Enhancing Editorial Tasks: A Case Study on Rewriting Customer Help Page Contents Using Large Language Models

Aleksandra Gabryszak[1], Daniel Röder[1], Arne Binder[1], Luca Sion[2]*, Leonhard Hennig[1]
[1]German Research Center for Artificial Intelligence (DFKI)
[2]Deutsche Telekom AG
{firstname.lastname}@dfki.de

Abstract

In this paper, we investigate the use of large language models (LLMs) to enhance the editorial process of rewriting customer help pages. We introduce a German-language dataset comprising Frequently Asked Question-Answer pairs, presenting both raw drafts and their revisions by professional editors. On this dataset, we evaluate the performance of four large language models (LLM) through diverse prompts tailored for the rewriting task. We conduct automatic evaluations of content and text quality using ROUGE, BERTScore, and ChatGPT. Furthermore, we let professional editors assess the helpfulness of automatically generated FAQ revisions for editorial enhancement. Our findings indicate that LLMs can produce FAQ reformulations beneficial to the editorial process. We observe minimal performance discrepancies among LLMs for this task, and our survey on helpfulness underscores the subjective nature of editors' perspectives on editorial refinement.

1 Introduction

In this paper, we evaluate the suitability of large language models to support the editorial process of customer help pages. The continuous evolution of natural language processing (NLP) technologies, particularly exemplified by advanced models like GPT-4 (Team, 2023), presents exciting prospects for content management across various sectors. One area where these models hold promise is in the maintenance and enhancement of customer help pages, which serve as vital resources for addressing user queries and concerns related to products or services.

The editorial workflow for customer help pages necessitates precision, clarity, and relevance to ensure users can efficiently locate solutions. Traditionally, this workflow involves manual content creation, review, and updates by human editors. However, managing the volume of content and keeping information current pose significant challenges. Large language models offer a compelling opportunity to enhance and expedite these editorial processes, potentially boosting efficiency and responsiveness to user needs.

Our objective is to explore practical applications of large language models in supporting essential editorial tasks for customer help pages. We will investigate how these models can contribute to content creation and quality control. By evaluating the advantages and constraints of incorporating such models into the editorial workflow, we aim to provide insights into their feasibility and effectiveness within customer support operations. This evaluation is essential for understanding how large language models can impact the scalability and responsiveness of customer help services in the digital era. The main contributions of this paper are:

1. Providing a dataset of FAQ question-answer pairs for testing editorial rewriting process,
2. Comparison of several LLMs on the task of FAQ rewriting,
3. Automatic assessment of content and verbal quality of automatically rewritten FAQ texts,
4. Manual error analysis of hallucinations,
5. Evaluation conducted by human experts on the helpfulness of machine-generated text reformulations in the editorial process.

2 Related work

The application of LLMs for rewriting texts covers a variety of text generation tasks, such as summarizing (Jin et al., 2024), text simplification (Tan et al., 2024), style transfer (Pu and Demberg, 2023) or query rewriting (Ma et al., 2023). The evaluation datasets often cover only one of those tasks,

*The opinions expressed in this article are the author's own and do not necessarily represent the views of Deutsche Telekom AG.

Proceedings of the 17th International Natural Language Generation Conference, pages 402–411
September 23–27, 2024.

however multi-purpose benchmarks have started emerged in recent years.

Dwivedi-Yu et al. (2022) created EditEval, an instruction-based suite that leverages high-quality existing and new datasets to automatically assess editing capabilities, including enhancing text fluency and clarity, as well as rewriting to simplify, neutralize, or update content. It covers various text types such as Wikipedia articles, Wikinews, news articles, and scientific publications from arXiv. The benchmark is provided with results of baselines, which use greedy decoding and do not perform any task-specific fine-tuning or in-context learning. The authors evaluate various LLMs using zero-shot prompting. The evaluation reveals that most baseline models lag behind the supervised state-of-the-art, especially in tasks like neutralizing and updating information. The analysis also indicates that commonly used metrics for editing tasks do not always correlate well, and optimizing for the highest-performing prompts does not necessarily ensure robustness across different models.

Shu et al. (2023) created a benchmark OpenRewriteEval by collecting human-generated text rewrites with natural language instructions. The benchmark is designed for testing cross-sentence rewrite of various types, such as text formality, expansion, conciseness, paraphrasing, tone and style transfer. The authors also developed RewriteLM, an instruction-tuned large language model designed for cross-sentence text rewriting. The model undergoes supervised fine-tuning and reinforcement learning (RL). For instruction tuning, edits from Wikipedia are extracted and filtered, and the associated edit summary of the revision is used as a proxy for the instructions. Additionally, to diversify the dataset a synthetic set of instructions is generated using chain-of-thought prompting and post-processing. The authors tested RewriteLM on EditEval and OpenRewriteEval and compared the results against a set of models, including various PaLM variants, LLama, Alpaca, GPT-3, InsGPT.

Zhu et al. (2023) addresses the problem of impracticality of large language models for the rewriting task on mobile-device due to models size. The authors recognize that developing a smaller, effective language model for text rewriting is challenging due to the need to balance size with maintaining capabilities, which requires expensive data collection. To tackle the challenge, a new instruction tuning method for mobile text rewriting models is introduced, generating high-quality training data without human labeling. A heuristic reinforcement learning framework improves performance without preference data. For the assesment of mobile text rewriting tasks a benchmark MessageRewriteEval is introduced. Empirical tests show the on-device model outperforms current state-of-the-art models while being much smaller.

3 Task and Data

3.1 Task definition

In our experimental setup, we aim to automatically transform raw versions of FAQ help pages into polished, easily readable texts for customers. The reformulation task involves transforming a text that may contain potential orthographic errors, complicated or unclear structure, too technical or unnecessary information into a text containing the most relevant information written in a stylistically clear way understandable for a broad range of customers. The task is an automatic version of editing process that is typically performed manually by professional editors at companies.

3.2 Data

The data was provided by Deutsche Telekom AG (DT), a large German telecommunications company. The corpus comprises 56 question-answer pairs addressing potential customer inquiries across various topics, including additional SIM cards, Netflix subscriptions, relocation, changing mobile service providers, house connection orders, hardware order and delivery status, and fixed-line internet and TV setup. The data is in German language.

For each FAQ pair, a raw input is provided by specialized departments, and a rewritten gold output is crafted by a professional editor of DT. The final dataset also includes LLM generated FAQ-pairs, which we created as described in section 4. Table 1 illustrates examples of input text and a gold reference text by professional editors, as well as an example LLM-generated rewritten version. The dataset is publicly available.[1]

4 Text Generation Pipeline

In the following we describe the text generation pipeline using various LLMs and prompt types.

input text	Q: *Do I have a special termination right if I move abroad?* *A: Yes. If you move abroad, you have a special termination right according to § 60 TKG (Telecommunications Act) with a legal notice period of one month. We require one of the following documents as proof: deregistration certificate from the residents' registration office, employer certificate, deployment order (for soldiers). Feel free to use the landline or mobile termination form and attach the proof as an attachment.*
reformulation crafted by professional editor	*Q: Do I have a special termination right if I move abroad?* *A: Yes, when moving abroad, you have a special termination right according to § 60 TKG (Telecommunications Act) with a legal notice period of 1 month. Simply use the online termination form for landline or mobile services. If you have a short stay abroad, you can suspend your contract.*
reformulation generated by LLM	*Q: Can I terminate my contract early if I move abroad?* *A: Yes, you can terminate your contract early if you move abroad. You have one month to do so. Please send us proof, such as a deregistration certificate, along with the completed termination form for landline or mobile services.*

Table 1: Example of FAQ pair: an input text, a gold reference text by professional editors and LLM-generated outputs after rewriting of the input. The texts are English translations of original German data.

	GPT	Llama2	Leo-Mistral
top_p	1	0.95	0.95
temperature	0.7	0.8	0.8
max_new_tokens	unlimited	2000	2000
max-input-length	-	2200	2200
max-total-tokens:	-	4096	4096

Table 2: Hyperparameters configured for selected LLMs used in the FAQ rewriting task: GPT-3.5 Turbo, GPT-4, LLama2-UP, LLama2-OA, and Leo-Mistral.

4.1 Pipeline and LLMs

For our experiment, we implemented a generation pipeline based on LangChain[2] to evaluate different large language models: OpenAI's GPT-3.5 Turbo[3] and GPT-4[4], two instruction-tuned variants of Llama2-70B fine-tuned on OpenAssistant (Llama2-OA)[5] and Orca-/Alpaca-style (Llama2-UP)[6] data respectively, and EM German Leo Mistral (Leo-Mistral)[7]. We ran the AWQ-quantized version of the open source models via HuggingFace's Text Generation Inference library[8]. Models were selected based on their performance on German-language text at the time of the experiments, and to include both proprietary and open-source models. Table 2 shows the hyperparameters for running the text generation experiments. We used default hyperparameter values as given by their API for OpenAI's models. For the open-source models, we used default parameter values from the LangChain implementation, except for the parameter temperature, which we set to 0.8 following Meister et al. (2022). For the open source models, we also increased the server-side maximum input length and number of new tokens, to be able to process the few-shot prompts.

4.2 Prompts

We defined mandatory and optional prompt components, which then were combined to prompt variants of different complexity.

Prompt components We designed various prompt components, as shown in Table 3, which are then used to build different prompt variants. The mandatory prompt components are the system prompt, base prompt and output format instruction. *System prompt* contains general information about wording style and role of the LLM model as editor for help texts for the telecommunication company website. *Base prompt* gives a direct instruction to reformulate FAQ. It explains the input structure as being a question-answer pair on a technical topic, provides one original question-answer pair and asks for its transformation. *Output format instruction* asks for three different reformulation suggestions being returned in a JSON format. The optional prompt components are additional instructions how to solve the task and examples of reformulations. The *Step-by-step "chain-of-thought" instruction* has proven to be a successful strategy, enabling LLMs to provide more precise answers. This approach is often implemented as a straightforward instruction within the prompt (see e.g. (Kojima et al., 2022) and the GPT-4 Techni-

[1] https://github.com/DFKI-NLP/faq-rewrites-llms
[2] https://www.langchain.com/
[3] gpt-3.5-turbo-0613
[4] gpt-4-0613
[5] https://hf.co/TheBloke/Llama-2-70B-OASST-1-200-AWQ
[6] https://hf.co/TheBloke/Upstage-Llama-2-70B-instruct-v2-AWQ
[7] https://hf.co/TheBloke/em_german_leo_mistral-AWQ
[8] https://github.com/huggingface/text-generation-inference

component type	component text
system prompt	You are a helpful editor of Deutsche Telekom help pages. You write help texts for customers who use the organizations products. Use simple, understandable language and shorten complicated or overly long questions and answers. Avoid negations. Use examples when appropriate
base prompt	Input: An Original Question and Answer (Q|A), consisting of one specific, detailed question and a technical, detailed one answer. Goal: Transform the Original Q|A into a Gold Q|A. The gold question should be more general and understandable to a wider audience. The gold answer should be simplified, clear and direct, focusing on the answering the question from the customer's perspective. Input: Original Question: {prompt_question} Original Answer: {prompt_answer}
json output	Generate up to 3 variants and return them in the following JSON format (Note: xxx is a wildcard). [{{'question': xxx, 'answer': xxx}}, {{'question': xxx, 'answer': xxx}}, {{'question': xxx, 'answer': xxx}}]. Please give me the reformulations in the given format without any further comment.
step-by-step*	Think step by step.
explicit instruction*	Instructions: 1. Analyze the original Q|A to identify the core of the question and the most important information in the answer. 2. Rephrase the question to make it more general and inclusive. Avoid overly specific or technical terms and make sure it is understandable to a broad audience. 3. If necessary, include helpful resources or links that may provide the reader with additional information or support. 4. Ensure the reworded Gold Q|A is clear, concise and customer-centric.
example integration*	Example input: Original Question: {orig_question} Original Answer: {orig_answer} Expected output: {{'question': {gold_question}, ' answer':{gold_answer}}}

Table 3: Prompt components for FAQ rewriting (the optional components are marked with *). The original prompts are in German and have been translated into English for readability.

prompt name	prompt components
zeroshot	system prompt, base prompt, json output
zeroshot step-by-step	zeroshot + step-by-step instruction
zeroshot instruction	zeroshot + explicit instruction
fewshot	system prompt, base prompt, json output, examples
fewshot step-by-step	fewshot + step-by-step instruction
fewshot instruction	fewshot + explicit instruction

Table 4: Prompt variants for FAQ reformulation

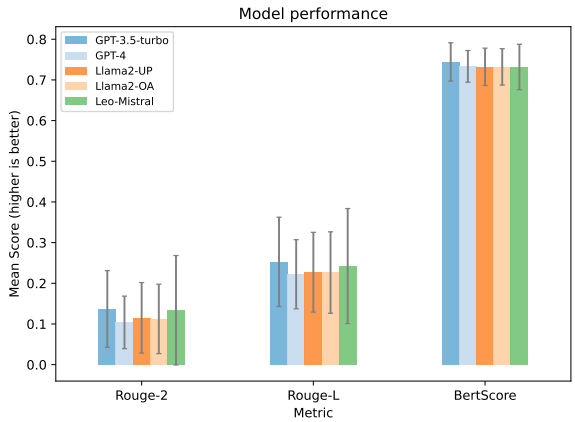
cal Report (Team, 2023)). Alternatively, *explicit instructions* can be integrated into the prompt that outline the work steps described in more detail. *Example integration* was designed to help the model to better understand the task.

Prompt variants The described prompt components are combined to create prompt variants of varying complexity, as shown in Table 4. The basic *zeroshot* prompt consists of the system prompt and a user prompt built from the base prompt and the output format instruction. The basic *fewshot* prompt consists of the system prompt and a user prompt built from the base prompt, the output format instruction and two reformulation examples. The fewshot samples are selected dynamically based on their semantic similarity to the input sample. For this sake existing samples are added to a dense search index using a BERT-like encoder model (Zhang et al., 2023). Additional prompt variants are formed by combining the basic prompts with the additional instructions or examples: *zeroshot-stepbystep*, *zeroshot-instruct*, *fewshot-stepbystep* and *fewshot-instruct*.

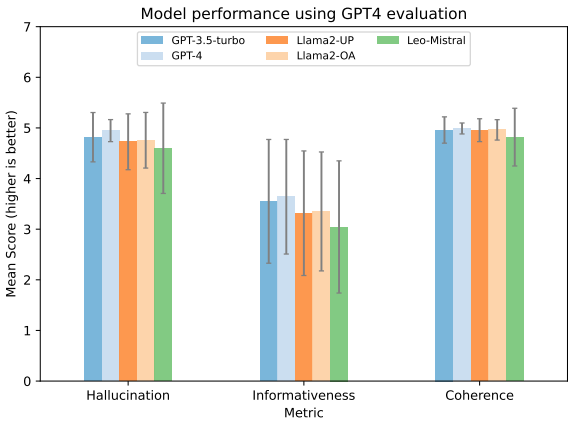
5 Automatic Text Evaluation

Evaluation with ROUGE and BERTScore First we analyzed the generated texts using ROUGE (Lin, 2004)[9], a traditional n-gram-based text similarity metric, and BERTScore (Zhang et al., 2020)[10], a metric relying on dense vector embeddings to approximate the semantic similarity between generated text and the groundtruth. Figure 1a

[9]https://huggingface.co/spaces/evaluate-metric/rouge
[10]https://huggingface.co/spaces/evaluate-metric/bertscore

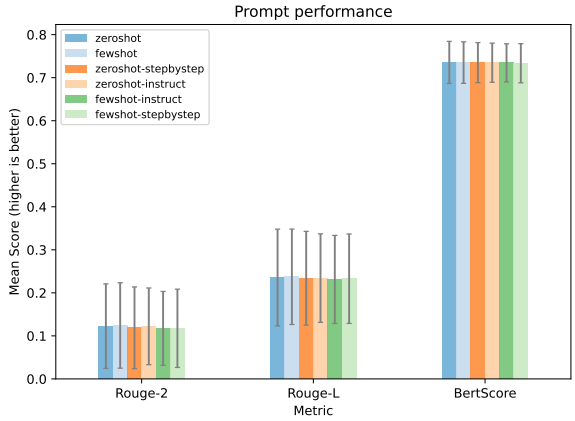


(a) Model evaluation with ROUGE and BERTScore

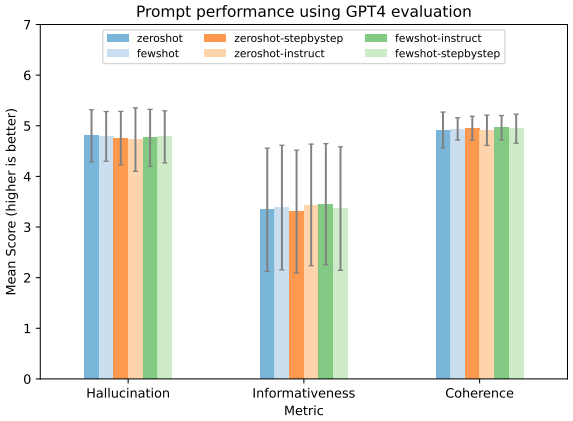


(b) Model evaluation with GPT4

Figure 1: Automatic evaluation of models performance. The error bars indicate the 95% confidence interval.



(a) Prompt variant evaluation with ROUGE and BERTScore



(b) Prompt variant evaluation with GPT4

Figure 2: Automatic evaluation of prompts performance. The error bars indicate the 95% confidence interval.

shows the performance of the models, averaged over the prompt variants. GPT-3.5-Turbo achieved slightly better values across all metrics than the other models, followed by Leo-Mistral (all metrics except for Bert_F1). However, the differences between the models are not significant, as they each fall within the 95% confidence intervals.

Also, in terms of the prompt variants, there is no clearly superior variant; all 6 variants perform roughly equally well (see Figure 2a). Therefore, it can only be said here that in terms of the automatic metrics, the precise formulation of the prompt - with or without examples, with or without instructions - did not have a major effect on the output, and roughly equally good suggestions were generated.

The small differences between the prompts could be due to the brevity of the input and generated texts, already well formulated input and also the inability of word overlap based metric to capture differences. There is still a lack of metrics to effectively measure the quality of rewriting short texts. On average, the input texts were 86 words long, and the generated FAQ texts ranged from 42 to 56 words. Leo-Mistral produced the shortest, while GPT-3.5-Turbo produced the longest question-answer pairs. However, the length of the generated texts does not correlate (Pearson correlation co-efficient $r = -0.074$ for Rouge-2, $r = -0.029$ for Rouge-L) or only weakly (Bert-F1, $r = 0.316$) with the scores achieved, so a model that generates longer texts does not necessarily perform better.

Evaluation with GPT-4 We followed the work of Wang et al. (2023) and utilized GPT-4 to score the output texts on a Likert scale of 1-5 stars using the evaluation prompts listed in Table 5. Figure 1b shows the performance of the models based on GPT-4 evaluations of the criteria *hallucinations*, *information content* and *coherence*. The highest rated

Hallucination System prompt: You are a system checking whether text B, which is a reformulation of an input text A, contains hallucinations as understood in context of text generation, i.e if text B contains information which is not supported by text A. Note, that omitting information in text B is not considered as hallucination; therefore do not lower the score if information are only omitted in text B!!!. Please score text B regarding hallucinations with one to five stars, where one star means text B contains many hallucinated information not contained in input text A and five stars mean text B contains no hallucinations when compared to input text A. I expect an answer in format: Score: "the score (e.g 3 stars)" Explanation: "hallucinated text parts or "no hallucinations" if the score is 5 stars" User prompt: Text A: {raw tex} Text B: {automatically generated reformulation} Response: Score: {rating on scale 1-5 stars} Explanation: {score explanation}
Coherence System prompt: You are a system checking whether the given text is coherent, i.e. whether the ideas, sentences, and paragraphs are logically and smoothly connected, making the text easy to understand and follow. A coherent text flows naturally and is organized in a way that allows readers or listeners to grasp the relationships between its various parts.. Please score a given text regarding coherence with one to five stars, where one star means text is very incoherent and five stars mean text has perfect coherence. I expect an answer in format: Score: the score (e.g 3 stars) Explanation: explanation of the score or "very coherent " if the score is 5 stars User prompt: Text automatically generated reformulation Response: Score: {rating on scale 1-5 stars} Explanation: {score explanation}
Informativeness System prompt: You are a system checking whether the text B contains all the information from Text A. Please score text B regarding informativeness with one to five stars, where one star means text B is much less informative then text A and five stars mean text B is as informative as text A. I expect an answer in format: Score: the score (e.g 3 stars) Explanation: explanation of the score or "very informative" if the score is 5 stars User prompt: Text A: reference text Text B: {automatically generated reformulation} Response: Score: {rating on scale 1-5 stars} Explanation: {score explanation}

Table 5: Templates of evaluation prompts fed to ChatGPT 4 as well as its responses.

model for all three criteria is GPT-4 itself,[11] although the differences are not very large (< 0.4 for hallucinations, < 0.6 for informativeness, < 0.1 for coherence). Leo-Mistral consistently achieves the worst score, and shows the highest variance. GPT-3.5-Turbo outperforms the two Llama2 models in relation to hallucinations and information content. The occurrence of hallucinations is hardly criticized at all, all models achieve relatively high scores of $>= 4.6$ here. Coherence is also judged to be very good across all generations. For the informativeness criterion, the GPT-4-based judge often assigns lower scores, with an average score of 3.4. However, one must take into account that some of the reference texts were supplemented by editors with information that was not contained in the source text, and that many of the generated texts are significantly shortened versions of the original texts due to the formulation of the task prompt. GPT-4 evaluates very strictly here, and even small errors lead to downgrades. Figure 2b shows the performance of the models by prompt type. No significant differences can be observed between the prompt variants, which is likely due to the short length of the output texts, resulting in low differentiability.

In addition, we also manually analyzed the existing hallucinations. A manual inspection of 100 hallucinations showed that the main error of GPT-4 was considering information omissions as hallucinations despite the very explicit instruction, that omissions should not lower the score. The problem was mitigated by the fact that the score of a text omitting some information was always > 4.5, i.e. the texts were scored as near perfect. The remaining hallucinations can be categorized into 4 types: 1) contradiction of the source text 2) topic shift 3) over-specification with extra information and 4) over-generalization of information. Examples of the hallucination types are given in Table 6.

6 Human Editors Evaluation of Helpfulness of FAQ-Rewriting

In addition to automated assessments, we conducted a manual evaluation of helpfulness of the

[11] Note that recent research has shown that GPT-based evaluation may prefer its own generations over that of other models (Panickssery et al., 2024).

category	input text	generated text
contradiction	Can a customer who has already booked a Netflix option on the mobile side additionally book the free option? Unfortunately, the option cannot be exchanged for a free one.	Can I book the free option in addition to the Netflix option on my mobile side? Yes, you can book the free option in addition to the Netflix option on your mobile side.
topic shift	How long is my old connection available?	What is the minimum contract term for my connection?
over-specification	The duration depends on individual circumstances and your on-site wishes	The duration depends on various factors, such as the number of customers in the region, local building regulations, and necessary installation work.
over-generalization	Do I have a special termination right if I move abroad?	Do I have a special termination right if I move?

Table 6: Hallucinations examples. The texts are English translations of original German data.

automatic FAQ-Rewriting, a study conducted by expert editors of DT.

We created a set of evaluation tuples consisting of an input text and two reformulations (a reference text and an LLM text or two texts generated by different LLMs). Due to constrained availability of professional editors, we limited our study to covering three LLM-prompt type combinations: the best two GPT models and the best open-source model, Leo-Mistral, using their respective best prompt variant according to BERTScore. The final selection includes GPT-3.5-Turbo-Zeroshot, GPT-4-Fewshot-Instruct, and Leo-Mistral-Fewshot. Given the selection, we created all possible combinations of an input text with two reformulations, including 3 model-model pairs and 3 model-human pairs. We then randomly selected 120 pairs for our study (20 for each combination), applying one constraint: for the LLM-generated texts, we considered only the best text out of three based on the BERTscore.

Seven editors of different professional experience levels were tasked with evaluating those 120 pairs of reformulation suggestions. Each editor assessed a random set of 30 pairs, with 90 pairs receiving evaluations from two annotators. The editors were prompted to address the following three questions:

1. Which reformulation of the input is superior: Version 1 or Version 2? (Please express a preference whenever possible). Response options included: Version 1, Version 2, or no preference.
2. On a scale, how much revision would be necessary for the better of the two suggestions to render an acceptable text? Answer choices ranged from: not at all, slightly, moderately, strongly, entirely.
3. Would the superior suggestion aid your work (e.g., save time)? Response options were limited to: yes or no.

Analysis of the first question revealed that when comparing a gold reference with a machine-generated text, editors favored the automatically generated suggestion in 41.9% of cases, while in 3.8% of cases, it was deemed equivalent to the gold reference. Notably, a slight preference for GPT-4 emerged when examining the distribution of models that most frequently outperformed the gold reference (see Figure 3).

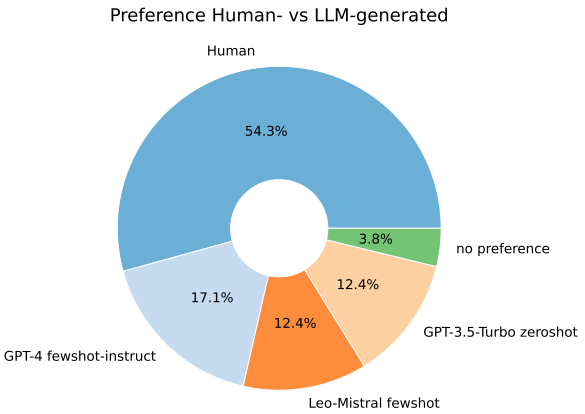


Figure 3: The analyzed preference distribution for all evaluated pairs of suggestions, where one of the suggestions was a human-written reference FAQ.

Next, we analyzed to what extent the editors rated their preferred suggestions as worthy of improvement. The editors were asked to rate the better suggestions on a scale: not at all, slightly, moderately, strongly, entirely. We mapped the ratings to numerical values from 1 (entirely) to 5 (not at

preferred model	mean score
no preference	2.39
GPT-3.5-Turbo zeroshot	3.37
GPT4 fewshot-instruct	3.63
Gold reference	3.72
Leo-Mistral fewshot	4.03

Table 7: Average results regarding the question NR. 2: On a scale, how much revision would be necessary for the better of the two suggestions to render an acceptable text? (1 = entirely, 5 = not at all)

annotator	helpful
A	100.00%
B	80.00%
C	66.67%
D	56.67%
E	46.67%
F	17.24%
G	0.00%

Table 8: Results for individual editors regarding the question NR. 3: Would the superior suggestion aid your work (e.g., save time)? Response options are 'yes' or 'no'.

all), so that a higher value reflects better quality of the texts. The results are presented in Table 7. The Leo-Mistral model received the highest overall rating in the evaluation, meaning that if the model was selected as the preferred model, the suggestion would need the least amount of modification. However, it should be noted that Leo-Mistral was the least frequently chosen as the preferred model overall. Gold references were rated with an average score of 3.72, GPT-4 with 3.63, and GPT-3.5-Turbo with 3.37. This indicates that even the gold references were often judged to be improvable. When analyzing the ratings, strong differences among the editors should be taken into account. For instance, one annotator stated, that 56.7% of the better suggestions (including automatically generated texts) do not need any reformulations while according to another annotator none of the texts were perfect, not even the gold references. We observed that the more experienced editors were much more critical of all texts.

The final question aimed to determine whether the editors perceive any advantage in using text suggestions. Overall, in 52% of all instances, a suggestion was deemed helpful for their work. When considering only instances where a machine-generated text was chosen as the better suggestion or no preference was indicated, the question was answered affirmatively in 48% of cases. It should be noted, however, that there are significant differences among individual editors: for example, one editor never found a suggestion helpful for editorial work, whereas other editor rated a suggestion as advantageous for the work process in all instances (see Table 8).

The agreement between the responses of the editors is rather weak. For example, there was agreement regarding question 1 in only 49% of cases, question 2 in 19% of cases, and question 3 in 39% of instances. We additionally measured the inter-annotator agreement using Krippendorff's alpha, first pairwise between annotators and then as the mean of these scores, obtaining overall values of $\alpha_{q1,nominal} = 0.103$, $\alpha_{q2,ordinal} = -0.252$, $\alpha_{q3,nominal} = -0,250$. The results suggest a high subjectivity of editors regarding the editorial process.

7 Conclusion

Our study explores the effectiveness of large language models in supporting the editorial process of rewriting customer help pages. We introduce a dataset containing Frequently Asked Question-Answer pairs, comprising raw drafts and their revisions by professional editors. Through various prompts tailored for the rewriting task, we evaluate the performance of four LLMs. Using ROUGE, BERTScore, and ChatGPT, we conduct automatic assessments of content and text quality. Additionally, we design an evaluation of the helpfulness of automatically generated FAQ revisions for editorial work, conducted by professional editors. Our findings demonstrate that LLMs can generate helpful FAQ reformulations for the editorial process. However, minimal performance differences were observed among LLMs for this task, and our survey on helpfulness highlights the subjective nature of editors' perspectives on editorial refinement. In our future work, we aim to explore additional editorial tasks, such as rephrasing texts to align with the editorial style guide or generating "metatexts" (teaser headlines, teaser texts, titles) for advisory articles.

Acknowledgments

We would like to thank the anonymous reviewers for their valuable feedback on our work. This work has been supported by the German Ministry of Education and Research (BMBF) as part of the project TRAILS (01IW24005).

Limitations

The work described in this paper is limited by being conducted using only a single, small dataset of question-answer pairs written by technical experts, and customer-friendly versions of these created by professional editors. Any conclusions drawn from the comparison of different models, as well as the user preference study, may not necessarily generalize to other text rewriting tasks, especially those involving more complex texts. In addition, since we relied on commercial APIs (in the case of OpenAI), it may be difficult to reproduce our results as OpenAI introduces better models and phases out the models we used in this study. While we experimented with different prompt variants, an exhaustive search for optimal prompts was not feasible, therefore, presented results may misrepresent the true task performance of each model. The GPT-based evaluation may also not reflect the true task performance, as recent research has shown that GPT-based evaluation may prefer its own generations over that of other models (Panickssery et al., 2024).

Ethical Considerations

The collected corpus is made freely available to the community. The corpus, as well as the human judgements in the preference study, were provided by professional editors of Deutsche Telekom AG, a large telecommunications company, as part of their regular task assignments. This research work aims to support editors, not to replace them. According to the vision of the company involved, the editors still need to approve and take responsibility for the content. Other than these, this study does not involve special ethical considerations. The research was conducted transparently, free from bias and in compliance with applicable laws and regulations. The use of AI models and data is intended to foster a deeper understanding of AI-generated content, with the goal of promoting responsible use and technological innovation.

References

Jane Dwivedi-Yu, Timo Schick, Zhengbao Jiang, Maria Lomeli, Patrick Lewis, Gautier Izacard, Edouard Grave, Sebastian Riedel, and Fabio Petroni. 2022. Editeval: An instruction-based benchmark for text improvements. arXiv.

Hanlei Jin, Yang Zhang, Dan Meng, Jun Wang, and Jinghua Tan. 2024. A comprehensive survey on process-oriented automatic text summarization with exploration of llm-based methods. *CoRR*, abs/2403.02901.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *ArXiv*, abs/2205.11916.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. 2023. Query rewriting in retrieval-augmented large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5303–5315, Singapore. Association for Computational Linguistics.

Clara Meister, Tiago Pimentel, Gian Wiher, and Ryan Cotterell. 2022. Locally typical sampling. *Transactions of the Association for Computational Linguistics*, 11:102–121.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. 2024. LLM Evaluators Recognize and Favor Their Own Generations.

Dongqi Pu and Vera Demberg. 2023. Chatgpt vs human-authored text: Insights into controllable text summarization and sentence style transfer. In *Annual Meeting of the Association for Computational Linguistics*.

Lei Shu, Liangchen Luo, Jayakumar Hoskere, Yun Zhu, Canoee Liu, Simon Tong, Jindong Chen, and Lei Meng. 2023. Rewritelm: An instruction-tuned large language model for text rewriting. In *AAAI Conference on Artificial Intelligence*.

Keren Tan, Kangyang Luo, Yunshi Lan, Zheng Yuan, and Jinlong Shu. 2024. An LLM-enhanced adversarial editing system for lexical simplification. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1136–1146, Torino, Italia. ELRA and ICCL.

OpenAI Team. 2023. Gpt-4 technical report.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is ChatGPT a good NLG evaluator? a preliminary study. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 1–11, Singapore. Association for Computational Linguistics.

Peitian Zhang, Shitao Xiao, Zheng Liu, Zhicheng Dou, and Jian-Yun Nie. 2023. Retrieve anything to augment large language models. *ArXiv*, abs/2310.07554.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yun Zhu, Yinxiao Liu, Felix Stahlberg, Shankar Kumar, Yu hui Chen, Liangchen Luo, Lei Shu, Renjie Liu, Jindong Chen, and Lei Meng. 2023. Towards an on-device agent for text rewriting. *ArXiv*, abs/2308.11807.