DR GENRÉ: Reinforcement Learning from Decoupled LLM Feedback for Generic Text Rewriting

Yufei Li[1*] **John Nham**[2] **Ganesh Jawahar**[2] **Lei Shu**[2] **David Uthus**[2]
Yun-Hsuan Sung[2] **Chengrun Yang**[2] **Itai Rolnick**[2] **Yi Qiao**[3] **Cong Liu**[1]
[1]UC Riverside [2]Google DeepMind [3]Google Cloud

Abstract

Generic text rewriting is a prevalent large language model (LLM) application that covers diverse real-world tasks, such as style transfer, fact correction, and email editing. These tasks vary in rewriting objectives (e.g., factual consistency vs. semantic preservation), making it challenging to develop a unified model that excels across all dimensions. Existing methods often specialize in either a single task or a specific objective, limiting their generalizability. In this work, we introduce a generic model proficient in *factuality*, *stylistic*, and *conversational* rewriting tasks. To simulate real-world user rewrite requests, we construct a conversational rewrite dataset, CHATREWRITE, that presents "natural"-sounding instructions, from raw emails using LLMs. Combined with other popular rewrite datasets, including LONGFACT for the factuality rewrite task and REWRITELM for the stylistic rewrite task, this forms a broad benchmark for training and evaluating generic rewrite models. To align with task-specific objectives, we propose DR GENRÉ, a **D**ecoupled-**r**eward learning framework for **Gen**eric **re**writing, that utilizes objective-oriented reward models with a task-specific weighting. Evaluation shows that DR GENRÉ delivers higher-quality rewrites across all targeted tasks, improving objectives including instruction following (*agreement*), internal consistency (*coherence*), and minimal unnecessary edits (*conciseness*).

1 Introduction

Text rewriting is a fundamental NLP task with applications spanning style transfer (Shu et al., 2024), summarization, and fact correction (Wei et al., 2024). Existing models often specialize in a particular transformation type, such as paraphrasing (Siddique et al., 2020), sentence fusion (Mallinson et al., 2022), or focus on optimizing a specific objective, such as syntactic overlap with reference texts, during post-training (Shen et al., 2017; Hu et al., 2017). This specialization limits their applicability in real-world scenarios where diverse rewriting capabilities are required. In this work, we address the challenges of building a generic rewrite model capable of handling multiple rewriting tasks, including fact correction, style transfer, and conversational rewriting.

Versatile rewriting tasks. *Factuality rewrite* involves correcting content that contains factual errors in initial responses generated by a large language model (LLM) responding to fact-seeking prompts on open-ended topics (Hu et al., 2023). The rewrite model takes the prompt, initial response, and critique outputs (e.g., span-level checks from autoraters), and generates revised responses that correct non-factual claims, maintain global coherence (a challenge for standard post-training methods, as shown in Table 3), and minimally edit the factual claims. *Stylistic rewrite* focuses on transforming a source text into another style without introducing new information (Shu et al., 2024), including formalization, paraphrasing, and summarization (Li et al., 2018; Zhang et al., 2020). While this task has been studied extensively in the in-context learning (ICL) setting (Brown et al., 2020; Raffel et al., 2020), few-shot methods struggle to follow user-specified instructions accurately (as shown in Table 4). Both factuality and stylistic rewrite expose limited user applicability. To mimic real-world user requests, we introduce a new rewriting task—*conversational rewrite*—which addresses the need for modifying specific parts of a text based on user instructions that may lack detailed context. For instance, a user might say, "This email is a bit dry; let's celebrate our success! Add some enthusiastic phrases like 'We nailed it!'". We construct a benchmark dataset, CHATREWRITE, through multi-turn in-

[*]Work done as a research intern at Google DeepMind. Correspondence: Yufei Li (yli927@ucr.edu), John Nham (jnham@google.com), Ganesh Jawahar (ganayu@google.com).

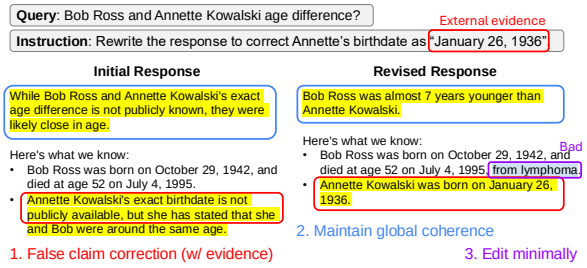

Figure 1: An example illustrating the three objectives for generic text rewriting: Agreement: Follow the rewrite instruction (e.g., false claim correction). Coherence: Maintain global coherence across revisions (e.g., updating the age difference to maintain logical flow). Conciseness: Avoid unnecessary edits (e.g., specifying Bob's death place "lymphoma" is irrelevant).

struction prompting (see Table 1) and revised conversation generation using LLMs.

Multi-objective intrinsic. From analyzing these tasks, we decouple three major objectives that are preferred for a high-quality rewrite, as shown in Figure 1. *Agreement*: the revised response should strictly follow the rewrite instruction, such as accurately correcting false claims (e.g., revising "Annette Kowalski's exact birthdate is not publicly available..." to "Annette Kowalski was born on January 26, 1936"). *Coherence*: the revised response should maintain global coherence after the desired corrections are made at a local level, such as modifying contradictory contents (e.g., changing "Bob Ross and Annette Kowalski's exact age difference is not publicly known..." to "Bob Ross was almost 7 years younger than Annette Kowalski"). *Conciseness*: the revision should avoid unnecessary edits and make minimal changes, such as updating the age difference calculation without altering unrelated details (e.g., "from lymphoma").

A generic framework. To build a versatile rewrite model, we propose DR GENRÉ, a **D**ecoupled-**r**eward learning framework for **Gen**eric **re**writing. At the starting supervised fine-tuning (SFT) stage, we train a student model on a mixture of the three task-specific datasets. During reward modeling, due to limited human preferences, we distill both the agreement and coherence preferences from a teacher LLM to derive task-agnostic, objective-oriented reward models (RMs) (Stiennon et al., 2020; Lee et al., 2023), and employ rule-based edit ratio as the conciseness reward. At the reinforcement learning (RL) stage, we compute the final preference for a input rewrite prompt and its corresponding rewritten response by performing a task-specific weighting of objective-oriented reward models. DR GENRÉ offers a fine-grained control over the alignment direction by adjusting the weights according to specific task requirements. Our contributions are threefold:

- We introduce conversational rewrite—a new rewriting task that is more challenging yet user-applicable—and create a dataset CHATREWRITE for benchmarking.
- We propose DR GENRÉ, a post-training framework for generic text rewriting, and establish robust baselines using few-shot LLMs, SFT, and single-reward RL across the three tasks.
- Experiments show that weighted decoupled rewards offer enhanced control over the alignment direction, leading to improved performance across multiple rewriting objectives.

2 Related Work

Text rewriting. Existing research on rewriting often focus on a particular set of rewriting tasks, including factuality correction and style transfer, aiming to improve accuracy, tone, or coherence.

Factuality rewrite addresses factual inaccuracies in generated responses where merely prompting LLMs such as GPT-4 (Han et al., 2024) cannot promise fact-related (e.g., dates, locations, statistics) accuracy in generation (Li et al., 2023), especially when dealing with open-ended or fact-seeking prompts. Existing methods often leverage external knowledge bases (Shen et al., 2017), fact-checking modules (Hu et al., 2017), or post-editing strategies (Hu et al., 2023), though balancing correctness with minimal edits remains challenging.

Style transfer, covering tasks like paraphrasing (May, 2021), formalization (Rao and Tetreault, 2018; Li et al., 2024b), and elaboration (Iv et al., 2022), adapts tone without altering meaning. Recent methods like RewriteLM (Shu et al., 2024) extend rewriting to broader domains but are limited by single reward granularity. Our work unifies these objectives, leveraging decoupled rewards to handle diverse rewriting tasks while ensuring instruction adherence, coherence, and minimal edits.

Data augmentation with LLMs. Leveraging LLMs for data augmentation has emerged as a widely adopted approach to enhance model performance by generating synthetic data for training. Early works (He et al., 2020; Huang et al., 2023) focused on augmenting data distributions to improve performance, while more recent advance-

Raw generated email	**Rewrite instruction** (Raw→Specific→Natural)
Dear [Customer Name], We hope this email finds you well. We're writing to you today to let you know about a new way we're improving our social media presence. We've created a new account on [social media platform], and we'd love for you to follow us! We'll be using this account to share news about our company, our products, and our industry. We'll also be posting photos and videos, and we'll be running contests and giveaways. We hope you'll join us on [social media platform]! We're looking forward to connecting with you there. Sincerely, [Your Name]	**Raw**: Specify the social media platform and highlight the specific benefits for customers. **Specific**: Specify the social media platform as TikTok and highlight the specific benefits for customers such as exclusive behind-the-scenes content, early access to product launches, and the chance to win prizes in contests and giveaways. **Natural**: Instead of just saying [social media platform], say we're now on TikTok! Also, let's tell them about the cool stuff they'll find there, like exclusive behind-the-scenes content, early access to new products, and even the chance to win prizes!

Table 1: An example of multi-turn instruction generation from our CHATREWRITE dataset. The instruction is first specified with more details, and then reorganized and expressed in a more human-like speaking style.

ments, such as PEER (Schick et al., 2023), demonstrated the effect of infilling missing data with synthetic samples. Methods like Self-Instruct (Wang et al., 2023) bootstrap instructions and model outputs to boost task-specific accuracy. Further research (Li et al., 2024a; Han et al., 2024) highlights the capability of LLMs to produce high-quality augmented data for diverse downstream tasks. Building upon these principles, our work uses LLMs to synthesize CHATREWRITE—a more challenging conversational rewrite dataset that aligns with real-world user-LLM interaction scenarios.

LLM-as-a-judge (AutoRater). LLMs, with their superiority in language understanding and knowledge integration, have been widely used as AutoRaters to judge the quality of generated responses due to high costs of human evaluation (Vu et al., 2024). In some benchmarks, LLM agents even exhibit better annotation capabilities than humans (Wei et al., 2024). We use Gemini-1.5-Ultra (Team et al., 2023) to auto-evaluate agreement, coherence and compare pairwise responses in their instruction satisfaction granularity.

Reinforcement Learning from AI feedback (RLAIF). RL from human feedback (RLHF) is a technique that combines RL with human evaluations to fine-tune LLMs (Christiano et al., 2017). In RLHF, a LLM generates outputs that are assessed by human evaluators, who provide feedback indicating preferences or ratings based on certain criteria (e.g., helpfulness, correctness, style). This feedback is used to train a RM that predicts the human-provided scores. The LLM is then fine-tuned using RL algorithms like Proximal Policy Optimization (PPO) (Schulman et al., 2017), optimizing the rewards to improve alignment with human preferences (Raffel et al., 2020). However, collecting large-scale human feedback is resource-intensive, motivating the exploration of RLAIF (Lee et al., 2024) such as leveraging LLMs to simulate human evaluations and derive reward preferences.

Dataset	**Size**	**IR len**	**RR len**	**ER**
LongFact	21,294	316	296	0.281
RewriteLM	29,985	131	127	0.729
ChatRewrite	82,290	108	111	0.650

Table 2: Statistics of the three datasets. "IR len" and "RR len" denote the lengths of initial and revised responses. "ER" denotes the edit ratio.

3 Dataset Generation

We describe the construction of our dataset, covering the three distinct tasks. Its statistics is shown in Table 2. Each dataset is generated through structured prompting that ensures high-quality rewrites tailored to specific objectives.

3.1 Factuality Rewrite

We follow LONGFACT (Wei et al., 2024) to generate factually rich, multi-faceted responses. First, we prompt LLMs to generate fact-seeking queries, such as "What happened in the first modern Olympics?". For each query, we obtain an initial response (IR) from LLMs, which is then passed to a critic model[1] that provides span-level factuality critiques and suggested revisions (see Table 9 in Appendix B). To construct the dataset, we integrate the query, IR, and critique outputs to prompt LLMs (see Table 11 in Appendix C) and gener-

[1]We employ SAFE (Wei et al., 2024) which enables external fact-checking calls from LLMs.