Jiaxin Huang, Shixiang Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2023. Large language models can self-improve. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1051–1068.

Robert Iv, Alexandre Passos, Sameer Singh, and Ming-Wei Chang. 2022. Fruit: Faithfully reflecting updated information in text. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3670–3686.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, et al. 2023. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Ren Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, et al. 2024. Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback. In *Forty-first International Conference on Machine Learning*.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: a simple approach to sentiment and style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1865–1874.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. Halueval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464.

Yichuan Li, Kaize Ding, Jianling Wang, and Kyumin Lee. 2024a. Empowering large language models for textual data augmentation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12734–12751, Bangkok, Thailand. Association for Computational Linguistics.

Zelong Li, Wenyue Hua, Hao Wang, He Zhu, and Yongfeng Zhang. 2024b. Formal-llm: Integrating formal language and natural language for controllable llm-based agents. *arXiv preprint arXiv:2402.00798*.

Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. 2022. EdiT5: Semi-autoregressive text editing with t5 warm-start. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2126–2138, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Philip May. 2021. Machine translated multilingual sts benchmark dataset.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may i introduce the gyafc dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 129–140.

Eric Sven Ristad and Peter N Yianilos. 1998. Learning string-edit distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(5):522–532.

Timo Schick, A Yu Jane, Zhengbao Jiang, Fabio Petroni, Patrick Lewis, Gautier Izacard, Qingfei You, Christoforos Nalmpantis, Edouard Grave, and Sebastian Riedel. 2023. Peer: A collaborative language model. In *The Eleventh International Conference on Learning Representations*.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pages 4596–4604. PMLR.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. *Advances in neural information processing systems*, 30.

Lei Shu, Liangchen Luo, Jayakumar Hoskere, Yun Zhu, Yinxiao Liu, Simon Tong, Jindong Chen, and Lei Meng. 2024. Rewritelm: An instruction-tuned large language model for text rewriting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18970–18980.

AB Siddique, Samet Oymak, and Vagelis Hristidis. 2020. Unsupervised paraphrasing via deep reinforcement learning. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1800–1809.

Hrituraj Singh, Gaurav Verma, Aparna Garimella, and Balaji Vasan Srinivasan. 2021. Drag: Director-generator language modelling framework for non-parallel author stylized rewriting. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 863–873.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Katherine Tian, Eric Mitchell, Huaxiu Yao, Christopher D Manning, and Chelsea Finn. 2024. Fine-tuning language models for factuality. In *The Twelfth International Conference on Learning Representations*.

Ricardo Vilalta and Youssef Drissi. 2002. A perspective view and survey of meta-learning. *Artificial intelligence review*, 18:77–95.

Tu Vu, Kalpesh Krishna, Salaheddin Alzubi, Chris Tar, Manaal Faruqui, and Yun-Hsuan Sung. 2024. Foundational autoraters: Taming large language models for better automatic evaluation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17086–17105, Miami, Florida, USA. Association for Computational Linguistics.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, and Quoc V Le. 2024. Long-form factuality in large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 80756–80827. Curran Associates, Inc.

Akhila Yerukola, Xuhui Zhou, Elizabeth Clark, and Maarten Sap. 2023. Don't take this out of context!: On the need for contextual models and evaluations for stylistic rewriting. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11419–11444.

Yi Zhang, Tao Ge, and Xu Sun. 2020. Parallel data augmentation for formality style transfer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3221–3228.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

# A Experimental Details

We describe the hardware setup, training configurations, and dataset weighting strategies used in different phases of our experiments, covering Supervised Fine-Tuning (SFT), Reward Modeling, and Reinforcement Learning (RL).

**Hardware setup.** All training experiments were conducted on 64 Tensor Processing Units (TPU) chips per phase:

- SFT & Reward modeling: TPU V3.
- RL fine-tuning: TPU V4.

For inference, we use a temperature of 1.0 with top-K sampling (K=40).

**Supervised fine-tuning (SFT).** We fine-tune the base PaLM 2-S model on our dataset mixture using Adafactor (Shazeer and Stern, 2018) with the following configuration:

- Batch size: 64.
- Max training steps: 1000.
- Learning rate: 1e-5.
- Dropout: 0.1.
- Max context length: 2048.
- Max decoding length: 1024.

**Reward modeling.** We train reward models on preference data collected from LLM comparisons using the following setup:

- Batch size: 64.
- Max training steps: 5000.
- Learning rate: 3e-3.
- Dropout: 0.05.
- Max context length: 1280.
- Max decoding length: 1024.
- Optional Z_loss: 1e-2.

**Reinforcement learning (RL) fine-tuning.** For policy optimization, we employ PPO with dynamically weighted multi-objective rewards. The policy and value functions are optimized separately:

- Batch size: 64.
- Max training steps: 3000 (with a warm-up phase of the first 100 steps where we train only value functions and freeze policy).
- Learning rate: 1e-7 for policy and 1e-5 for value.
- Dropout: None.
- Max context length: 2048.
- Max decoding length: 1024.

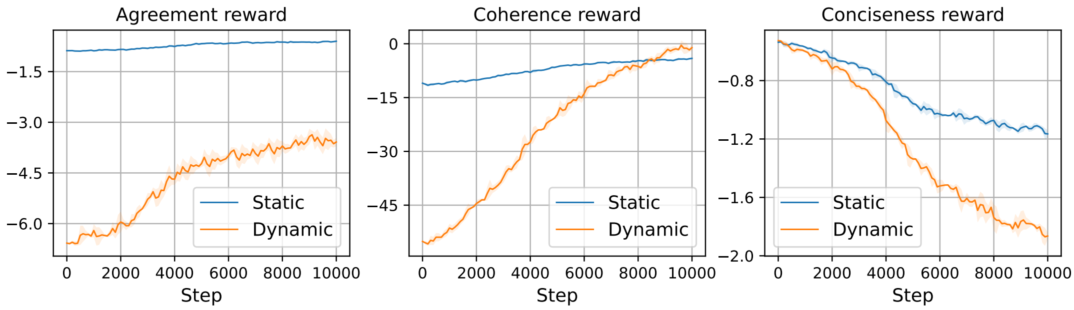


Figure 4: Reward learning curves during RL fine-tuning under static and dynamic weighting.

**Dataset weighting strategy.** We assign different dataset weights based on task-specific objectives to balance training across agreement, coherence, and edit conciseness.

For DR GENRÉ-static (task-agnostic), we use a fixed weighting of $w_1 = 9/16$, $w_2 = 2/16$, $w_3 = 5/16$ for agreement, coherence, and conciseness.

For DR GENRÉ (task-specific), we empirically set:

- LONGFACT: $w_1 = 8/16$, $w_2 = 6/16$, $w_3 = 2/16$.
- REWRITELM: $w_1 = 3/9$, $w_2 = 4/9$, $w_3 = 2/9$.
- CHATREWRITE: $w_1 = 9/16$, $w_2 = 5/16$, $w_3 = 2/16$.

The dynamic weighting scheme ensures that different datasets prioritize their most relevant rewrite objectives, allowing for more effective RL fine-tuning.

**Baseline selection.** In our experiments, we focus on three major categories of baselines: ICL-based, SFT-based, and RL-based methods. There are some existing works in factual or stylistic rewriting focus on either direct editing heuristics or single-objective models that do not fit well with our multi-objective formulation.

For factual rewriting, works like knowledge-grounded editing rely on retrieval-based fact verification or human annotations (Tian et al., 2024), whereas our approach optimizes for factuality without requiring explicit retrieval.

For stylistic rewriting, previous works often rely on large supervised datasets for a single transformation (e.g., formality change, politeness adjustment, style matching (Singh et al., 2021)) or context integration (Yerukola et al., 2023), whereas our model generalizes across multiple stylistic transformations.

Even if the above task-specific approaches perform well in their domain, they do not necessarily generalize across diverse rewriting tasks, making them less suitable as baselines in our setting.

**Static and dynamic weights.** Figure 4 represents the reward curves of the RL fine-tuning phase (DR GENRÉ-static and DR GENRÉ). Dynamic RL (DR GENRÉ) adapts objectives over time, focusing more on agreement and coherence, and less on the conciseness. In contrast, static RL (DR GENRÉ-static) is more stable (balanced) but less optimized learning across objectives. Dynamic RL exhibits stronger overall improvement rates across all three objectives, confirming its ability to adjust to task needs more effectively.

## B Generated Examples

This section presents qualitative examples of factuality and stylistic rewrite cases generated by DR GENRÉ, illustrating its ability to correct errors while preserving coherence and adhering to task-specific instructions.

**Factuality rewrite.** The first example in Table 8 showcases a factuality rewrite on the topic of the United States' involvement in the East Asia Summit (EAS). The initial response contains several factual inaccuracies, such as "Incorrectly stating that the U.S. has been involved in the EAS since its inception in 2005 (corrected to 2011)". The critique outputs (highlighted in red for incorrect and blue for revised content) pinpoint these errors, allowing DR GENRÉ to generate a factually accurate response while maintaining internal coherence. We also show an example of critique outputs from SAFE (Wei et al., 2024) in Table 9, where for each span, the outputs contain a revision (from external fact-checking calls) and a reason. Compared to the initial response, the revised version (i) corrects all factual errors without introducing unnecessary modifications, (ii) preserves the original structure and key ideas, and (iii) Improves clarity by streamlining redundant information (e.g., simplifying the