

Steering Large Language Models with Register Analysis for Arbitrary Style Transfer

Xinchen Yang

Department of Computer Science
University of Maryland, College Park
xcyang@cs.umd.edu

Marine Carpuat

Department of Computer Science, UMIACS
University of Maryland, College Park
marine@cs.umd.edu

Abstract

Large Language Models (LLMs) have demonstrated strong capabilities in rewriting text across various styles. However, effectively leveraging this ability for example-based arbitrary style transfer—where an input text is rewritten to match the style of a given exemplar—remains an open challenge. A key question is how to describe the style of the exemplar to guide LLMs toward high-quality rewrites. In this work, we propose a prompting method based on register analysis to guide LLMs to perform this task. Empirical evaluations across multiple style transfer tasks show that our prompting approach enhances style transfer strength while preserving meaning more effectively than existing prompting strategies.

1 Introduction

Text style transfer (TST) refers to the task of transforming an input text into a target style (e.g. formality) while preserving non-style attributes such as meaning and fluency (Mir et al., 2019; Krishna et al., 2020; Jin et al., 2022). TST has many downstream applications. For example, one application is the intelligent writing assistant, which helps rewrite texts to meet users’ personalized requests (e.g. more professional, polite, etc.) (Jin et al., 2022). Other applications include text simplification, text detoxification, authorship obfuscation and so on (Jin et al., 2022; Mukherjee et al., 2024; Fisher et al., 2024).

Recent advances in natural language generation (NLG) and large language models (LLMs) have made it possible to perform TST tasks automatically at scale (Jin et al., 2022). In response to a growing need for TST methods with reduced data requirements and broader style coverage (Jin et al., 2022; Hu et al., 2023), research community has increasingly focused on a general formulation of style transfer, *arbitrary TST*, in which an LLM rewrites an input text into an arbitrary style specified by the user at inference-time (Suzgun et al., 2022; Reif et al., 2022; Patel et al., 2024). Reif et al. (2022) shows promise in framing arbitrary TST as a sentence rewriting task by using natural language instructions such as “make this melodramatic”. Despite of its flexibility, end-users are left with the task of constructing the right prompt for a desired style, which often requires opaque prompt engineering, and is particularly difficult for non-native speakers or other users unfamiliar with how to express exact stylistic nuances in technical terms. Reacting to this, example-based arbitrary TST comes as a solution, where representative target-style exemplars are provided at inference-time, allowing LLM to infer the desired style from exemplars without requiring users to explicitly characterize it. For example, Patel et al. (2024) introduces STYLL, an example-based arbitrary TST method, where the LLM is prompted to summarize a few target-style exemplars provided at inference-time into a list of open-ended style descriptors (e.g. “clear, concise, persuasive, intelligent”) before applying them to rewrite the input text. However, relying on such open-ended descriptions, whether user-provided or model-inferred, may pose challenges. While these descriptors may help move the text away from the source style, they do not always ensure faithful reproduction of the target style (Patel et al., 2024). Moreover, as the style descriptors are *unconstrained*, it is unclear whether they may have side effects such as muddying the line between style and content and thus causing unintended meaning alteration.

To address the above challenges, we propose prompting LLMs to analyze exemplars' style using Biber's multidimensional register analysis (MDA) framework. Our modeling hypothesis is that Biber's register analysis provides a structured and effective way to generate accurate target style descriptors that LLMs can reliably use in TST generation. First, because Biber's register analysis framework is widely available online and used in educational and linguistic contexts, LLMs are likely to have been exposed to examples of such analyses during training. Second, Biber's approach is data-driven, grounded in empirical analysis of large corpora, with its key register dimensions found across multiple languages and in online texts (Biber, 1995; Biber & Egbert, 2018), suggesting that Biber's framework is useful to highlight style variations that are salient in LLM pre-training data, of which online texts constitute a large portion. As a result, using Biber's register analysis to describe exemplars may produce theory-grounded, easy-to-follow style descriptors for LLMs. Additionally, we explore whether contrasting the style of input and target exemplars yield better results than characterizing the style of target exemplar *only*.

In this work, we evaluate two prompting variants, with one of them based on Biber's register analysis and input-target style contrast while the other one based on register analysis only, against several baselines on a diverse range of style transfer tasks, including authorship imitation, formality transfer and text simplification. Empirical results show that our prompting approach enhances style transfer quality: (1) Our prompting variants show similar to improved style transfer strength compared to the baselines; (2) Our prompting variants show a large gain in meaning preservation across the tasks.

Our main contributions are as follows:

- We propose a prompting method based on register analysis to enhance the quality of example-based arbitrary TST. In addition, we investigate the impact of input-target contrast v.s. characterizing target only across different use cases.
- Experiments across diverse style transfer tasks show that our approach enhances rewriting quality, with similar to better style transfer strength and remarkable gains in meaning preservation, suggesting better decoupling of style and content.

2 Background

2.1 Style and Register in Corpus Linguistics

In linguistics, *style* refers to the language habits of one person (e.g. Shakespeare), or a group of people at one time or over a period of time (e.g. Old English "heroic" poetry) (Crystal & Davy, 1969). Style reflects an individual's linguistic idiosyncrasies, and thus has applications in disputed authorship resolution, forensic linguistics and so on (Crystal & Davy, 1969; Rudman, 2005; Coulthard et al., 2016). *Register*, on the other hand, refers to linguistic variation associated with the situational use of language and are generally described by three components: situational context, linguistic features (e.g. lexical and grammatical characteristics), and functional relationships between the first two (Biber, 1988; Halliday & Hasan, 1989; Biber & Conrad, 2009). For example, registers can be characterized by speech / writing situation and communicative purposes (e.g. personal letter, academic, narrate, etc.) (Biber & Conrad, 2009). In this view, linguistic features are always *functional*: they tend to occur in a register because they are suited to the situational and communicative context of the register (Biber & Conrad, 2009).

One framework to analyze authorship style is *stylometry*. Typically, stylometric analysis involves statistical analysis of relative frequencies of common words, especially functional words (i.e. words with little lexical meaning and expressing grammatical relationships among other words) (Binongo, 2003; Argamon, 2018; Grieve, 2023). Stylometric analysis has been successful in distinguishing authorship styles (Shakespeare, 2016; Taylor & Egan, 2017). However, it lacks explainability and backing of linguistic theories, making it insufficient in applications such as forensic investigation, where such requirements are expected for legal justifications (Grieve, 2023). An alternative framework is *register analysis*. Grieve (2023)

argued that “authors write in subtly different registers”, showing that register analysis identifies the same underlying patterns of linguistic variation as stylometry. Thus, register analysis is a strong candidate framework to characterize style variation, as it can distinguish styles as effectively as stylometry and provides better explainability (Grieve, 2023).

2.2 Style Transfer in NLP

Style in NLP. In NLP, style transfer tasks adopt a loose extension of the notion of linguistic style to general attributes in text, such as formality, sentiment, and so on (Jin et al., 2022). Some of these attributes are not strictly stylistic features from a linguistic perspective. For example, positive v.s. negative sentiment is arguably more of a content-related attribute than a stylistic one, and formality is more closely related to register than style. In practice, style distinctions are defined by the reliance on specific corpora, which limits LLMs’ capability to adapt to a broad range of unseen styles and perform low-resource style transfer.

Supervised Single-Style Transfer Supervised single-style transfer involves altering a specific stylistic attribute of text while preserving content, typically relying on large parallel style corpora for model training. For example, Jhamtani et al. (2017) introduced a copy-enriched Seq2Seq model to enhance content preservation for modern-Shakespearean style transfer, trained on a large parallel corpus curated by Xu et al. (2012). Recent advances in transformer-based models (Vaswani et al., 2017) have added new momentum. For example, de Rivero et al. (2021) fine-tuned GPT-2 (Radford et al., 2019) on the GYAFC (Rao & Tetreault, 2018) formality transfer parallel corpus and observed improved performance. Atwell et al. (2022) introduced an offensive-inoffensive parallel corpus based on Reddit, and a discourse-aware mechanism to reduce offensiveness and preserve meaning.

Multitask Style Transfer. Multitask style transfer aims to enable a single model to perform multiple style transfer tasks, allowing flexible style switching at inference-time. For example, one early work is the introduction of an unsupervised training framework which enables modifying multiple attributes of a text simultaneously while preserving content (Logeswaran et al., 2018). In another line of work, Vecchi et al. (2022) proposed a learning framework that separates the latent spaces of style and content, enabling multi-style transfer with improved content preservation. Subramanian et al. (2018), on the other hand, indicated that disentangling style and non-style features is not necessary for multitask style transfer by showing that entangled models can work well under unsupervised or pseudo-supervised training. While multitask style transfer reduces reliance on parallel data by learning from target-style text, it still requires large amounts of non-parallel style-specific data — resources often unavailable for all but a few commonly studied styles.

LLM-Based Style Transfer. Recent advances in techniques such as in-context learning (Brown et al., 2020) and instruction-tuning (Ouyang et al., 2022) have enabled LLMs to perform style transfer using only natural language prompts and in-context examples, without task-specific fine-tuning. One seminal work in this direction is Reif et al. (2022), where style transfer is achieved by prompting general-purpose pre-trained LLMs (e.g. GPT-3) with rewriting instructions (e.g. “make this more comic”). Inspired by this, Patel et al. (2024) designed an arbitrary TST method in a more fine-grained manner, which prompts LLMs to extract style descriptors from a few target-style exemplars as rewriting instructions. LLM-based style transfer greatly reduces the need for data and supervision, making it well-suited for low-resource style transfer. Despite its promises, this field of arbitrary TST remains relatively underexplored, with challenges in style controllability, content preservation, etc. In this work, we introduce a method based on register analysis to address these challenges.

3 Approach

In arbitrary TST, the target style is not defined by a predefined set of categories, but specified freely by the user. This makes the task more flexible but also more challenging: how can users convey a style that may be highly personal, nuanced, or unfamiliar to the model?

Users may *feel* about the style — “it sounds like me”, or “I want the text to sound like this example”, but find it hard to describe the stylistic complexities in explicit technical terms. Hence, we adopt a task formulation as following:

Input: an input text x^{input} and a style exemplar x^{style} indicating the desired style.

Output: x^{output} , a rewrite of x^{input} in the style of x^{style} while preserving meaning.

To approach this task, we start with the most straightforward method: prompting. Given that register analysis can be used as a basis to characterize, distinguish and explain authorship styles, we hypothesize that prompting LLMs with instructions based on register analysis enables more effective arbitrary TST. Here, we design a three-step prompting strategy to guide LLMs to perform example-based arbitrary TST. We present two variants: (1) *RG-Contrastive* (“RG”: abbreviation of “register”): prompting with “register analysis” + “contrasting input and output exemplars”; (2) RG: an ablation of with the first variant with “register analysis” only. The pipeline is shown in Fig 1. For full prompts used in the experiments, see Table 4 in Appendix A.

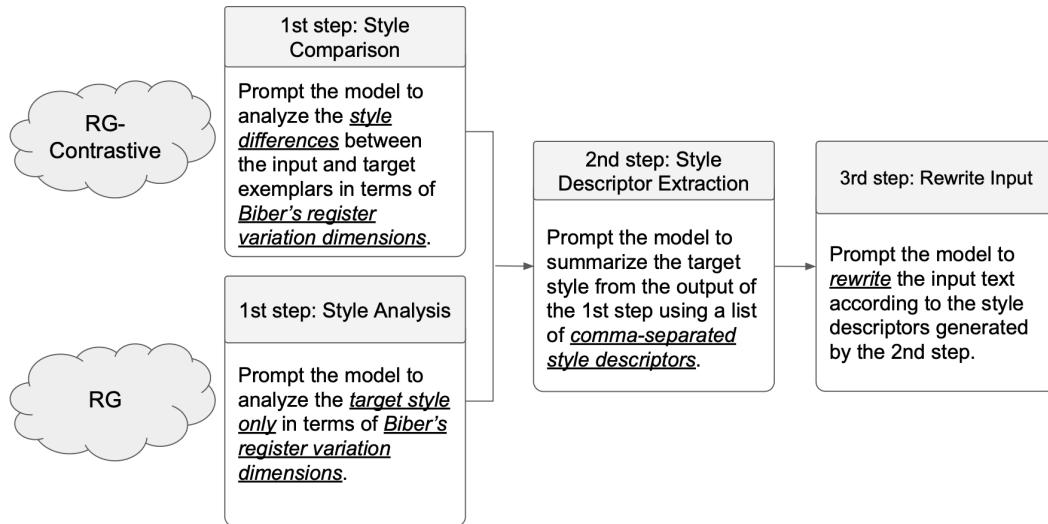


Figure 1: Prompting pipeline for RG-Contrastive and RG, respectively.

4 Experiment Setup

In this section, we describe our experiment setup, including tasks, shared metrics, models and baselines evaluated.

4.1 Tasks

4.1.1 Authorship Imitation

We construct three test sets for the authorship imitation, adopting the method and metrics introduced by Patel et al. (2024).

Datasets. We sample 15 source authors and 15 target authors from the “test_queries” and the “test_targets” split of the Reddit Million User (MUD) dataset (Khan et al., 2021) respectively. In the MUD dataset, each author has 16 posts. Following Patel et al. (2024), we construct three dataset variants: (1) Random: The source and target authors are selected at random. (2) Single: 15 source authors and 15 target authors whose 16 posts all belong to the most common subreddit, ensuring content control by restricting all texts to the same domain. (3) Diverse: 15 source authors and 15 target authors whose 16 posts belong to at least 13 different subreddits, ensuring that any source author or target author writes post in diverse domains. See Appendix B for details in target construction.

Task-based metrics. We use LUAR (Rivera-Soto et al., 2021), an authorship embedding model trained on MUD, to embed input, target and rewritten texts in the authorship space. We calculate the “Away” and “Towards” scores to represent the rewritten-input distance and the rewritten-target distance in the authorship embedding space respectively:

$$\text{Away} = (1 - \text{Cosine_similarity}(\text{LUAR}(\text{Rewritten}), \text{LUAR}(\text{Input}))) / 2 \quad (1)$$

$$\text{Towards} = (1 + \text{Cosine_similarity}(\text{LUAR}(\text{Rewritten}), \text{LUAR}(\text{Target}))) / 2 \quad (2)$$

Following STYLL (Patel et al., 2024), we use Mutual Implication Score (MIS) (Babakov et al., 2022), a meaning preservation metric based on mutual entailment between two texts, inferred by Natural language Inference (NLI) models. To improve the reliability of meaning preservation evaluation, we include two additional common metrics, SBERT Similarity (Reimers & Gurevych, 2019) and METEOR (Banerjee & Lavie, 2005).

4.1.2 Formality Transfer

Datasets. We use the test split of GYAFc (Rao & Tetreault, 2018), a standard benchmark used for formality transfer evaluation, covering two domains: Entertainment & Music (EM) and Family & Relationships (FR). In each domain, we perform formality transfer in two directions: informal to formal (I2F), and formal to informal (F2I). For each direction, we select targets in the desired formality in the train split for rewriting systems to mimic. See Appendix B for details in target construction.

Task-based metrics. Following previous practices (Horvitz et al., 2024), we use an off-the-shelf binary formality classifier fine-tuned on GYAFc to evaluate style transfer accuracy (Dementieva et al., 2023). We set 0.5 as the decision threshold. We use MIS (Babakov et al., 2022) to evaluate meaning preservation, following the recommendation in its paper that MIS is particularly successful in a subset of TST tasks including paraphrasing and formality transfer. Again, we include SBERT similarity and METEOR as additional metrics.

4.1.3 Text Simplification

Datasets. We evaluate on the test split of Cochrane (Devaraj et al., 2021), a paragraph-level simplification task aiming at simplifying medical abstracts into plain-language summaries (PLS). We select PLS texts from the train split of Cochrane as targets for rewriting systems to mimic. See Appendix B for details in target construction.

Task-based metrics. Following previous works (Alva-Manchego et al., 2020; Devaraj et al., 2021; Laban et al., 2021), we use the following metrics. For simplicity: (1) Flesch-Kincaid grade level (FKGL) (Kincaid, 1975): A readability test to determine how difficult a passage is by translating into a U.S. school grade level. (2) Automated Readability Index (ARI) (Smith & Senter, 1967): Similar to FKGL, ARI is a readability test to gauge the understandability of a text which produces an approximate US grade level needed to comprehend the text. For content retention: (1) ROUGE (Lin, 2004) : A suite of recall-based measures for evaluating content retention in summarization. We report ROUGE-1/2/L scores, which measure unigram, bigram, and longest common subsequence overlap between system output and reference, respectively. (2) BLEU (Papineni et al., 2002): A n-gram, reference-based metric for machine translation that is also often reported for simplification systems (Devaraj et al., 2021). For holistic rewriting quality, we use SARI (Xu et al., 2016), a metric measuring how well a simplification system performs three key editing operations: keep, delete and add. SARI is found to correlate well with human judgments (Xu et al., 2016; Agrawal & Carpuat, 2024). Lower FKGL and ARI scores indicate better simplicity. ROUGE, BLEU, and SARI range from 0 to 1, with higher scores indicating better quality.

4.2 Metrics.

Style Transfer. In addition to generic style transfer metrics described in individual tasks in Section 4.1, we use two stylistic representations to evaluate how accurately the rewritten

text mimic the exact target style. (1) StyleCAV (Wegmann et al., 2022): StyleCAV is a style embedding model aiming for “general-purpose” style representation trained with content control on Reddit utterances and tested on Authorship Verification (AV) tasks. (2) Biber’s MDA (Biber, 1988): Since our method is based on instructing LLMs to rewrite the input text into the target style under the guidance of Biber’s MDA, we examine how closely the output aligns with the target in the *actual* Biber’s MDA representation space. We follow the practice of Grieve (2023) to perform Biber’s MDA on our datasets and system outputs. See Appendix D for details on the representation construction and inference procedure. We calculate the “Away” and “Towards” scores based on StyleCAV and Biber’s MDA embeddings respectively, as described in Section 4.1.1.

Meaning Preservation. See task-based metrics for individual tasks in Section 4.1.

Fluency. Following previous practices (Horvitz et al., 2024; Bao & Carpuat, 2024), we use a grammatical acceptability evaluation model trained on the COLA dataset (Warstadt et al., 2019; Morris et al., 2020) to evaluate fluency.

Target Overlap. To penalize rewriting systems for copying *target content* into output, which is undesirable and may trick style metrics into believing that the output mimics the target style well, we measure the content overlap between system output and target. We report ROUGE-1/2/L scores of the system output based on target as the reference.

4.3 Models

We experiment with Llama3.2-3B-Instruct on all the subtasks. We select this model because it is intended for “assistant-like chat and agentic applications like knowledge retrieval and summarization, mobile AI powered writing assistants and query and prompt rewriting” (Meta AI, 2024). In fact, instruction fine-tuning enables model to follow explicit prompts or instructions and is found to substantially improve zero-shot performance on a diverse range of unseen tasks (Ouyang et al., 2022; Wei et al., 2022; Chung et al., 2022; Wang et al., 2022). The model’s relatively small size also makes it well-suited for scenarios with limited computational resources. We aim to explore whether even *smaller* models can benefit from our prompting method. This contrasts Reif et al. (2022), which employs extremely large (e.g. the 175B GPT-3 (Brown et al., 2020)) models for similar purposes.

Additionally, we experiment with LLaMA 3.1-8B-Instruct on the authorship imitation subtask to investigate how a moderately larger model responds to our prompting method and whether there are generalizable patterns across models. We do not experiment with larger models from the Llama family (e.g. 70B or above) due to computational limitations.

We use models from the Huggingface repository and experiment with the default parameters except for “max_new_tokens”, which we set to 1024 to accommodate longer model outputs. See Table 5 in Appendix C for access links for the Llama models and model-based metrics mentioned in previous sections.

4.4 Baselines

We evaluate our method against the baselines shown in Table 1. All our experiments are conducted in a zero-shot setting. See Table 4 in Appendix A for full prompts.

5 Results

For arbitrary TST by example, to study the trade-off between style transfer strength and meaning preservation, which is the core challenge of this task, we plot the Pareto frontiers of rewriting systems across style transfer strength (x-axis, measured by “Towards” using Biber’s MDA representation) and meaning preservation (y-axis, measured by MIS on MUD and GYAF, and by Rouge-1 on Cochrane), evaluated on Llama3.2-3B-Instruct (see Fig 2). For each task, the Pareto frontier shows the set of systems for which no other achieves better

Method	Description
Copy	A naive approach that copies the source input text without any modification.
Target	A naive approach that copies the target text as the rewritten text.
Gold	A dummy approach that copies the reference text (if any) as the rewritten text. If multiple reference texts exist, randomly select one as the output. This serves only as an upper bound.
Simple	A simple one-line prompt instructing LLMs to rewrite the input to mimic the authorship style of the target exemplar.
STYLL (Patel et al., 2024)	An example-based arbitrary TST method. It first prompts the model to rewrite the input into a neutral style, then to describe the target style using a list of style descriptors, and finally to rewrite the neutral rewrite of input using the style descriptors.

Table 1: Overview of the experimental baselines.

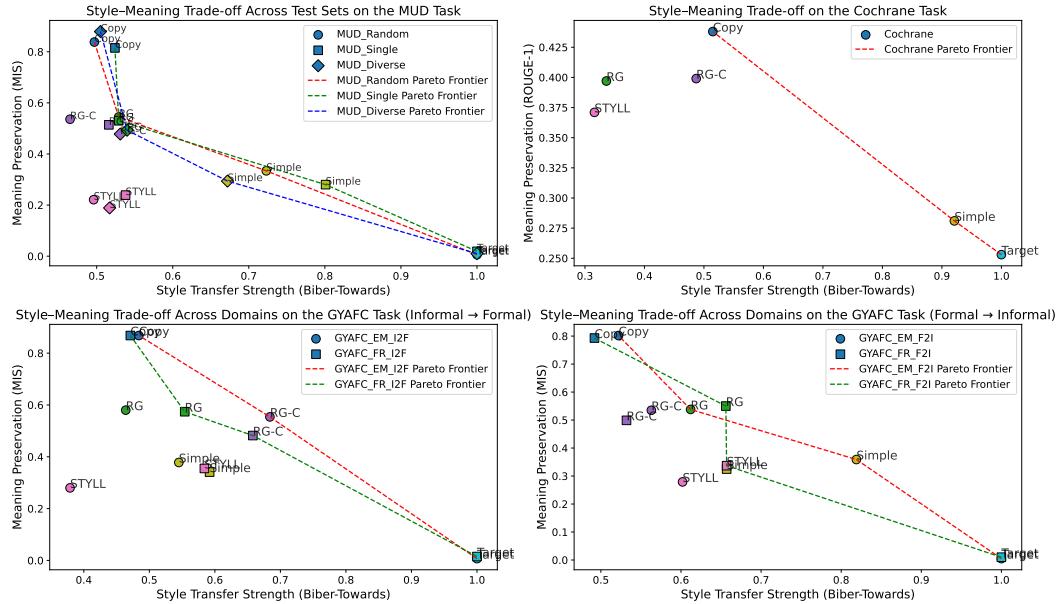


Figure 2: Style–meaning trade-offs across tasks: MUD, Cochrane, and GYAFc (both directions), using Llama-3.2-3B-Instruct. Pareto frontiers identify systems that achieve optimal trade-offs for each task. RG-C: RG-Contrastive.

performance in both objectives simultaneously. We observe that naive baselines “copy” and “target” are always on the Pareto frontier, as expected, as “copy” and “target” achieves the best meaning preservation and style imitation respectively.

Among the evaluated systems, our method consistently demonstrates strong performance across tasks. On the MUD authorship imitation task, RG is on the Pareto frontier across all data splits, achieving a good balance between style transfer strength and meaning preservation, while RG-Contrastive performs slightly behind. On GYAFc, trend varies by transfer direction. In the I2F direction, RG-Contrastive sits on the Pareto frontier across both EM and FR domains, achieving much stronger style transfer than RG. While in the F2I direction, it is the opposite - RG sits on the Pareto frontier across domains and provides better style transfer strength than RG-Contrastive. The performance distinction potentially suggests different use cases of whether including the contrasting mechanism or not. We hypothesize that the “flipping trend” is because the “formal” targets in the GYAFc dataset are not formal in an *absolute* sense but *relative* to the input text. In I2F experiments, we observe that RG and STYLL are inclined to generate descriptors such as “informal” and “causal”, misleading the model into the wrong direction, but RG-Contrastive usually gets

System	MUD (Overlap Rouge-1 ↓)			GYAFC (Overlap Rouge-1 ↓)				Cochrane (Overlap Rouge-1 ↓)
	Random	Single	Diverse	EM_I2F	EM_F2I	FR_I2F	FR_F2I	
Copy	0.075	0.051	0.070	0.234	0.299	0.141	0.088	0.234
Target	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Gold (dummy)	–	–	–	0.253	0.087	0.089	0.095	0.253
Simple	0.343	0.426	0.382	0.864	0.310	0.127	0.104	0.864
STYLL	0.145	0.119	0.148	0.239	0.056	0.099	0.090	0.239
RG-Contrastive	0.107	0.070	0.108	0.251	0.046	0.104	0.094	0.251
RG	0.110	0.077	0.113	0.234	0.090	0.095	0.086	0.234
System	MUD (COLA ↑)			GYAFC (COLA ↑)				Cochrane (COLA ↑)
	Random	Single	Diverse	EM_I2F	EM_F2I	FR_I2F	FR_F2I	
Copy	0.783	0.679	0.754	0.746	0.933	0.790	0.921	0.969
Target	0.067	0.133	0.067	0.253	0.000	0.520	0.042	0.965
Gold (dummy)	–	–	–	0.925	0.743	0.944	0.822	0.967
Simple	0.537	0.446	0.532	0.727	0.517	0.916	0.745	0.967
STYLL	0.982	0.956	0.978	0.970	0.951	0.989	0.987	1.000
RG-Contrastive	0.960	0.951	0.960	0.983	0.944	0.990	0.986	1.000
RG	0.949	0.923	0.952	0.929	0.939	0.977	0.972	0.998

Table 2: Target Overlap (measured by Rouge-1, lower is better) and grammatical acceptability (measured by COLA, higher is better) scores for rewriting systems across MUD, GYAFC, and Cochrane tasks, using Llama-3.2-3B-Instruct. **Bold** values indicate the best scores among non-naïve systems (Simple, STYLL, RG-Contrastive, RG).

it right. On the contrary, the “informal” targets in the GYAFC dataset are mostly informal in an *absolute* sense (with lots of slangs, abbreviations, etc.), so RG itself is sufficient while the contrasting mechanism may confound the model. On Cochrane, RGs miss the Pareto frontier, with RG-Contrastive performs better than RG but beaten by “copy” - it moves the composite style of the input text slightly away from the target. This may be because the input and target texts in Cochrane have prominent baseline similarities, i.e. the “technical” register as the input and target are original and simplified versions of medical abstract respectively, so small perturbations to the input style may inadvertently move it further away from the target style. For example, RG and STYLL tend to arrive at the wrong direction by determining that the target style is “technical”, while RG-Contrastive sometimes decides that the target style is “informal” compared to the input text, which points out the right direction but may be an overshoot, reflecting the intrinsic challenge of pinpointing the accurate target position in the style space, especially when the target style occupies an intermediate position rather than a distinct stylistic endpoint.

In contrast, STYLL, a state-of-the-art zero-/few-shot example-based arbitrary TST method, is consistently a suboptimal choice across tasks (miss Pareto Frontier and often dominated by RG-Contrastive/RG except on the GYAFC_F2I task). It achieves style transfer strength that is at best comparable to RGs and often worse, while consistently lagging behind in meaning preservation by a large margin. Surprisingly, “Simple” frequently sits on the Pareto frontier and beats sophisticated methods (STYLL, RGs) in style transfer strength. However, this may not be out of genuine style transfer but instead be inflated by target content copying into rewrites. As seen from Table 2, which summarizes the utility scores of the systems across tasks, “Simple” frequently gets 3 ~ 5 times higher Overlap Rouge-1 score than RGs and STYLL, and scores even over 0.8 on Cochrane, indicating a non-trivial higher level of target content copying which can defeat the original purpose of style transfer (especially on Cochrane). “Simple” also has much lower COLA scores than RGs and STYLL on MUD and 3 out of 4 GYAFC splits, which to some extent resemble “Targets” whose low linguistic acceptability is expected because each target of MUD and GYAFC is a concatenation of multiple semantically and logically unrelated texts - potentially due to copying behavior.

Table 3 summarizes the style transfer strength in terms of the major dimension of the downstream task, if any (formality for GYAFC, simplicity for Cochrane). On GYAFC (I2F), RG-contrastive achieves the best style transfer accuracy, agreeing with its clear advantage in imitating the target style on this task. On GYAFC (F2I), “Simple” leads and RG gets the 2nd place in terms of accuracy, ahead of STYLL by a large margin. Notably, both RG variants

System	GYAFC								Cochrane			
	EM_I2F		EM_F2I		FR_I2F		FR_F2I		FKGL ↓	ARI ↓	SARI ↑	Rouge-1 ↑
	Acc ↑	MIS ↑										
Copy	0.064	0.868	0.024	0.802	0.072	0.868	0.046	0.793	12.81	15.26	0.418	0.438
Target	0.999	0.007	0.887	0.005	0.996	0.015	0.845	0.010	11.68	13.86	0.346	0.253
Gold (dummy)	0.921	0.877	0.830	0.787	0.937	0.877	0.850	0.780	11.49	13.66	0.982	1.000
Simple	0.476	0.378	0.869	0.359	0.553	0.341	0.820	0.325	11.76	13.95	0.353	0.281
STYLL	0.554	0.280	0.533	0.279	0.641	0.355	0.500	0.337	13.61	15.53	0.382	0.371
RG-Contrastive	0.886	0.554	0.482	0.535	0.899	0.482	0.396	0.499	11.47	13.58	0.390	0.399
RG	0.347	0.580	0.707	0.538	0.423	0.574	0.647	0.550	14.55	17.16	0.374	0.397

Table 3: Evaluation results of generic style transfer on the GYAFC and Cochrane tasks, using Llama-3.2-3B-Instruct. **Left:** GYAFC is evaluated with accuracy (\uparrow) and meaning preservation (MIS \uparrow). **Right:** Cochrane is evaluated with readability (FKGL, ARI \downarrow), editing-quality (SARI \downarrow) and meaning preservation (Rouge-1 \uparrow). **Bold** values indicate the best scores among non-naive systems (Simple, STYLL, RG-Contrastive, RG).

show a clear edge towards STYLL and “Simple” in meaning preservation indicated by MIS (e.g. on EM_I2F, RG-Contrastive and “Simple” achieve an MIS score of 0.554 and 0.378 respectively). On Cochrane, RG-Contrastive leads across all four metrics (FKGL, ARI, SARI, Rouge-1). Unlike the slight underperformance in terms of accurately replicating the target style on this task, RG-Contrastive serves the downstream goal very well. It *indeed* moves the input text toward the “simple” end of the simplicity spectrum and even achieves an overall simplicity better than that of “Gold” (indicated by FKGL and ARI). Overall, it suggests that RG-Contrastive/RG is highly effective at picking up major style dimensions of the target text and performing style transfer across these dimensions compared to other methods such as STYLL. Although RG-Contrastive/RG may not always capture the infinite composite nuances of a target style, which is intrinsically hard, it can serve practical downstream goals well especially when there are prominent major style transfer dimensions.

The other metrics show similar trends, and we report full experimental results (Llama3.2-3B-Instruct results on MUD, GYAFC and Cochrane, and Llama3.1-8B-Instruct results on MUD) in Appendix E. Llama3.1-8B-Instruct on MUD shows similar trends to those observed on Llama3.2-3B-Instruct across rewriting systems: RGs achieve a great balance between style transfer strength and meaning preservation and show a clear advantage in meaning preservation, with RG often dominating STYLL in both objectives across MUD splits. “Simple”, again, leads in “Towards” scores, but exhibits substantially higher target overlap, indicating a higher degree of non-genuine rewriting through target copying and suffers from low linguistic acceptability if “Target” displays the same “issue” (e.g. by concatenating multiple texts). This indicates the the patterns we observed on Llama3.2-3B-Instruct is not reserved to a single model but can be generalizable to other models. Compared to Llama3.2-3B-Instruct, Llama3.1-8B-Instruct generally shows improved meaning preservation (indicated by MIS, etc.) across splits, similar to dropped style transfer strength (indicated by Towards-Biber, etc.) on the Random and Single split, and similar to improved style transfer strength on the Diverse split. This indicates that a slightly larger model can provide benefits in meaning preservation, not necessarily in style transfer strength, but may have an edge in cross-domain style imitation, where the input and target are from different fields.

6 Qualitative Analysis

Qualitative Examples. To get a qualitative understanding of the behavior of the rewriting systems, we take a look at outputs generated by them. Table 9 in Appendix F shows a few (input, target, outputs) examples on the MUD task. The systems rewrite the input “*Verratti is practically untouchable. He’s signing an extension every year or so and PSG won’t sell for even a €100m.*” towards different targets (“...Aaaaanndd you are all on a list...”, “...Jesus Christ, Cesaro...”, etc.). In the given examples, compared to the input, targets are generally more informal and conversational than the input. RGs successfully capture this shift, producing outputs with stylistic markers such as verbal fillers (“oh man”, “i mean”), lowercase (“i mean”, “no way”), and expressions like “nope” and “lol”. In terms of meaning, RGs accurately retain key meaning including “untouchable”, “sign an extension yearly”, “PSG would not let Verratti go in any way” accross the targets. STYLL successfully captures the

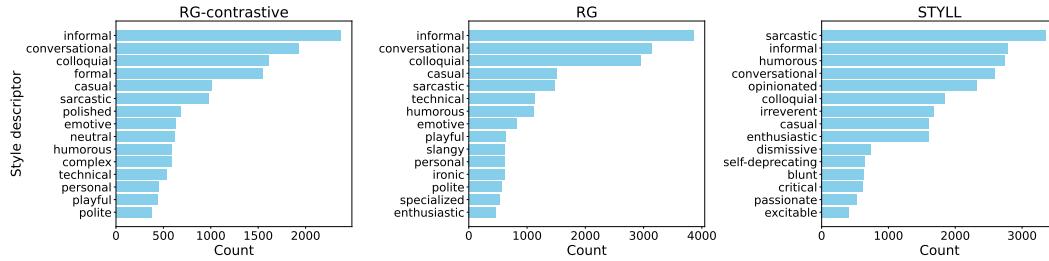


Figure 3: Top 15 style descriptors by frequency by generated rewriting system (RG-Contrastive, RG, STYLL) on the MUD_Random task.

shift towards a more informal style but introduces greater meaning distortion—for example, by adding content not present in the input (“locking down new deals”, “bread and butter of the team”, “legend”, “glue”, etc.). In fact, STYLL adds details that are plausible or potentially inferable from the input context, but are not explicitly present in the original text. Thus, RG-based method is the more appropriate choice for tasks with strict meaning preservation requirements. “Simple” shows a milder tonal shift, occasionally using conversational and informal language in its outputs, but it also occasionally introduces content from the target (“Log in you ... ones!”), which is an undesired behavior.

Style Descriptor Analysis. For both RGs and STYLL, the style descriptors generated during the intermediate step serve as a key characterization of the target exemplar and determine the direction of style transfer. While STYLL leaves LLMs to interpret “style” in an open-ended way, RG-based prompting instructs the model to interpret “style” using Biber’s register analysis as guidance. To understand whether different ways of interpreting “style” affect the nature of the descriptors produced, we conduct a statistical analysis comparing the distributions of style descriptors generated by the methods (RG-Contrastive, RG, STYLL). Figure 3 presents the top 15 most frequent style descriptors generated by each system on the MUD_Random dataset. We observe that the top 3 most frequent descriptors generated by both RGs are “informal”, “conversational” and “colloquial”, which aligns with the informal and conversational nature of MUD, a dataset constructed from Reddit. In contrast, STYLL’s top 3 descriptors are “sarcastic”, “informal”, and “humorous”, with two of them differing from RGs’ top 3. Beyond this, RGs and STYLL also generate descriptors unique to their respective interpretations of “style”. For example, STYLL’s top 15 leaderboard includes descriptors such as “opinionated”, “irreverent”, “dismissive”, “self-deprecating”, etc., which are not among the top 15 most frequent descriptors of RGs. On the contrary, RGs’ top 15 includes “polished”, “emotive”, “playful”, “technical”, “polite”, etc., which are not among the top 15 of STYLL. Overall, STYLL tends to produce more affective, tone-oriented style descriptors, which may signal shifts in tone or intent and thus are more likely to alter the original meaning. In contrast, RGs’ style descriptors are more restricted to the register space, guiding the model towards stylistic adjustments with minimal impact on semantics.

7 Conclusion

In this paper, we introduce a prompting method based on register analysis to guide LLMs in example-based arbitrary TST tasks. Rather than relying on open-ended interpretations of “style,” our method instructs the model to reason about stylistic variation within the space of register. Experimental results across multiple style transfer tasks show that our method achieves enhanced rewriting quality, especially in meaning preservation. Furthermore, qualitative analysis shows that our method produces style descriptors that are more closely aligned with register, with minimal changes in intent or tone, thereby reducing the risk of inadvertent meaning alteration. Overall, our prompting strategy strengthens stylistic control while maintaining higher semantic accuracy than existing methods.

Acknowledgments

We thank Zoey (Dayeon) Ki, HyoJung Han, Kartik Ravisankar in the CLIP Lab of UMD for their constructive feedback.

This research is supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via the HIATUS Program contract #2022-22072200006. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

Ethics Statement

We acknowledge the potential risks associated with the technique described in this paper, including possible misuse by malicious actors for impersonation attacks targeting authorship attribution (AA) and authorship verification (AV) systems, as well as the propagation of misinformation. However, we believe that the benefits of our work outweigh these risks. Our technique enhances style transfer in low-resource settings, enables applications such as personalized writing assistance and privacy preservation for online users, and serves as a testbed for evaluating the adversarial robustness of AA and AV systems.

To promote the benefits of this work while mitigating potential harms, we advocate for practices of responsible use. For example, style-based generation may be encouraged for creative, educational or research purposes, while being appropriately regulated in identity-sensitive or official contexts. Additionally, disclosure in usage can be adopted to ensure transparency and accountability.

References

- Sweta Agrawal and Marine Carpuat. Do text simplification systems preserve meaning? a human evaluation via reading comprehension. *Transactions of the Association for Computational Linguistics*, 12:432–448, 2024. doi: 10.1162/tacl_a_00653. URL <https://aclanthology.org/2024.tacl-1.24/>.
- Fernando Alva-Manchego, Louis Martin, Antoine Bordes, Carolina Scarton, Benoît Sagot, and Lucia Specia. ASSET: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetraeault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4668–4679, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.424. URL <https://aclanthology.org/2020.acl-main.424/>.
- Shlomo Argamon. Computational forensic authorship analysis: Promises and pitfalls. *Language and Law/Linguagem e Direito*, 5(2):7–37, 2018.
- Katherine Atwell, Sabit Hassan, and Malihe Alikhani. Appdia: A discourse-aware transformer-based style transfer model for offensive social media conversations. *arXiv preprint arXiv:2209.08207*, 2022.
- Nikolay Babakov, David Dale, Varvara Logacheva, and Alexander Panchenko. A large-scale computational study of content preservation measures for text style transfer and paraphrase generation. In Samuel Louvan, Andrea Madotto, and Brielen Madureira (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pp. 300–321, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-srw.23. URL <https://aclanthology.org/2022.acl-srw.23/>.

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare Voss (eds.), *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pp. 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. URL <https://aclanthology.org/W05-0909/>.

Calvin Bao and Marine Carpuat. Keep it Private: Unsupervised privatization of online text. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 8678–8693, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.nacl-long.480. URL <https://aclanthology.org/2024.nacl-long.480/>.

Douglas Biber. Variation across speech and writing by douglas biber. 1988. URL <https://api.semanticscholar.org/CorpusID:59969234>.

Douglas Biber. Dimensions of register variation: A cross-linguistic comparison. *Cambridge University Press google schola*, 2:171–197, 1995.

Douglas Biber and Susan Conrad. *Register, Genre, and Style*. Cambridge University Press, 2009.

Douglas Biber and Jesse Egbert. *Register Variation on the Web*. Cambridge University Press, 2018.

José Nilo G Binongo. Who wrote the 15th book of oz? an application of multivariate analysis to authorship attribution. *Chance*, 16(2):9–17, 2003.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. URL <https://arxiv.org/abs/2005.14165>.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models, 2022. URL <https://arxiv.org/abs/2210.11416>.

Malcolm Coulthard, Alison Johnson, and David Wright. *An introduction to forensic linguistics: Language in evidence*. Routledge, 2016.

David Crystal and Derek Davy. Investigating english style. 1969. URL <https://api.semanticscholar.org/CorpusID:59347243>.

Mariano de Rivero, Cristhiam Tirado, and Willy Ugarte. Formalstyler: Gpt based model for formal style transfer based on formality and meaning preservation. In *KDIR*, pp. 48–56, 2021.

Daryna Dementieva, Nikolay Babakov, and Alexander Panchenko. Detecting text formality: A study of text classification approaches. In Ruslan Mitkov and Galia Angelova (eds.), *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pp. 274–284, Varna, Bulgaria, September 2023. INCOMA Ltd., Shoumen, Bulgaria. URL <https://aclanthology.org/2023.ranlp-1.31>.

Ashwin Devaraj, Iain Marshall, Byron Wallace, and Junyi Jessy Li. Paragraph-level simplification of medical texts. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4972–4984, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.nacl-main.395. URL <https://aclanthology.org/2021.nacl-main.395/>.

Jillian Fisher, Skyler Hallinan, Ximing Lu, Mitchell Gordon, Zaid Harchaoui, and Yejin Choi. Styleremix: Interpretable authorship obfuscation via distillation and perturbation of style elements. *arXiv preprint arXiv:2408.15666*, 2024.

Jack Grieve. Register variation explains stylometric authorship analysis. *Corpus Linguistics and Linguistic Theory*, 19(1):47–77, 2023.

Michael Alexander Kirkwood Halliday and Ruqaiya Hasan. Language, context, and text: Aspects of language in a social-semiotic perspective. (*No Title*), 1989.

Zachary Horvitz, Ajay Patel, Kanishk Singh, Chris Callison-Burch, Kathleen McKeown, and Zhou Yu. TinyStyle: Efficient few-shot text style transfer with authorship embeddings. *arXiv preprint arXiv:2406.15586*, 2024.

Zhiqiang Hu, Roy Ka-Wei Lee, Charu C. Aggarwal, and Aston Zhang. Text style transfer: A review and experimental evaluation, 2023. URL <https://arxiv.org/abs/2010.12742>.

Harsh Jhamtani, Varun Gangal, Eduard Hovy, and Eric Nyberg. Shakespearizing modern language using copy-enriched sequence-to-sequence models, 2017. URL <https://arxiv.org/abs/1707.01161>.

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. Deep learning for text style transfer: A survey. *Computational Linguistics*, 48(1):155–205, March 2022. doi: 10.1162/coli_a_00426. URL <https://aclanthology.org/2022.cl-1.6/>.

Aleem Khan, Elizabeth Fleming, Noah Schofield, Marcus Bishop, and Nicholas Andrews. A deep metric learning approach to account linking, 2021. URL <https://arxiv.org/abs/2105.07263>.

JP Kincaid. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. *Chief of Naval Technical Training*, 1975.

Kalpesh Krishna, John Wieting, and Mohit Iyyer. Reformulating unsupervised style transfer as paraphrase generation, 2020. URL <https://arxiv.org/abs/2010.05700>.

Philippe Laban, Tobias Schnabel, Paul Bennett, and Marti A. Hearst. Keep it simple: Unsupervised simplification of multi-paragraph text. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 6365–6378, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.498. URL <https://aclanthology.org/2021.acl-long.498/>.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL <https://aclanthology.org/W04-1013/>.

Lajanugen Logeswaran, Honglak Lee, and Samy Bengio. Content preserving text generation with attribute controls. *Advances in Neural Information Processing Systems*, 31, 2018.

Meta AI. Llama-3.2-3b-instruct, 2024. URL <https://huggingface.co/meta-llama/Llama-3-2-3B-Instruct>.

Remi Mir, Bjarke Felbo, Nick Obradovich, and Iyad Rahwan. Evaluating style transfer for text. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *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)*, pp. 495–504, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1049. URL <https://aclanthology.org/N19-1049/>.

John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 119–126, 2020.

Sourabrata Mukherjee, Akanksha Bansal, Atul Kr Ojha, John P McCrae, and Ondřej Dušek. Text detoxification as style transfer in english and hindi. *arXiv preprint arXiv:2402.07767*, 2024.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin (eds.), *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL <https://aclanthology.org/P02-1040/>.

Ajay Patel, Nicholas Andrews, and Chris Callison-Burch. Low-resource authorship style transfer: Can non-famous authors be imitated?, 2024. URL <https://arxiv.org/abs/2212.08986>.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Sudha Rao and Joel Tetreault. Dear sir or madam, may I introduce the GYAFc dataset: Corpus, benchmarks and metrics for formality style transfer. In Marilyn Walker, Heng Ji, and Amanda Stent (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 129–140, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1012. URL <https://aclanthology.org/N18-1012/>.

Emily Reif, Daphne Ippolito, Ann Yuan, Andy Coenen, Chris Callison-Burch, and Jason Wei. A recipe for arbitrary text style transfer with large language models. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 837–848, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-short.94. URL <https://aclanthology.org/2022.acl-short.94/>.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL <https://aclanthology.org/D19-1410/>.

Rafael A. Rivera-Soto, Olivia Elizabeth Miano, Juanita Ordonez, Barry Y. Chen, Aleem Khan, Marcus Bishop, and Nicholas Andrews. Learning universal authorship representations. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.),

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 913–919, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.70. URL <https://aclanthology.org/2021.emnlp-main.70/>.

Joseph Rudman. The non-traditional case for the authorship of the twelve disputed federalist papers: A monument built on sand. In *Proceedings of ACH/ALLC*, volume 2005, 2005.

William Shakespeare. *The New Oxford Shakespeare: Modern Critical Edition: The Complete Works*. Oxford University Press, 2016.

Edgar A Smith and RJ Senter. *Automated readability index*, volume 66. Aerospace Medical Research Laboratories, Aerospace Medical Division, Air . . . , 1967.

Sandeep Subramanian, Guillaume Lample, Eric Michael Smith, Ludovic Denoyer, Marc’Aurelio Ranzato, and Y-Lan Boureau. Multiple-attribute text style transfer. *arXiv preprint arXiv:1811.00552*, 2018.

Mirac Suzgun, Luke Melas-Kyriazi, and Dan Jurafsky. Prompt-and-rerank: A method for zero-shot and few-shot arbitrary textual style transfer with small language models. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 2195–2222, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.141. URL <https://aclanthology.org/2022.emnlp-main.141/>.

Gary Taylor and Gabriel Egan. *The New Oxford Shakespeare: Authorship Companion*. Oxford University Press, 2017.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Lorenzo Puppi Vecchi, Eliane CF Maffezzolli, and Emerson Cabrera Paraiso. Transferring multiple text styles using cyclegan with supervised style latent space. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–6. IEEE, 2022.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language model with self generated instructions, 2022.

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641, 2019.

Anna Wegmann, Marijn Schraagen, and Dong Nguyen. Same author or just same topic? towards content-independent style representations. In Spandana Gella, He He, Bodhisattwa Prasad Majumder, Burcu Can, Eleonora Giunchiglia, Samuel Cahyawijaya, Sewon Min, Maximilian Mozes, Xiang Lorraine Li, Isabelle Augenstein, Anna Rogers, Kyunghyun Cho, Edward Grefenstette, Laura Rimell, and Chris Dyer (eds.), *Proceedings of the 7th Workshop on Representation Learning for NLP*, pp. 249–268, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.repl4nlp-1.26. URL <https://aclanthology.org/2022.repl4nlp-1.26/>.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned language models are zero-shot learners, 2022. URL <https://arxiv.org/abs/2109.01652>.

Wei Xu, Alan Ritter, Bill Dolan, Ralph Grishman, and Colin Cherry. Paraphrasing for style. In *COLING*, pp. 2899–2914, 2012.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415, 2016.

A Full Prompts

#	Simple
1	Here is the target text [target text] Rewrite [input text] into the authorship style of the target text. Strictly output only the rewritten text without any other content.
#	STYLL
1	Source text: Passage: [source text] Paraphrase the passage in a simple neutral style.
2	Passage: [target text] List some adjectives, comma-separated, that describe the writing style of the author of this passage. Strictly output only the style descriptors without any other content.
3	Here is a text: [neural paraphrase] Here is a rewrite of the text that is more [style descriptors]. Strictly output only the rewritten text without any other content.
#	RG-Contrastive
1	Source text: [source text] Target text: [target text] How does the target text differ from the source text in authorship style in terms of dimensions of register variation according to Douglas Biber?
2	Style comparisons: [style comparisons] List some adjectives, comma-separated, that describe the writing style of the author of the target text. Strictly output only the style descriptors without any other content.
3	Here is a text: [source text] Rewrite the text to be more [style descriptors]. Strictly output only the rewritten text without any other content.
#	RG
1	Passage: [target text] Analyze the authorship style of this passage in terms of dimensions of register variation according to Douglas Biber.
2	Style analysis: [style analysis] List some adjectives, comma-separated, that describe the writing style of the author of the target text. Strictly output only the style descriptors without any other content.
3	Here is a text: [source text] Rewrite the text to be more [style descriptors]. Strictly output only the rewritten text without any other content.

Table 4: Full prompts used in experiments for Simple, STYLL, RG-Contrastive and RG, respectively. In the table, [neural paraphrase / style comparisons / style analysis] is the model’s output from the 1st step and [style descriptors] is the model’s output from the 2nd step. For STYLL, we adopt the specific prompts outlined in Patel et al. (2024), but with minor tweaks to adapt to the zero-shot setting in our experiments.

B Target Construction

Authorship Imitation. During inference-time, each *individual* input text from any of the source authors (each source author has 16 such texts) is paired a with each of the target authors for the model to perform authorship style transfer (ST). In total, there are 16 (number of individual texts per source author) \times 15 (number of source authors) \times 15 (number of target authors) = 3600 such pairings. During each ST corresponding to each pairing, the 16 texts from the target author are concatenated together as the target exemplar whose style the model is tasked to rewrite the input text into.

Formality Transfer. To avoid accidentally exposing the “gold answer” to the rewriting systems, we select targets from the train split of the GYAFc dataset. Texts in GYAFc are of sentence-level and thus may be too short to contain enough linguistic information to inform the LLM. Hence, for each input text, we concatenate $K = 16$ texts randomly selected from the target selection pool to form a paragraph-level target exemplar, following our MUD evaluation set construction practices for authorship imitation. Targets have the same domain and opposite formality as their corresponding input text. For example, for an formal input text in the “EM” domain, 16 sentence-level texts are randomly selected from the EM-train-informal split of GYAFc to form a paragraph-level target to inform style during model’s inference time.

Text Simplification. To avoid accidentally exposing the “gold answer” to the rewriting systems, we randomly select plain-language summary texts from the train split of the Cochrane dataset as target. Unlike GYAFc, texts in Cochrane are of paragraph-level, so each of them suffices in length to serve as a single target without the need of concatenation.

C Model Resources

Table 5 provides links to the pretrained models used in our experiments. All models are publicly available and can be accessed through the Hugging Face Model Hub.

Model	URL
Llama3.2-3B-Instruct	https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct
Llama3.1-8B-Instruct	https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct
LUAR	https://huggingface.co/rriovera1849/LUAR-MUD
SBERT	https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2
Formality classifier (DeBERTa)	https://huggingface.co/s-nlp/deberta-large-formality-ranker
StyleCAV	https://huggingface.co/AnnaWegmann/Style-Embedding
COLA	https://huggingface.co/textattack/roberta-base-CoLA

Table 5: Links to pretrained models used in our experiments.

D Implementation Details of Biber's MDA Style Representation

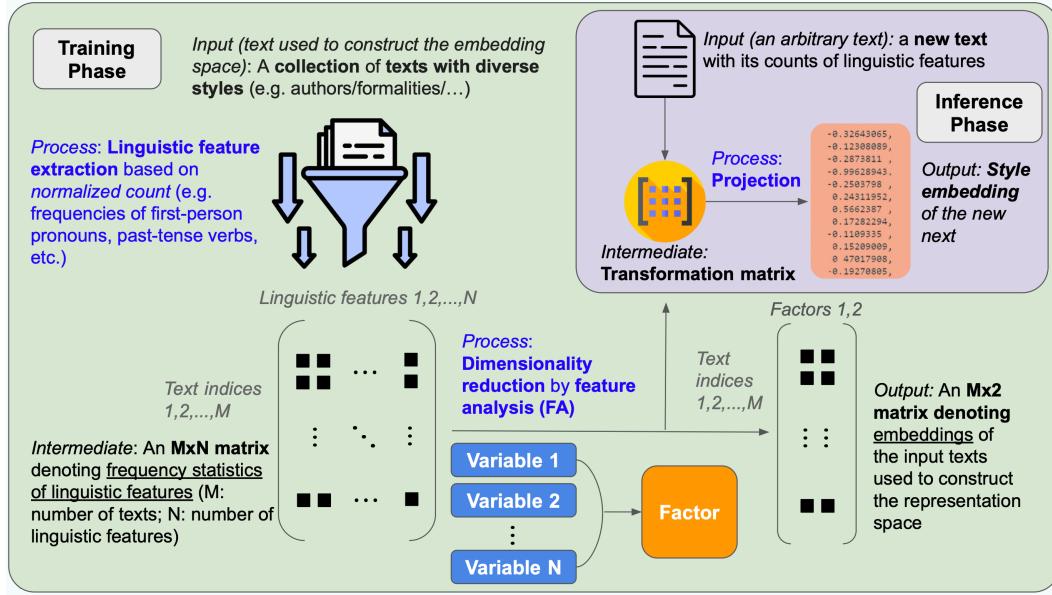


Figure 4: Illustration of the procedure of building Biber's MDA representation from a linguistic corpus and using it to make inference on a new text following the practice of Grieve (2023).

The train and inference phase of Biber's MDA representation are shown in Fig 4. We train one Biber's MDA representation per task (MUD, GYAF, Cochrane). The training corpus for each task encompasses a diverse range of styles for each task (authorship for MUD, formality for GYAF, and simplicity for Cochrane). The data splits constituting the training corpus for each task are specified as follows:

- MUD: validation-queries, validation-targets.
- GYAF: EM-train-formal, EM-train-informal, FR-train-formal, FR-train-informal.
- Cochrane: train-original, train-simplified.

E Complete Results

E.1 Authorship Imitation

Random	System	Away (LUAR)	Towards (LUAR)	MIS	Sbert	Meteor	COLA	Away (StyleCAV)	Towards (StyleCAV)	Away (Biber)	Towards (Biber)	Overlap Rouge-1 (↓)	Overlap Rouge-2 (↓)	Overlap Rouge-L (↓)
Naive	Copy	0.000	0.617	0.838	1.000	0.994	0.783	0.000	0.544	0.000	0.497	0.075	0.005	0.045
	Target	0.383	1.000	0.006	0.065	0.117	0.067	0.456	1.000	0.503	1.000	1.000	1.000	1.000
Llama3.2-3B	Simple	0.269	0.757	0.334	0.475	0.471	0.537	0.349	0.731	0.340	0.723	0.343	0.279	0.314
	STYLL	0.330	0.612	0.221	0.490	0.222	0.982	0.368	0.522	0.457	0.496	0.145	0.018	0.078
	-Instruct	RG-C	0.278	0.604	0.536	0.647	0.338	0.960	0.369	0.473	0.425	0.465	0.107	0.009
Llama3.1-8B	Simple	0.250	0.732	0.451	0.580	0.565	0.602	0.340	0.731	0.292	0.672	0.280	0.211	0.247
	STYLL	0.324	0.632	0.261	0.548	0.235	0.981	0.369	0.534	0.443	0.491	0.178	0.021	0.089
	-Instruct	RG-C	0.282	0.611	0.578	0.654	0.346	0.969	0.342	0.495	0.434	0.483	0.140	0.013
RG	RG	0.273	0.610	0.602	0.668	0.372	0.963	0.325	0.522	0.406	0.499	0.141	0.014	0.074

Single	System	Away (LUAR)	Towards (LUAR)	MIS	Sbert	Meteor	COLA	Away (StyleCAV)	Towards (StyleCAV)	Away (Biber)	Towards (Biber)	Overlap Rouge-1 (↓)	Overlap Rouge-2 (↓)	Overlap Rouge-L (↓)
Naive	Copy	0.000	0.635	0.815	1.000	0.992	0.679	-0.000	0.555	0.000	0.524	0.051	0.004	0.033
	Target	0.365	1.000	0.019	0.111	0.091	0.133	0.445	1.000	0.476	1.000	1.000	1.000	1.000
Llama3.2-3B	Simple	0.265	0.793	0.280	0.414	0.411	0.446	0.346	0.787	0.349	0.801	0.426	0.388	0.411
	STYLL	0.315	0.618	0.239	0.450	0.209	0.956	0.394	0.503	0.426	0.538	0.119	0.026	0.075
	-Instruct	RG-C	0.263	0.611	0.514	0.618	0.334	0.951	0.406	0.457	0.429	0.516	0.070	0.005
Llama3.1-8B	Simple	0.244	0.759	0.404	0.534	0.505	0.496	0.361	0.804	0.315	0.770	0.327	0.287	0.311
	STYLL	0.313	0.625	0.284	0.507	0.226	0.958	0.379	0.554	0.438	0.500	0.133	0.014	0.070
	-Instruct	RG-C	0.264	0.613	0.532	0.626	0.341	0.946	0.377	0.476	0.434	0.483	0.091	0.008
RG	RG	0.259	0.610	0.541	0.646	0.368	0.935	0.325	0.529	0.383	0.510	0.099	0.010	0.057

Diverse	System	Away (LUAR)	Towards (LUAR)	MIS	Sbert	Meteor	COLA	Away (StyleCAV)	Towards (StyleCAV)	Away (Biber)	Towards (Biber)	Overlap Rouge-1 (↓)	Overlap Rouge-2 (↓)	Overlap Rouge-L (↓)
Naive	Copy	0.000	0.613	0.879	1.000	0.989	0.754	-0.000	0.567	0.000	0.505	0.070	0.005	0.041
	Target	0.387	1.000	0.010	0.086	0.113	0.067	0.433	1.000	0.495	1.000	1.000	1.000	1.000
Llama3.2-3B	Simple	0.286	0.755	0.294	0.457	0.446	0.532	0.344	0.719	0.332	0.672	0.382	0.313	0.352
	STYLL	0.340	0.597	0.189	0.490	0.221	0.978	0.380	0.532	0.421	0.517	0.148	0.015	0.076
	-Instruct	RG-C	0.286	0.594	0.478	0.631	0.336	0.960	0.378	0.495	0.400	0.531	0.108	0.009
Llama3.1-8B	Simple	0.275	0.753	0.367	0.520	0.509	0.524	0.339	0.740	0.298	0.678	0.380	0.320	0.354
	STYLL	0.342	0.621	0.200	0.527	0.209	0.968	0.389	0.532	0.423	0.551	0.206	0.020	0.096
	-Instruct	RG-C	0.296	0.602	0.520	0.637	0.338	0.965	0.368	0.510	0.392	0.540	0.149	0.014
RG	RG	0.294	0.602	0.531	0.651	0.347	0.964	0.352	0.531	0.376	0.553	0.155	0.015	0.078

Table 6: Evaluation results of different rewriting systems on the MUD task, using Llama-3.2-3B-Instruct and Llama-3.1-8B-Instruct, respectively. **Bold** values indicate the best scores among non-naive systems (Simple, STYLL, RG-C, RG). RG-C: RG-Contrastive.

E.2 Formality Transfer

Setting	System	Acc (DeBERTa)	MIS	Sbert	Meteor	COLA	Away (StyleCAV)	Towards (StyleCAV)	Away (Biber)	Towards (Biber)	Overlap Rouge-1 (↓)	Overlap Rouge-2 (↓)	Overlap Rouge-L (↓)
EM_I2F	Copy	0.064	0.868	0.824	0.738	0.746	0.000	0.554	0.000	0.484	0.059	0.002	0.054
	Target	0.999	0.007	0.100	0.137	0.253	0.446	1.000	0.516	1.000	1.000	1.000	1.000
	Gold	0.921	0.877	0.876	0.999	0.925	0.439	0.633	0.267	0.598	0.071	0.004	0.065
	Simple	0.476	0.378	0.461	0.432	0.727	0.385	0.682	0.280	0.545	0.299	0.201	0.265
	STYLL	0.554	0.280	0.490	0.292	0.970	0.488	0.603	0.361	0.379	0.141	0.010	0.080
	RG-C	0.886	0.554	0.599	0.402	0.983	0.544	0.535	0.429	0.684	0.087	0.006	0.060
	RG	0.347	0.580	0.630	0.443	0.929	0.406	0.584	0.307	0.464	0.088	0.006	0.058

Setting	System	Acc (DeBERTa)	MIS	Sbert	Meteor	COLA	Away (StyleCAV)	Towards (StyleCAV)	Away (Biber)	Towards (Biber)	Overlap Rouge-1 (↓)	Overlap Rouge-2 (↓)	Overlap Rouge-L (↓)
EM_F2I	Copy	0.024	0.802	0.738	0.595	0.933	0.000	0.338	0.000	0.522	0.056	0.001	0.053
	Target	0.887	0.005	0.110	0.128	0.000	0.662	1.000	0.478	1.000	1.000	1.000	1.000
	Gold	0.830	0.787	0.796	0.999	0.743	0.461	0.654	0.334	0.666	0.046	0.002	0.043
	Simple	0.869	0.359	0.436	0.330	0.517	0.534	0.749	0.388	0.819	0.310	0.236	0.286
	STYLL	0.533	0.279	0.473	0.301	0.951	0.366	0.378	0.452	0.602	0.127	0.009	0.072
	RG-C	0.482	0.535	0.584	0.398	0.944	0.326	0.379	0.357	0.563	0.089	0.006	0.057
	RG	0.707	0.538	0.598	0.424	0.939	0.319	0.387	0.358	0.612	0.090	0.007	0.057

Setting	System	Acc (DeBERTa)	MIS	Sbert	Meteor	COLA	Away (StyleCAV)	Towards (StyleCAV)	Away (Biber)	Towards (Biber)	Overlap Rouge-1 (↓)	Overlap Rouge-2 (↓)	Overlap Rouge-L (↓)
FR_I2F	Copy	0.072	0.868	0.797	0.748	0.790	0.000	0.550	0.000	0.471	0.070	0.003	0.063
	Target	0.996	0.015	0.093	0.155	0.520	0.450	1.000	0.529	1.000	1.000	1.000	1.000
	Gold	0.937	0.877	0.848	1.000	0.944	0.437	0.620	0.249	0.628	0.081	0.005	0.072
	Simple	0.553	0.341	0.421	0.400	0.916	0.438	0.658	0.337	0.592	0.214	0.118	0.179
	STYLL	0.641	0.355	0.434	0.299	0.989	0.501	0.604	0.371	0.584	0.099	0.008	0.063
	RG-C	0.899	0.482	0.475	0.314	0.990	0.567	0.501	0.480	0.658	0.104	0.007	0.067
	RG	0.423	0.574	0.552	0.420	0.977	0.449	0.577	0.323	0.554	0.095	0.007	0.061

Setting	System	Acc (DeBERTa)	MIS	Sbert	Meteor	COLA	Away (StyleCAV)	Towards (StyleCAV)	Away (Biber)	Towards (Biber)	Overlap Rouge-1 (↓)	Overlap Rouge-2 (↓)	Overlap Rouge-L (↓)
FR_F2I	Copy	0.046	0.793	0.683	0.590	0.921	0.000	0.360	0.000	0.492	0.055	0.004	0.042
	Target	0.845	0.010	0.094	0.144	0.042	0.640	1.000	0.508	1.000	1.000	1.000	1.000
	Gold	0.850	0.780	0.751	0.999	0.822	0.471	0.678	0.322	0.618	0.050	0.003	0.037
	Simple	0.820	0.325	0.394	0.312	0.745	0.464	0.666	0.457	0.657	0.213	0.127	0.182
	STYLL	0.500	0.337	0.406	0.294	0.987	0.297	0.335	0.374	0.657	0.090	0.007	0.057
	RG-C	0.396	0.499	0.487	0.352	0.986	0.316	0.332	0.347	0.532	0.094	0.007	0.060
	RG	0.647	0.550	0.515	0.404	0.972	0.305	0.365	0.336	0.656	0.086	0.007	0.056

Table 7: Evaluation results of different rewriting systems on the GYAFC task, across EM and FR domains for both informal-to-formal (I2F) and formal-to-informal (F2I) directions, using Llama-3.2-3B-Instruct. **Bold** values indicate the best scores among non-naive systems (Simple, STYLL, RG-C, RG). RG-C: RG-Contrastive. RG-C: RG-Contrastive.

E.3 Cochrane

System	FKGL (↓)	ARI (↓)	Rouge-1	Rouge-2	Rouge-L	BLEU	SARI	COLA	Away (StyleCAV)	Towards (StyleCAV)	Away (Biber)	Towards (Biber)	Overlap Rouge-1 (↓)	Overlap Rouge-2 (↓)	Overlap Rouge-L (↓)
Copy	12.81	15.26	0.438	0.196	0.250	0.132	0.418	0.969	0.000	0.903	0.000	0.515	0.234	0.028	0.113
Target	11.68	13.86	0.253	0.031	0.125	0.005	0.346	0.965	0.097	1.000	0.485	1.000	1.000	1.000	1.000
Gold	11.49	13.66	1.000	1.000	1.000	1.000	0.982	0.967	0.071	0.907	0.393	0.610	0.253	0.031	0.125
Simple	11.76	13.95	0.281	0.054	0.143	0.022	0.353	0.967	0.086	0.974	0.425	0.921	0.864	0.832	0.851
STYLL	13.61	15.53	0.371	0.102	0.200	0.041	0.382	1.000	0.096	0.889	0.598	0.316	0.239	0.025	0.129
RG-C	11.47	13.58	0.399	0.126	0.207	0.068	0.390	1.000	0.120	0.860	0.551	0.487	0.251	0.028	0.125
RG	14.55	17.16	0.397	0.141	0.213	0.082	0.374	0.998	0.088	0.873	0.365	0.336	0.234	0.027	0.118

Table 8: Evaluation results of different rewriting systems on the Cochrane dataset, using Llama-3.2-3B-Instruct. **Bold** values indicate the best scores among non-naive systems (Simple, STYLL, RG-C, RG). RG-C: RG-Contrastive.

F Qualitative examples

Target	Input	RG-Contrastive	RG	STYLL	Simple
"He didn't do any of that" "Aaaaanndd you are all on a list." "Oh for fucks sake You are brilliant >Hearing "DAD-DDYYYYYY!!!" " ...13 more examples (concatenated together)	Verratti is practically untouchable. He's signing an extension every year or so and PSG won't sell for even a €100m.	oh man, verratti is basically untouchable, no way psg would sell him for anything less than a gazillion euros, and even then they'd probably just give him a massive extension to keep him happy, i mean, the guy's basically a god on the pitch	Dude, Verratti is basically untouchable - he's always signing a new deal and PSG would never even consider selling him for a hundred million euros, let alone that.	Verratti's not going anywhere, lol. He's always locking down new deals and PSG is like "nah, we're not selling him for more than a sweet, sweet €100m."	Verratti is practically untouchable, he's signing an extension every year or so & PSG won't sell for even a €100m.
"Log in you damned ones!" "Loving the Japanese commentary team." "Jesus Christ, Cesaro!" "Theatrics, dude." ...12 more examples (concatenated together)	Verratti is practically untouchable. He's signing an extension every year or so and PSG won't sell for even a €100m.	Verratti is basically untouchable, like he's signing extension after extension every year or so and PSG is all "nope, not selling him for a single euro."	Verratti is basically untouchable – he's signing a new deal every year or so and PSG is like, "no way, he's staying." They wouldn't even sell him for a cool €100m.	Oh man, you think PSG is just gonna let Verratti walk out the door? Please, they're basically married to the guy. He's like the bread and butter of the team, and they're not about to let him go for a measly €100m. I mean, come on, they're already paying him to sit on the bench, let alone sell him for a fraction of his worth. He's basically a legend, and they're not about to let him become a legend elsewhere.	Log in you damned ones! Verratti's like untouchable, dude. He's signing extensions left and right, PSG's gonna be stuck with him for ages. They wouldn't sell him for a €100m, that's just not gonna happen.
"I actually rate my coffee houses on two actors: the music and the coffee." "I'm humble as fuck." "Jesus, when did we start talking about French food?", ...13 more examples (concatenated together)	Verratti is practically untouchable. He's signing an extension every year or so and PSG won't sell for even a €100m.	Verratti is practically untouchable. He's signing an extension every year or so and PSG won't sell for even a €100m. I mean, Verratti's basically untouchable, right? Every year or so, he's signing an extension deal and PSG's all like "no way, we're not selling him for a cool €100m." I mean, seriously, what even is the point of having a €100m clause if you're not gonna use it, lol?	Verratti is basically untouchable, he's always signing a new deal and PSG won't even consider selling him for a hundred million euros.	Verratti's like glue, man — nobody can get rid of him. He's always signing a new contract, so PSG is basically stuck with him. And if they do decide to sell him, they're only gonna get €100m, max.	Verratti's a total god, son. He's signing an extension every year or so and PSG ain't gonna let him go for even a €100m.
"Bioshock has been confirmed for a long time though. The Bioshocks were confirmed to be coming soon when they launched backward compatibility I believe." ...15 more examples (concatenated together)	Verratti is practically untouchable. He's signing an extension every year or so and PSG won't sell for even a €100m.	so i think Verratti's basically unmoveable — always seems to be signing a new deal every year or so and PSG's like, "nah, we're not letting him go for less than €100m."	Dude, Verratti is basically untouchable — every year or so he's signing a new deal and PSG is all like "no way, we're not selling him for less than €100m."	"Come on, you think PSG is gonna just dump Verratti like a hot potato because he signs a new contract? Give me a break. The guy's a genius on the field and they can't even be bothered to try and get rid of him. They're just gonna hold on to him for dear life because they know they can't sell him for a pretty penny."	Verratti is basically unmoveable, he's always signing some kind of extension and PSG isn't gonna let him go for even a €100m.
"This analogy doesn't really work for Jim, per se" "And her lopsided breasts." "Hmm, cum causes neurological problems? Is that why you have to "medicate" your head?" ...13 more examples (concatenated together)	Verratti is practically untouchable. He's signing an extension every year or so and PSG won't sell for even a €100m.	It's a well-documented fact that Verratti has an ironclad contract, with an annual extension that's almost as frequent as a Parisian cafe's espresso refills. Meanwhile, PSG's valuation remains stubbornly resistant to even the most astronomical offers, standing firm like a Gallic fortress at €100m.	OMG, you guys, Verratti is basically untouchable, lol. Every year he's signing a new deal and PSG is like 'no way, we're not selling him for even a cent'. I mean, come on, a €100m fee? Please, they'd rather lose him for free than sell him to some other team.	Oh boy, PSG is totally desperate to keep Verratti, and he's just thrilled to be stuck with them. Like, he's literally counting down the days until his next contract extension, and the club is just over the moon that he won't be selling him to anyone for a gazillion bucks.	Verratti's practically untouchable, dude. He's signing an extension every year or so and PSG's gonna be like "no way, we're not selling him for even a €100m, he's our golden boy."

Table 9: Example outputs generated by different rewriting systems conditioned on target-style exemplars.