

Improving Synthetic Data Generation with LLMs through Strategic Comparisons

Yao Rong¹, Shuo Yang², Gjergji Kasneci², Enkelejda Kasneci²

¹ Rice University, ² Technical University of Munich

yao.rong@rice.edu, {shuo.yang, gjergji.kasneci, enkelejda.kasneci}@tum.de

Abstract

Large Language Models (LLMs) have shown advanced capabilities in tasks like counterfactual generation and style transfer using prompt strategies. However, previous strategies lacked detailed instructions, limiting effectiveness. To address this, we introduce Compare&Generate, an algorithm inspired by human comparison, where minimal instructions lead to substantial learning. Specifically, our method incorporates an objective function that quantitatively assesses alignment with the task goal and the content relevance in the output. Then, it constructs comparison pairs based on generation assessments and prompts the model to reconsider how to optimize its output. Through comparison, the model focuses on the critical aspects of the task objective and refines its outputs accordingly. We benchmark our method with single-instruction as well as iterative refinement approaches across three natural language generation tasks. Experimental results show that our approach outperforms other related methods; for instance, it surpasses its single-instruction base by 17% and a state-of-the-art refinement approach by 7% on IMDB datasets in generated label accuracy, highlighting the effectiveness of using comparisons in prompts to enhance LLMs.

Introduction

Large Language Models (LLMs) demonstrate promising results in tackling diverse tasks ranging from natural language understanding to generation. Especially in generation tasks, these models show high proficiency and creativity (Yang et al. 2024). Natural language generation tasks include converting input texts into new sequences, such as style transfer (Jin et al. 2022; Reif et al. 2022), and counterfactual augmentation (Li et al. 2023; Chen et al. 2023b). These tasks are valuable for various applications. For instance, counterfactual augmentation increases data samples and reduces data annotation costs, which can improve the training of smaller language models. Moreover, style transfer allows chatbots to communicate with diverse user groups using appropriate tones.

LLMs always have a comprehensive understanding of the input contents and task requirements, making it possible to solve generation tasks without explicit fine-tuning. For example, providing models with several examples can complete various tasks such as sentiment analysis, natural language

inference, and style transfer (Min et al. 2022; Reif et al. 2022; Yang et al. 2024). Furthermore, algorithms that enhance the model’s reasoning, such as Chain-of-Thought (CoT) demonstrating some example reasoning steps, (Wei et al. 2022), can significantly improve performance. However, these single-instruction algorithms rely heavily on the sufficiency and quality of examples, and may not consistently deliver optimal performance. To improve this, recent works (Madaan et al. 2024; Shinn et al. 2023; Pryzant et al. 2023; Chen et al. 2023a) suggest integrating feedback into prompts to improve the output iteratively. While these works use natural language as feedback, Yang et al. (2023) propose Optimization by PROMpting (OPRO) to generate objective values for each solution. These values are then used by LLMs as optimizers to refine the next generations, implicitly summarizing characteristics across multiple solutions. However, the challenge lies in effectively communicating instructions to refine outputs when designing prompts.

To enhance the effectiveness of prompts to guide the model in generating better outputs, we utilize the power of *comparison*. Comparison is an important scheme in humans’ learning process (Rittle-Johnson and Star 2011; Christie 2022). By comparing, we can capture *substantial* information with *minimal* instructions. Inspired by its efficiency, we introduce a novel output refinement algorithm named Compare&Generate (C&G), which leverages strategic comparisons in the prompts to provide the model with extensive information in a concise format. To build comparisons, the best examples from previous generations are selected to guide the model along the desired trajectory. Furthermore, our algorithm explicitly prompts the model to think about why the better example stands out and then generate a new output following the good characteristics. Compared to prior works that utilize prompts as instructions, our method has the advantage of concretizing the optimization aspects and communicating efficiently with the model.

Figure 1 illustrates the workflow of our C&G compared to CoT and OPRO. CoT breaks down the task into steps without refining the output, whereas OPRO includes “solution-value” pairs in its prompts. Our algorithm first compares two selected sentences from the previous generation, and prompts the model with the results of this comparison. This enables the model to generate a new output by *analyzing* the reasons one example outperforms another.

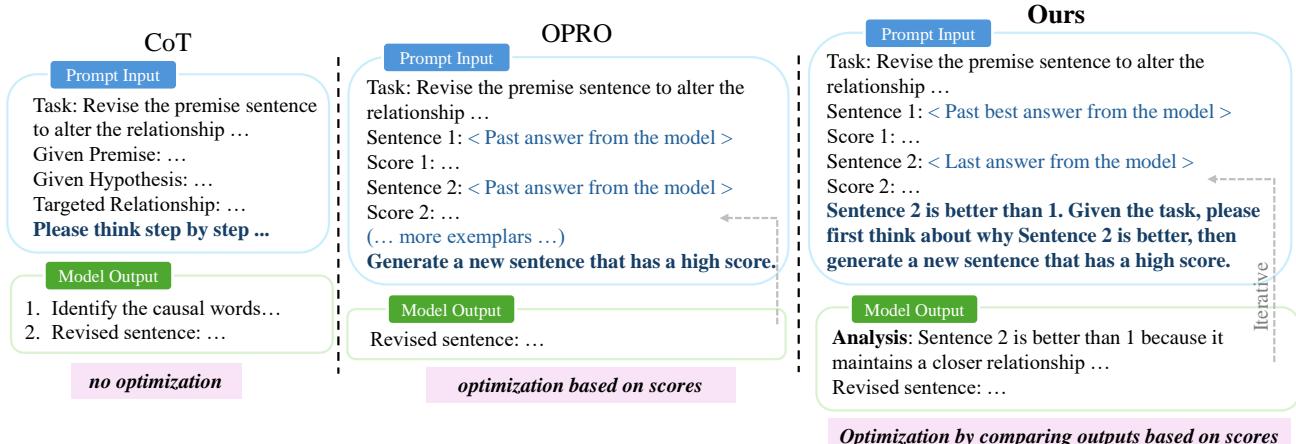


Figure 1: Illustration of using our Compare&Generate compared to other methods CoT (Wei et al. 2022) and OPRO (Yang et al. 2023). Our method improves the output by comparing, analyzing, and progressively refining previous results to achieve a superior score.

To summarize, our contributions are as follows:

- We introduce a novel algorithm, Compare&Generate, designed for various tasks including counterfactual generation and style transfer. This algorithm enhances the model’s ability to reflect and optimize outputs by incorporating strategic comparisons in the prompts.
- Our method utilizes a comparison mechanism based on an objective function. The objective function is designed to evaluate task alignment and relevance to the input content.
- We benchmark our method with other prompt strategies across five datasets. Our approach achieves state-of-the-art performance in generating outputs aligned with task goals, demonstrating effectiveness and generalizability.

Our method does not require additional knowledge or augmentation, enabling seamless integration into various applications with minimal effort. In essence, our results highlight that our method is able to enhance the model’s capability for thoughtful reflection with fewer instructions, and it has consistent stable improvement across multiple backbone models and datasets.

Approach: Compare&Generate

We introduce our method, a novel strategy that refines model output by prompting the model with a pair of examples, along with feedback that enables the model to reflect and improve its output.

The formal setting of our methodology is defined as follows: Given a task description $T(x)$ such as “generate a sentence that changes the sentiment for the sentence x ”, where x is an individual sentence, our method works as a generator of prompts. These prompts guide the given LLM to produce an answer, denoted as \hat{y} . Our method leads the LLM to optimize the \hat{y} in an iterative manner as follows:

$$p_t = \phi((\hat{y}_m, \sigma(\hat{y}_m)), (\hat{y}_{t-1}, \sigma(\hat{y}_{t-1})), T(x)) \quad (1)$$

$$\hat{y}_t = \text{LLM}(p_t), \quad (2)$$

where $\sigma(\cdot)$ is an objective function that assesses the generation, and $\phi(\cdot)$ is the function that utilizes two generated

answers from previous steps and the given LLM model to construct the prompt p_t for the current step. We introduce the design of $\phi(\cdot)$ and $\sigma(\cdot)$, and the workflow of the whole algorithm in the following subsections.

Comparison Prompt Construction

$\sigma(\cdot)$ first evaluates the two given potential answers. For a given \hat{y}_i , this objective function decomposes the task goal and returns a quantitative result to enable the comparison. The objective function evaluates two aspects of \hat{y}_i : (1) **Alignment** sa_i : How well the output matches the task’s goals, ensuring that the generation effectively serves the intended purpose; (2) **Relevance** sr_i : How closely the output relates to the input, ensuring the answer is close to the provided query. Alignment is a score obtained by querying the LLM model on how \hat{y}_i corresponds to the task. Relevance is calculated by a (dis)similarity metric, such as Levenshtein distance, to measure the differences between \hat{y}_i and x . sa_i and sr_i are used to construct the prompt, i.e., an overall score s is calculated as the weighted sum of these variables with α as the weight.

Given a pair of previous \hat{y} and their corresponding sa and sr , $\phi(\cdot)$ constructs the comparison prompt. This assists the model in formulating thoughts to optimize the output. The overview of the p_t is shown in Figure 2. More details of $\sigma(\cdot)$ and p_t can be found in Appendix “Implementation Details”.

Iterative Optimization

C&G optimizes generation output based on previous outputs. Specifically, it utilizes the generation from the previous step ($t-1$) and the best generation before $t-1$ (the output with the highest objective function score, denoted by index m , to construct the comparison prompt for the current step (t)¹. This approach enables the model to learn and explore aspects that optimize the objective function, leading to improved

¹In cases where a lower score indicates better quality, the output with the lowest score will be selected. For simplicity, we refer to this selection criterion as the “highest score”.

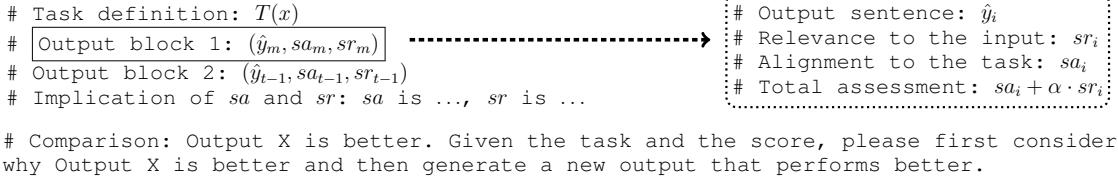


Figure 2: Overview of the prompt constructed by C&G, which incorporates a comparison of two generations.

| | | GPT-3.5 | | | | | Llama3-8B | | | | |
|------|-------------|--------------------|------------------|------------------|-----------------|--------------|--------------------|------------------|------------------|-----------------|--------------|
| | | SBLEU \downarrow | PPL \downarrow | Cont. \uparrow | Acc. \uparrow | H \uparrow | SBLEU \downarrow | PPL \downarrow | Cont. \uparrow | Acc. \uparrow | H \uparrow |
| SNLI | zero-shot | 17.7 | 139 | 0.75 | 0.62 | 0.68 | 20.0 | 103 | 0.71 | 0.54 | 0.61 |
| | few-shot | 16.6 | 100 | 0.74 | 0.59 | 0.66 | 24.2 | 121 | 0.71 | 0.59 | 0.64 |
| | CoT | 20.8 | 119 | 0.70 | 0.48 | 0.57 | 20.5 | 100 | 0.71 | 0.59 | 0.64 |
| | Self-Re | 22.5 | 48 | 0.62 | 0.57 | 0.60 | 25.6 | 43 | 0.53 | 0.81 | 0.64 |
| | OPRO | 19.9 | 85 | 0.76 | 0.60 | 0.67 | 22.7 | 96 | 0.75 | 0.51 | 0.61 |
| | Ours | 20.3 | 97 | 0.78 | 0.66 | 0.72 | 22.8 | 100 | 0.77 | 0.55 | 0.64 |
| MNLI | zero-shot | 6.2 | 112 | 0.79 | 0.52 | 0.63 | 6.6 | 93 | 0.74 | 0.48 | 0.58 |
| | few-shot | 6.4 | 118 | 0.78 | 0.47 | 0.59 | 6.8 | 96 | 0.76 | 0.46 | 0.57 |
| | CoT | 6.6 | 145 | 0.75 | 0.31 | 0.44 | 7.2 | 138 | 0.78 | 0.39 | 0.52 |
| | Self-Re | 25.0 | 53 | 0.56 | 0.38 | 0.45 | 9.8 | 44 | 0.51 | 0.72 | 0.60 |
| | OPRO | 6.1 | 90 | 0.75 | 0.56 | 0.64 | 7.1 | 95 | 0.75 | 0.44 | 0.55 |
| | Ours | 6.3 | 108 | 0.77 | 0.56 | 0.65 | 6.5 | 103 | 0.77 | 0.49 | 0.60 |

Table 1: Comparison of counterfactual data generation on SNLI and MNLI datasets. Harmonic mean (**H**) is highlighted in gray, with the best results in bold.

generation. Our algorithm optimizes the output iteratively. To stop the search, we apply an early stop strategy. If the objective function score does not increase through I_r steps, the generation terminates and returns the generated output with the highest score throughout the trajectory. A maximal step size I_m is set to avoid an endless loop. Finally, the generation with the highest score will be returned. The algorithm of C&G can be found in Algorithm 1 in Appendix.

Experiments

Evaluation Tasks

Counterfactual Data Generation for Natural Language Inference Natural Language Inference models assess the logical relationship between a premise and a hypothesis, i.e., entailment, contradiction, and neutral. In this paper, we employ the two most popular datasets, SNLI (Bowman et al. 2015) and MNLI (Williams, Nangia, and Bowman 2018) for counterfactual data generation. Concretely, we ask the model to revise the premise sentence to change the relationship between it and the hypothesis sentence to a targeted one.

Style Transfer The style transfer task aims to transfer the language style or the sentiment of a sentence while preserving its semantic content. Specifically, we adopt SST-2 (Socher et al. 2013) and IMDB (Maas et al. 2011) datasets for transferring a remark from positive to negative or vice versa. We also study the performance of our method on the text style transfer task. It facilitates the customization and personalization of text styles to meet the needs of different user groups or

application scenarios. We use Grammarly’s Yahoo Answers Formality Corpus (Rao and Tetreault 2018), a parallel corpus of informal and formal text, to assess our model for rewriting sentence styles.

Evaluation Metrics

To comprehensively assess the quality of the generated data, we consider two main aspects of the generated samples: goal-oriented and text-oriented evaluation. Goal-oriented metrics evaluate whether the model generates the data that accomplishes the task and is still relevant to the original data. Specifically, we use the following metrics: Content preservation (**Cont.**): It quantifies the fidelity of the generated text in the semantic meaning of the input text. We deploy BERTScore between the generated text and the original text, or human-annotated text if available, to measure the content preservation. Generation Accuracy (**Acc.**): This metric assesses how often the generated data matches the targeted label. Specifically, we use models that pre-trained on larger datasets for each task to evaluate the accuracy of the generated data. Text-oriented metrics assess whether the generated data mimics human-like quality and has the potential to substitute for human-generated data. Concretely, we use the following metrics: Diversity (**SelfBLEU**): This evaluates whether the model tries to use diverse words as humans do. We use SelfBLEU (Zhu et al. 2018) to represent it following (Chen et al. 2023b). Fluency (**PPL**): it suggests how well the model captures the underlying language structure, indicating that a proficient English speaker could write the generated text.

| | | GPT-3.5 | | | | | Llama3-8B | | | | |
|-------|-------------|--------------------|------------------|------------------|-----------------|--------------|--------------------|------------------|------------------|-----------------|--------------|
| | | SBLEU \downarrow | PPL \downarrow | Cont. \uparrow | Acc. \uparrow | H \uparrow | SBLEU \downarrow | PPL \downarrow | Cont. \uparrow | Acc. \uparrow | H \uparrow |
| SST-2 | zero-shot | 8.5 | 411 | 0.77 | 0.74 | 0.76 | 9.3 | 202 | 0.71 | 0.84 | 0.77 |
| | few-shot | 8.6 | 283 | 0.77 | 0.78 | 0.77 | 8.9 | 205 | 0.72 | 0.83 | 0.77 |
| | CoT | 8.1 | 268 | 0.78 | 0.75 | 0.76 | 9.5 | 195 | 0.71 | 0.83 | 0.77 |
| | Self-Re | 30.7 | 60 | 0.59 | 0.85 | 0.70 | 28.2 | 50 | 0.53 | 0.85 | 0.66 |
| | OPRO | 7.8 | 170 | 0.77 | 0.75 | 0.76 | 9.8 | 143 | 0.71 | 0.86 | 0.77 |
| | Ours | 8.0 | 238 | 0.80 | 0.82 | 0.80 | 8.3 | 184 | 0.74 | 0.87 | 0.80 |
| IMDB | zero-shot | 19.1 | 42 | 0.84 | 0.74 | 0.79 | 24.9 | 39 | 0.83 | 0.92 | 0.87 |
| | few-shot | 18.1 | 43 | 0.89 | 0.65 | 0.75 | 25.4 | 49 | 0.69 | 0.80 | 0.74 |
| | CoT | 15.1 | 49 | 0.85 | 0.73 | 0.79 | 24.8 | 39 | 0.83 | 0.92 | 0.87 |
| | Self-Re | 23.4 | 33 | 0.80 | 0.79 | 0.80 | 36.4 | 25 | 0.65 | 0.68 | 0.67 |
| | OPRO | 16.8 | 41 | 0.78 | 0.80 | 0.79 | 30.4 | 32 | 0.77 | 0.94 | 0.84 |
| | Ours | 19.9 | 39 | 0.84 | 0.82 | 0.83 | 25.1 | 37 | 0.82 | 0.97 | 0.89 |
| GYAFC | zero-shot | 9.5 | 81 | 0.77 | 0.42 | 0.54 | 15.4 | 74 | 0.70 | 0.77 | 0.73 |
| | few-shot | 10.6 | 88 | 0.76 | 0.56 | 0.64 | 16.5 | 90 | 0.67 | 0.76 | 0.71 |
| | CoT | 10.5 | 86 | 0.76 | 0.52 | 0.62 | 14.6 | 102 | 0.69 | 0.55 | 0.61 |
| | Self-Re | 11.4 | 93 | 0.71 | 0.78 | 0.74 | 17.4 | 91 | 0.61 | 0.86 | 0.71 |
| | OPRO | 9.2 | 90 | 0.73 | 0.72 | 0.72 | 12.8 | 112 | 0.63 | 0.83 | 0.71 |
| | Ours | 11.1 | 78 | 0.72 | 0.78 | 0.75 | 11.6 | 93 | 0.69 | 0.87 | 0.77 |

Table 2: Comparison of style transfer on SST-2, IMDB and GYAFC. Harmonic mean (**H**) is highlighted in gray, with the best results in bold.

We use the perplexity measure by GPT-2 (117M). Content preservation and accuracy are two main evaluation metrics across all NLG tasks, thus we use their harmonic mean (**H**) to represent the overall performance ²

Quantitative Results

NLI Counterfactual Generation From the results demonstrated in Table 1, we see that our C&G achieves the highest harmonic mean of accuracy and content preservation using both backbones. Compared to other refinement algorithms, our C&G consistently enhances output by optimizing outcomes to closely align with past best results, unlike OPRO. Furthermore, Self-Refine, which has less constraints on feedback or refinement criteria, tends to produce less accurate text by adding hallucinated content. In particular, it generates longer text than other methods (cf. Figure 4), leading to lower perplexity due to the averaging effect over more words and reduced content preservation. Compared to Llama3, GPT-3.5 demonstrates better performance on both datasets when using our method, suggesting that a more powerful model has advantages in understanding the tasks and analyzing the output. We further illustrate this improvement by employing GPT-4 as the backbone in Appendix.

Style Transfer Table 2 presents the results across three style transfer datasets. C&G consistently outperforms others in terms of harmonic mean. Notably, among the single-

instruction methods, few-shot learning shows stable effectiveness across all datasets. Since we used few-shot learning in the initial setup of our method, we observed a significant improvement in performance. For example, our approach exceeds the baseline by 17% in transfer accuracy on the IMDB dataset. However, the CoT does not consistently yield benefits, as the stepwise reasoning poses challenges for data generation. When using Llama3, our approach generates the most diverse data compared to all baseline methods. Notably, Llama3 outperforms GPT-3.5 in transferring labels, while GPT-3.5 is good at content preservation. In the ST task, which is more straightforward and requires less logical reasoning, Llama3 and GPT-3.5 demonstrate satisfactory performance across various tasks. Qualitative results are in the Appendix.

Conclusion

In this paper, we propose a novel algorithm named Compare&Generate for data generation tasks utilizing LLMs. Our method belongs to the iterative refinement algorithms, and it improves the feedback of the model by utilizing strategic comparisons. In particular, our method utilizes an objective function that quantitatively evaluates the relevance of the content and its alignment with the task objectives of the output. Based on the objective scores, it constructs comparison pairs in the prompt and encourages the model to rethink how to enhance its output. We compare our method against state-of-the-art methods across three natural language generation tasks. Our results show that our method surpasses other baselines in the harmonic mean of content preservation and generation accuracy. This highlights the effectiveness of using an objective function to construct comparisons in prompts, enhancing the performance of LLMs.

²We use the generated data for augmentation, and the test accuracy with augmentation aligns with the H-mean. This confirms that the H-mean can represent the quality of the generated data in practical usage.

References

Akbik, A.; Bergmann, T.; Blythe, D.; Rasul, K.; Schweter, S.; and Vollgraf, R. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics (demonstrations)*, 54–59.

Babakov, N.; Dale, D.; Gusev, I.; Krotova, I.; and Panchenko, A. 2023. Don’t Lose the Message While Paraphrasing: A Study on Content Preserving Style Transfer. In Métais, E.; Meziane, F.; Sugumaran, V.; Manning, W.; and Reiff-Marganiec, S., eds., *Natural Language Processing and Information Systems*, 47–61. Cham: Springer Nature Switzerland. ISBN 978-3-031-35320-8.

Borisov, V.; Sessler, K.; Leemann, T.; Pawelczyk, M.; and Kasneci, G. 2022. Language Models are Realistic Tabular Data Generators. In *The Eleventh International Conference on Learning Representations*.

Bowman, S. R.; Angeli, G.; Potts, C.; and Manning, C. D. 2015. A large annotated corpus for learning natural language inference. In Márquez, L.; Callison-Burch, C.; and Su, J., eds., *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 632–642. Lisbon, Portugal: Association for Computational Linguistics.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS ’20. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781713829546.

Chen, X.; Lin, M.; Schaerli, N.; and Zhou, D. 2023a. Teaching Large Language Models to Self-Debug. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Chen, Z.; Gao, Q.; Bosselut, A.; Sabharwal, A.; and Richardson, K. 2023b. DISCO: Distilling Counterfactuals with Large Language Models. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Christie, S. 2022. Why play equals learning: Comparison as a learning mechanism in play. *Infant and Child Development*, 31(1): e2285.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR*, abs/1810.04805.

Jin, D.; Jin, Z.; Hu, Z.; Vechtomova, O.; and Mihalcea, R. 2022. Deep learning for text style transfer: A survey. *Computational Linguistics*, 48(1): 155–205.

Laurer, M.; Van Atteveldt, W.; Casas, A.; and Welbers, K. 2024. Less annotating, more classifying: Addressing the data scarcity issue of supervised machine learning with deep transfer learning and BERT-NLI. *Political Analysis*, 32(1): 84–100.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Li, Y.; Xu, M.; Miao, X.; Zhou, S.; and Qian, T. 2023. Large language models as counterfactual generator: Strengths and weaknesses. *arXiv preprint arXiv:2305.14791*.

Maas, A. L.; Daly, R. E.; Pham, P. T.; Huang, D.; Ng, A. Y.; and Potts, C. 2011. Learning Word Vectors for Sentiment Analysis. In Lin, D.; Matsumoto, Y.; and Mihalcea, R., eds., *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 142–150. Portland, Oregon, USA: Association for Computational Linguistics.

Madaan, A.; Tandon, N.; Gupta, P.; Hallinan, S.; Gao, L.; Wiegreffe, S.; Alon, U.; Dziri, N.; Prabhumoye, S.; Yang, Y.; et al. 2024. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36.

Min, S.; Lyu, X.; Holtzman, A.; Artetxe, M.; Lewis, M.; Hajishirzi, H.; and Zettlemoyer, L. 2022. Rethinking the Role of Demonstrations: What Makes In-Context Learning Work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 11048–11064.

Olausson, T. X.; Inala, J. P.; Wang, C.; Gao, J.; and Solar-Lezama, A. 2023. Is Self-Repair a Silver Bullet for Code Generation? In *The Twelfth International Conference on Learning Representations*.

Pryzant, R.; Iter, D.; Li, J.; Lee, Y.; Zhu, C.; and Zeng, M. 2023. Automatic Prompt Optimization with “Gradient Descent” and Beam Search. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 7957–7968.

Rao, S.; and Tetreault, J. 2018. Dear Sir or Madam, May I Introduce the GYAFc Dataset: Corpus, Benchmarks and Metrics for Formality Style Transfer. In Walker, M.; Ji, H.; and Stent, A., eds., *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 129–140. New Orleans, Louisiana: Association for Computational Linguistics.

Reif, E.; Ippolito, D.; Yuan, A.; Coenen, A.; Callison-Burch, C.; and Wei, J. 2022. A Recipe for Arbitrary Text Style Transfer with Large Language Models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 837–848.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Rittle-Johnson, B.; and Star, J. R. 2011. The power of comparison in learning and instruction: Learning outcomes sup-

ported by different types of comparisons. In *Psychology of learning and motivation*, volume 55, 199–225. Elsevier.

Shinn, N.; Cassano, F.; Gopinath, A.; Narasimhan, K.; and Yao, S. 2023. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36.

Socher, R.; Perelygin, A.; Wu, J.; Chuang, J.; Manning, C. D.; Ng, A.; and Potts, C. 2013. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In Yarowsky, D.; Baldwin, T.; Korhonen, A.; Livescu, K.; and Bethard, S., eds., *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1631–1642. Seattle, Washington, USA: Association for Computational Linguistics.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Williams, A.; Nangia, N.; and Bowman, S. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In Walker, M.; Ji, H.; and Stent, A., eds., *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 1112–1122. New Orleans, Louisiana: Association for Computational Linguistics.

Wu, Y.; Gardner, M.; Stenetorp, P.; and Dasigi, P. 2022. Generating Data to Mitigate Spurious Correlations in Natural Language Inference Datasets. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2660–2676. Dublin, Ireland: Association for Computational Linguistics.

Yang, C.; Wang, X.; Lu, Y.; Liu, H.; Le, Q. V.; Zhou, D.; and Chen, X. 2023. Large Language Models as Optimizers. In *The Twelfth International Conference on Learning Representations*.

Yang, J.; Jin, H.; Tang, R.; Han, X.; Feng, Q.; Jiang, H.; Zhong, S.; Yin, B.; and Hu, X. 2024. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Transactions on Knowledge Discovery from Data*, 18(6): 1–32.

Yang, S.; and Kasneci, G. 2024. Is Crowdsourcing Breaking Your Bank? Cost-Effective Fine-Tuning of Pre-trained Language Models with Proximal Policy Optimization. *arXiv preprint arXiv:2402.18284*.

Yuan, W.; Cho, K.; and Weston, J. 2023. System-level natural language feedback. *arXiv preprint arXiv:2306.13588*.

Zhang*, T.; Kishore*, V.; Wu*, F.; Weinberger, K. Q.; and Artzi, Y. 2020. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.

Zhu, Y.; Lu, S.; Zheng, L.; Guo, J.; Zhang, W.; Wang, J.; and Yu, Y. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, 1097–1100.

Related Work

Single-instruction Algorithms for LLM Enhancement. The increasing model sizes of LLMs make the cost of fine-tuning prohibitive, significantly limiting their applicability. As an alternative to fine-tuning, Brown et al. (2020) introduced few-shot learning, which incorporates manually crafted examples into prompts, guiding models to accurately capture human intent and implement generation. Additionally, Wei et al. (2022) proposed Chain-of-thought prompting, which involves adding reasoning steps to prompts to require LLMs to reflect on the original task, thereby achieving more accurate and reasonable generation. However, these methods use unified prompts for all input samples without refining the output. In contrast, prompts incorporating customized feedback have shown their effectiveness in complex tasks such as code debugging (Madaan et al. 2024; Shinn et al. 2023; Olausson et al. 2023; Chen et al. 2023a) and math reasoning (Madaan et al. 2024; Shinn et al. 2023). In this work, we follow the concept of providing individualized instructions to refine output for natural language generation tasks.

Using Prompts for Refinement. Recent works further improve the output of LLMs by utilizing feedback within prompts to refine their outputs. For instance, (Yuan, Cho, and Weston 2023) utilize human feedback, while (Madaan et al. 2024; Shinn et al. 2023; Pryzant et al. 2023; Olausson et al. 2023; Chen et al. 2023a) use the LLMs to generate feedback, which is less costly. Moreover, (Pryzant et al. 2023) uses LLMs to generate textual gradients to create more effective prompts for solving tasks. Commonly, this feedback is in the form of natural language. Beyond using only natural language feedback, Yang et al. (2023) propose to use an LLM as an optimizer. In this framework, an evaluator assigns numerical scores to outputs, which are then integrated into prompts to let LLMs generate new outputs. This method allows LLMs to implicitly aggregate common characteristics among solutions. However, this method does not give concrete aspects that the model should optimize. Our method refines outputs by comparing and analyzing previous generations to leverage the strengths of earlier generations. In this manner, our approach contextualizes the aspects for improvement in prompts to guide the model towards achieving optimal results.

Data Generation Using LLMs. The use of LLMs for generating data, such as counterfactuals or contrastive examples, becomes a significant application in data augmentation (Chen et al. 2023b; Wu et al. 2022; Yang and Kasneci 2024; Borisov et al. 2022). Using LLMs for data generation is cost-effective, and the augmented data enhances the training process. For instance, incorporating counterfactual data into training sets helps mitigate shortcut problems and improves model robustness (Wu et al. 2022). In (Borisov et al. 2022), a GPT-2 model generates synthetic tabular data by first converting tabular data into textual data using a template. (Yang and Kasneci 2024) employs clustering and noise injection methods, such as n-gram level edits, the addition or deletion of negation words, and reordering of sentences, to generate negative examples. These examples are then utilized to fine-tune an LM, enabling it to produce textual counterfactuals based on predefined templates. Chen et al. (2023b) deploys an LM

to generate counterfactual data by completing the sentences with words masked out. A neural syntactic parser (Akbik et al. 2019) is employed to determine words to be perturbed. However, these methods utilize pre-defined rules and potentially limit the algorithm’s generality. Compared to these approaches, our method does not require extra fine-tuning and can be directly used in various data generation tasks.

Implementation Details of C&G

To create a comparison prompt, we first calculate objective scores for each solution. The objective function comprises two components: the alignment score (sa) and the relevance score (sr). In this context, we transform the objective function into a loss function, where minimizing losses corresponds to enhanced generation. For the computation of sa , we use the template given in Figure 2 bottom, as the performance of LLMs in classification tasks is outstanding (Yang et al. 2024). When the label aligns with the target, sa is assigned a 0; otherwise, it is set to 1. To assess sr , we use the Levenshtein distance between the two sentences divided by the length of the original input sentence. The weighted factor α is set to 0.1. Figure 3 provides the template for the comparison prompt for sentiment transfer tasks; First, we give a task explanation, followed by the values calculated by the objective function. In the process of iterative optimization, we set the early stop for the duration of plateau Ir to 5 and the maximal step size I_m to 10. We use the few-shot learning prompt to initialize the sentence for comparison. Our code is available at <https://anonymous.4open.science/r/CG-LLM-4BBB/README.md>.

Algorithm 1 shows the algorithm of our C&G. Now, we show templates we use for our C&G on different datasets. Figure 5 shows the template for sentiment transfer datasets SST-2 and IMDB, Figure 6 for the natural language inference datasets SNLI and MNLI, and Figure 7 for the style transfer dataset GYAFC. Please note that the function of computing sa and sr used in C&G are not used in the evaluation metrics, ensuring fair comparisons.

Implementation Details of Baselines

As single-instruction baselines, we use zero-shot, few-shot, and CoT, which are commonly used in NLG tasks. We align our experimental settings with those given by Li et al. (2023), for instance, the choice and the number of examples in these methods. As refinement baselines, we deploy Self-Refine (Self-Re) (Madaan et al. 2024) and Optimization by PROmpting (OPRO) (Yang et al. 2023). Following (Madaan et al. 2024), we set the number of iterations to 4 and use similar prompts to request natural language feedback and refinement. To compare our method with OPRO fairly, we use the same objective function to calculate the score for the OPRO prompts and set the number of iterations to 10. GPT-3.5 and Llama3-8B are backbone language models for all algorithms in the comparison in the next section.

Figure 8, Figure 9, Figure 10 and Figure 11 demonstrates the template on two NLI datasets for few-shot learning, CoT, Self-Refine and OPRO, respectively. For sentiment transfer datasets SST-2 and IMDB, we use the templates for few-shot

```

# Template for  $\phi(\cdot)$ .
Task Definition: Revise a given sentence with minimal changes to alter its sentiment
polarity.
Given Sentence: input sentence
Target Sentiment Polarity: targeted sentiment
Revised Sentence 1: revised sentence 1
Distance to the Given Sentence: Levenshtein distance
Distance to the Target Sentiment Polarity: 0 or 1
Loss: the weighted sum of the two distances
Revised Sentence 2: revised sentence 2
Distance to the Given Sentence: Levenshtein distance
Distance to the Target Sentiment Polarity: 0 or 1
Loss: the weighted sum of the two distances
The loss contains two parts: Distance to the Given Sentence and the distance to the Target
Sentiment Polarity.
Given the task and loss definition, please first think why the Sentence Nr is better than
Sentence Nr., and give the analysis. Then, generate a new Revised Sentence that minimizes the
loss.
Revised Sentence:

# Template for computing sa.
Given Sentence: "The movie is the best that I have ever seen."
Sentiment: "positive"
# One more example ...
Complete the "Sentiment:" by imitating the given demonstration.
Given Sentence: input sentence
Sentiment:

```

Figure 3: Template used for constructing comparison prompts. Text in black represents the fixed template, while text in blue indicates the input information needed for a specific data sample. The template on the bottom is to verify whether the generated sentence aligns with the task goal (*sa*).

learning in Figure 12, and CoT in Figure 13. Self-refine templates are in Figure 14 and Figure 15. Figure 16 demonstrates the template for OPRO.

Datasets

Statistics of datasets used is demonstrated in Table 3. As computing with LLMs can be costly in time and resources, we randomly sample 510 from the training set from each dataset as our testbed to ensure efficiency. The same training sets are used for different methods for fair comparisons.

Implementation Details of Evaluation Metrics

Transfer Accuracy. To compute the transfer accuracy, we use models that are trained on larger datasets and yield satisfactory performance on the test datasets. Specifically, we use DeBERTa-v3 model (Laurer et al. 2024) trained on SNLI with a test accuracy of 90.9%, and a large BART model (Lewis et al. 2020) trained on MNLI with a test accuracy of 91.0% to evaluate the transfer accuracy of the generated data. On SST-2 and IMDB datasets, we deploy the BERT model (Devlin et al. 2018), which achieves the accuracy of 92.8% and 88.17% on SST-2 and IMDB test sets, respectively. On GYafc, we use the classifier trained on GYafc (Babakov et al. 2023), which has an accuracy of 90.9%.

Content Preservation. BERTScore computes the similarity of two sentences as a sum of cosine similarities between their tokens’ embeddings (Zhang* et al. 2020). To compute the embeddings, we use a pre-trained Sentence-BERT model (Reimers and Gurevych 2019).

Results with GPT4

Table 5 lists the results using GPT4 on various datasets. Similar to the results in the main paper, our algorithm consistently achieves the best performance in various metrics, especially in the H mean of content preservation and generation accuracy. Moreover, using GPT4 outperforms the other two backbones in the main paper, indicating that advanced capabilities in understanding and analyzing can further improve generation quality.

Extended Tables

In this section, we show the extended tables, Table 8 and Table 6, with the mean \pm standard deviation.

Qualitative Results

Figure 17 presents three examples, each from tasks in counterfactual generation, sentiment transfer, and style transfer. We demonstrate one iteration of our algorithm, illustrating its effectiveness in optimizing for different objectives. In the SNLI example, the first revised sentence adeptly shifts the relationship from “Entailment” to “Contradiction” by modifying the content from “riding a bicycle” to “walking.” The second revised sentence, however, makes minimal changes and consequently fails to change the relationship. The model’s analysis captures this reason, generating a revised premise that not only changes the relationship but also maintains the content, such as “pushing a bicycle.”

The sentiment transfer example on SST-2 illustrates a case where the revised sentences maintain the same value in the

| Task | Datasets | Train | Test |
|------|--|---------|--------|
| NLI | SNLI (Bowman et al. 2015) | 550,152 | 10,000 |
| | MNLI (Williams, Nangia, and Bowman 2018) | 392,702 | 10,000 |
| | SST-2 (Socher et al. 2013) | 6,920 | 1,821 |
| ST | IMDB (Maas et al. 2011) | 366,466 | 1,000 |
| | GYAFC (Rao and Tetreault 2018) | 52,429 | 2,498 |

Table 3: Statistics of datasets used.

Algorithm 1: C&G Algorithm.

Input: A task $T(x)$, an LLM , max.iteration I_m , early stop iteration I_r .
Output: \hat{y} .
Initialization: $\hat{y}_0, \hat{y}_1 = LLM(p_0)$, $\hat{y}_m \leftarrow \hat{y}_0$, $I_b = 0$, $C = 1$, a list for generation $L = [\hat{y}_0, \hat{y}_1]$, and a list for scores $L_s = []$.
initialize scores for \hat{y}_m .
 $sa_m, sr_m = \sigma(\hat{y}_m)$, $s_m = sa_m + \alpha \cdot sr_m$.
Add s_m into L_s .
for $t \in [1, 2, \dots, I_m]$ **do**
 $sa_{t-1}, sr_{t-1} = \sigma(\hat{y}_{t-1})$,
 $s_{t-1} = sa_{t-1} + \alpha \cdot sr_{t-1}$.
 Add s_{t-1} into L_s .
 # compare scores and record the better one.
 if $s_m \leq s_{t-1}$ **then**
 $s_m \leftarrow s_{t-1}$, $I_b \leftarrow \text{index}(s_m)$, $C = 0$.
 else
 $C \leftarrow C + 1$
 end if
 if $C > I_r$ **then**
 # early stop.
 Break the loop.
 end if
 # construct p_t using the template in Figure 2.
 $p_t = \phi((\hat{y}_m, sa_m, sr_m),$
 $(\hat{y}_{t-1}, sa_{t-1}, sr_{t-1}), T(x))$.
 $\hat{y}_t = LLM(p_t)$.
 Add \hat{y}_t into L .
end for
According to L, L_s , return \hat{y} with the highest score.

objective function. Both do not successfully flip the sentiment to positive but make only minor modifications. Our model reveals that “a more significant change is necessary to alter the sentiment,” which also suggests that the model inherently balances the trade-off between maintaining faithfulness to the original sentence and transferring the label. Therefore, the model generates a sentence with changing the structure, and clearly adopts a positive tone by “a step above.”

An extended figure of Figure 17 is demonstrated in Figure 18. In the style transfer task, the objective function is crafted as a score that the model aims to maximize, continually enhancing the sentence’s informality. In this example, although both sentences achieve the same informality score—determined by the count of informal words—the first sentence more closely aligns with the semantic essence of the original expression. The model states “The phrase ‘crushes you’ more effectively captures the essence of ‘breaks your

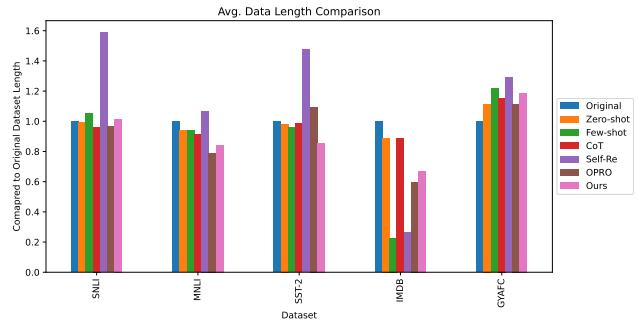


Figure 4: Average length of generated data.

Table 4: Computational infrastructure details.

| Device Attribute | Value |
|--------------------------|-------------|
| Computing infrastructure | GPU |
| GPU model | NVIDIA A100 |
| GPU number | 1 |
| CUDA version | 12.3 |

heart’ than ‘hits you hard.’ ” Thus, the model improves the semantic similarity while adding another informal expression “heartbreaker pro”.

Figure 19 and Figure 20 show the analysis from Llama3 and its generation, while Figure 21 and Figure 22 demonstrate the thorough thoughts of GPT4.

Using Generated Data as Augmentation

Generated counterfactual samples can be used to augment the dataset. Thus, we use the accuracy of a trained smaller model, a BART model (Lewis et al. 2020), on the augmented training dataset to indicate the overall quality. In Table 6 and Table 7, we list the test accuracy when using the generation from different algorithms as data augmentation. Concretely, we use the original data and its generated counterparts from our sampled training set to train a BART model. The model is trained using the AdamW optimizer with a learning rate of $2e^{-5}$, and the training ends after 50 epochs. The trained model is tested on the original test set on the SNLI and MNLI datasets. From the results, we observe that the H mean corresponds to the test accuracy, indicating that the usage of the H mean can represent the quality of generated data in data augmentation.

Length of Generated Data

We show the comparison of data length in Figure 4. From the comparison, we see that Self-Refine often generates data that are longer than the original data. For instance, on SNLI it generates almost 1.6 times longer text than other methods.

Computational Infrastructure Details

All experiments in this paper are conducted on the device given in Table 4.

Ethical Statement

In this research, our goal is to improve the LLMs' capabilities to solve natural language generation tasks. We believe that by enhancing the accessibility, acceptability, and user-friendliness of AI, we can better tap into its potential to assist humans. We do not foresee any negative societal impacts from our work.

Declaration

Throughout this project's writing process, we used ChatGPT only for spellchecking and grammar improvement. After using these tools, we thoroughly reviewed and re-edited the content. We take full responsibility for the final outcome.

Template for $\phi(\cdot)$.

Task Definition: Revise a given sentence with minimal changes to alter its sentiment polarity.

Given Sentence: *input sentence*

Target Sentiment Polarity: *targeted sentiment*

Revised Sentence 1: *revised sentence 1*

Distance to the Given Sentence: *Levenshtein distance*

Distance to the Target Sentiment Polarity: *0 or 1*

Loss: *the weighted sum of the two distances*

Revised Sentence 2: *revised sentence 2*

Distance to the Given Sentence: *Levenshtein distance*

Distance to the Target Sentiment Polarity: *0 or 1*

Loss: *the weighted sum of the two distances*

The loss contains two parts: Distance to the Given Sentence and the distance to the Target Sentiment Polarity.

Revised Sentence *Nr.* is better than *Nr.*

Or Revised Sentence 1 and 2 are equally bad.

Given the task and loss definition, please first think why the Sentence *Nr* is better than Sentence *Nr.*, and give the analysis. Then, generate a new Revised Sentence that minimizes the loss.

Or: Given the task and loss definition, please first think about why the two sentences have high losses and give the analysis. Then, generate a new Revised Sentence that minimizes the loss.

Revised Sentence:

Template for computing *sa*.

Given Sentence: "The movie is the best that I have ever seen."

Sentiment: "positive"

Given Sentence: "This movie is quite boring to me."

Sentiment: "negative"

Complete the "Sentiment:" by imitating the given demonstration.

Given Sentence: *input sentence*

Sentiment:

Figure 5: Template of C&G on SST-2 and IMDB. Text in black represents the fixed template, while text in blue indicates the input information needed for a specific data sample. The template on the bottom is to verify whether the generated sentence aligns with the task goal (*sa*).

Template for $\phi(\cdot)$.

Task Definition: Revise the premise sentence, using minimal changes, to alter the relationship between it and the hypothesis sentence to either entailment, contradiction, or neutral.

Given Premise Sentence: *input premise*

Given Hypotheses Sentence: *input hypothesis*

Target Relationship: *targeted relationship*

Revised Premise Sentence 1: *revised sentence 1*

Distance to the Given Premise Sentence: *Levenshtein distance*

Distance to the Target Relationship: *0 or 1*

Loss: *the weighted sum of the two distances*

Revised Premise Sentence 2: *revised sentence 1*

Distance to the Given Premise Sentence: *Levenshtein distance*

Distance to the Target Relationship: *0 or 1*

Loss: *the weighted sum of the two distances*

The loss contains two parts: The loss contains two parts: Distance to the Given Premise Sentence and the distance to the Target Relationship.

Revised Premise *Nr.* is better than *Nr.*.

Or: Revised Premise 1 and 2 are equally bad.

Given the task and loss definition, please first think why the Sentence *Nr* is better than Sentence *Nr.*, and give the analysis. Then, generate a new Revised Sentence that minimizes the loss.

Or: Given the task and loss definition, please first think about why the two sentences have high losses and give the analysis. Then, generate a new Revised Sentence that minimizes the loss.

Revised Sentence:

Template for computing *sa*.

Premise: "The marathon runner crossed the finish line in under three hours."

Hypothesis: "The marathon runner trained in high altitude conditions."

Relationship: Neutral

Premise: "The cafe was bustling with customers all day."

Hypothesis: "The cafe was closed yesterday."

Relationship: Contradiction

Premise: "The astronomer observed the comet through the telescope."

Hypothesis: "The comet was visible from Earth." Relationship: Entailment

Premise: *input premise*

Hypothesis: *input hypothesis*

Complete this sentence: The relationship between the Given Premise and Given

Hypothesis is

Figure 6: Template of C&G on SNLI and MNLI. Text in black represents the fixed template, while text in blue indicates the input information needed for a specific data sample. The template on the bottom is to verify whether the generated sentence aligns with the task goal (*sa*).

Template for $\phi(\cdot)$.

Task Definition: Please transfer the given sentence to an informal one.

Given Sentence: *input sentence*

Revised Sentence 1: *revised sentence 1*

Semantic Similarity to the Given Sentence: *sentence similarity*

Informality Score: *a positive natural number*

Score: *the weighted sum of the two scores*

Revised Sentence 2: *revised sentence 1*

Semantic Similarity to the Given Sentence: *sentence similarity*

Informality Score: *a positive natural number*

Score: *the weighted sum of the two scores*

The score contains two parts: Semantic similarity between Revised Sentence and Given Sentence, and the informality of Revised Sentence.

Revised Sentence *Nr.* is better than *Nr.*.

Or Revised Sentence 1 and 2 are equally bad.

Given the task and score definition, please first think why the Sentence *Nr* is better than Sentence *Nr.*, and give the analysis. Then, generate a new Revised Sentence that maximizes the score.

Or: Given the task and score definition, please first think about why the two sentences have high losses and give the analysis. Then, generate a new Revised Sentence that maximizes the score.

Revised Sentence:

Template for computing *sa*.

Please score the question based on the number of informal expressions in the sentence. Sentence: "The In-Laws movie isn't a holiday movie, but it's okay."

Informality Score: 1

Sentence: "the movie The In-Laws not exactly a holiday movie but funny and good!"

Informality Score: 4

Sentence: "The In-Laws movie isn't a holiday flick, but it's alright."

Informality Score: 2

Complete "Informality Score:" by imitating the given demonstration.

Sentence: *input sentence*

Informality Score:

Figure 7: Template of C&G on GYAF. Text in black represents the fixed template, while text in blue indicates the input information needed for a specific data sample. The template on the bottom is to verify whether the generated sentence aligns with the task goal (*sa*).

| | SBLEU | PPL | Cont. | Acc. | H-Mean |
|--------------|-------------|------------|-------------|-------------|--------|
| SNLI | | | | | |
| zero-shot | 0.240±0.003 | 111.4±2.35 | 0.715±0.004 | 0.771±0.003 | 0.741 |
| few-shot | 0.232±0.007 | 109.1±1.98 | 0.763±0.001 | 0.837±0.005 | 0.798 |
| CoT | 0.232±0.001 | 124.1±3.14 | 0.728±0.002 | 0.667±0.006 | 0.696 |
| OPRO | 0.225±0.006 | 99.0±1.34 | 0.805±0.001 | 0.843±0.004 | 0.823 |
| Ours | 0.210±0.002 | 104.6±0.98 | 0.823±0.002 | 0.859±0.001 | 0.841 |
| MNLI | | | | | |
| zero-shot | 0.178±0.018 | 105.4±2.01 | 0.701±0.003 | 0.586±0.008 | 0.638 |
| few-shot | 0.250±0.021 | 106.1±0.98 | 0.787±0.002 | 0.714±0.004 | 0.748 |
| CoT | 0.062±0.006 | 174.1±1.53 | 0.781±0.003 | 0.529±0.006 | 0.631 |
| OPRO | 0.071±0.006 | 105.7±1.56 | 0.778±0.002 | 0.764±0.003 | 0.771 |
| Ours | 0.066±0.004 | 106.3±1.20 | 0.807±0.001 | 0.743±0.002 | 0.773 |
| SST-2 | | | | | |
| zero-shot | 0.080±0.003 | 218.3±2.68 | 0.758±0.004 | 0.888±0.005 | 0.818 |
| few-shot | 0.102±0.004 | 158.3±1.43 | 0.743±0.002 | 0.867±0.002 | 0.800 |
| CoT | 0.078±0.005 | 201.5±2.34 | 0.719±0.001 | 0.888±0.004 | 0.795 |
| OPRO | 0.082±0.002 | 189.4±3.56 | 0.768±0.002 | 0.849±0.005 | 0.806 |
| Ours | 0.056±0.003 | 181.3±2.56 | 0.789±0.002 | 0.880±0.002 | 0.832 |
| IMDB | | | | | |
| zero-shot | 0.080±0.003 | 42.5±0.98 | 0.905±0.000 | 0.958±0.005 | 0.931 |
| few-shot | 0.076±0.009 | 54.3±1.21 | 0.864±0.002 | 0.922±0.004 | 0.892 |
| CoT | 0.077±0.002 | 66.4±1.58 | 0.845±0.004 | 0.912±0.002 | 0.877 |
| OPRO | 0.102±0.010 | 41.3±1.20 | 0.908±0.001 | 0.914±0.004 | 0.911 |
| Ours | 0.076±0.005 | 40.1±0.92 | 0.917±0.001 | 0.970±0.002 | 0.943 |

Table 5: Comparison of generated data using GPT-4 as the backbone.

| | Llama3-8B | | GPT-3.5 | | GPT4 | |
|-------------|-------------|------|-------------|------|-------------|------|
| | Acc. | H | Acc. | H | Acc. | H |
| zero-shot | 0.535±0.107 | 0.61 | 0.536±0.013 | 0.68 | 0.567±0.002 | 0.74 |
| few-shot | 0.530±0.008 | 0.64 | 0.535±0.010 | 0.66 | 0.571±0.004 | 0.80 |
| CoT | 0.529±0.006 | 0.64 | 0.493±0.014 | 0.57 | 0.565±0.002 | 0.70 |
| OPRO | 0.525±0.002 | 0.61 | 0.541±0.004 | 0.67 | 0.589±0.005 | 0.82 |
| Ours | 0.565±0.003 | 0.65 | 0.573±0.008 | 0.72 | 0.592±0.004 | 0.84 |

Table 6: Accuracy of using generated data for data augmentation on SNLI.

| | Llama3-8B | | GPT-3.5 | | GPT4 | |
|-------------|-------------|------|-------------|------|-------------|------|
| | Acc. | H | Acc. | H | Acc. | H |
| zero-shot | 0.541±0.007 | 0.58 | 0.575±0.003 | 0.63 | 0.632±0.004 | 0.64 |
| few-shot | 0.541±0.004 | 0.57 | 0.564±0.006 | 0.59 | 0.671±0.007 | 0.75 |
| CoT | 0.508±0.005 | 0.52 | 0.515±0.004 | 0.44 | 0.621±0.009 | 0.63 |
| OPRO | 0.523±0.002 | 0.55 | 0.574±0.005 | 0.64 | 0.677±0.006 | 0.77 |
| Ours | 0.577±0.006 | 0.60 | 0.588±0.002 | 0.65 | 0.683±0.001 | 0.77 |

Table 7: Accuracy of using generated data for data augmentation on MNLI.

Template for few-shot learning.

Task Definition: Revise the premise sentence, using minimal changes, to alter the relationship between it and the hypothesis sentence to either entailment, contradiction, or neutral.

Instruction: This process consists of two steps. The first step is to identify the words in the given premise sentence that have the highest potential to change the relationship with the hypothesis sentence after substitution, known as the causal words. The second step is to select appropriate replacement words for the causal words that will change the relationship with the hypothesis sentence to the desired relationship, either entailment, contradiction, or neutral.

Given Premise Sentence: "A group of men riding bicycles in a line."

Given Hypothesis Sentence: "The men riding together."

Current Relationship between the premise sentence and the hypothesis sentence:
"Entailment"

Target Relationship: "Contradiction"

Generated Premise Sentence: "A group of men walking separately in different directions."

Target Relationship: "Neutral"

Generated Premise Sentence: "A group of men riding bicycles in various directions."

Based on the given task definition and instruction, complete the following text by imitating the given demonstration.

Given Premise Sentence: *input premise*

Given Hypothesis Sentence: *input hypothesis*

Current Relationship between the premise sentence and the hypothesis sentence: *current relationship*

Target Relationship: *target relationship*

Generated Premise Sentence:

Figure 8: Few-shot learning template on SNLI and MNLI. Text in black represents the fixed template, while text in blue indicates the input information needed for a specific data sample.

| | | GPT-3.5 | | | | Llama3-8B | | | |
|------|-------------|--------------------|-------------------|-------------------|-------------------|--------------------|-------------------|-------------------|-------------------|
| | | SBLEU \downarrow | PPL \downarrow | Cont. \uparrow | Acc. \uparrow | SBLEU \downarrow | PPL \downarrow | Cont. \uparrow | Acc. \uparrow |
| SNLI | zero-shot | 0.177 \pm 0.009 | 139.0 \pm 0.93 | 0.750 \pm 0.006 | 0.617 \pm 0.002 | 0.200 \pm 0.002 | 102.5 \pm 1.23 | 0.711 \pm 0.001 | 0.543 \pm 0.001 |
| | few-shot | 0.166 \pm 0.003 | 100.0 \pm 0.68 | 0.744 \pm 0.004 | 0.583 \pm 0.003 | 0.241 \pm 0.005 | 120.88 \pm 1.03 | 0.714 \pm 0.002 | 0.594 \pm 0.004 |
| | CoT | 0.208 \pm 0.007 | 118.9 \pm 1.03 | 0.704 \pm 0.003 | 0.478 \pm 0.006 | 0.205 \pm 0.009 | 100.7 \pm 0.32 | 0.709 \pm 0.004 | 0.592 \pm 0.003 |
| | Self-Re | 0.225 \pm 0.009 | 47.8 \pm 0.62 | 0.617 \pm 0.002 | 0.570 \pm 0.008 | 0.256 \pm 0.002 | 43.4 \pm 1.24 | 0.528 \pm 0.002 | 0.812 \pm 0.002 |
| | OPRO | 0.199 \pm 0.003 | 84.9 \pm 1.98 | 0.763 \pm 0.002 | 0.605 \pm 0.005 | 0.227 \pm 0.009 | 96.3 \pm 1.14 | 0.749 \pm 0.001 | 0.510 \pm 0.003 |
| | Ours | 0.203 \pm 0.003 | 96.8 \pm 0.91 | 0.778 \pm 0.010 | 0.664 \pm 0.003 | 0.228 \pm 0.003 | 100.3 \pm 2.55 | 0.773 \pm 0.003 | 0.549 \pm 0.003 |
| MNLI | zero-shot | 0.062 \pm 0.003 | 112.3 \pm 0.54 | 0.787 \pm 0.002 | 0.522 \pm 0.002 | 0.066 \pm 0.002 | 93.4 \pm 1.53 | 0.741 \pm 0.002 | 0.478 \pm 0.003 |
| | few-shot | 0.064 \pm 0.000 | 113.7 \pm 4.43 | 0.778 \pm 0.004 | 0.474 \pm 0.002 | 0.068 \pm 0.000 | 95.6 \pm 0.72 | 0.761 \pm 0.001 | 0.456 \pm 0.004 |
| | CoT | 0.066 \pm 0.006 | 145.9 \pm 0.98 | 0.753 \pm 0.001 | 0.308 \pm 0.003 | 0.072 \pm 0.003 | 138.2 \pm 1.03 | 0.783 \pm 0.002 | 0.393 \pm 0.001 |
| | Self-Re | 0.250 \pm 0.021 | 53.1 \pm 0.58 | 0.562 \pm 0.001 | 0.395 \pm 0.015 | 0.098 \pm 0.002 | 44.3 \pm 2.01 | 0.514 \pm 0.005 | 0.723 \pm 0.003 |
| | OPRO | 0.061 \pm 0.006 | 90.3 \pm 0.34 | 0.749 \pm 0.005 | 0.561 \pm 0.003 | 0.071 \pm 0.004 | 94.5 \pm 6.13 | 0.754 \pm 0.006 | 0.443 \pm 0.004 |
| | Ours | 0.063 \pm 0.009 | 108.29 \pm 0.92 | 0.772 \pm 0.001 | 0.564 \pm 0.003 | 0.065 \pm 0.000 | 103.6 \pm 3.33 | 0.781 \pm 0.003 | 0.492 \pm 0.001 |

Table 8: Comparison of counterfactual data generation on SNLI and MNLI datasets. Harmonic mean (**H**) is highlighted in gray, with the best results in bold.

Template for CoT

Task Definition: Revise the premise sentence, using minimal changes, to alter the relationship between it and the hypothesis sentence to either entailment, contradiction, or neutral.

Instruction: This process consists of two steps. The first step is to identify the words in the given premise sentence that have the highest potential to change the relationship with the hypothesis sentence after substitution, known as the causal words. The second step is to select appropriate replacement words for the causal words that will change the relationship with the hypothesis sentence to the desired relationship, either entailment, contradiction, or neutral.

Demonstration:

Given Premise Sentence: "A group of men riding bicycles in a line."

Given Hypothesis Sentence: "The men riding together."

Current Relationship between the premise sentence and the hypothesis sentence:
"Entailment"

Target Relationship: "Contradiction"

Causal Words Identification: "riding bicycles", "a line".

Causal Words Replacement: "walking separately", "different directions".

Generated Premise Sentence: "A group of men walking separately in different directions."

Target Relationship: "Neutral"

Causal Words Identification: "riding bicycles", "a line".

Causal Words Replacement: "riding bicycles", "different directions".

Generated Premise Sentence: "A group of men riding bicycles in various directions."

Based on the given task definition and instruction, complete the following text by imitating the given demonstration.

Given Premise Sentence: *input premise*

Given Hypothesis Sentence: *input hypothesis*

Current Relationship between the premise sentence and the hypothesis sentence: *current relationship*

Target Relationship: *target relationship*

Causal Words Identification:

Causal Words Replacement:

Generated Premise Sentence:

Figure 9: CoT template on SNLI and MNLI. Text in black represents the fixed template, while text in blue indicates the input information needed for a specific data sample.

Template for Feedback in Self-Refine.

Given Premise Sentence: A group of men riding bicycles in a line.

Given Hypothesis Sentence: The men riding together.

Revised Premise Sentence: The men are riding bicycles individually, spaced far apart from each other.

Why does this Revised Premise Sentence not establish a clear Contradiction to the Given Hypothesis Sentence?

Feedback: The Revised Premise Sentence does not contradict the Given Hypothesis because both can be true simultaneously. The Hypothesis suggests the men are "riding together," which can imply a shared activity rather than close proximity. The Revised Premise indicates they are "spaced far apart," but this doesn't negate the possibility of them riding together in a general sense. Given Premise Sentence: *input premise*

Given Hypothesis Sentence: *input hypothesis*

Revised Premise Sentence: *revised premise*

Why does this Revised Premise Sentence not establish a clear *target relationship* to the Given Hypothesis Sentence?

Feedback:

Template for Refine in Self-Refine.

Given Premise Sentence: A group of men riding bicycles in a line.

Given Hypothesis Sentence: The men riding together.

Revised Premise Sentence: The men are riding bicycles individually, spaced far apart from each other.

Feedback: The Revised Premise Sentence does not contradict the Given Hypothesis because both can be true simultaneously. The Hypothesis suggests the men are "riding together," which can imply a shared activity rather than close proximity. The Revised Premise indicates they are "spaced far apart," but this doesn't negate the possibility of them riding together in a general sense.

Okay, let's try again. Rewrite the Premise Sentence to have a very clear Contradiction to the Given Hypothesis Sentence using the feedback above.

New Revised Premise Sentence: The men are not riding together; each man is cycling alone on a different route.

Given Premise Sentence: *input premise*

Given Hypothesis Sentence: *input hypothesis*

Revised Premise Sentence: *revised premise*

Feedback: *feedback from last step*

Okay, let's try again. Rewrite the Premise Sentence to have a very clear *target relationship* to the Given Hypothesis Sentence using the feedback above.

New Revised Premise Sentence:

Figure 10: Self-Refine template of on SNLI and MNLI. Text in black represents the fixed template, while text in blue indicates the input information needed for a specific data sample.

Template for OPRO.

Task Definition: Revise the premise sentence, using minimal changes, to alter the relationship between it and the hypothesis sentence to either entailment, contradiction, or neutral.

Given Premise Sentence: *input premise*

Given Hypotheses Sentence: *input hypothese*

Target Relationship: *targeted relationship*

Revised Premise Sentence: *revised sentence 1*

Distance to the Given Premise Sentence: *Levenshtein distance*

Distance to the Target Relationship: *0 or 1*

Loss: *the weighted sum of the two distances*

Examples from previous steps...

The loss contains two parts: The loss contains two parts: Distance to the Given Premise Sentence and the distance to the Target Relationship.

Given the task and loss definition, please generate a new Revised Sentence that minimizes the loss.

Revised Sentence:

Figure 11: OPRO template of on SNLI and MNLI. Text in black represents the fixed template, while text in blue indicates the input information needed for a specific data sample.

Template for few-shot learning.

Task Definition: Revise a given sentence with minimal changes to alter its sentiment polarity.

Instruction: This process consists of two steps. The first step is to identify the words in the given sentence that have the highest potential to change the sentiment polarity after substitution, known as the causal words. The second step is to select appropriate replacement words for the causal words that will change the sentiment polarity of the sentence to the desired polarity.

Demonstration:

Given Sentence: "The movie is the best that I have ever seen."

Current Sentiment Polarity: "positive"

Target Sentiment Polarity: "negative"

Revised Sentence: "The movie is the worst that I have ever seen."

Based on the given task definition and instruction, complete the following text by imitating the given demonstration.

Given Sentence: *input sentence*

Current Sentiment Polarity: *current sentiment*

Target Sentiment Polarity: *targeted sentiment*

Revised Sentence:

Figure 12: Few-shot learning template on SST-2 and IDMB. Text in black represents the fixed template, while text in blue indicates the input information needed for a specific data sample.

Template for few-shot learning.

Task Definition: Revise a given sentence with minimal changes to alter its sentiment polarity.

Instruction: This process consists of two steps. The first step is to identify the words in the given sentence that have the highest potential to change the sentiment polarity after substitution, known as the causal words. The second step is to select appropriate replacement words for the causal words that will change the sentiment polarity of the sentence to the desired polarity.

Demonstration:

Given Sentence: "The movie is the best that I have ever seen."

Current Sentiment Polarity: "positive"

Target Sentiment Polarity: "negative"

Causal Words Identification: The sentiment polarity "positive" depends on words "best".

Causal Words Replacement: To change the sentiment polarity of the given sentence from "positive" to "negative", causal words "best" are replaced by "worst".

Revised Sentence: "The movie is the worst that I have ever seen."

Based on the given task definition and instruction, complete the following text by imitating the given demonstration. Please think step by step.

Given Sentence: *input sentence*

Current Sentiment Polarity: *current sentiment*

Target Sentiment Polarity: *targeted sentiment*

Causal Words Identification:

Causal Words Replacement:

Revised Sentence:

Figure 13: CoT template on SST-2 and IMDB. Text in black represents the fixed template, while text in blue indicates the input information needed for a specific data sample.

| | | GPT-3.5 | | | | Llama3-8B | | | |
|-------------|-----------|--------------------|-------------------|-------------------|-------------------|--------------------|------------------|-------------------|-------------------|
| | | SBLEU \downarrow | PPL \downarrow | Cont. \uparrow | Acc. \uparrow | SBLEU \downarrow | PPL \downarrow | Cont. \uparrow | Acc. \uparrow |
| SST-2 | zero-shot | 0.085 \pm 0.008 | 410.8 \pm 5.12 | 0.772 \pm 0.004 | 0.742 \pm 0.003 | 0.093 \pm 0.005 | 202.4 \pm 2.01 | 0.714 \pm 0.004 | 0.842 \pm 0.005 |
| | few-shot | 0.086 \pm 0.006 | 282.7 \pm 2.02 | 0.769 \pm 0.003 | 0.778 \pm 0.003 | 0.089 \pm 0.002 | 204.7 \pm 7.50 | 0.721 \pm 0.001 | 0.831 \pm 0.002 |
| | CoT | 0.081 \pm 0.012 | 267.5 \pm 6.71 | 0.776 \pm 0.004 | 0.752 \pm 0.005 | 0.095 \pm 0.002 | 194.6 \pm 1.58 | 0.709 \pm 0.007 | 0.829 \pm 0.004 |
| | Self-Re | 0.307 \pm 0.002 | 59.6 \pm 1.03 | 0.593 \pm 0.003 | 0.848 \pm 0.006 | 0.282 \pm 0.018 | 50.4 \pm 1.43 | 0.534 \pm 0.003 | 0.849 \pm 0.003 |
| | OPRO | 0.078 \pm 0.009 | 169.7 \pm 1.68 | 0.738 \pm 0.003 | 0.753 \pm 0.004 | 0.098 \pm 0.009 | 143.2 \pm 0.93 | 0.708 \pm 0.005 | 0.863 \pm 0.003 |
| Ours | | 0.082 \pm 0.009 | 237.54 \pm 2.14 | 0.799 \pm 0.001 | 0.824 \pm 0.002 | 0.083 \pm 0.005 | 184.1 \pm 3.95 | 0.743 \pm 0.003 | 0.872 \pm 0.001 |
| IMDB | zero-shot | 0.191 \pm 0.010 | 41.8 \pm 0.47 | 0.842 \pm 0.002 | 0.743 \pm 0.005 | 0.249 \pm 0.032 | 39.4 \pm 0.48 | 0.832 \pm 0.003 | 0.917 \pm 0.005 |
| | few-shot | 0.181 \pm 0.008 | 43.4 \pm 0.35 | 0.888 \pm 0.007 | 0.648 \pm 0.012 | 0.254 \pm 0.012 | 49.3 \pm 1.44 | 0.668 \pm 0.004 | 0.803 \pm 0.002 |
| | CoT | 0.151 \pm 0.007 | 48.8 \pm 0.74 | 0.848 \pm 0.003 | 0.729 \pm 0.005 | 0.248 \pm 0.011 | 39.4 \pm 0.89 | 0.833 \pm 0.004 | 0.916 \pm 0.005 |
| | Self-Re | 0.234 \pm 0.008 | 33.3 \pm 0.50 | 0.801 \pm 0.004 | 0.778 \pm 0.007 | 0.364 \pm 0.003 | 24.6 \pm 0.87 | 0.653 \pm 0.002 | 0.678 \pm 0.008 |
| | OPRO | 0.168 \pm 0.005 | 41.4 \pm 0.07 | 0.778 \pm 0.003 | 0.801 \pm 0.005 | 0.304 \pm 0.012 | 32.4 \pm 1.23 | 0.773 \pm 0.003 | 0.939 \pm 0.005 |
| Ours | | 0.199 \pm 0.005 | 38.6 \pm 0.87 | 0.844 \pm 0.002 | 0.824 \pm 0.005 | 0.251 \pm 0.003 | 37.4 \pm 0.83 | 0.822 \pm 0.002 | 0.969 \pm 0.007 |
| GYAFC | zero-shot | 0.095 \pm 0.003 | 81.2 \pm 1.33 | 0.774 \pm 0.003 | 0.423 \pm 0.006 | 0.154 \pm 0.009 | 74.2 \pm 1.34 | 0.703 \pm 0.002 | 0.774 \pm 0.007 |
| | few-shot | 0.106 \pm 0.011 | 88.4 \pm 2.31 | 0.758 \pm 0.004 | 0.564 \pm 0.005 | 0.165 \pm 0.002 | 90.2 \pm 3.32 | 0.667 \pm 0.005 | 0.763 \pm 0.003 |
| | CoT | 0.105 \pm 0.009 | 86.0 \pm 0.99 | 0.759 \pm 0.001 | 0.518 \pm 0.002 | 0.146 \pm 0.008 | 101.8 \pm 2.06 | 0.692 \pm 0.003 | 0.554 \pm 0.003 |
| | Self-Re | 0.114 \pm 0.007 | 92.8 \pm 1.38 | 0.708 \pm 0.002 | 0.781 \pm 0.005 | 0.174 \pm 0.006 | 91.0 \pm 1.53 | 0.614 \pm 0.001 | 0.862 \pm 0.005 |
| | OPRO | 0.092 \pm 0.008 | 90.3 \pm 1.25 | 0.732 \pm 0.004 | 0.718 \pm 0.005 | 0.128 \pm 0.007 | 111.7 \pm 3.09 | 0.631 \pm 0.002 | 0.830 \pm 0.004 |
| Ours | | 0.111 \pm 0.003 | 78.4 \pm 1.43 | 0.722 \pm 0.002 | 0.783 \pm 0.001 | 0.116 \pm 0.004 | 93.2 \pm 0.75 | 0.692 \pm 0.004 | 0.872 \pm 0.004 |

Table 9: Comparison of style transfer on SST-2, IMDB and GYAFc. Harmonic mean (**H**) is highlighted in gray, with the best results in bold.

Template for Feedback in Self-Refine.

Given Sentence: If you ever stayed at the Trop, you may have noticed that it's not quite up to the standards of other Vegas hotels. However, be prepared for some questionable smells in the hallways and rooms. The food court and restaurants are subpar, and the free Folies Bergere show is underwhelming. The rooms have a retirement community feel to them. Overall, it's not the best option, but it may work in a pinch.

Revised Sentence: If you're considering a stay at the Trop, it has a unique charm that some of the more polished Vegas hotels might lack. While the vintage ambiance of the hallways and rooms may evoke a nostalgic feel, it adds to the character of the place. The food court and restaurants offer simple, no-frills dining options, and the complimentary Folies Bergere show provides a classic Vegas entertainment experience. The rooms are comfortably old-fashioned, reminiscent of a cozy retirement community. Overall, the Trop is a satisfactory choice if you're looking for a more laid-back and budget-friendly alternative in Vegas.

Why is this revised sentence not very positive? Please give feedback.

Feedback: The revised sentence might not seem very positive because it uses phrases like "vintage ambiance" and "comfortably old-fashioned," which can imply outdatedness rather than charm. Additionally, terms like "no-frills" and "cozy retirement community" may suggest simplicity and old age, not luxury or modern appeal.

Given Sentence: If you're looking for a unique and affordable experience in Vegas, the Trop may be the perfect place for you. The hallways and rooms have a charming and cozy feel...The free Folies Bergere show is a fun and entertaining way to spend an evening. Overall, it's a great value and an enjoyable stay.

Revised Sentence: If you're considering a budget stay in Vegas, the Trop might initially seem appealing. However, the hallways and rooms feel outdated rather than charming, and the coziness can come off as cramped. While the Folies Bergere show is free, it might be disappointing for those expecting a high-quality Vegas performance. Overall, despite its affordability, the experience leaves much to be desired.

Why is this revised sentence not very negative? Please give feedback.

Feedback: The review isn't very negative because it still mentions positive aspects, such as the initial appeal of affordability and the inclusion of a free show. To convey a stronger negative sentiment, it should focus more on the shortcomings and omit any positive framing. Given Sentence: a terrific date movie , whatever your orientation.

Given Sentence: *input sentence*

Revised Sentence: *revised sentence*

Why is this review not very *target sentiment*? Please give feedback.

Feedback:

Figure 14: Self-Refine (Feedback) template on SST-2 and IDMB. Text in black represents the fixed template, while text in blue indicates the input information needed for a specific data sample.

Template for Refine in Self-Refine.

Feedback: The review is not very negative because it lacks specific details or explanations to support the claim that the movie is terrible. To provide a more negative review, specific reasons or examples should be given to justify the negative assessment. Okay, let's try again. Rewrite this sentence to have a very "negative" sentiment using the feedback above. New Revised Sentence: Given Sentence: If you ever stayed at the Trop, you may have noticed that it's not quite up to the standards of other Vegas hotels. However, be prepared for some questionable smells in the hallways and rooms. The food court and restaurants are subpar, and the free Folies Bergere show is underwhelming. The rooms have a retirement community feel to them. Overall, it's not the best option, but it may work in a pinch.

Revised Sentence: If you ever wondered where the magic of Vegas crawled into a hole to rot, look no further. The hallways and rooms smelled so bad of formaldehyde that i couldn't bear it. Their food court and restaurants are terrible. The free Folies Bergere show is lame. The rooms look like a retirement community from florida. I highly recommend avoiding the Trop like the plague.

Feedback: The rewritten sentence is "Very negative". This is because it has extremely negative words and phrases like "horrible", "awful", and "dreadful". The output is extremely vile and negative. This review is already as negative as it can get. Excellent work! But there is always room for more. Make it even more negative. Try again!

Feedback: The Revised Premise Sentence does not contradict the Given Hypothesis because both can be true simultaneously. The Hypothesis suggests the men are "riding together," which can imply a shared activity rather than close proximity. The Revised Premise indicates they are "spaced far apart," but this doesn't negate the possibility of them riding together in a general sense.

Okay, let's try again. Rewrite this sentence to have a very negative sentiment using the feedback above.

New Revised Sentence: If you ever wondered where the hell is the bottom of the barrel, look no further. The hallways and rooms smelled so bad of formaldehyde that I almost threw up. Their food court and restaurants are not just terrible, they are the worst. The free Folies Bergere show is lame and disgusting. The rooms look like a retirement community from florida. I highly recommend avoiding the Trop like the plague. You will regret it if you don't.

Given Sentence: *input sentence*

Revised Sentence: *revised sentence*

Feedback: *feedback from the last step*

Okay, let's try again. Rewrite this sentence to have a very *target sentiment* sentiment using the feedback above.

New Revised Premise Sentence:

Figure 15: Self-Refine (Refine) template on SST-2 and IDMB. Text in black represents the fixed template, while text in blue indicates the input information needed for a specific data sample.

Template for OPRO.

Task Definition: Revise a given sentence with minimal changes to alter its sentiment polarity.

Given Sentence: *input sentence*

Target Sentiment Polarity: *targeted sentiment*

Revised Sentence 1: *revised sentence 1*

Distance to the Given Sentence: *Levenshtein distance*

Distance to the Target Sentiment Polarity: *0 or 1*

Examples from previous steps...

The loss contains two parts: Distance to the Given Sentence and the distance to the Target Sentiment Polarity.

Given the task and loss definition, please generate a new Revised Sentence that minimizes the loss.

Revised Sentence:

Figure 16: OPRO template on SST-2 and IDMB. Text in black represents the fixed template, while text in blue indicates the input information needed for a specific data sample.

Result on SNLI

Given Premise Sentence: *A person riding a bicycle down a red ramp.*

Given Hypothesis Sentence: *A person riding a bicycle fast down a red ramp.*

Target Relationship: *Contradiction*

Revised Sentence 1: *A person walking slowly down a red ramp.*

Distance to the Given Premise Sentence: *0.333*

Distance to the Target Relationship: *0*

Loss: *0.0333*

Revised Sentence 2: *A person riding a blue bicycle down a red ramp.*

Distance to the Given Premise Sentence: *0.111*

Distance to the Target Relationship: *1*

Loss: *1.0111*

Analysis: Sentence 1 is better than Sentence 2 because Sentence 1 maintains the key element of a person moving down a ramp, albeit walking slowly, which directly contradicts the hypothesis of riding a bicycle fast. To minimize the loss, a new Revised Premise Sentence could be: "A person pushing a broken bicycle down a red ramp."

Result on SST-2

Given Sentence: *it's not as awful as some of the recent hollywood trip tripe... but it's far from a groundbreaking endeavor.*

Target Sentiment Polarity: *positive*

Revised Sentence 1: *it's not as bad as some of the recent hollywood trip tripe... but it's far from a groundbreaking endeavor.*

Distance to the Given Sentence: *0.048*

Distance to the Target Sentiment Polarity: *1*

Loss: *1.0048*

Revised Sentence 2: *it's not as impressive as some of the recent hollywood trip tripe... but it's far from a groundbreaking endeavor.*

Distance to the Given Sentence: *0.048*

Distance to the Target Sentiment Polarity: *1*

Loss: *1.0048*

Analysis: The two revised sentences have high losses because they only made minimal changes to the original sentence without effectively altering the sentiment polarity to positive. To minimize the loss, a more significant change is needed to shift the sentiment towards a positive direction. Revised Sentence: "While it may not reach groundbreaking status, it's a step above the recent Hollywood trip tripe."

Figure 17: Qualitative results of C&G on SNLI, and SST-2. The text in blue is provided by our C&G for comparison, while text in violet marks the output from the model. Underlined words show the improvement.

Result on GYAFc

Given Sentence: If that breaks your heart, then I'm a good heart breaker.

Revised Sentence 1: If that crushes you, then I'm a pro at breaking hearts.

Semantic Similarity to the Given Sentence: 0.813

Informality Score: 2.0

Score: 2.0813

Revised Sentence 2: If that hits you hard, then I'm a pro at breaking hearts.

Semantic Similarity to the Given Sentence: 0.793

Informality Score: 2.0

Score: 2.0793

Analysis: Sentence 1 is better than Sentence 2 because it maintains a closer semantic similarity to the given sentence while being informal. The use of "crushes you" captures the essence of "breaks your heart" more effectively than "hits you hard" in Sentence 2. New Revised Sentence: "If that crushes you, then I'm a heartbreaker pro."

Figure 18: Example of C&G using GPT-3.5 as the backbone on GYAFc. The text in blue is provided by our C&G for comparison, while text in violet marks the output from the model. Underlined words show the improvement.

Result on IMDB

Task Definition: Revise a given sentence with minimal changes to alter its sentiment polarity.

Given Sentence: The actors play wonderfully, especially Kenneth Branagh himself. It's good that Robin Williams got the comedy role of Osiric, otherwise it could be a bit strange to see him in such a production. It is really great that Kenneth decided to use the fullest version of the text, this happens definitely not too often... Thanks to that the viewers can see the whole, not the chosen - by the director - parts. Also - thank God that the film is in a classical form; NO to surrealistic fanfaberries ! Although "Tytus Andronicus" was impressive nevertheless, but still Hamlet is a different story, at least that's my point of view.

Target Sentiment Polarity: negative

Revised Sentence 1: The actors play disappointingly, especially Kenneth Branagh himself. It's unfortunate that Robin Williams got the comedy role of Osiric, otherwise it could be a bit strange to see him in such a production. It is really disappointing that Kenneth decided to use a watered-down version of the text, this happens definitely too often... Luckily, the viewers are only shown the chosen parts - by the director - and not the whole. Also, thank goodness the film is in a surrealistic form; YES to experimental storytelling! Although.

Distance to the Given Sentence: 0.47706422018348627

Distance to the Target Sentiment Polarity: 1

Loss: 1.0477064220183485

Revised Sentence 2: The actors play unconvincingly, especially Kenneth Branagh himself. It's unfortunate that Robin Williams was miscast in the comedy role of Osiric, which makes the production feel awkward. The text is adapted in a way that's overly familiar, a common mistake. The viewers are only shown the director's chosen parts, which is a shame. Thankfully, the film's surrealistic approach is a refreshing change from the usual; however, it's not enough to redeem the overall mediocrity. Hamlet is a different story, and I'm underwhelmed.

Distance to the Given Sentence: 0.8440366972477065

Distance to the Target Sentiment Polarity: 0

Loss: 0.08440366972477065

Distance to the Given Sentence: 0.08152173913043478

Distance to the Target Sentiment Polarity: 1

Loss: 0.008152173913043478

The loss contains two parts: The distance to the Given Sentence and the distance to the Target Sentiment Polarity.

Revised Sentence 2 is better than 1.

Given the task and loss definition, please first think why the Sentence 1 is better than Sentence 2, and give the analysis. Then, generate a new Revised Premise Sentence that minimizes the loss.

Sentence 2 is better than Sentence 1 because it has a lower distance to the target sentiment polarity (0) compared to Sentence 1 (1). This means that Sentence 2 is more aligned with the target sentiment polarity of "negative".

To generate a new revised sentence that minimizes the loss, I will analyze the strengths and weaknesses of Sentence 2. The strengths are:

* It has a clear negative sentiment towards the film, with phrases such as "unconvincingly", "miscast", "awkward", "overly familiar", "common mistake", "shame", "mediocrity", and "underwhelmed".

* It provides specific criticisms of the film, such as the casting of Robin Williams and the adaptation of the text.

The weaknesses are:

* The sentence is a bit lengthy and could be condensed for better flow.

* Some of the phrases, such as "refreshing change" and "not enough to redeem", could be rephrased for a more consistent negative tone.

With these points in mind, I propose the following revised sentence:

The actors' performances are unconvincing, with Kenneth Branagh being particularly disappointing. Robin Williams' miscasting in the comedy role of Osiric makes the production feel awkward and forced. The text is overly familiar, a common mistake that results in a lackluster film. The director's chosen parts are all we get to see, which is a shame. The surrealistic approach is a failed experiment, and the film's mediocrity is underwhelming. Hamlet is a different story, and I'm left unimpressed.

This revised sentence maintains the negative sentiment of Sentence 2 while addressing the weaknesses mentioned above. It is more concise and has a more consistent tone throughout. The distance to the given sentence is likely to be lower, and the distance to the target sentiment polarity remains at 0.

Figure 19: Example of C&G using Llama3 as the backbone on IMDB. The text in **blue** is provided by our C&G for comparison, while text in **violet** marks the output from the model.

#Result on GYAFc

Task Definition: Please transfer the given sentence to an informal one.

Given Sentence: The children had boy puppets that were dressed sloppily.

Revised Sentence 1: The kids had these ridiculously sloppy boy puppets that looked like they'd been put together by a toddler having a meltdown.

Semantic Similarity to the Given Sentence: 0.855998158454895

Informality Score: 5.0

Score: 5.0855998158454895

Revised Sentence 2: The kids had these super sloppy boy puppets that looked like they'd been through a messy playdate.

Semantic Similarity to the Given Sentence: 0.8576760292053223

Informality Score: 5.0

Score: 5.0857676029205322

The score contains two parts: Semantic similarity between the Revised Sentence and the Given Sentence and the informality of the Revised Sentence.

Revised Sentence 2 is better than 1. Given the task and score definition, please first think why the Sentence 2 is better than Sentence 1. Generate a new Revised Sentence that maximizes the score.

Based on the scores, Revised Sentence 2 is better than Sentence 1 because it has a slightly higher semantic similarity to the original sentence (0.8576760292053223 vs 0.855998158454895) and the same informality score of 5.0.

To generate a new Revised Sentence that maximizes the score, I'll analyze what makes Sentence 2 better. It seems that the addition of the phrase "through a messy playdate" adds a more vivid and relatable description of the puppets' appearance, which increases the semantic similarity to the original sentence. It also maintains the same level of informality as Sentence 1.

With this in mind, I'll try to improve upon Sentence 2 by adding more descriptive language while keeping the same level of informality. Here's a new Revised Sentence:

Revised Sentence: "The kids had these ridiculously sloppy boy puppets that looked like they'd been through a chaotic craft session gone wrong."

This sentence maintains the same level of informality as Sentence 2, but adds more descriptive language to increase the semantic similarity to the original sentence. The phrase "chaotic craft session gone wrong" adds a more vivid and relatable description of the puppets' appearance, which should increase the score.

Figure 20: Example of C&G using Llama3 as the backbone on GYAFc. The text in **blue** is provided by our C&G for comparison, while text in **violet** marks the output from the model.

#Result on MNLI

Task Definition: Revise the premise sentence, using minimal changes, to alter the relationship between it and the hypothesis sentence to either entailment, contradiction, or neutral.

Given Premise Sentence: Significant findings and recommendations are those matters that, if not corrected, could affect the results of the auditors' work and users' conclusions about those results.

Given Hypothesis Sentence: The findings need to remain incorrect to improve the results.

Target Relationship: Entailment

Revised Premise Sentence 1: Significant findings and recommendations are those matters that, if not corrected, could improve the results of the auditors' work and users' conclusions about those results.

Distance to the Given Premise Sentence: 0.004

Distance to the Target Relationship: 1

Loss: 1.0004

Revised Premise Sentence 2: Significant findings and recommendations are those matters that, if not corrected, will improve the results of the auditors' work and users' conclusions about those results.

Distance to the Given Premise Sentence: 0.008

Distance to the Target Relationship: 1

Loss: 1.0008

The loss contains two parts: Distance to the Given Premise Sentence and the distance to the Target Relationship.

Revised Premise 1 is better than 2. Given the task and loss definition, please first think why the Sentence 1 is better than Sentence 2, and give the analysis. Then, generate a new Revised Premise Sentence that minimizes the loss. Please put the new revised sentence in [].

Revised Premise Sentence 1 is better than Sentence 2 because it makes a smaller change to the original premise sentence while still attempting to alter the relationship to entailment. The change from "could affect" to "could improve" is minimal and keeps the structure and meaning of the sentence closer to the original. In contrast, Sentence 2 changes "could affect" to "will improve," which is a more significant alteration and deviates further from the original premise. To achieve the target relationship of entailment with minimal changes, we need to ensure that the revised premise clearly supports the hypothesis that "The findings need to remain incorrect to improve the results." This means the premise should indicate that not correcting the findings leads to an improvement.

Here is a revised premise sentence that aims to minimize the loss:

[Significant findings and recommendations are those matters that, if left incorrect, could improve the results of the auditors' work and users' conclusions about those results.]

This revision maintains the structure of the original sentence while altering the relationship to entailment by suggesting that leaving the findings incorrect could lead to an improvement.

Figure 21: Example of C&G using GPT-4 as the backbone on MNLI. The text in **blue** is provided by our C&G for comparison, while text in **violet** marks the output from the model.

Result on IMDB

Task Definition: Revise a given sentence with minimal changes to alter its sentiment polarity.

Given Sentence: This is a thriller with a good concept, good acting, good photography and good intentions all around, but which is confused and disjointed in execution.
>
Garcia stars as John Berlin, an L.A. forensic detective who has moved to a small California town at the behest of a friend of his on the force there. He soon becomes involved in the investigation of an unsolved murder which leads to his theorizing about the existence of a serial killer whom no one else believes in. The known victim is theorized to be blind, which leads to a romance with a blind girl - believed to be a witness - at a nearby school for the blind.
>
Despite a basically intriguing story there were too many quantum leaps and plot holes in this movie where I found myself wondering, 'how the hell did we wind up here?' or 'how did we find this out?' I found it confusing and disjointed, despite the good acting, etc. John Malkovich has a small part toward the end as an F.B.I. investigator out to get Berlin.
>
Not recommended.

Target Sentiment Polarity: positive

Revised Sentence 1: This is a thriller with a brilliant concept, excellent acting, stunning photography and noble intentions all around, but which is complex and multifaceted in execution.
>
Garcia stars as John Berlin, an L.A. forensic detective who has moved to a small California town at the behest of a friend of his on the force there. He soon becomes involved in the investigation of an unsolved murder which leads to his theorizing about the existence of a serial killer whom no one else believes in. The known victim is theorized to be blind, which leads to a romance with a blind girl - believed to be a witness - at a nearby school for the blind.
>
Despite a basically intriguing story there were many unexpected twists and turns in this movie where I found myself wondering, 'how the hell did we wind up here?' or 'how did we find this out?' I found it intriguing and multifaceted, despite the excellent acting, etc. John Malkovich has a small part toward the end as an F.B.I. investigator out to get Berlin.
>
Highly recommended

Distance to the Given Sentence: 0.08152173913043478

Distance to the Target Sentiment Polarity: 1

Loss: 0.008152173913043478

Revised Sentence 2: This is a thriller with a brilliant concept, excellent acting, stunning photography, and noble intentions all around, but which is complex and multifaceted in execution.
>
Garcia stars as John Berlin, an L.A. forensic detective who has moved to a small California town at the behest of a friend of his on the force there. He soon becomes involved in the investigation of an unsolved murder which leads to his theorizing about the existence of a serial killer whom no one else believes in. The known victim is theorized to be blind, which leads to a romance with a blind girl - believed to be a witness - at a nearby school for the blind.
>
Despite a basically intriguing story, there were many unexpected twists and turns in this movie where I found myself eagerly wondering, 'how did we wind up here?' or 'how did we find this out?' I found it intriguing and multifaceted, enhanced by the excellent acting, etc. John Malkovich has a small part toward the end as an F.B.I. investigator out to get Berlin.
>
Highly recommended.

Distance to the Given Sentence: 0.11956521739130435

Distance to the Target Sentiment Polarity: 0

Loss: 0.011956521739130435

Distance to the Given Sentence: 0.08152173913043478

Distance to the Target Sentiment Polarity: 1

Loss: 0.008152173913043478

The loss contains two parts: The distance to the Given Sentence and the distance to the Target Sentiment Polarity.

Revised Sentence 1 is better than 2.

Given the task and loss definition, please first think why the Sentence 1 is better than Sentence 2, and give the analysis. Then, generate a new Revised Premise Sentence that minimizes the loss.

Analysis:

Revised Sentence 1 is better than **Revised Sentence 2** primarily because it maintains a closer resemblance to the original sentence while effectively altering the sentiment polarity to positive. Here are the key points of analysis:

1. **Sentiment Polarity:**

- Both sentences successfully shift the sentiment from negative to positive by using words like "brilliant," "excellent," "stunning," and "noble."

- Sentence 1 uses "unexpected twists and turns" which is a more neutral to positive phrase compared to Sentence 2's "eagerly wondering," which might still carry a slight sense of confusion.

2. **Minimal Changes:**

- Sentence 1 makes fewer changes to the original sentence structure and wording, maintaining the original flow and meaning more closely.

- Sentence 2 introduces additional words like "eagerly," which slightly deviates from the original sentence's structure.

3. **Clarity and Consistency:**

- Sentence 1 maintains clarity and consistency by using "complex and multifaceted" instead of "confusing and disjointed," which aligns well with the positive sentiment.

- Sentence 2, while also positive, slightly overcomplicates the sentence with additional adjectives and adverbs.

New Revised Sentence:

This is a thriller with a brilliant concept, excellent acting, stunning photography, and noble intentions all around, but which is complex and multifaceted in execution.
>
Garcia stars as John Berlin, an L.A. forensic detective who has moved to a small California town at the behest of a friend of his on the force there. He soon becomes involved in the investigation of an unsolved murder which leads to his theorizing about the existence of a serial killer whom no one else believes in. The known victim is theorized to be blind, which leads to a romance with a blind girl - believed to be a witness - at a nearby school for the blind.
>
Despite a basically intriguing story, there were many unexpected twists and turns in this movie where I found myself captivated, wondering, 'how did we wind up here?' or 'how did we find this out?' I found it intriguing and multifaceted, enhanced by the excellent acting, etc. John Malkovich has a small part toward the end as an F.B.I. investigator out to get Berlin.
>
Highly recommended.

Figure 22: Example of C&G using GPT-4 as the backbone on IMDB. The text in **blue** is provided by our C&G for comparison, while text in **violet** marks the output from the model.