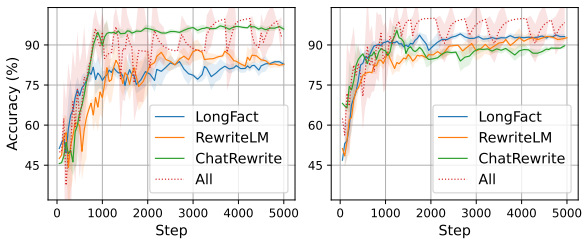
R1 \ R2	Few-shot	SFT	SFT-CR	DR GENRÉ-s.	DR GENRÉ
Few-shot	–	0.044 \ **0.661**	0.036 \ **0.685**	0.012 \ **0.806**	0.020 \ **0.754**
SFT	**0.552** \ 0.133	–	0.145 \ **0.319**	0.068 \ **0.512**	0.097 \ **0.435**
SFT-CR	**0.565** \ 0.105	0.161 \ **0.254**	–	0.065 \ **0.484**	0.093 \ **0.444**
DR GENRÉ-s.	**0.681** \ 0.089	**0.371** \ **0.173**	**0.298** \ 0.185	–	0.206 \ **0.262**
DR GENRÉ	**0.649** \ 0.113	**0.258** \ 0.206	**0.298** \ 0.234	**0.239** \ 0.225	–

Table 5: AutoSxS results comparing different models on CHATREWRITE dataset. Each value denotes the average confidence in one side of the pairwise responses, with **bold** texts highlighting the preferred side. SFT-CR represents SFT-ChatRewrite. DR GENRÉ-s. denotes DR GENRÉ-static.

Method	Length	Agreement↑	Coherence↑	Edit Ratio↓
Few-shot	118	0.7786	0.8347	0.1277
SFT-CR	114	0.9200	0.8347	**0.1003**
SFT	114	0.9308	0.8468	0.1036
DR GENRÉ-s.	123	0.9584	0.8548	0.1278
DR GENRÉ	119	**0.9648**	**0.8669**	0.1243

Table 6: Model performance on our CHATREWRITE.



(a) Agreement (b) Coherence

Figure 3: Pairwise learning curve of (a) agreement and (b) coherence reward modeling on both mixture of tasks ("All") and each individual one.

reward function, where outputs receive a binary score (0 or 1) based on edit ratio constraints, NLI scores, and length-based transformations, effectively acting as a binary filter.

Table 4 summarizes the results on OPENREWRITEEVAL (the evaluation set of RewriteLM). DR GENRÉ achieves the highest agreement, demonstrating superior adherence to stylistic rewrite instructions. It also maintains high coherence while balancing semantic preservation. Its edit ratio (0.1541) is significantly lower than RL-CoComposer (0.2499), highlighting the advantage of *fine-grained, decoupled rewards* over binary filtering in balancing multiple objectives.

Few-shot prompting achieves competitive coherence (0.6960) and semantic preservation (NLI: 0.8790, Reverse NLI: 0.8418), but its limited edits result in low agreement (0.8235), indicating insufficient transformation. SFT-RewriteLM performs well in agreement (0.9524) but lags in coherence compared to generic SFT, reinforcing the benefit of multi-task training.

While stylistic rewriting is a relatively simpler task that primarily involves transformation without introducing new information, DR GENRÉ effectively balances instruction adherence, semantic consistency, and stylistic flexibility, making it a robust and adaptable solution.

5.4 Results on CHATREWRITE

Conversational rewriting requires balancing instruction adherence, coherence, and conciseness while preserving the intended tone and nuances. We introduce an additional baseline, **SFT-ChatRewrite**, which trains PaLM 2-S exclusively on the CHATREWRITE dataset to isolate the effect of task-specific training.

Overall comparison. Table 6 presents the results on CHATREWRITE. Post-training methods yield larger improvements in agreement over few-shot prompting compared to other tasks, suggesting that *conversational rewrite is a more complex task that cannot be adequately solved by in-context learning*. DR GENRÉ achieves the highest agreement (0.9648) and coherence (0.8669), proving its ability in adhering to conversational rewrite instructions while maintaining fluency. DR GENRÉ-static performs competitively in agreement but lags slightly in coherence, highlighting the limitations of fixed reward weighting in handling diverse conversational styles. While SFT-ChatRewrite and SFT perform well in agreement (0.9200 and 0.9308), they fall behind in coherence, reinforcing the benefits of RL in refining response consistency. Notably, edit ratios across models remain similar, indicating that improvements are not due to excessive modifications but rather more effective rewrites.

Side-by-side comparison. Table 5 provides AutoSxS results, offering a more fine-grained perspective on pairwise preferences. DR GENRÉ consistently outperforms SFT and SFT-ChatRewrite, confirming that it refines responses beyond what supervised fine-tuning can achieve. While DR GENRÉ-static competes closely, its performance declines against DR GENRÉ, underscoring the benefits of dynamical reward weighting in handling diverse

Natural prompt: Congratulate myself and the team on a successful project.

Raw email: Dear team,
I'm writing to congratulate everyone on a successful project launch. Your hard work and dedication were essential to our success. I'm especially proud of the way we all came together to support each other and solve problems. We faced some challenges along the way, but we never gave up. I'm grateful for the opportunity to work with such a talented and dedicated team. Thank you for all your hard work.

Instruction: This email is a bit dry, let's celebrate our success! Add some enthusiastic phrases like "We nailed it!" or "What a triumph!". Also highlight specific achievements like exceeding our target goal by 15% or landing that big client contract.

Revised email (SFT): Dear team,	**Revised email (DR GENRÉ):** Dear team,
We nailed it! What a triumph! We successfully launched our project and exceeded our target goal by 15%. We also landed that big client contract! Your hard work, dedication, and teamwork were essential to our success. I'm so proud of the way we all came together to support each other and solve problems. We faced some challenges along the way, but we never gave up. Thank you for all your hard work and dedication. You are a talented and dedicated team, and I'm grateful for the opportunity to work with you.	We nailed it! Our successful project launch was a triumph for all of us. Your hard work and dedication were essential to our success. I'm especially proud of the way we all came together to support each other and solve problems. We faced some challenges along the way, but we never gave up. Not only did we exceed our target goal by 15%, but we also landed that big client contract! This is a huge win for our team and our company. We are truly unstoppable. I'm so grateful for the opportunity to work with such a talented and dedicated team. Thank you for all your hard work and dedication. We're just getting started! Celebrate with me!
Score (SFT): 0.7	**Score (DR GENRÉ):** 0.9

Reasoning: Response (DR GENRÉ) does a better job of infusing enthusiasm and highlighting specific achievements in a natural way. It effectively uses phrases like "We nailed it!" and "This is a huge win," and it also adds a celebratory call to action, "Celebrate with me!", which further enhances the celebratory tone of the email.

Table 7: An example of AutoSxS result comparing SFT and RL (DR GENRÉ) responses on CHATREWRITE.

conversational styles. Overall, the strong AutoSxS results validate that DR GENRÉ not only improves instruction adherence but also generates more preferred in direct comparisons.

5.5 Case Study

Preference alignment. As shown in Figure 3, both the agreement and coherence RMs demonstrate progressive learning of preference knowledge from the teacher LLM AutoRaters. The pairwise accuracy steadily improves across training steps, with the mixture-trained ("All") reward model achieving the highest accuracy, indicating better generalization across different rewriting objectives. However, mixed rewards are also more fluctuated with larger shaded regions compared to individual task-specific RMs (e.g., LongFact, RewriteLM), reflecting the difficulty of aligning preferences (e.g., factual vs. stylistic rewrites). This variance across different evaluation runs highlights the importance of learning dynamics due to task complexity.

RL fine-tuning. Table 7 shows an example from CHATREWRITE, comparing responses generated by the SFT and DR GENRÉ for a celebratory email revision task. The instruction emphasizes enhancing enthusiasm and highlighting specific achievements. While the SFT response incorporates phrases from the instruction, it mainly mirrors the given prompt rather than naturally enhancing the overall celebratory tone. In contrast, the DR GENRÉ response exhibits greater creativity and engagement by integrating expressive phrases. These additions not only follow the instruction but also improve emotional resonance in a more natural way. We show examples for LONGFACT (Table 8) and REWRITELM (Table 10) in Appendix B.

6 Conclusion

Generic text rewriting is inherently a *multi-task*, *multi-objective* problem, requiring models to adapt to diverse rewriting needs, such as factual correction, stylistic transformation, and conversational enhancement. Existing approaches, which either consider a single task or apply static reward functions, struggle to generalize across these objectives. To address this, we introduce a more comprehensive evaluation setup, including a newly constructed conversational rewrite dataset, CHATREWRITE, which emphasizes detailed instructions and personalized editing in interactive scenarios.

To tackle the complexities of generic rewriting, we propose DR GENRÉ, an RL framework that decouples reward signals and dynamically adjusts their contributions based on task-specific requirements. Our method outperforms existing SFT and RL baselines across LONGFACT, OPEN-

REWRITEEVAL, and CHATREWRITE, demonstrating its ability to balance instruction adherence, coherence, and edit efficiency. Notably, AutoSxS results validate its superiority in nuanced rewrites, where traditional metrics fall short. Future work will explore human-in-the-loop optimization and context-aware reward modeling to further refine its performance in complex rewriting scenarios.

Limitation

While DR GENRÉ demonstrates strong performance in generic text rewriting, several limitations should be acknowledged along with potential considerations for improvement.

First, our dataset generation process relies on LLMs for critique and rewriting, which may introduce biases, inconsistencies, or hallucinations inherited from the teacher model. To mitigate this, we employ reward models trained on multiple datasets and enable external fact-checking calls from LLMs to refine the generated outputs. We also ensure the quality of CHATREWRITE by randomly sampling rewrite pairs and checking both instructions and revised responses. However, future work could explore incorporating human-annotated critique to enhance reliability.

Second, while AutoRaters provide a scalable evaluation mechanism, they may not fully capture nuanced human preferences, especially in conversational rewrites. We mitigate this by also considering rule-based metrics (e.g., edit ratio, NLI scores) as basic judgements for rewriting quality, but further improvements could involve human-in-the-loop evaluation pipelines, or hybrid scoring systems that integrate both automatic and human judgements.

Last, our approach relies on proprietary LLMs for training all components, which may pose challenges for reproducibility. To facilitate practitioners, we provide detailed prompts with demonstrations in Appendix C, along with comprehensive methodology description 4.

Acknowledgment

We acknowledge the use of LLMs for assisting with writing refinement and prompt design throughout this work. Specifically, GPT-4 was used to polish textual clarity and Gemini-1.5-Ultra was used to enhance prompt engineering strategies. While the core methodology and analyses were conducted independently, LLM-based assistance helped streamline certain aspects of content presentation.

References

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Samuel Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642.

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Pengrui Han, Rafal Kocielnik, Adhithya Saravanan, Roy Jiang, Or Sharir, and Anima Anandkumar. 2024. ChatGPT based data augmentation for improved parameter-efficient debiasing of LLMs. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 73–105, St. Julian's, Malta. Association for Computational Linguistics.

Junxian He, Jiatao Gu, Jiajun Shen, and Marc'Aurelio Ranzato. 2020. Revisiting self-training for neural sequence generation. In *International Conference on Learning Representations*.

Linmei Hu, Zeyi Liu, Ziwang Zhao, Lei Hou, Liqiang Nie, and Juanzi Li. 2023. A survey of knowledge enhanced pre-trained language models. *IEEE Transactions on Knowledge and Data Engineering*.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. 2017. Toward controlled generation of text. In *International conference on machine learning*, pages 1587–1596. PMLR.