Evaluating Style-Personalized Text Generation: Challenges and Directions

Anonymous ACL submission

Abstract

With the surge of large language models (LLMs) and their ability to produce customized output, style-personalized text generation–"write like me"–has become a rapidly growing area of interest. However, style personalization is highly specific, relative to every user, and depends strongly on the pragmatic context, which makes it uniquely challenging. Although prior research has introduced benchmarks and metrics for this area, they tend to be non-standardized and have known limitations (e.g., poor correlation with human subjects). LLMs have been found to not capture author-specific style well, it follows that the metrics themselves must be scrutinized carefully. In this work we critically examine the effectiveness of the most common metrics used in the field, such as BLEU, embeddings, and LLMs-as-judges. We evaluate these metrics using our proposed style discrimination benchmark, which spans eight diverse writing tasks across three evaluation settings: *domain discrimination*, *authorship attribution*, and *LLM-generated personalized vs non-personalized discrimination*. We find strong evidence that employing ensembles of diverse evaluation metrics consistently outperforms single-evaluator methods, and conclude by providing guidance on how to reliably assess style-personalized text generation.

1 Introduction

Large language models have many capabilities that have powered their end-user adoption. One core feature is their ability to perform writing assistance in multiple domains, such as journalism (Diakopoulos, 2019); legal services (Magesh et al., 2024); academic writing (Khalifa and Albadawy, 2024; Nguyen et al., 2024); to name a few. They have also gained traction in more personal tasks, including drafting emails (Li et al., 2025), résumés and cover letters (Zinjad et al., 2024), and social media posts (Long et al., 2023; Jain et al., 2023). What joins these application domains together is their need to have some level of style personalization; and, in particular, style-personalized text generation (SPTG; Mysore et al. 2025). Even though SPTG is neither new nor exclusive to LLMs, it has not been until the latter's advent that this technology has been ubiquitous.

SPTG evaluation is complex: it is susceptible to user preferences, specificity, and shifts in the pragmatic context. Contemporary SPTG is even more challenging; thanks to the capabilities of LLMs, it is now *long-form* and *open-ended*. There is also a scarcity of high-quality, high-volume datasets. All of this, in turn, makes SPTG notoriously hard to evaluate. In response, the research community has pushed to develop new metrics (Li et al., 2024a; Alhafni et al., 2024) and benchmarks (Salemi et al., 2023; Kumar et al., 2024). However, these measurement practices are disjoint and lack standardization. Existing works rely on metrics such as n-gram overlap metrics (e.g., BLEU (Papineni et al., 2002)), embeddings, and LLMs-as-judges (Gu et al., 2024; Li et al., 2024b), even though they have known limitations (e.g, Reiter 2018; De Wynter 2025a). Since LLMs themselves are known to be ineffective at emulating author-specific content (Bhandarkar et al., 2024), it raises the question as to whether the metrics used were truly measuring what was intended; or, better yet, how can the measurement process in SPTG be improved. For this, there is a need to evaluate **to what extent these metrics are effective at measuring contemporary SPTG**.

In this work, we evaluate the style discrimination capabilities of various metrics alongside two newer paradigms: style-aware embeddings (Wegmann et al., 2022; Patel et al., 2025) and LLM-as-judge approaches. Our focus is on open-ended, long-form SPTG. To resemble real-world conditions, we define a low-resource style discrimina-

tion setting, where a limited amount of reference style text (fewer than 1,500 words) is available. Our analysis spans eight diverse writing tasks and three evaluation settings, including ensembles of these metrics. To our knowledge, our work is the first to critically evaluate common and contemporary style-personalized text generation (SPTG) metrics. Our main contributions are as follows:

1. We introduce a diverse evaluation benchmark spanning eight diverse writing tasks and three evaluation settings to rigorously assess the style discrimination task[1].
2. We conduct a comprehensive assessment of three major metric paradigms for SPTG: ngram-based metrics, style-embeddings, and LLM-as-judges, highlighting their respective strengths and limitations.
3. We provide empirical evidence that ensembles of weak evaluation metrics can produce more robust judgments than any individual metric alone.
4. We analyze the practical challenges of obtaining high-quality human annotations for SPTG evaluation.

2 Background

2.1 Traditional (non-LLM) measurements

SPTG builds upon early work in text style transfer. For a comprehensive overview of this area, see the surveys by Hu et al. (2022); Mukherjee et al. (2024); and Liang et al. (2024). Text style transfer is usually evaluated in three dimensions: fluency, style accuracy, and content preservation (Briakou et al., 2021). There are a variety of metrics to assess style accuracy, ranging from simple n-gram overlap measures (e.g. `BLEU`; `ROUGE`, Lin 2004; `METEOR`, Banerjee and Lavie 2005), to more sophisticated approaches. The most common, non-LLM-based metrics are the Wasserstein distance over style-specific lexicons (Pele and Werman, 2009; Mir et al., 2019) and style-specific classifier scores (Fu et al., 2018; Krishna et al., 2020; Jangra et al., 2022). Most of these efforts focused on sentence-level style transfer, which **does not align with the demands of contemporary SPTG** in long-form, low-resource scenarios for two main reasons: (a) they typically require larger parallel or pseudo-parallel datasets, which are not available in contemporary SPTG tasks; and (b) their style assessments are designed for short, isolated sentences and therefore cannot capture higher-level stylistic phenomena critical for long-form style discrimination, such as discourse structure, narrative coherence, and co-reference style.

More generally, it is crucial to note that **n-gram overlap measures have several shortcomings**, such as not being fully applicable to text generation (Liu et al., 2016), nor correlating well with human judgments (Gehrmann et al., 2023; Reiter, 2018). These issues also exist in style transfer, and range from being prone to exploitation–for example, by sampling random sentences from the target distribution to inflate style-accuracy scores (Krishna et al., 2020)–to having weak correlations with human judgments (Pang and Gimpel, 2018). However, **ensembling various metrics has been found to improve performance**, at least in machine translation (Stefanik et al., 2021). In our work, we integrate this latter finding within our evaluations.

2.2 LLM-based generation and measurement

LLMs can perform long-form text generation from minimal grounding data, which makes measurement of LLM-based systems challenging, chiefly because of their sensitivity to the prompt (De Wynter, 2025a). **LLM-based SPTG** is no exception; however, its evaluation is further complicated by the coexistence of **legacy n-gram metrics and contemporary LLM-as-judge approaches**, which were designed under different assumptions about how generation quality should be measured, especially for long-form text. Works relying on n-gram-based methods are ubiquitous. For example, previous work applying LLMs to SPTG in prompt rewriting (Li et al., 2024a); steerage through inference (Zhang et al., 2025); and retrieval-augmented personalization (Mysore et al., 2023) were measured with (in order) `BLEU` and `ROUGE-1,2,L`; `ROUGE-1,2` and `BertScore-F1` (Zhang et al., 2020); and `ROUGE-L` and `METEOR`. Benchmarks such as LaMP (Salemi et al., 2023) and LongLaMP (Kumar et al., 2024) also use `ROUGE-1,L`; with `METEOR` added to the latter. Alternate approaches use embedding-based approaches (Alhafni et al., 2024; Horvitz et al., 2024; Shang et al., 2019); policy optimization (Liu et al., 2024b); and LLMs-as-judges (Ostheimer et al., 2024; Lai et al., 2023; Logacheva et al., 2022). Nonetheless, an evaluation by Bhandarkar et al. (2024) revealed **significant limitations on an**

[1]The code will be made available in the camera-ready draft.

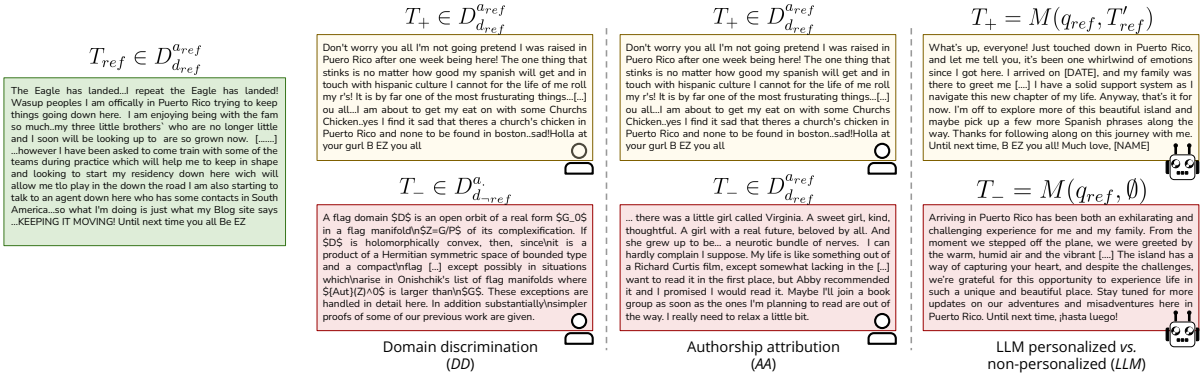


Figure 1: Illustration of the style discrimination task. Given a reference text $T_{ref} \in D_{d_{ref}}^{a_{ref}}$ (texts written by author a_{ref} in domain d_{ref}), the goal is to identify which candidate text (T_+ or T_-) is stylistically closer. The three panels show evaluation settings obtained by varying author/domain constraints: domain discrimination (*DD*), authorship attribution (*AA*), and LLM generation, where candidates are produced by a model M with or without style grounding.

LLM's ability to emulate author-specific styles, thus suggesting that the measurements above were insufficient. Their own evaluation metrics were five variants on authorship attribution classifiers in *high-resource* texts (at least 50,000 words per author), and cannot be adapted to the low-resource SPTG setting in our work, which uses 100-1000 words for style discrimination.

Extending metrics to long-form SPTG is non-trivial, as several evaluation dimensions commonly used in text style transfer become less informative in the context of modern LLMs For instance, fluency is no longer a meaningful discriminative dimension, as contemporary LLM-generated outputs in our setting are generally fluent. Similarly, content preservation has been extensively studied in prior work (Lyu et al., 2024) and is not the primary challenge in contemporary SPTG. On the other hand, the difficulty of measuring LLMs, the limitations around known metrics, and the findings on the capabilities of LLMs to emulate style call for a deeper understanding of *what* is measured in SPTG. We take a pragmatic approach to it and focus solely on style accuracy.

3 Problem Statement

For an author a who wrote texts in the domain d, let D_d^a denote the set of said texts. Then, given a reference text $T_{ref} \in D_{d_{ref}}^{a_{ref}}$, and two candidate texts $T_+ \in D_{d_1}^{a_1}$ and $T_- \in D_{d_2}^{a_2}$, the *style discrimination task* aims to identify which candidate is stylistically closer to the reference. The choice of a_1, d_1 and a_2, d_2 alters the evaluation setting (e.g., authorship attribution will consider $d_1 = d_2$ but $a_1 \neq a_2$), and are defined in Section 4. Assuming an underlying style distribution S_x for each text T_x, such that $S_{ref} \approx S_+$ and $S_{ref} \not\approx S_-$, the goal is to learn a discriminator $f : (T_{ref}, T_+, T_-) \mapsto \{+, -\}$ that predicts whether T_+ or T_- is more similar in style to T_{ref}. This framing is thus a binary classification problem to enable robust style measurement. Hence, we reduce this problem to determining if the metrics *can* measure style accuracy, rather than to what extent.

4 Experimental Settings

4.1 Datasets

Source dataset selection. We selected eight publicly available NLP datasets encompassing diverse writing tasks and domains. They were chosen based on the following criteria: (i) they must include multiple writing instances produced by the same author, (ii) they must represent realistic and well-defined writing tasks, and (iii) they must be openly accessible and licensed for research use. The datasets span diverse writing tasks, namely, creative writing (*short stories*, Carney and Robertson 2019; *lyrics*, Edenbd 2020) formal writing (*Reuters news*, Lewis 1987; *Enron emails*, Klimt and Yang 2004; *ArXiv scientific abstracts*, arXiv.org submitters 2024), and informal writing (*blogs*, Schler et al. 2006; *Amazon food reviews*, McAuley and Leskovec 2013; *Reddit microblogs*, Patel et al. 2022).

Domain-level down-sampling. We randomly sampled instances from each dataset based on the number of tokens written by each author. We used

Table 1: Statistics for our extended style personalization evaluation benchmark across the three evaluation settings - *domain discrimination*, *authorship attribution*, and *LLM-generated personalized vs non-personalized*, 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 Generated*
Amazon	100	157/267/388	161/267/263	267/272/300
ArXiV	100	144/224/447	144/224/229	224/406/505
Blogs	95	173/329/373	164/329/349	329/369/396
Enron	100	123/328/520	119/328/334	328/193/219
Lyrics	95	210/278/495	211/278/288	278/308/315
Reddit	56	87/38/410	93/38/38	38/222/261
Reuters	49	363/506/461	366/506/507	506/416/456
Short Stories	41	903/1289/267	911/1289/1364	1289/543/515

the `tiktoken` library[2] to determine the number of tokens in each article. For each author a in domain d, we build $\mathcal{D}_d^a$ by randomly sampling the articles written by the author. We set an upper limit of 500 tokens for the style reference text (D_d^a), and randomly selected target articles with 200 to 700 tokens. *Reddit* and *short stories* domains had different token lengths due to their non-standard text lengths. *Reddit* domain instances were too short, and hence we removed the lower limit for the target article selection. The instances in the *short stories* domain were too long, and we increased the upper limit of grounding articles to 1500 tokens and target articles to 2000 tokens. After this sampling step, we obtained 50 instances for *Reddit*, *Reuters*, and *short stories* domains, and 100 instances for the remaining domains. We removed a few examples from the dataset where the personalized text generation model did not yield an appropriate response[3], leading to a final dataset comprised of 636 instances. Table 1 summarizes the overall statistics of the evaluation benchmark.

4.2 Evaluation settings

To assess style discrimination under different conditions (see Figure 1 for example), we define three evaluation settings based on the source of the candidate set $\{T_+, T_-\}$. For a given reference text $T_{ref} \in D_{d_{ref}}^{a_{ref}}$, the candidates T_+ and T_- are randomly sampled as follows-

Domain discrimination (*DD*): sample $T_+ \sim D_{d_{ref}}^{a_{ref}}$ with $T_+ \neq T_{ref}$ and $T_- \sim D_{d_{\neg ref}}^{a_{\cdot}}$.

Authorship attribution (*AA*): sample $T_+ \sim D_{d_{ref}}^{a_{ref}}$ with $T_+ \neq T_{ref}$ and $T_- \sim D_{d_{ref}}^{a_{\neg ref}}$.

LLM personalized versus non-personalized (*LLM*): suppose a generative model maps an author's queries (prompts) and sample texts to output texts, $M\colon \mathbf{Q}_d^a \times D_{d'}^{a'} \rightarrow D_d^a$. Then a query reconstruction function $f_{query} : D_d^a \rightarrow \mathbf{Q}_d^a$ is an approximate inverse function $f_{query} \approx M^{-1}$ that takes in outputs and returns queries. f_{query} was implemented using an LLM as a meta-prompting approach (De Wynter et al., 2024). To resemble human-written queries, we used queries from `WildChat` (Zhao et al., 2024). For that, we randomly selected 50,000 English-language user interactions and performed two consecutive filtering steps using text classification: i) discarding all the user prompts that do not correspond to writing requests, and ii) discarding the writing prompts of irrelevant writing tasks for our dataset domains. The filtered user queries are used to generate the reconstructed user query q_{ref}. Using a generation model M, we then generate $T_+ = M(q_{ref}, T'_{ref})$ and $T_- = M(q_{ref}, \emptyset)$, where $T'_{ref} \sim D_{d_{ref}}^{a_{ref}}$ such that $T'_{ref} \neq T_{ref}$. See Appendices A for detailed call parameters and C for the prompts used.

4.3 Evaluation metrics

N-gram overlap-based evaluation metrics. We explore three widely adopted n-gram evaluation metrics: `BLEU`, `ROUGE`, and `METEOR`. For a n-gram similarity measure $Sim\colon D_d^a \times D_{d'}^{a'} \rightarrow [0, 1]$, we obtain the binary label as $\mathbb{1}\left[Sim(T_{ref}, T_+) > Sim(T_{ref}, T_-)\right]$.

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

[3] We removed a few examples (specifically, 5 from *blogs*, 5 from *lyrics*, 1 from *Reuters* and 1 from *short stories*) because the generation model used to generate T_+ or T_- for the LLM evaluation setting did not generate responses due to its guardrailed output for plausible copyrighted input.

Style Embedding-based evaluation metrics. We evaluate `Wegmann`[4] (Wegmann et al., 2022), and `StyleDistance` (Patel et al., 2025) style embeddings. Using cosine similarity between the embeddings, we assign the binary label as $\mathbb{1}[Sim_{cos}(\vec{T_{ref}}, \vec{T_{+}}) > Sim_{cos}(\vec{T_{ref}}, \vec{T_{-}})]$.

LLM-as-judge evaluation metrics. We evaluate a range of open-source models, including `Ministral-3B`, `Llama-3.1-8B` (Grattafiori et al., 2024), `Mistral-24B`, `Qwen3-32B` (Yang et al., 2025), `DeepSeek-V3`[5] (Liu et al., 2024a). We also evaluate OpenAI's closed-source models `o4-mini` and `gpt-4.1` (Hurst et al., 2024), covering multiple model families and parameter sizes. See Appendices C and A for the evaluation prompt and call parameters, respectively.

4.4 Ensemble of evaluation metrics

We evaluate several ensembles of evaluation metrics as follows:

- ρ_{ngram}: ensembling over n-gram evaluation metrics.
- $\rho_{\neg LLM}$: ensembling over all non-LLM evaluation metrics.
- ρ_{LLM}: ensembling over LLM-as-judge evaluation metrics.
- ρ_{OS}: ensembling over all open-source evaluation metrics.
- ρ_{all}: ensembling over all evaluation metrics detailed in Section 4.3.

We explore two ensembling strategies: majority voting (MV) and performance-weighted voting (PWV). PWV ensemble strategy was implemented by using the normalized aggregate of individual metric scores with a threshold of 0.5 to obtain the binary decision.

5 Results

5.1 Performance of individual evaluation metrics

Table 2 summarizes the performance of individual evaluation metrics on our proposed evaluation benchmark. We observe notable trends in metric performance across different evaluation strategies and paradigms, which we discuss next.

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

[5]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 *DD*, *AA*, and *LLM*. Scores in **bold** and underline indicate the best and the second best metrics for each evaluation paradigm respectively.

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**
`ROUGE-1`	0.858	0.695	0.613	0.722
`ROUGE-2`	0.803	0.643	**0.634**	0.693
`ROUGE-L`	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**

Performance across evaluation settings Across nearly all evaluation metrics, performance was highest in the *DD* setting, followed by *AA*, and lowest in *LLM*. Over four metrics achieved over 0.9 accuracy in *DD*, with the highest performance of 0.961 accuracy by `gpt-4.1`. Comparatively, `gpt-4.1` achieved 0.807 accuracy for *AA* and 0.678 accuracy in *LLM*. On average, accuracy dropped by -16.2% from *DD* to *AA*, and -7.5% from *AA* to *LLM*.

Performance across evaluation metrics. Close-sourced LLM-as-judge metrics achieved the highest accuracy scores, with `gpt-4.1` attaining the highest overall accuracy (0.815), followed by `o4-mini` (0.755). Comparatively, the best performing n-gram metric, `BLEU`, achieved 0.733 accuracy, and the best performing style embedding metric, `StyleDistance`, achieved 0.722 accuracy. Across the three evaluation settings, the performance decline from *DD* to *LLM* was the highest for style embedding metrics (-38.3%), moderate for n-gram-based metrics (-26.7%), and least for LLM-as-judge metrics (-11.3%).

Model size and access type influenced performance: larger and closed-source language models consistently outperformed smaller models. The three smallest models evaluated had random or near-random performance. See Appendix B for a distribution of the LLM-as-a-judge responses.

5.2 Performance of ensemble of metrics

Table 3 presents the results of the best-performing ensemble combination across different metrics.

Table 3: Accuracy of best performing ensemble across the three evaluation settings. Scores in **bold** and underline indicate the best and the second best metrics for each evaluation paradigm respectively.

	Metric	Ensemble Composition	*DD*	*AA*	*LLM*	Mean
Individual	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**
MV	ρ_{ngram}	BLEU ROUGE2 ROUGEL	0.866	0.704	0.654	0.742
	$\rho_{\neg LLM}$	BLEU ROUGE2 ROUGEL Wegmann StyleDistance	0.914	0.725	0.682	0.774
	ρ_{LLM}	Qwen3-32B o4-mini gpt-4.1	0.937	0.781	0.640	0.786
	ρ_{OS}	BLEU rougeL Wegmann StyleDistance DeepSeek-V3	0.925	0.750	0.684	0.786
	ρ_{all}	BLEU ROUGEL Wegmann StyleDistance gpt-4.1	**0.962**	**0.786**	**0.697**	**0.815**
PWV	ρ_{ngram}	BLEU ROUGE1 ROUGE2 ROUGEL	0.862	0.714	0.654	0.743
	$\rho_{\neg LLM}$	BLEU ROUGE2 Wegmann StyleDistance	0.948	0.722	0.670	0.780
	ρ_{LLM}	Qwen3-32B DeepSeek-V3 o4-mini gpt4.1	0.945	0.772	0.662	0.793
	ρ_{OS}	BLEU ROUGEL Wegmann StyleDistance Qwen3-32B DeepSeek-V3	0.937	0.750	0.682	0.790
	ρ_{all}	BLEU ROUGE1 StyleDistance gpt-4.1	**0.967**	**0.802**	**0.693**	**0.821**

Ensembles generally outperformed their individual constituents; for example, the $\rho_{\neg LLM}$ majority-voting ensemble achieves +5.6% higher accuracy than its best-performing constituent, BLEU. Furthermore, *PWV* ensembles consistently yielded marginal improvements to their *MV* counterparts. The most effective ensembles exhibited a uniform distribution over different evaluation paradigms. See Figure 2 for a depiction of the pairwise disagreement on evaluation metrics in *AA*.

Table 4: Accuracy of various LLM-as-judge prompting strategies. P_{orig}, P_{struct}, and P_{binary} are the original, structured-parseable, and binary prompts, respectively. $P_X + APO$ is the prompt after APO on P_X.

Prompt Strategy	*DD*	*AA*	*LLM*	Mean
P_{orig}	0.279	0.383	0.476	0.379
$P_{orig} + APO$	0.504	0.463	0.448	0.472
P_{struct}	0.160	0.238	0.467	0.288
$P_{struct} + APO$	0.448	0.477	0.488	0.471
P_{binary}	**0.526**	0.491	0.514	**0.510**
$P_{binary} + APO$	0.453	**0.525**	**0.539**	0.506

5.3 Ablation: LLM-as-judges prompting

To examine the robustness of LLM-as-judge as an evaluation metric, we conducted a prompt sensitivity analysis over several prompt formulations. Since LLM-based evaluation can be sensitive to prompt wording, relying on a single prompt risks conflating metric performance with prompt-specific artifacts. Specifically, we further explored three prompt variants: (1) a parseable structured output format (P_{struct}), (2) a binary judgment format (P_{binary}), and (3) automatically optimized prompts obtained via Automatic Prompt Optimization (APO). For APO, we used the strategy proposed by Pryzant et al. (2023) to optimize the prompts using natural language gradients. For this we separated out a development set of 64 instances (8 instances per domain), and report the evaluation scores on the remaining 572 instances. We limited this exploration to the Llama-3.1-8b model for all evaluation settings, and leave the exploration over other models to future work. The results are in Table 4. P_{struct} consistently yielded worse performance than the original prompting strategy (P_{orig}), with a -24% decrease in overall performance. Comparatively, P_{binary} achieved +34.5% higher overall accuracy than P_{orig}. Using APO helped improve the performance of P_{orig} and P_{struct} by +24.5% and +63.5%, respectively, while yielding similar performance to P_{binary}. See Appendix C for the prompts used.

5.4 Ablation: human agreement

Although the focus of our work is automated metrics, we conducted a brief study to evaluate inter-annotator agreement (IAA) in human annotators (Appendix E). The goal was to better understand the relationship of SPTG characteristics across human judgments. We found that humans had high IAA as measured by the agreement ratio on content (0.779), moderate-to-high on style (0.641), and low on copy-editing (usefulness of the text; 0.400) preferences, respectively.

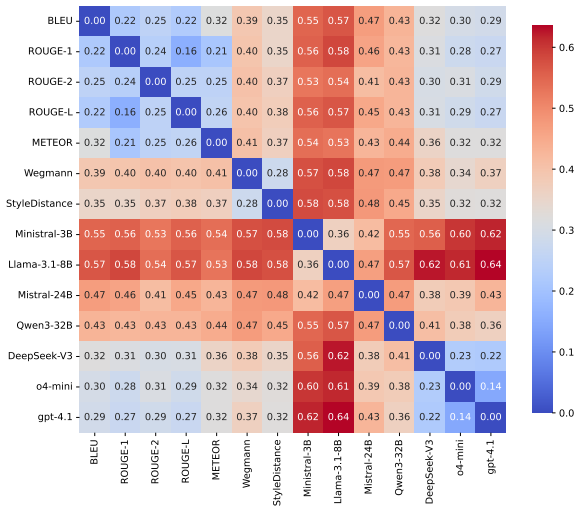


Figure 2: Pairwise disagreement of evaluation metrics for the *authorship attribution* evaluation setting. It can be seen from the figure that evaluation metrics across different evaluation paradigms have higher disagreement compared to metrics within the same evaluation paradigms. For example, even though `ROUGE-1` and `StyleDistance` achieve the same overall performance score of 0.722 (see Table 2); their disagreement score is 0.35, compared to the lower disagreement scores of `ROUGE-1` and `BLEU` (0.22) and `StyleDistance` and `Wegmann` (0.28).

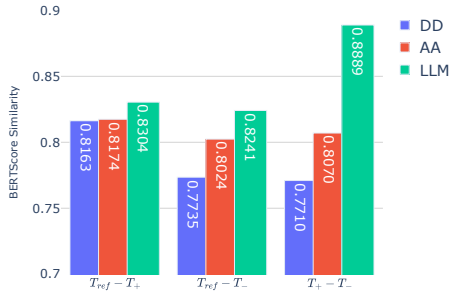


Figure 3: Average pairwise `BertScore` (Zhang et al., 2020) similarity across T_{ref}, T_+, and T_-. Semantic overlap between the candidate set T_+, T_- is significantly higher for the *LLM* compared to other two.

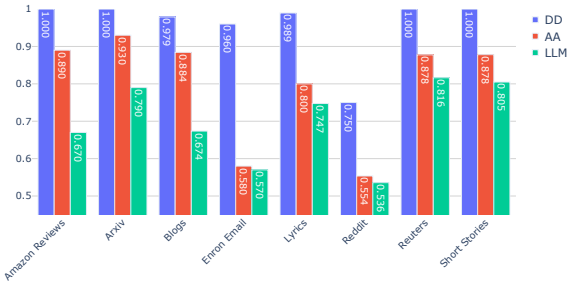


Figure 4: Accuracy of $\rho_{all}(PWV)$ across all domains. *Enron emails* and *Reddit microblogs* achieve lowest accuracy for AA and LLM evaluation settings, marginally outperforming the `Random` baseline.

6 Discussion

***LLM* as a challenging evaluation setting.** We attribute the lower performance of evaluation metrics in the LLM evaluation setting to two main factors: (a) both T_+ and T_- are generated by the same model for the same query (q_{ref}), differing only in the inclusion of reference style text for T_+. This is perhaps expected, as generation models exhibit inherent stylistic biases (Reinhart et al., 2025), and hence the outputs may share overlapping stylistic features, complicating discrimination. Also, (b), unlike DD and AA, where the input triplet $\{T_{ref}, T_+, T_-\}$ differs in content, all texts in the LLM setting correspond to the same writing task, making style evaluation inherently more difficult. As shown in Figure 3, the *LLM* setting also exhibits a substantial increase in semantic similarity as per its `BertScore` between T_+ and T_-. This further highlights the challenge of distinguishing stylistic differences in this scenario.

Analysis of best-performing ensembles To assess the contribution of each constituent metric, we computed the agreement between $\rho_{all}(PWV)$ and its individual components: `BLEU` (0.679), `ROUGE1` (0.675), `StyleDistance` (0.664), and `gpt-4.1` (0.780). All metrics contributed approximately equally, and the higher agreement of `gpt-4.1` can be attributed to the PWV strategy.

After investigating domain-wise accuracy for $\rho_{all}(PWV)$ (Figure 4), we observed that the *Reddit* subset yielded the poorest performance, even in DD. This is likely due to the short length of reference texts (Table 1). The accuracy further drops to 0.536 in the *LLM* setting, barely surpassing the `Random 100` baseline. A similar trend is in the *Enron* emails, which also contain short texts and substantial content overlap across authors, reflecting a homogeneous corporate email style. In other domains within *LLM*, there was a clear divide between standard writing domains (*arXiv*, *Reuters*, *short stories*) and informal domains (*Amazon food reviews*, *blogs*), with roughly a 0.1 accuracy gap. Manual inspection showed that generated responses often failed to capture certain stylistic features, such as ellipses (...), non-standard capitalization, and niche slang (see, e.g., Figure 17).

Unexpected strength of n-gram metrics. Contrary to the evidence from prior work in text style

transfer (Pang and Gimpel, 2018), our findings revealed that *n*-gram-based evaluation metrics performed well in the low-resource style discrimination task. We hypothesize that this could be due to the difference in evaluation setup, and attribute this to their evaluation focusing on sentence-level style transfer with limited reference n-grams. This, in turn, prevented the n-gram overlap metrics from capturing broader, more frequent, and nuanced stylistic features within longer spans of text. Our quantitative investigation of different writing domains supports this, as the metrics performed worse for domains with smaller reference texts compared to domains with longer reference texts.

Explaining performance decline across settings in style embedding-based metrics. From Table 2, we observe that the performance drop from the DD setting to the LLM setting is the highest for style embedding-based metric (-38.3%) compared to n-gram (-26.7%) and LLM-as-judge metrics (-11.3%). Both style embedding frameworks, `Wegmann` and `StyleDistance`, are trained with a contrastive learning objective to separate texts with varying style. This training paradigm leads to a significantly higher performance in the DD task. However, since the writing query for both candidates in the LLM setting is the same, these metrics fail to distinguish between the reference and candidate texts. This is further evident when we look at the statistics of the cosine similarity across the three evaluation settings. The mean and standard deviation for each evaluation setting are: DD : `Wegmann`$(\mu = 0.69, \sigma = 0.47)$, `StyleDistance`$(\mu = 0.19, \sigma = 0.13)$, AA : `Wegmann`$(\mu = 0.19, \sigma = 0.45)$, `StyleDistance`$(\mu = 0.05, \sigma = 0.12)$, and LLM : `Wegmann`$(\mu = 0.05, \sigma = 0.25)$, `StyleDistance`$(\mu = 0.01, \sigma = 0.05)$. We observe that the difference $Sim_{cos}(\vec{T_{ref}}, \vec{T_{+}}) - Sim_{cos}(\vec{T_{ref}}, \vec{T_{-}})$ is the highest for the DD setting, then AA, and then LLM; indicating that the style embeddings are least confident in their style discrimination predictions in LLM.

Limitations of LLM-as-judge Our analysis of different prompting strategies to improve LLM-as-judges yielded positive, yet inconsistent, outcomes. Positive outcomes were due to APO boosting accuracy by about +25%. However, remark that this, as an improvement, could be a false positive: the performance for in *AA* for the P_{struct} and P_{binary} prompts are 0.238 and 0.491, respectively, providing close to random performance in their output space. Hence, APO on these prompts also does not improve accuracy significantly, if at all. This is expected, since adaptive prompting strategies like APO have been shown to be prone to overfitting to the training instances (De Wynter, 2025b). Limiting the model output to a binary choice, instead of allowing responses like "Both" and "None", yielded about a +35% increase, although it did not improve the model's reasoning capability in the style discrimination task. It then follows that LLMs-as-judges are, as it is well-known, difficult to reliably apply to this task.

IAA Human Analysis The agreement amongst humans revealed that style preferences are more subjective than content preferences, which aligns with previous work (Salemi et al., 2025). Likewise, style-personalized preferences vary across individuals and per person. This can be seen in the low agreement found in copy-editing when compared to that of the other two preference tasks.

7 Conclusion

SPTG evaluation in the LLM era is challenging due to the dependency on pragmatics and the complexities associated with non-standardized evaluation metrics, intricate annotation requirements, and data scarcity.

Our experiments revealed several limitations in both widely adopted and newly investigated evaluation paradigms. Namely, **individual metrics did not have good agreement** in settings: although they were most effective at distinguishing texts across domains, they were only moderately effective at within-domain authorship attribution, and least effective at distinguishing personalized and non-personalized LLM-generated text. However, **ensembling metrics** matched or outperformed their individual constituents.

Based on our findings, we argue that more careful scrutiny is required when evaluating SPTG. This is reinforced by our findings in the human IAA ablation study: due to pragmatic dependencies, **the complexity of measuring SPTG is exacerbated by non-standardized, low-agreement metrics**. It then follows that emphasis should be on developing reliable metrics that better adjust to the *problem*, rather than the dataset evaluated, in order to properly measure SPTG and make research comparable.

Limitations

While we introduce a comprehensive evaluation benchmark for assessing style-personalized text generation, our work has three primary limitations. **First**, we focus exclusively on the binary classification task of discriminating between two candidate responses given a reference text, leaving the development of fine-grained and explainable evaluation systems to future work. We argued, however, that our focus was on correlation between metrics, and a more fine-grained measure would make agreement more complex and less reliable. **Second**, we only explore two ensembling approaches: majority voting and performance-weighted voting, leaving exploration of other ensembling techniques like likelihood-based voting, threshold-tuned voting and meta-ensemble learners to future work. Our findings did support that simple ensembling techniques performed better than other metrics, and thus we argue that more effective ensembles can only benefit SPTG measurement. **Third**, the binary task formulation does not fully capture the first-person perspective inherent to style-personalized text generation. It is difficult to build a diverse, user-specific first-person evaluation benchmark; especially when it is necessary that it be out-of-domain to counteract any potential memorization by LLMs. Instead, however, we adopted a set of simpler style discrimination tasks to evaluate both existing and proposed metrics.

Ethics

Our work is an analysis of the measurements for SPTG. We are unaware of any potential misuse of our work, although we recognize that we could not have covered all possible scenarios. All the datasets we collected were licensed under CC BY 4.0, CC BY-SA 4.0, CC0, and Apache Version 2.0. We do not release any content from these datasets, and will only release a reproducibility script for future research work in the camera-ready version. For our human annotation study, we recruited participants through a professional crowdsourcing platform. All participants were over 18 years old, and prior to the annotation study their consent was collected electronically. The annotators were compensated at an hourly rate of 12 USD and were advised to take breaks as needed to maintain focus and comfort throughout the task. This study was approved by the institutional review board of the authors' organization(s).

References

Nader Akoury, Shufan Wang, Josh Whiting, Stephen Hood, Nanyun Peng, and Mohit Iyyer. 2020. Storium: A dataset and evaluation platform for machine-in-the-loop story generation. *arXiv preprint arXiv:2010.01717*.

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, St. Julians, Malta. Association for Computational Linguistics.

arXiv.org submitters. 2024. arXiv dataset.

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, Ann Arbor, Michigan. Association for Computational Linguistics.

Avanti Bhandarkar, Ronald Wilson, Anushka Swarup, and Damon Woodard. 2024. Emulating author style: a feasibility study of prompt-enabled text stylization with off-the-shelf LLMs. In *Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)*, pages 76–82.

Eleftheria Briakou, Sweta Agrawal, Ke Zhang, Joel Tetreault, and Marine Carpuat. 2021. A review of human evaluation for style transfer. In *Proceedings of the First Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 58–67, Online. Association for Computational Linguistics.

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

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

Edenbd. 2020. 150k lyrics labeled with Spotify Valence.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2023. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *Journal of Artificial Intelligence Research*, 77:103–166.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. Falte: A toolkit for fine-grained annotation for long text evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 351–358.

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*.

Zachary Horvitz, Ajay Patel, Kanishk Singh, Chris Callison-Burch, Kathleen McKeown, and Zhou Yu. 2024. TinyStyler: Efficient few-shot text style transfer with authorship embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13376–13390, Miami, Florida, USA. Association for Computational Linguistics.

Zhiqiang Hu, Roy Ka-Wei Lee, Charu C Aggarwal, and Aston Zhang. 2022. Text style transfer: A review and experimental evaluation. *ACM SIGKDD Explorations Newsletter*, 24(1):14–45.

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.

Anubhav Jangra, Preksha Nema, and Aravindan Raghuveer. 2022. T-STAR: Truthful style transfer using amr graph as intermediate representation. *arXiv preprint arXiv:2212.01667*.

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.

Kalpesh Krishna, John Wieting, and Mohit Iyyer. 2020. Reformulating unsupervised style transfer as paraphrase generation. *arXiv preprint arXiv:2010.05700*.

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

Huiyuan Lai, Antonio Toral, and Malvina Nissim. 2023. Multidimensional evaluation for text style transfer using ChatGPT. *arXiv preprint arXiv:2304.13462*.

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.

Xun Liang, Hanyu Wang, Yezhaohui Wang, Shichao Song, Jiawei Yang, Simin Niu, Jie Hu, Dan Liu, Shunyu Yao, Feiyu Xiong, et al. 2024. Controllable text generation for large language models: A survey. *arXiv preprint arXiv:2408.12599*.

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. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Shuai Liu, Shantanu Agarwal, and Jonathan May. 2024b. Authorship style transfer with policy optimization. *arXiv preprint arXiv:2403.08043*.

Varvara Logacheva, Daryna Dementieva, Irina Krotova, Alena Fenogenova, Irina Nikishina, Tatiana Shavrina, and Alexander Panchenko. 2022. A study on manual and automatic evaluation for text style transfer: The case of detoxification. In *Proceedings of the 2nd Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 90–101.

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*.

Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. 2024. Towards faithful model explanation in nlp: A survey. *Computational Linguistics*, 50(2):657–723.

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.

Remi Mir, Bjarke Felbo, Nick Obradovich, and Iyad Rahwan. 2019. Evaluating style transfer for text. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 495–504, Minneapolis, Minnesota. Association for Computational Linguistics.

Sourabrata Mukherjee, Mateusz Lango, Zdenek Kasner, and Ondrej Dušek. 2024. A survey of text style transfer: Applications and ethical implications. *arXiv preprint arXiv:2407.16737*.

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.

Phil Ostheimer, Mayank Nagda, Marius Kloft, and Sophie Fellenz. 2024. Text style transfer evaluation using large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15802–15822, Torino, Italia. ELRA and ICCL.

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. 2025. StyleDistance: Stronger content-independent style embeddings with synthetic parallel examples. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8662–8685, Albuquerque, New Mexico. Association for Computational Linguistics.

Ofir Pele and Michael Werman. 2009. Fast and robust earth mover's distances. In *2009 IEEE 12th international conference on computer vision*, pages 460–467. IEEE.

Reid Pryzant, Dan Iter, Jerry Li, Yin Lee, Chenguang Zhu, and Michael Zeng. 2023. Automatic prompt optimization with "gradient descent" and beam search. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7957–7968, Singapore. Association for Computational Linguistics.

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.

Ehud Reiter. 2018. A structured review of the validity of BLEU. *Computational Linguistics*, 44(3):393–401.

Alireza Salemi, Julian Killingback, and Hamed Zamani. 2025. Expert: Effective and explainable evaluation of personalized long-form text generation. *arXiv preprint arXiv:2501.14956*.

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.

Mingyue Shang, Piji Li, Zhenxin Fu, Lidong Bing, Dongyan Zhao, Shuming Shi, and Rui Yan. 2019. Semi-supervised text style transfer: Cross projection in latent space. *arXiv preprint arXiv:1909.11493*.

Michal Stefanik, Vít Novotný, and Petr Sojka. 2021. Regressive ensemble for machine translation quality evaluation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1041–1048, Online. Association for Computational Linguistics.

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

Adrian de Wynter. 2025a. Awes, laws, and flaws from today's LLM research. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12834–12854, Vienna, Austria. Association for Computational Linguistics.

Adrian de Wynter. 2025b. Is in-context learning learning? *arXiv preprint arXiv:2509.10414*.

Adrian de Wynter, Xun Wang, Qilong Gu, and Si-Qing Chen. 2024. On meta-prompting. *arXiv preprint arXiv:2312.06562*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Jinghao Zhang, Yuting Liu, Wenjie Wang, Qiang Liu, Shu Wu, Liang Wang, and Tat-Seng Chua. 2025. Personalized text generation with contrastive activation steering. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7128–7141, Vienna, Austria. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

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 Call Parameters

For our meta-prompting approach we used `gpt4o` for both f_{query} and M, with temperature set to one. The calls were made in August 2024. For our LLMs-as-judges, we used `Ministral-3B`, `Llama-3.1-8B` (Grattafiori et al., 2024), `Mistral-Small-24B-Instruct-2501`, `Qwen3-32B` (Yang et al., 2025), `DeepSeek-V3-0324` (Liu et al., 2024a), `o4-mini` and `gpt-4.1` (Hurst et al., 2024). For classification, we set the temperature to zero whenever possible; and the maximum output token length to 16192 tokens or their maximum allowable[6]. For APO, we used the implementation by (Pryzant et al., 2023) with beam size 4, training batch size 64, search depth 6 and keep the rest of the hyperparameters the same as the original paper. All temperatures in the algorithm were set to 0.7, except during the classification phase, where we set it to zero. We use the same training and test sets across all experiments.

B LLM-as-a-Judge Responses

The class distribution (evaluation metric responses) for LLMs-as-judges are in Figure 5. It can be seen that the model performance increases with increase in number of parameters consistently across all three evaluation settings. Smaller models like `Ministral-3b` and `Llama-3.1-8b` generate a higher degree of indifferent responses (including responses like "Both" and "None") compared to larger models. We can also note that the misclassification error rate increases from the DD setting to the AA setting, and is the highest for the LLM setting.

C Prompts

C.1 Data Generation

The prompts for filtering are in Figures 13 and 14). The prompt to generate the reconstructed user query q_{ref} is in Figure 10). We provide the respective outcomes of filtration steps on the subset of `WildChat` user queries in Figures 15 and 16.

Finally, prompts for the texts to be annotated are in Figures 11 and 12 for the prompts.

[6]We set a high maximum output token length to accommodate thinking model responses, and used the same value for non-thinking models for comparable response.

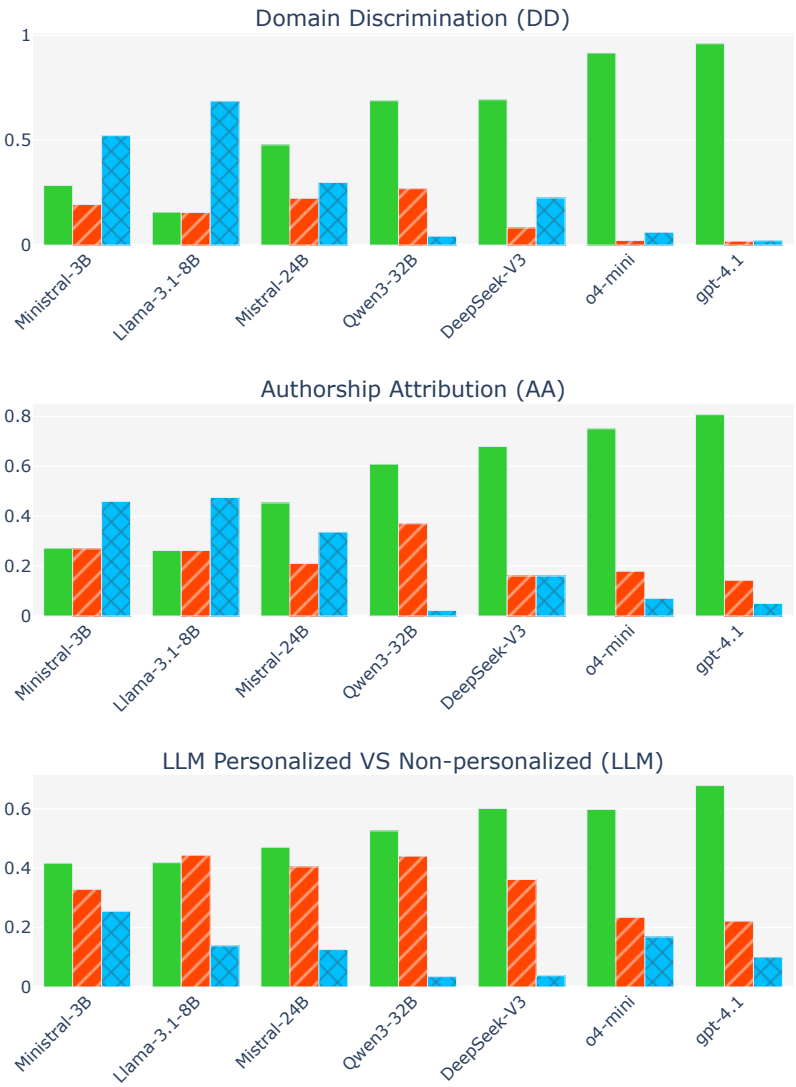


Figure 5: Distribution of LLM-as-judge evaluation metric responses across different evaluation settings. Correct classification is denoted by solid green bars, misclassification is denoted by diagonal red bars, and indifferent by cross-hatch blue bars. A response is marked "Indifferent" if it does not respond with T_+ or T_+, including other responses like "Both" and "None" (see Figure 7).

C.2 Evaluation

The prompt used for judging is in Figure 7. For exploration of different prompting strategies (Section 5.3), we use the prompts in Figures 7, 8, and 9 as P_{orig}, P_{struct} and P_{binary} respectively.

D Dataset Statistics

The data statistics for our filtered data are in Figures 15 and 16. From the first filtration step, we filter out 17% of the non-writing queries, and 26% of queries marked as not safe for work (NSFW), keeping 27,567 queries for the second filtration step. We select the top 40 most frequent labels from the second filtration step, and retain the queries that resemble the writing tasks of our style discrimination dataset. When selecting seed user queries for the query reconstruction function, f_{query}, we map *Amazon reviews* and *arXiv abstracts* domains to [REPORT] queries, *blog* domain to [BLOG] queries, *Enron emails* domain to [EMAIL] queries, *lyrics* domain to [LYRICS] queries, *short stories* domain to [STORY] and [PLAY], *Reddit microblogs* domain to [ESSAY] queries, and *Reuters news articles* to [ARTICLE] and [REPORT] queries.

We provide an example from our dataset for the *blog* domain in Figure 17, as well as the licensing details of the source datasets in Table 5.

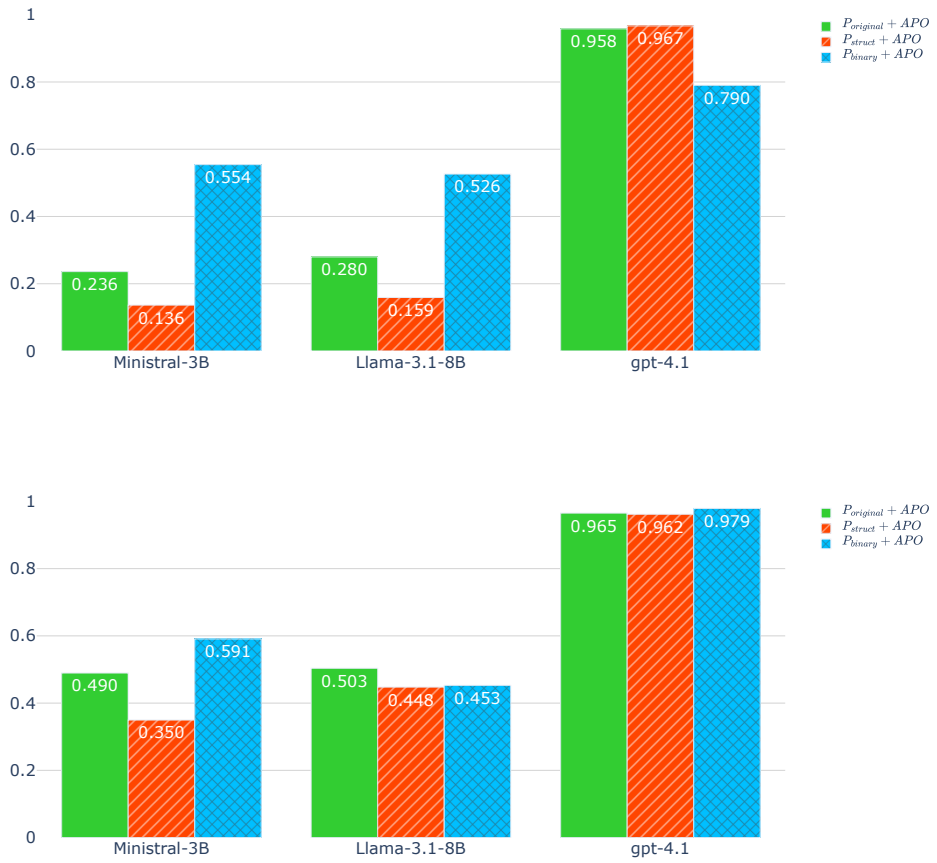


Figure 6: Ablation study results for LLM-as-judges prompting for the *domain discrimination (DD)* evaluation setting for `Ministral-3B`, `Llama-3.1-8b`, and `gpt-4.1` models. Top: LLM-as-judge results for different prompting strategies P_{orig} (solid green bars), P_{struct} (diagonal red bars), and P_{binary} (cross-hatch blue bars). Bottom: automatic prompt optimization LLM-as-judge results.

Table 5: Licensing status of individual data source.

Dataset Name	License
Amazon	CC0: Public Domain
ArXiV	CC0: Public Domain
Blogs	Unspecified[7]
Enron	CC-BY-3
Lyrics	CC BY-SA 4.0
Reddit	Apache License, Version 2.0
Reuters	CC BY 4.0
Short Stories	CC BY 4.0

E Challenges in Collecting Human Annotations for Long-Form SPTG

As mentioned in passing, SPTG is difficult to measure, in part due to the scarcity of data as well as the *personal* nature of the task. While the main body of our work focused on *automated* measures, we also conducted a brief two-part, side-by-side study on human preferences. On the first part, we measured human preference on *content preference* (which generated content better responded to the prompt); *style preference* (which content better adjusted to the style of the grounding text); and *copy-editing* (which of the outputs would be most useful). The second part was identical to the first, including the same annotators, although performed a month after the study.

[7]The Blog Authorship Corpus website states, "*The corpus may be freely used for non-commercial research purposes*".

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 7: Zero-shot evaluation prompt used to evaluate all LLM-as-judge metrics (Section 4.3). We refer to this prompt as P_{orig} in Section 5.3.

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', and provide your answer in the following JSON format: {{"answer": "<your_answer_here>"}}.

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

Figure 8: Zero-shot evaluation prompt (P_{struct}) used to evaluate parseable structured output in Section 5.3.

We found that, as measured by the agreement ratio, although human annotators had high agreement on content preference (0.779 IAA) and moderate-to-high on style preference (r. 0.641), their agreement on copy-editing preference was low (r. 0.400). Remark that it is *comparatively* low, since the label set is a five-label set and thus chance agreement would be 0.2. The second study reproduced the same ordering, albeit with slightly different numbers (0.736, 0.615, and 0.385 respectively).

This revealed two key findings: first, given that content preference was higher than style preference (and more reproducible), **style preferences are more subjective than content preferences**. This aligns with previous work on SPTG evaluation (Salemi et al., 2025). Second, **style-personalized preferences vary across individuals and per-person**. This can be seen in the low agreement in copy-editing in both rounds. Both takeaways highlight one of the core difficulties of long-form SPTG, and reinforce our emphasis on the need for better, standardized measurements: human annotation, *especially when the annotators are not the authors*, is unpredictable. Having non-standardized metrics with low agreement amongst each other only makes the problem–and the comparability of works–more difficult.

Next we describe our study in detail.

First Part: Corpus Annotation We collected labels for SPTG emphasizing three key aspects:

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. Please restrict your answer to one of these options only: 'A' or 'B', and provide your answer in the following JSON format: {{"answer": "<your_answer_here>"}}.

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

Figure 9: Zero-shot evaluation prompt (P_{binary}) used to evaluate binary output in Section 5.3.

<|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 10: Query reconstruction prompt in chat markup format.

content preference, style preference, and copy-editing preference. For each labeling task, the annotators were given a reference text for a given author's writing style (denoted "|*Grounding Text*|" in the example below), a user query (r. "|*User Input*|"), and two generated responses based on the query: one personalized and another non-personalized. To avoid positional bias, each generated output was randomly marked as "|*Generated Output A*|" or "|*Generated Output B*|".

The task involved answering each of the three key aspects for our labeling task. They were asked, in order, for content preference

> *Which generated output is better able to cover the content being asked by the |User Input|?*
> *(a) |Generated Output A|*
> *(b) |Generated Output B|*
> *(c) Both generated responses have similar content*
> *(d) None of the generated responses are able to generate a useful response for the |User Input|,*

for style preference

> *Which generated output is better able to reflect the writing style of |Grounding Text|?*
> *(a) |Generated Output A|*
> *(b) |Generated Output B|*
> *(c) Both generated responses have similar writing style to |Grounding Text|*
> *(d) None of the generated responses reflect the writing style to |Grounding Text|*
> *(e) I cannot make out the difference in the writing styles ,*

and, finally, for copy-editing preference

> *If you were asked to write a response towards |User Input|, which generated output(s) would you use to develop an answer?*
> *(a) |Generated Output A|*
> *(b) |Generated Output B|*
> *(c) I will use a combination of both of them, taking inspirations for certain aspects from each*

Table 6: Average Krippendorff's alpha agreement scores between the automatic evaluation metrics and the decisive human annotations, where the annotators explicitly selected T_+ or T_-.

Metric	α
BLEU	0.0452 ± 0.0341
ROUGE-1	0.0459 ± 0.0281
ROUGE-2	0.0783 ± 0.0379
ROUGE-L	**0.1028** ± 0.0374
METEOR	0.0127 ± 0.0150
Wegmann	0.0047 ± 0.0097
StyleDistance	0.0532 ± 0.0445
Ministral-3B	0.0056 ± 0.0024
Llama-3.1-8B	0.0152 ± 0.0303
Mistral-24B	-0.0029 ± 0.0059
Qwen3-32B	0.0280 ± 0.0206
DeepSeek-V3	0.0472 ± 0.0166
o4-mini	0.0618 ± 0.0425
gpt-4.1	0.0777 ± 0.0092
ρ_{all}^{PWV}	0.0842 ± 0.0389

generated response.
(d) I will use either of the two generated outputs.
(e) I will not use any of the generated responses and rather write a response from scratch.

Our corpus was comprised of 396 instances, spanning 75 instances from *Amazon reviews*, 71 instances from *blogs*, 68 instances from *lyrics*, 42 instances from *Reddit microblogs*, 35 instances from *Reuters news articles*, and 30 instances from *short stories*[8]. Each instance was labeled by three annotators.

The IAA scores for the content preference was 0.779, for style preference was 0.641, and for copy-editing preference was 0.400. We provide the distribution of response choices in Figure 18. We found that the annotators had difficulty distinguishing between the personalized and non-personalized responses, often labeling "Both" as the most prominent response for all three annotation dimensions. We explain this in two ways: (1) there is high content overlap between the personalized and non-personalized generations (recall the high semantic overlap in the *LLM* setting from Figure 3); and (2) annotating long-form text was difficult for the annotators. Obtaining high-quality long-form annotations is known to be complex (Goyal et al., 2022; Akoury et al., 2020). Our study aligns with these findings, although focuses on long-form SPTG outputs.

While the collected annotations were highly skewed toward indecisive labels, we still compared human annotations with automatic evaluation metrics by computing Krippendorffs α between metric predictions and human judgments. To ensure meaningful comparisons, we restricted this analysis to decisive instances, where annotators explicitly selected one of the two candidate outputs. Due to the small number of such instances[9], average agreement between automatic metrics and human judgments was generally low. Nevertheless, the relative agreement trends largely mirror the findings from the main results in Tables 2 and 3, with the exception of ROUGE-L, which exhibits higher agreement than the ensemble ρ_{all}^{PWV} (see Table 6).

Second Part: Reproducibility Study Given the agreement scores and the takeaways on annotation difficulty, we carried out a second annotation round in order to evaluate the reproducibility of our work. For this, we conducted another study with the same annotators one month after the first. We used the same annotators under the rationale that, although none of the annotators had authored the texts, they would be most familiar with the existing corpus. This study was conducted over a smaller subset of the original corpus, comprising 129 instances. We provide the agreement statistics in Figure 19. Same as the first study, we found that the annotators were most confident in their content preferences (0.736 IAA) than on the style preferences (r. 0.615). Likewise, the copy-editing preferences had the lowest agreement, with about 0.385 agreement across the two annotation studies.

[8] *ArXiv scientific abstracts* were excluded from our annotation study due to the domain-specific expertise required.

[9] The three annotators selected decisive responses for 91, 35, and 43 instances respectively out of the 396 total instances from the first annotation study.

<|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 11: Personalized text generation prompt used to generate T_+ for *LLM* evaluation setting 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 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 12: Non-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 13: 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 14: Classification prompt used to obtain the writing task corresponding to the filtered English writing queries from the WildChat dataset (Zhao et al., 2024).

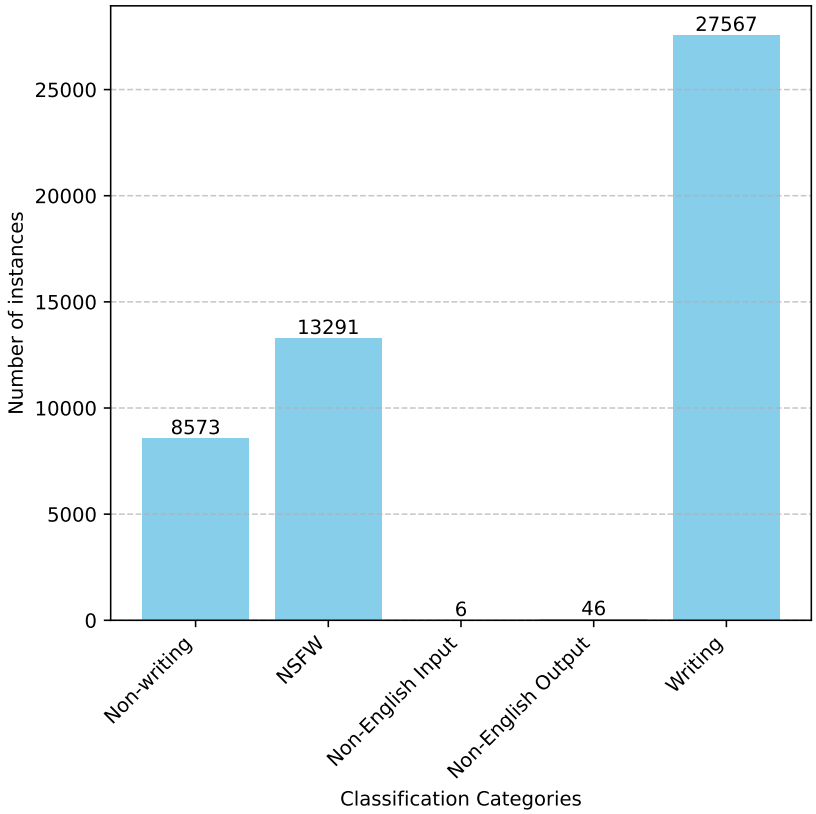


Figure 15: WildChat user query classification statistics for writing task identification over randomly selected 50,000 instances.

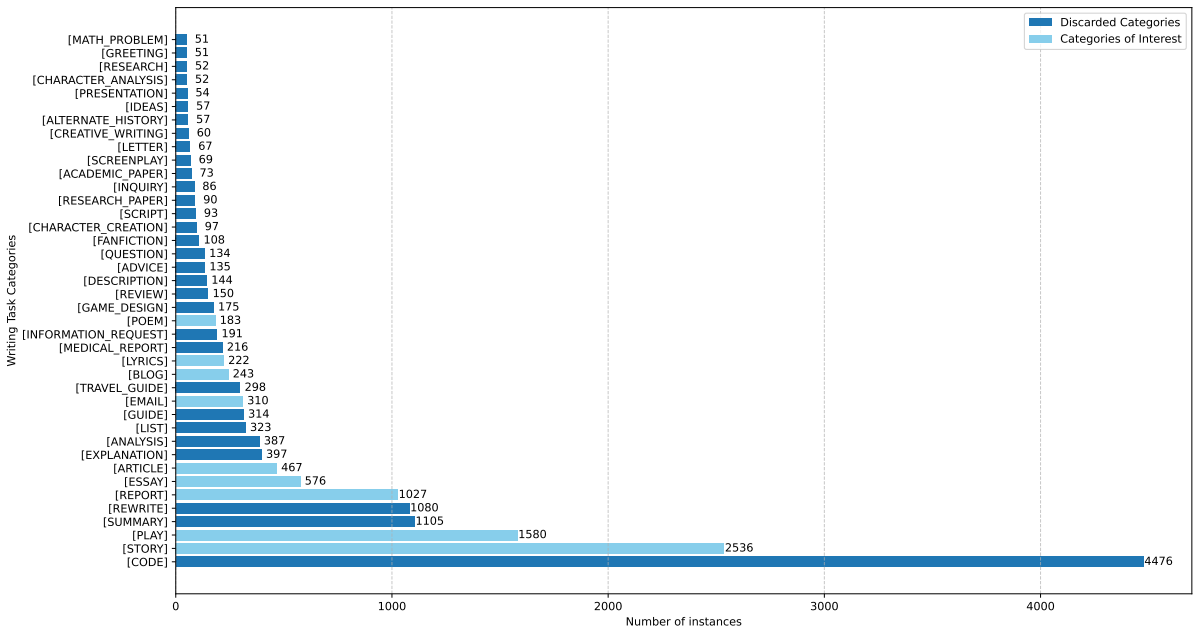


Figure 16: WildChat user query classification statistics for categorizing the writing queries obtained from Figure 15.

The Eagle has landed...I repeat the Eagle has landed! Wasup peoples I am offically in Puerto Rico trying to keep things going down here. I am enjoying being with the fam so much..my three little brothers` who are no longer little and I soon will be looking up to are so grown now. One is 19 (turning 20 soon) Mario, the middle one is 17 John, and the youngest one is 15 Nathaniel which I can't even try to type the way you prnounce his name in spanish. So yea wow..me in Puerto Rico finally the culture here of course is so different! [.....] ....of course the number one job is to play ball but at the present time they are only allowing natives of Puerto Rico to play...basically meaning not me :(haha...however I have been asked to come train with some of the teams during practice which will help me to keep in shape and looking to start my residency down here wich will allow me tlo play in the down the road I am also starting to talk to an agent down here who has some contacts in South America...so what I'm doing is just what r says ...KEEPING IT MOVING! Until next time you all Be EZ

T_{ref}

Wasup to my small audience...I'm peacing out. Finally got that ticket to Puerto Rico and I'm headed out offically today Tues July 21 as I am still up doing some last minute cleaning and packing! Well wish me the best and I hope to have internet access over there so I can keep updating my blog for the couple of people that actually read it and or leave comments ha its all good ! Anyway gotta keep it moving B EZ you all Jes

Don't worry you all I'm not going pretend I was raised in Puero Rico after one week being here! The stinks is no matter how good my spanish will get and in touch with hispanic culture I cannot for the life of me roll my r's! It is by far one of the most frusturating things [.....] I probably won't ever give in to drinking beer ...especially if I made it through college without drinking it...and yes I have played beer pong with Coke! Well thats it for now you all...I am about to get my eat on with some Churchs Chicken..yes I find it sad that theres a church's chicken in Puerto Rico and none to be found in boston..sad!Holla at your gurl B EZ you all

D^a_{blog}

Write a personal and engaging blog post about my experience arriving in Puerto Rico and adjusting to the culture here. The post should reflect my emotions and observations, including details about my family, cultural differences, and challenges related to finding work. It should be written in a conversational and human-like style, avoiding repetitive phrases and unnatural sentences. Please make it detailed and authentic, capturing the essence of my journey and experiences.

q_{ref}

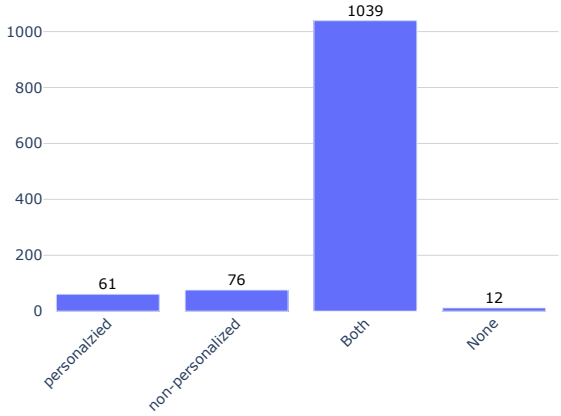
What's up, everyone! Just touched down in Puerto Rico, and let me tell you, it's been one whirlwind of emotions since I got here. I arrived on [DATE], and my family was there to greet me with open arms. It felt so good to see familiar faces after the long flight, but man, the culture shock is real! First off, Puerto Rico is absolutely gorgeous. The beaches, the vibrant colors of the buildings, and the music that seems to follow [....] My family has been amazing, helping me navigate everything from finding the best local spots to eat to understanding the nuances of Puerto Rican customs. The sense of community here is strong, and it's comforting to know I have a solid support system as I navigate this new chapter of my life. Anyway, that's it for now. I'm off to explore more of this beautiful island and maybe pick up a few more Spanish phrases along the way. Thanks for following along on this journey with me. Until next time, B EZ you all! Much love, [NAME]

T_+

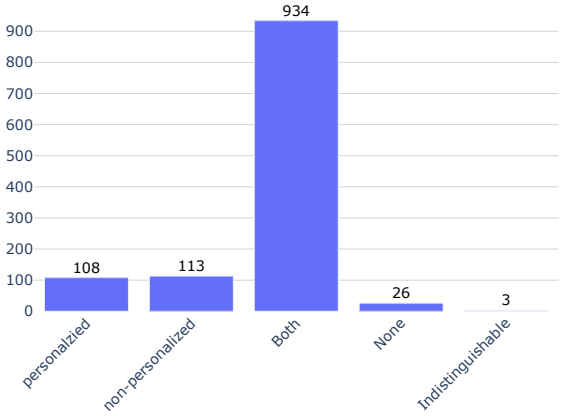
Arriving in Puerto Rico has been both an exhilarating and challenging experience for me and my family. From the moment we stepped off the plane, we were greeted by the warm, humid air and the vibrant colors of the island. The lush greenery and the beautiful coastline instantly made us feel like we were in paradise. Adjusting to the culture here has been an adventure in itself... [....] Family time has become even more precious, and we make it a point to explore new places on the island every weekend. Overall, our journey in Puerto Rico has been filled with highs and lows, but it's shaping up to be an unforgettable chapter in our lives. We're learning, growing, and adapting every day. The island has a way of capturing your heart, and despite the challenges, we're grateful for this opportunity to experience life in such a unique and beautiful place. Stay tuned for more updates on our adventures and misadventures here in Puerto Rico. Until next time, ¡hasta luego!

T_-

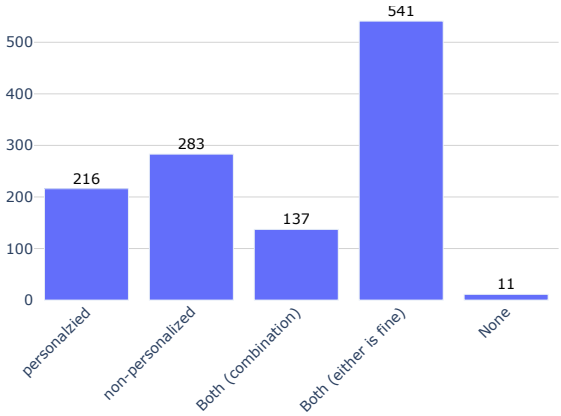
Figure 17: Example instance of our evaluation dataset from the *blogs* domain. The candidates T_+ and T_- are generated for the *LLM* evaluation setting, with T_+ denoting the personalized response and T_- denoting the non-personalized response.



(a) Response distribution for the content preference question "*Which generated output is better able to cover the content being asked by the |User Input|?*"



(b) Response distribution for the style preference question "*Which generated output is better able to reflect the writing style of |Grounding Text|?*"



(c) Response distribution for the copy-editing preference question "*If you were asked to write a response towards |User Input|, which generated output(s) would you use to develop an answer?*"

Figure 18: Distribution of annotation responses for content, style, and copy-editing preference questions.

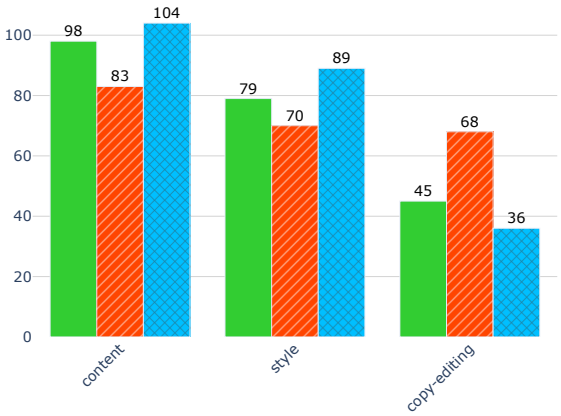


Figure 19: Agreement statistics for three annotators for the style, content and copy-editing preferences for the reproducibility study (out of 129 instances).