guiding text to reconstruct query corresponding to the reference text (T_{ref}). We randomly selected 50,000 user interactions with English language tag, and performed two consecutive filtering steps using text classification - i) removing all non-writing user queries (using Figure 10), and ii) removing writing queries of irrelevant writing tasks (using Figure 11). We provide the statistics of classification prompts in Figure 7. Once we obtain the seed user queries from the WildChat dataset (Zhao et al., 2024), we match each of our personalization datasets with the closest user queries and feed it to our user query reconstruction prompt along with the target article to obtain the reconstructed user query q_{ref} (see Fig. 8). We then use the style personalized text generation prompt and non-personalized text generation prompt to obtain T_+ and T_- respectively (see Figures 9 and 4).

[3]All the datasets we collected were licensed under CC BY 4.0, CC BY-SA 4.0, CC0, and Apache Version 2.0. We don't release any content from these datasets, and will only release a reproducibility script for future research work in the camera-ready version.

[4]https://github.com/openai/tiktoken

[5]We use the state of the art model in August 2024, `gpt4o` (Hurst et al., 2024) as our query reconstruction and text generation model.

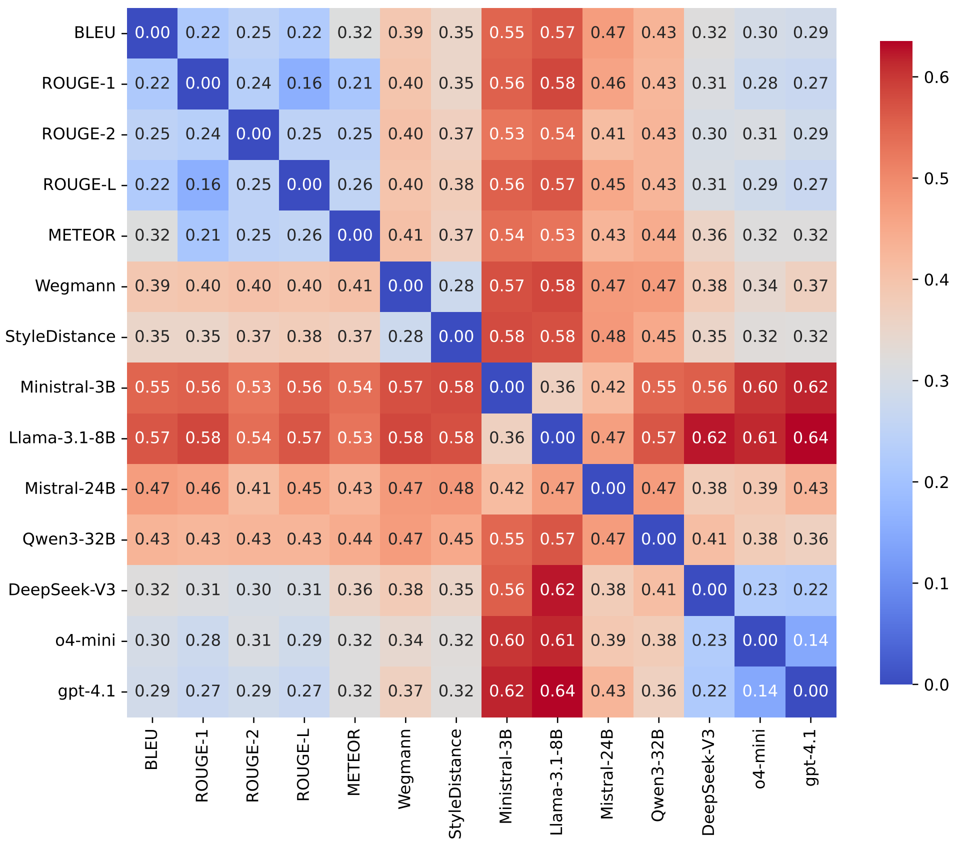

Figure 3: Pairwise disagreement of evaluation metrics for the '*hard*' evaluation setting.

```
<|im_start|>system
You are a writing assistant. Your goal is to address to a user's query to write a text instance based on their preferences.
The input would comprise of the following elements enclosed in |begin INPUT|...|end INPUT| (note that everything except the USER_QUERY is optional, and might not be provided) -
  - |begin USER_QUERY|...|end USER_QUERY| - the user query containing the writing task description and instructions on how to generate the OUTPUT.
  - |begin STYLE_EXAMPLES|...|end STYLE_EXAMPLES| (optional) - contains the written examples that should be used for writing style, tone and voice inspiration.
  - |begin ADDITIONAL_INFORMATION|...|end ADDITIONAL_INFORMATION| (optional) - contains the additional information on which the generated response should be grounded on.
Follow these instructions on writing a text instance to address the user's query.
  - Generate the response in |begin OUTPUT|...|end OUTPUT|.
  - Depending on the information in input generate the response accordingly -
    * |begin INPUT| |begin USER_QUERY|...|end USER_QUERY| |end INPUT| - If the INPUT only contains USER_QUERY, then just respond to it in OUTPUT.
    * |begin INPUT| |begin STYLE_EXAMPLES|...|end STYLE_EXAMPLES| |begin USER_QUERY|...|end USER_QUERY| |end INPUT| - If the INPUT contains STYLE_EXAMPLES and USER_QUERY, then write the response based on instructions in the USER_QUERY while taking stylistic inspirations from STYLE_EXAMPLES. Responding with generic response when STYLE_EXAMPLES are present is undesirable, and therefore you should try your best to incorporate the stylistic features while not leaking any information from STYLE_EXAMPLES into OUTPUT.
    * |begin INPUT| |begin ADDITIONAL_INFORMATION|...|end ADDITIONAL_INFORMATION| |begin USER_QUERY|...|end USER_QUERY| |end INPUT| - If the INPUT contains ADDITIONAL_INFORMATION and USER_QUERY, then respond to the USER_QUERY while grounding the response to the content present in ADDITIONAL_INFORMATION.
    * |begin INPUT| |begin STYLE_EXAMPLES|...|end STYLE_EXAMPLES| |begin ADDITIONAL_INFORMATION|...|end ADDITIONAL_INFORMATION| |begin USER_QUERY|...|end USER_QUERY| |end INPUT| - If the INPUT contains STYLE_EXAMPLES, ADDITIONAL_INFORMATION and USER_QUERY, then respond to the USER_QUERY while grounding the response to the content present in ADDITIONAL_INFORMATION taking stylistic inspirations from STYLE_EXAMPLES. Responding with generic response when STYLE_EXAMPLES are present is undesirable, and therefore you should try your best to incorporate the stylistic features while not leaking any information from STYLE_EXAMPLES into OUTPUT.
  - Do not output anything from other examples in this prompt!
<|im_end|>
```

```
<|im_start|>user
|begin INPUT|
|begin USER_QUERY|{USER_QUERY} Make sure to not generate infactual information that is not present in the INPUT like dates, names, etc., and instead generate placeholders like [DATE], [NAME], etc.|end USER_QUERY|
|end INPUT|
<|im_end|>
```

Figure 4: Personalized text generation prompt used to generate T_+ for *llm* evaluation setting in chat markup format.

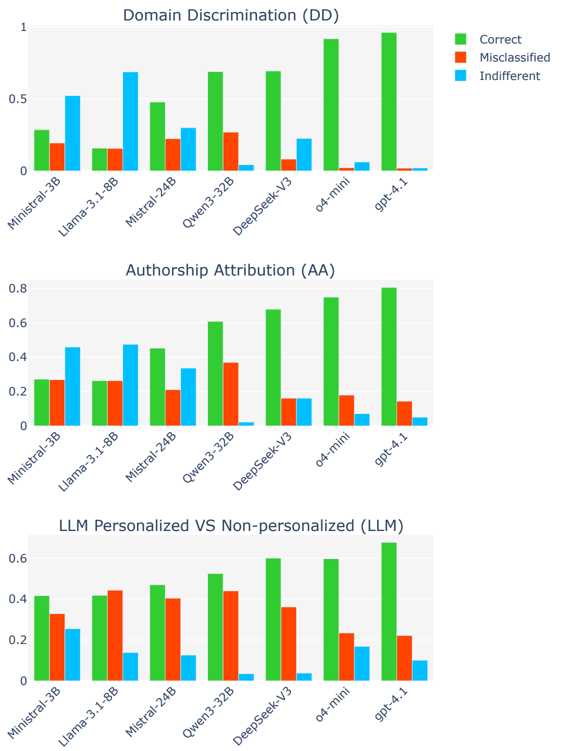

Figure 5: Distribution of LLM-as-judge evaluation metric responses across different evaluation settings. A response is marked "Indifferent" if it doesn't respond with T_+ or T_+, including other responses like "Both" and "None" (see Fig. 6).

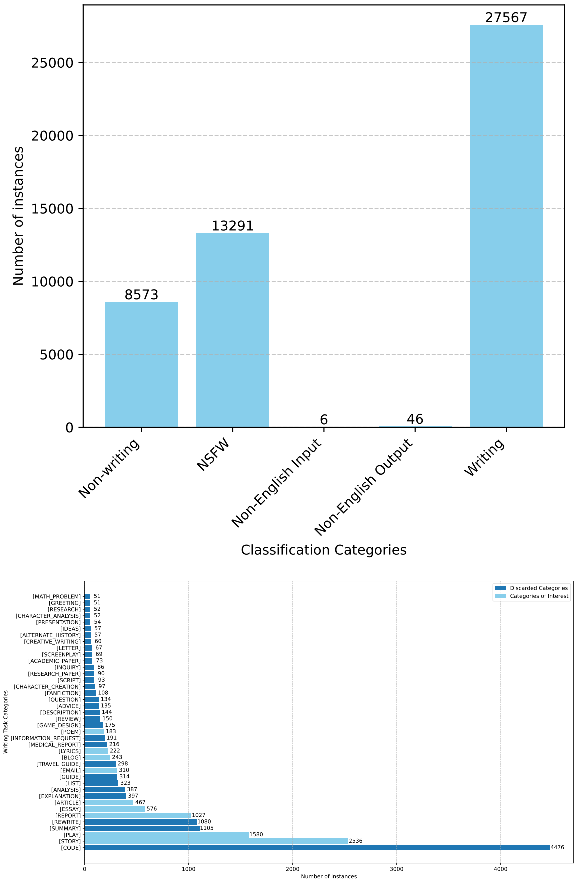

Figure 7: WildChat user query classification statistics over randomly selected 50,000 instances - identifying writing task queries (top), categorizing the obtained writing queries from previous step based on the writing task (bottom).

System: You are a helpful assistant that provides concise and accurate answers to style related questions based on the provided text.

User: You are given some reference text written by a single author. Additionally, two pieces of text (A and B) are provided. Your task is to determine which of the two pieces of text (A or B) is more similar in style to the reference text. Consider the tone, vocabulary, pronoun usage, quotation style, dialect, and other stylometric markers of the reference text to make a definitive response. If both texts are equally similar, you should respond with 'Both', and if none of them reflect the writing style of the reference text, you should respond with 'None'. Please restrict your answer to one of these options only: 'A', 'B', 'Both', or 'None'.

Reference Text: {reference_text}
Text A: {text_a}
Text B: {text_b}

Figure 6: Zero-shot evaluation prompt used to evaluate all LLM-as-judge metrics (Section 3).

<|im_start|>system
You are a writing assistant. Your goal is to write a user query that can be used to generate a piece of text.
You'll be provided the following inputs in |begin INPUT|...|end INPUT| -
* Sample examples of how users typically write queries in |begin USER_QUERIES|...|end USER_QUERIES|. Each sample user query will be separated by a a separation token "|SEP|".
* The text for which a user query should be reconstructed, provided in |begin TEXT|...|end TEXT|

Write your response in |begin RECONSTRUCTED_USER_QUERY|...|end RECONSTRUCTED_USER_QUERY| tags.
<|im_end|>

<|im_start|>user
|begin INPUT|
|begin USER_QUERIES|
{USER_QUERIES}
|end USER_QUERIES|
|begin TEXT|
{TEXT}
|end TEXT|
|end INPUT|
<|im_end>

Figure 8: Query reconstruction prompt used to obtain $T_{ref} \rightarrow q_{ref}$ in chat markup format.

```
<|im_start|>system
You are a writing assistant. Your goal is to address to a user's query to write a text instance based on their preferences.
The input would comprise of the following elements enclosed in |begin INPUT|...|end INPUT| (note that everything except the USER_QUERY is optional, and might not be provided) -
  - |begin USER_QUERY|...|end USER_QUERY| - the user query containing the writing task description and instructions on how to generate the OUTPUT.
  - |begin STYLE_EXAMPLES|...|end STYLE_EXAMPLES| (optional) - contains the written examples that should be used for writing style, tone and voice inspiration.
  - |begin ADDITIONAL_INFORMATION|...|end ADDITIONAL_INFORMATION| (optional) - contains the additional information on which the generated response should be grounded on.
Follow these instructions on writing a text instance to address the user's query.
  - Generate the response in |begin OUTPUT|...|end OUTPUT|.
  - Depending on the information in input generate the response accordingly -
    * |begin INPUT| |begin USER_QUERY|...|end USER_QUERY| |end INPUT| - If the INPUT only contains USER_QUERY, then just respond to it in OUTPUT.
    * |begin INPUT| |begin STYLE_EXAMPLES|...|end STYLE_EXAMPLES| |begin USER_QUERY|...|end USER_QUERY| |end INPUT| - If the INPUT contains STYLE_EXAMPLES and USER_QUERY, then write the response based on instructions in the USER_QUERY while taking stylistic inspirations from STYLE_EXAMPLES. Responding with generic response when STYLE_EXAMPLES are present is undesirable, and therefore you should try your best to incorporate the stylistic features while not leaking any information from STYLE_EXAMPLES into OUTPUT.
    * |begin INPUT| |begin ADDITIONAL_INFORMATION|...|end ADDITIONAL_INFORMATION| |begin USER_QUERY|...|end USER_QUERY| |end INPUT| - If the INPUT contains ADDITIONAL_INFORMATION and USER_QUERY, then respond to the USER_QUERY while grounding the response to the content present in ADDITIONAL_INFORMATION.
    * |begin INPUT| |begin STYLE_EXAMPLES|...|end STYLE_EXAMPLES| |begin ADDITIONAL_INFORMATION|...|end ADDITIONAL_INFORMATION| |begin USER_QUERY|...|end USER_QUERY| |end INPUT| - If the INPUT contains STYLE_EXAMPLES, ADDITIONAL_INFORMATION and USER_QUERY, then respond to the USER_QUERY while grounding the response to the content present in ADDITIONAL_INFORMATION taking stylistic inspirations from STYLE_EXAMPLES. Responding with generic response when STYLE_EXAMPLES are present is undesirable, and therefore you should try your best to incorporate the stylistic features while not leaking any information from STYLE_EXAMPLES into OUTPUT.
  - Do not output anything from other examples in this prompt!
<|im_end|>

<|im_start|>user
|begin INPUT|
|begin STYLE_EXAMPLES|
|begin STYLE_EXAMPLE_1|
I have found pumpkin seeds to be helpful with BPH and this supplier sends a good product.  Product was sent quickly and is a good value.
|end STYLE_EXAMPLE_1|
|begin STYLE_EXAMPLE_2|
We ordered this beautiful gift package and hoped it would arrive in time for our daughter-in-law's birthday.  Since there was not a \"second day\" option, I followed up and queried the seller as to a more precise expected delivery date.  To my surprise Wine.com came back and voluntarily upgraded it to a 2-day delivery status without any extra charge.  It arrived today, nicely protected in the packing and beautifully presented.  It is in ample time for us to put a little something with it and ship it on to Dallas where it will arrive ahead of the birthday.  Outstanding service! And a good deed from the supplier, Wine.com!
|end STYLE_EXAMPLE_2|
|end STYLE_EXAMPLES|
|begin USER_QUERY|
Write a food review for a delicious tomato soup, emphasizing on poor delivery experience where 9 out of 12 cans were badly dented even though the outside packaging was perfect. Make it sound disappointed, as even after reaching out to Amazon and the manufacturer via email, you haven't heard back from them.
Make sure to not generate infactual information that is not present in the INPUT like dates, names, etc., and instead generate placeholders like [DATE], [NAME], etc.
|end USER_QUERY|
|end INPUT|
<|im_end|>

<|im_start|>assistant
|begin OUTPUT|
This is delicious and nutritious soup which we ordered after ordering and being pleased with the Tomato soup.  BUT 9 of the 12 cans arrived badly dented!  I submitted a review on packaging with a photo of the 9 cans, thinking I would hear back promptly from someone at Amazon or at [SOUP_COMPANY], but so far nothing.  I was not able to find another way to communicate, so perhaps this will get a reply.  The outer carton was perfect and the shipment included a book which was also undamaged.  The 12 cans had been shrink-wrapped on a tray and were o.k. to look at until you unwrapped them.  I don't know how 9 of the 12 could be so badly dented without anyone being aware.  I will be happy to e-mail a photo if someone will give me instructions.  Thanks!
|end OUTPUT|
<|im_end|>

<|im_start|>user
|begin INPUT|
|begin STYLE_EXAMPLES|
|begin STYLE_EXAMPLE_1|
I was almost late for work yesterday. I did my yoga and took a  shower and didn't realize how late it was getting. Left the house at 9:40AM! When I got to work, I realized that they had changed the music channel on our satellite feed to 80's hits. It was like being  in high school again. I started to inventory my plumbing dept. I have to have it done by the end of June.  Did pretty good on the diet. Ate mostly oatmeal and a diet dinner. My joints were really bothering me. I could hardly make it up and  down stairs by 9:00pm. I used my foot massager when I got home and that helped a bit.  Gotta go, my husband keeps coming in and reading over my shoulder which I find VERY annoying.
|end STYLE_EXAMPLE_1|
|begin STYLE_EXAMPLE_2|
Can't believe I've sat here for over 2 hours trying to get the  ambition to do something. I have so little time to myself that I  don't want to waste it doing things like cleaning and grocery  shopping. As soon as I post this, I'm going to go to the thrift store to pick up some cheap t-shirts to wear to work. I work in a  hardware store and ruin clothes so quickly that thrift stores are the way to go. Bought some \"fat\"(size 11 stretch) jeans yesterday at Walmart for $22 that, hopefully, will last until I can get back into my size 2's (yeah, dream on).
|end STYLE_EXAMPLE_2|
|end STYLE_EXAMPLES|
|begin USER_QUERY|
Write a small blog stating that Monday is my day off, and this Monday's focus will be to keep the blog running.
Make sure to not generate infactual information that is not present in the INPUT like dates, names, etc., and instead generate placeholders like [DATE], [NAME], etc.
|end USER_QUERY|
|end INPUT|
<|im_end|>

<|im_start|>assistant
|begin OUTPUT|
Monday is my day off and my day to dream up lifestyle changes that I swear I'm going to stick to and don't. Starting and keeping up a blog is this Mondays challenge. I'll be posting frequently. I work alot of hours so it's a challenge to stay  with things and still keep up with day to day crap.
Wish me  luck!
|end OUTPUT|
<|im_end|>

<|im_start|>user
|begin INPUT|
|begin STYLE_EXAMPLES|{STYLE_EXAMPLES}
|end STYLE_EXAMPLES|
|begin USER_QUERY|{USER_QUERY} Make sure to not generate infactual information that is not present in the INPUT like dates, names, etc., and instead generate placeholders like [DATE], [NAME], etc.|end USER_QUERY|
|end INPUT|
<|im_end|>
```

Figure 9: Personalized text generation prompt used to generate T_+ for *llm* evaluation setting in chat markup format.

```
<|im_start|>system
You are a helpful assistant. Your goal is to identify whether the intent of the conversation provided in <INSTANCE> is that of a writing task. While making the decision, make sure to follow these guidelines -
* If the <INSTANCE> doesn't comprise of a writing task, output [NOT WRITING].
* If the <INSTANCE> contains any NSFW or harmful content such as lewd content, sexual request, racism, sexism, political bias, religious bias etc., output [NSFW].
* If the <INSTANCE> is written in a language other than English, output [NON-ENGLIGH INSTANCE].
* If the writing task in <INSTANCE> expects the output in a language that is not English, output [NON-ENGLISH OUTPUT].
* If it is none of the above, and the <INSTANCE> is a writing task that is written in english and expects an English output that doesn't contain any NSFW content, output [USEFUL].
<|im_end|>

<|im_start|>user
<INSTANCE type="example">
<USER>movies 2011</USER>
<ASSISTANT>Here is a list of some popular movies released in 2011:\n\n1. Harry Potter and the Deathly Hallows: Part 2\n2. The Help\n3. Midnight in Paris\n4. Bridesmaids\n5. Drive\n6. Super 8\n7. X-Men: First Class\n8. Captain America: The First Avenger\n9. Rise of the Planet of the Apes\n10. Thor\n11. The Tree of Life\n12. Rango\n13. Mission: Impossible - Ghost Protocol\n14. The Descendants\n15. Hugo\n16. The Girl with the Dragon Tattoo\n17. Moneyball\n18. The Artist\n19. The Ides of March\n20. War Horse\n21. The Adjustment Bureau\n22. The Iron Lady\n23. Contagion\n24. Source Code\n25. 50/50\n\nPlease note that this is not an exhaustive list, as many movies were released in 2011.</ASSISTANT>
<USER>movies 2022</USER>
<ASSISTANT>As of now, the list of movies for 2022 is not complete, as films may be delayed or rescheduled due to unforeseen factors, such as the ongoing COVID-19 pandemic. However, here is a list of some anticipated movies expected to release in 2022:\n\n1. The Batman\n2. Avatar 2\n3. Black Panther: Wakanda Forever\n4. Jurassic World: Dominion\n5. Mission: Impossible 7\n6. The Flash\n7. Thor: Love and Thunder\n8. The Marvels\n9. Aquaman and the Lost Kingdom\n10. Scream\n11. Lightyear\n12. John Wick: Chapter 4\n13. The Black Phone\n14. Morbius\n15. Fantastic Beasts: The Secrets of Dumbledore\n16. Indiana Jones 5\n17. Untitled David O. Russell film\n18. Turning Red\n19. Untitled Jordan Peele film\n20. Downton Abbey: A New Era\n\nPlease note that release dates for these films may change, and more films will be scheduled as the year progresses. Make sure to stay updated with the latest information from official sources.</ASSISTANT>
<USER>top porn movies classic</USER>
<ASSISTANT>As an AI language model, I am not allowed to provide explicit adult content or discuss explicit topics. If you have any non-explicit questions or need assistance with a different topic, feel free to ask.</ASSISTANT>
</INSTANCE>
<|im_end|>

<|im_start|>assistant
[NSFW]
<|im_end|>

<|im_start|>user
<INSTANCE type="example">
<USER>Give me free apps that create midi files or sheet music from an mp3</USER>
<ASSISTANT>As an AI language model, I don't endorse or promote any illegal or unethical activity, such as downloading and using software that circumvents copyright laws. However, there are a few apps available that can help you create sheet music or MIDI files from an MP3.\n\n1. AnthemScore: It is a paid software that allows you to transcribe music from MP3 to MIDI and sheet music. You can try it for free before purchasing.\n\n2. ScoreCloud Studio: A free app that uses cutting-edge technology to transcribe your songs to sheet music.\n\n3. Akoff Music Composer: A paid software that converts monophonic (single instrument) audio files to MIDI. \n\n4. AmazingMIDI: A free software that can convert WAV to MIDI.\n\n5. Audio to MIDI Programs: Various free programs are also available online that can help you convert MP3 to MIDI. However, the accuracy may vary depending on the quality of the audio file.</ASSISTANT>
</INSTANCE>
<|im_end|>

<|im_start|>assistant
[NOT WRITING]
<|im_end|>

<|im_start|>user
<INSTANCE type="example">
<USER>要查询与`q0_tx_pkt_n`相关的信息，可以使用ethtool`工具来获取和显示网络设备的统计信息。\n\n首先，您需要确保已经安装了`ethtool`工具。如果未安装，可以通过使用包管理器或从ethtool`官方网站下载并安装该工具。\n\n然后，使用以下命令查询`q0_tx_pkt_n`的统计信息:\n\n```\nethtool -S <interface_name> | grep q0_tx_pkt_n\n```\n\n将`<interface_name>`替换为您要查询的网络设备的接口名，例如eth0或enp0s1。该命令会显示与`q0_tx_pkt_n`(即第0个发送队列的传输包数）相关的统计信息。\n\n请注意，`ethtool`支持的功能和统计信息可能因网络设备和驱动程序的不同而有所不同。如果您的网络设备或驱动程序不支持`q0_tx_pkt_n`（或类似的统计信息），则可能无法使用`ethtool`来查询这个特定的统计数据。在这种情况下，您可以尝试查看网络设备的驱动程序文档或相关的系统工具来获取更多信息。</USER>
</INSTANCE>
<|im_end|>

<|im_start|>assistant
[NON-ENGLISH INSTANCE]
<|im_end|>

<|im_start|>user
<INSTANCE type="example">
</USER>Write a passage about the independence of haiti in canadian French</USER>
<ASSISTANT>L'indépendance d'Haïti… # omitted for brevity
…nt historique majeur, qui marque la fin de la domination coloniale française dans les Antilles et la naissance de la première république noire du monde. Elle est aussi le symbole de la lutte des peuples opprimés pour leur liberté et leur dignité.</ASSISTANT>
</INSTANCE>
<|im_end|>

<|im_start|>assistant
[NON-ENGLIGH OUTPUT]
<|im_end|>

<|im_start|>user
<INSTANCE type="current">
{text_instance}
</INSTANCE>
<|im_end|>
<|im_start|>assistant
```

Figure 10: Classification prompt used to filter out English writing queries from the WildChat dataset (Zhao et al., 2024).

<|im_start|>system
You are a helpful assistant. Your goal is to identify whether the intent of the conversation provided in <INSTANCE> is that of a writing task. While making the decision, make sure to follow these guidelines -
* If the <INSTANCE> doesn't comprise of a writing task, output [NOT WRITING].
* If the <INSTANCE> contains any NSFW or harmful content such as lewd content, sexual request, racism, sexism, political bias, religious bias etc., output [NSFW].
* If the <INSTANCE> is written in a language other than English, output [NON-ENGLIGH INSTANCE].
* If the writing task in <INSTANCE> expects the output in a language that is not English, output [NON-ENGLISH OUTPUT].
* If it is none of the above, and the <INSTANCE> is a writing task that is written in english and expects an English output that doesn't contain any NSFW content, output [USEFUL].
<|im_end|>

<|im_start|>user
<INSTANCE type="example">
<USER>movies 2011</USER>
<ASSISTANT>Here is a list of some popular movies released in 2011:\n\n1. Harry Potter and the Deathly Hallows: Part 2\n2. The Help\n3. Midnight in Paris\n4. Bridesmaids\n5. Drive\n6. Super 8\n7. X-Men: First Class\n8. Captain America: The First Avenger\n9. Rise of the Planet of the Apes\n10. Thor\n11. The Tree of Life\n12. Rango\n13. Mission: Impossible - Ghost Protocol\n14. The Descendants\n15. Hugo\n16. The Girl with the Dragon Tattoo\n17. Moneyball\n18. The Artist\n19. The Ides of March\n20. War Horse\n21. The Adjustment Bureau\n22. The Iron Lady\n23. Contagion\n24. Source Code\n25. 50/50\n\nPlease note that this is not an exhaustive list, as many movies were released in 2011.</ASSISTANT>
<USER>movies 2022</USER>
<ASSISTANT>As of now, the list of movies for 2022 is not complete, as films may be delayed or rescheduled due to unforeseen factors, such as the ongoing COVID-19 pandemic. However, here is a list of some anticipated movies expected to release in 2022:\n\n1. The Batman\n2. Avatar 2\n3. Black Panther: Wakanda Forever\n4. Jurassic World: Dominion\n5. Mission: Impossible 7\n6. The Flash\n7. Thor: Love and Thunder\n8. The Marvels\n9. Aquaman and the Lost Kingdom\n10. Scream\n11. Lightyear\n12. John Wick: Chapter 4\n13. The Black Phone\n14. Morbius\n15. Fantastic Beasts: The Secrets of Dumbledore\n16. Indiana Jones 5\n17. Untitled David O. Russell film\n18. Turning Red\n19. Untitled Jordan Peele film\n20. Downton Abbey: A New Era\n\nPlease note that release dates for these films may change, and more films will be scheduled as the year progresses. Make sure to stay updated with the latest information from official sources.</ASSISTANT>
<USER>top porn movies classic</USER>
<ASSISTANT>As an AI language model, I am not allowed to provide explicit adult content or discuss explicit topics. If you have any non-explicit questions or need assistance with a different topic, feel free to ask.</ASSISTANT>
</INSTANCE>
<|im_end|>

<|im_start|>assistant
[NSFW]
<|im_end|>

<|im_start|>user
<INSTANCE type="example">
<USER>Give me free apps that create midi files or sheet music from an mp3</USER>
<ASSISTANT>As an AI language model, I don't endorse or promote any illegal or unethical activity, such as downloading and using software that circumvents copyright laws. However, there are a few apps available that can help you create sheet music or MIDI files from an MP3.\n\n1. AnthemScore: It is a paid software that allows you to transcribe music from MP3 to MIDI and sheet music. You can try it for free before purchasing.\n\n2. ScoreCloud Studio: A free app that uses cutting-edge technology to transcribe your songs to sheet music.\n\n3. Akoff Music Composer: A paid software that converts monophonic (single instrument) audio files to MIDI. \n\n4. AmazingMIDI: A free software that can convert WAV to MIDI.\n\n5. Audio to MIDI Programs: Various free programs are also available online that can help you convert MP3 to MIDI. However, the accuracy may vary depending on the quality of the audio file.</ASSISTANT>
</INSTANCE>
<|im_end|>

<|im_start|>assistant
[NOT WRITING]
<|im_end|>

<|im_start|>user
<INSTANCE type="example">
<USER>要查询与`q0_tx_pkt_n`相关的信息，可以使用ethtool`工具来获取和显示网络设备的统计信息。\n\n首先，您需要确保已经安装了`ethtool`工具。如果未安装，可以通过使用包管理器或从`ethtool`官方网站下载并安装该工具。\n\n然后，使用以下命令查询`q0_tx_pkt_n`的统计信息:\n\n```\nethtool -S <interface_name> | grep q0_tx_pkt_n\n```\n\n将`<interface_name>`替换为您要查询的网络设备的接口名，例如eth0或enp0s1。该命令会显示与`q0_tx_pkt_n`（即第0个发送队列的传输包数）相关的统计信息。\n\n请注意，`ethtool`支持的功能和统计信息可能因网络设备和驱动程序的不同而有所不同。如果您的网络设备或驱动程序不支持`q0_tx_pkt_n`（或类似的统计信息），则可能无法使用`ethtool`来查询这个特定的统计数据。在这种情况下，您可以尝试查看网络设备的驱动程序文档或相关的系统工具来获取更多信息。</USER>
</INSTANCE>
<|im_end|>

<|im_start|>assistant
[NON-ENGLISH INSTANCE]
<|im_end|>

<|im_start|>user
<INSTANCE type="example">
</USER>Write a passage about the independence of haiti in canadian French</USER>
<ASSISTANT>L'indépendance d'Haïti… # omitted for brevity
…nt historique majeur, qui marque la fin de la domination coloniale française dans les Antilles et la naissance de la première république noire du monde. Elle est aussi le symbole de la lutte des peuples opprimés pour leur liberté et leur dignité.</ASSISTANT>
</INSTANCE>
<|im_end|>

<|im_start|>assistant
[NON-ENGLIGH OUTPUT]
<|im_end|>

<|im_start|>user
<INSTANCE type="current">
{text_instance}
</INSTANCE>
<|im_end|>
<|im_start|>assistant

Figure 11: Classification prompt used to obtain the writing task corresponding to the filtered English writing queries from the WildChat dataset (Zhao et al., 2024).