CLEAR: A Comprehensive Linguistic Evaluation of Argument Rewriting by Large Language Models

Anonymous ACL submission

Abstract

While LLMs have been extensively studied on general text generation tasks, there is less research on text rewriting, a task related to general text generation, and particularly on the behavior of models on this task. In this paper we analyze what changes LLMs make in a text rewriting setting. We focus specifically on argumentative texts and their improvement, a task named Argument Improvement (ArgImp). We present CLEAR: an evaluation pipeline consisting of 57 metrics mapped to four linguistic levels: lexical, syntactic, semantic and pragmatic. This pipeline is used to examine the qualities of LLM-rewritten arguments on a broad set of argumentation corpora and compare the behavior of different LLMs on this task and analyze the behavior of different LLMs on this task in terms of linguistic levels. By taking all four linguistic levels into consideration, we find that the models perform ArgImp by shortening the texts while simultaneously increasing average word length and merging sentences. Overall we note an increase in the persuasion and coherence dimensions.

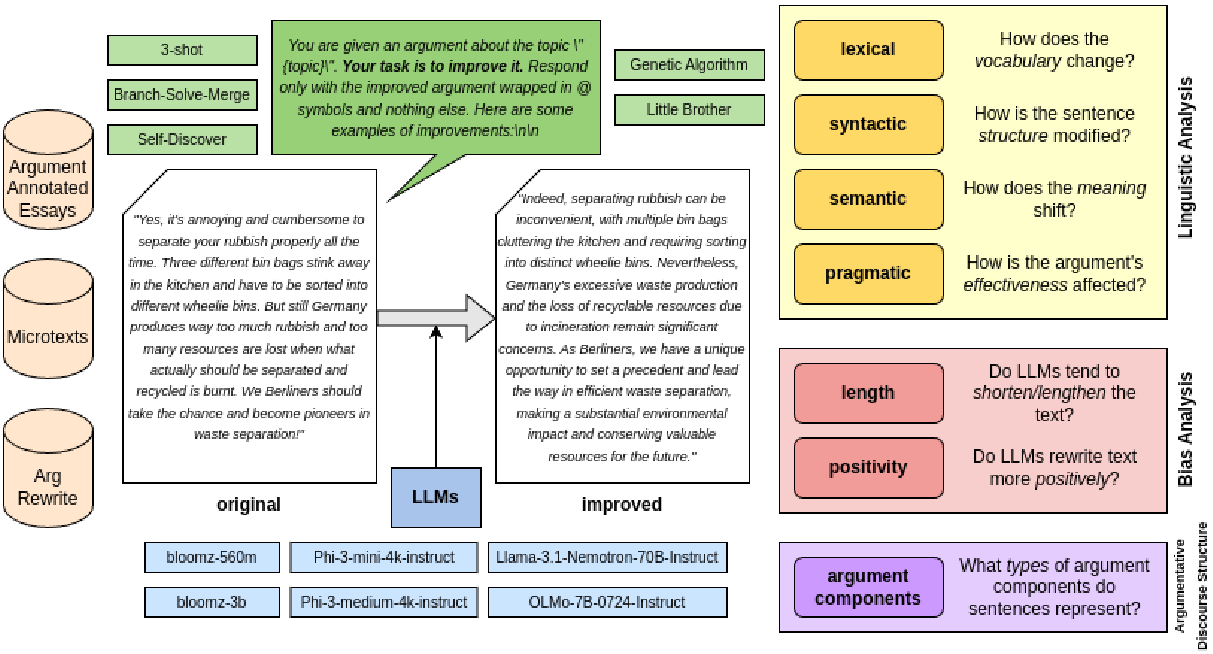
1 Introduction

Text rewriting is an important task in Natural Language Processing, with applications in style transfer (Fu et al., 2018; Hu et al., 2022; Reif et al., 2022; Riley et al., 2021), paraphrase generation (Zhou and Bhat, 2021; Li et al., 2018), and text simplification (Shardlow, 2014; Saggion and Hirst, 2017; Alva-Manchego et al., 2020), among others. It can be seen as a form of controllable text generation (Zhang et al., 2023b), where a given text is modified based on specific requirements, such as improving readability, accuracy, or suitability for a particular context (Dou et al., 2024). Recent advancements in large language models (LLMs) have shown promising performance on a wide range of text generation tasks, allowing them to refine text based on natural language instructions to produce high-quality rewrites (Shu et al., 2024).

A relevant but underexplored application of text rewriting is the task of ArgImp, i.e. rephrasing an argumentative text, with the objective of enhancing its overall quality. Arguments can be refined through various linguistic modifications, including lexical, syntactic, semantic, and pragmatic changes. LLMs have been increasingly studied in the domain of Computational Argumentation, with recent works showcasing their capabilities in the tasks of Argument Mining (Chen et al., 2024b; Abkenar et al., 2024), Argument Generation (Chen et al., 2024b; Eskandari Miandoab and Sarathy, 2024; Kao and Yen, 2024), and Argument Quality Assessment (Wachsmuth et al., 2024; Mirzakhmedova et al., 2024).

This work aims to bridge this gap by investigating the linguistic transformations performed by LLMs when prompted to improve an argumentative text. Specifically, we analyze how these models alter texts at four distinct linguistic levels: word choice *(lexical)*, sentence structure *(syntactic)*, meaning shifts *(semantic)*, and rhetorical effectiveness *(pragmatic)*. By systematically categorizing and evaluating these modifications, we aim to better understand the role of LLMs in ArgImp and their potential for enhancing argumentative writing (see Figure 1).

LLMs are known to exhibit biases in text generation settings (Oketunji et al., 2023). Due to a lack of research investigating LLMs in an ArgImp scenario, it is not clear what, if any, biases they exhibit in this setting. A so called verbosity bias can be observed in LLMs (Chen et al., 2024a; Zheng et al., 2023), as well as a positivity bias (Palmer and Spirling, 2023; Buhnila et al., 2025; Markowitz et al., 2024). These biases are of particular interest in an ArgImp setting. The models may consider longer texts as better, and thus make less changes, or shift the tone and inadvertently change the meaning of



You are given an argument about the topic \" {topic}\". **Your task is to improve it.** Respond only with the improved argument wrapped in @ symbols and nothing else. Here are some examples of improvements:\n\n

"Yes, it's annoying and cumbersome to separate your rubbish properly all the time. Three different bin bags stink away in the kitchen and have to be sorted into different wheelie bins. But still Germany produces way too much rubbish and too many resources are lost when what actually should be separated and recycled is burnt. We Berliners should take the chance and become pioneers in waste separation!"

"Indeed, separating rubbish can be inconvenient, with multiple bin bags cluttering the kitchen and requiring sorting into distinct wheelie bins. Nevertheless, Germany's excessive waste production and the loss of recyclable resources due to incineration remain significant concerns. As Berliners, we have a unique opportunity to set a precedent and lead the way in efficient waste separation, making a substantial environmental impact and conserving valuable resources for the future."

Argument Annotated Essays
Microtexts
Arg Rewrite
3-shot
Branch-Solve-Merge
Self-Discover
Genetic Algorithm
Little Brother
original
LLMs
improved
bloomz-560m
bloomz-3b
Phi-3-mini-4k-instruct
Phi-3-medium-4k-instruct
Llama-3.1-Nemotron-70B-Instruct
OLMo-7B-0724-Instruct
lexical
How does the vocabulary change?
syntactic
How is the sentence structure modified?
semantic
How does the meaning shift?
pragmatic
How is the argument's effectiveness affected?
Linguistic Analysis
length
Do LLMs tend to shorten/lengthen the text?
positivity
Do LLMs rewrite text more positively?
Bias Analysis
argument components
What types of argument components do sentences represent?
Argumentative Discourse Structure

Figure 1: Overview of our experimental setup for the task of ArgImp. We evaluate the quality of argumentative texts rewritten by LLMs prompted for improvement. We apply six models across five datasets (each revision of the ArgRewrite corpus is treated as a distinct dataset). The evaluation spans four linguistic levels, examines two types of biases, and compares the argumentative discourse structure of the original and improved texts.

the original argument. For this reason we include an investigation into length and positivity biases.

To investigate the behavior of LLMs in an ArgImp setting we have created an evaluation pipeline consisting of 57 metrics commonly used in natural language generation (NLG), named CLEAR[1]. These include scores that measure lexical, syntactic, semantic and pragmatic aspects of the texts. The focus of our work is on analyzing what changes the models make when used in an ArgImp setting, but the pipeline and approach are applicable to other text generation tasks as well. We applied five different prompting techniques to make LLMs write improved versions of arguments from the Microtexts (Peldszus and Stede, 2015) (both English and German), Argument Annotated Essays 2.0 (Stab and Gurevych, 2017) and ArgRewrite V.2 (Kashefi et al., 2022) corpora. To assess the effectiveness of these revisions, we evaluate the linguistic quality of the rewritten argumentative texts both quantitatively and qualitatively.

Our contributions are as follows: (i) a comprehensive pipeline for evaluating the output quality of text rewriting tasks, consisting of 57 different metrics[2]; (ii) a mapping of existing text generation metrics to four linguistic levels (Section 4); (iii) a measure of what transformations are performed in a reference-based text generation setting as well as well as a measure of what grammatical changes are made, both of which are part of the text generation pipeline; (iv) an analysis of LLM behavior on four different linguistic levels for the task of ArgImp (Section 5); and (v) an investigation of LLM biases in an ArgImp setting (Section 5.3/5.4).

2 Related Work

The capabilities of LLMs in the field of Computational Argumentation have been previously explored, particularly in the areas of Argument Mining (Chen et al., 2024b; Abkenar et al., 2024) and Argument Quality Assessment (Wachsmuth et al., 2024; Mirzakhmedova et al., 2024). Recent work has also made use of LLMs to generate and rephrase arguments and their components. For instance, Wang et al. (2025) and Skitalinskaya et al. (2023) have used LLMs in the context of claim optimization. Moreover, Ziegenbein et al. (2024) present a reinforcement learning-based approach for rewriting inappropriate argumentation in online discussions. With the objective of generating complete and balanced arguments, Hu et al. (2025)

[1] Comprehensive Linguistic Evaluation for Argument Rewriting

[2] To support transparency and reproducibility, the code will be released publicly upon acceptance.

use LLM agents to simulate a discussion among them and consolidate it into diverse and holistic arguments. Furthermore, Hu et al. (2024) introduce AMERICANO, a framework with agent interaction for argument generation. It incorporates an argument refinement module that evaluates and improves argument drafts based on feedback regarding their quality. El Baff et al. (2024) make use of LLMs to rewrite existing arguments to make them more appealing to readers of a certain political ideology.

3 Argument Improvement with LLMs

We aim to evaluate the quality of argumentative texts improved by LLMs. Argumentation occurs in various contexts; our work centers on the following setting: (i) We focus on *global argumentation* rather than local arguments. (ii) Our analysis is limited to *monological texts*, excluding dialogical debates. (iii) We primarily assess *intrinsic, i.e. text-focused, quality* rather than extrinsic reader-focused text effectiveness (Schriver, 1989). With our analysis we aim to answer the following research questions: (i) What changes on linguistic levels do LLMs make in an ArgImp setting? (ii) What biases do LLMs exhibit in an ArgImp setting? (iii) Do models of different sizes behave differently in an ArgImp setting?

3.1 Model Selection

We selected models of different families and sizes to provide a broad overview[3]. For our experiment we used bloomz-560m and bloomz-3b (Muennighoff et al., 2022), Phi-3-mini-4k-instruct and Phi-3-medium-4k-instruct (Abdin et al., 2024), OLMo-7B-0724-Instruct (Groeneveld et al., 2024) and Llama-3.1-Nemotron-70B-Instruct (Wang et al., 2024) (henceforth referred to as Llama 3.1).

3.2 Datasets

Our aim is to present results on a diverse set of datasets representing different settings. We use the well-known Argument Annotated Essays 2.0 corpus by Stab and Gurevych (2017), which consists of student-written essays, the Microtexts corpus (Peldszus and Stede, 2015) which consists of very short argumentative texts in English and German, and the ArgRewrite V.2 corpus by Kashefi et al. (2022), consisting of three revisions of essays by students.

Dataset metrics describing their core properties can be found in Table 4 in Appendix A.

3.3 Prompting Techniques

We use 3-shot prompting with demonstrations from the Argument Revision Corpus, Branch-Solve-Merge Saha et al. (2024), Self-Discover Zhou et al. (2024), Genetic Algorithm prompting as per Guo et al. (2024) and our own technique, called Little Brother, where the model is asked to give correcting feedback to an answer produced by its 'little brother'. The prompts used are included in Appendix B.

4 Evaluation Setup

4.1 Linguistic Analysis

We employ a wide range of NLG evaluation metrics (Schmidtova et al., 2024). Our selection aims to cover a broad spectrum of linguistic aspects to enable a comprehensive analysis of the modifications introduced by the models in our improvement setting. Following Akmajian et al. (2010), we manually mapped the metrics to their corresponding linguistic levels: 14 lexical, 22 semantic[4], 15 syntactic, 2 pragmatic as well as 4 argument components.

Lexical Analysis We analyze changes on the word level as well as word distribution. We use metrics such as the number of n-syllable words and readability scores. Our aim is to provide insight into how the vocabulary the models use changes in comparison to the original texts.

Syntactic Analysis To investigate structural modifications, we analyze the syntax of the sentences using dependency parse tags generated by spaCy (Honnibal et al., 2020). In that way, we aim to identify patterns in complex sentence structures and determine which types of complex clausal constructions are used more or less frequently in the improved versions of the argumentative texts. Moreover, we make use of BERTAlign (Liu and Zhu, 2022), a sentence alignment method originally developed for the task of machine translation. It is designed to align comparable sentences from source and target languages. We use these as part of a new text generation score. We align sentences from the

[3] All models are from the HuggingFace repository.

[4] The German sentiment scores are split into three probabilities. We consider them to be one score.

original texts with their corresponding improved versions. This allows us to categorize sentence transformations into several types and count their number: (i) *rephrase* and *copy* (1:1); (ii) *split* of an original sentence (1:m); (iii) *merge* of original sentences (n:1); (iv) *fusion* of original and improved sentences (n:m, where n and $m > 1$); (v) *deletion* of an original sentence (1:0), and; (vi) *addition* of a sentence in the improved text (0:1). This reveals what kind of modifications the models make as part of the text generation process.

Semantic Analysis To capture changes in meaning, we include a sentiment classifier, GRUEN score metrics (Zhu and Bhat, 2020), an automated metric based on grammaticality, non-redundancy, focus, structure and coherence of texts, and a discourse analysis based on Rhetorical Structure Theory (RST) (Mann and Thompson, 1988). We applied the parser from Feng and Hirst (2014). We decided against using the more recent approach by Maekawa et al. (2024) due to the significantly higher computational cost and only marginal performance gains[5]. We aim to capture both changes in the general tone and nuanced shifts in meaning resulting from LLMs' improvement.

Pragmatic Analysis We adopt the approach by Hu et al. (2024) to evaluate the texts' persuasiveness and coherence as key aspects of pragmatics. The prompts we used are based on their approach and can be found in Appendix B. The models are prompted to rate the texts with a focus on the two metrics. These metrics allow us to assess whether the improvements were successful or not, considering not only the individual changes but also the overall context of the argumentative texts. In that way, we measure the effectiveness of the communication in terms of both the texts' ability to persuade and their internal coherence within the given context.

4.2 Bias Analysis

It has been discussed that LLMs have both a length[6] (Chen et al., 2024a; Zheng et al., 2023) and a positivity bias (Palmer and Spirling, 2023; Buhnila et al., 2025; Markowitz et al., 2024). Length bias in this context refers to the LLM preferring longer texts. Positivity bias refers to the observation that LLM generated text tends to have a more positive tone than human-written texts. Verbosity bias is a relevant factor in our setting as the models may consider texts of certain lengths to be of higher quality, and perform less changes to improve them, regardless of the actual quality. Positivity bias may cause the models to shift the tone of the argument, and could change the meaning of the argument as a whole, i.e. shifting from arguing against a topic to arguing in favor. We investigate the presence of these biases by correlating the magnitude of changes made with the change in length as well as the sentiment of the original text.

4.3 Analysis of the Argumentative Discourse Structure

We classify each sentence into one of the following types of argument components: claim, premise, major claim, or none, to assess structural modifications. For the English datasets, we make use of an implementation of the best-performing approach proposed in Stab and Gurevych (2014) which achieves an accuracy of 0.77. Sazid and Mercer (2022) propose a more recent approach using deep learning models but do not significantly outperform Stab and Gurevych (2014). We are not aware of any more recent approaches or available implementations as an alternative. To make the CLEAR pipeline accessible to a wider range of users we use the approach by Stab and Gurevych (2014), acknowledging that higher performance may be possible by implementing a novel approach using LLMs.

For the German Microtexts, we apply the same classification, trained on the corpus introduced by Wambsganss et al. (2020b), achieving an accuracy of 0.65 as reported by Wambsganss et al. (2020a).

5 Results

Due to the large number of possible analyses[7] we focus on the most relevant combination. The most commonly used approach for LLMs is either zero-shot or few-shot prompting, with few-shot generally performing better (Brown et al., 2020). Based on public benchmarks, such as Chatbot Arena (Chiang et al., 2024), Llama 3.1 is the best performing LLM among our selection. We use the combination of both Llama 3.1 as well as the 3-shot prompting approach for a deeper analysis. Both Bloomz models generated very short, barely legible texts.

[5] 60.0 F1 for relation classification in Maekawa et al. (2024) vs. 58.2 Accuracy in Feng and Hirst (2014)

[6] Also referred to as 'verbosity bias'.

[7] Six models, five datasets (each revision of the ArgRewrite corpus is treated as its own dataset), five prompting techniques and four linguistic levels for a total of $6 * 5 * 5 * 4 = 600$.

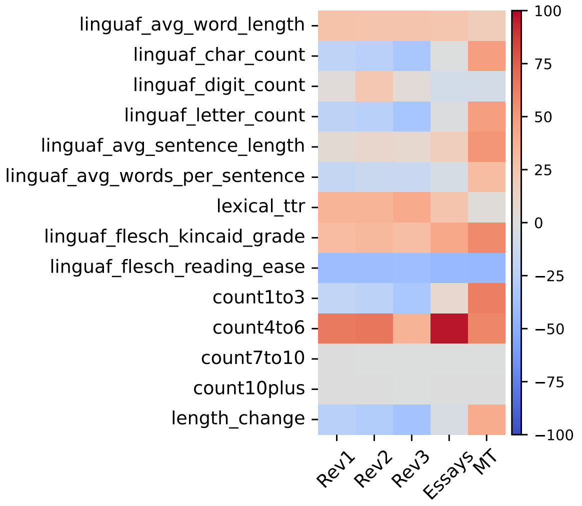


Figure 2: Changes on the lexical level for Llama 3.1. The 'count' rows refer to n-syllable words. Change is measured in %.

We omit them from the analysis. We perform a broad analysis across all other models by analyzing the scores, and further include detailed results of a manual analysis on a sample of 10 texts per dataset.

5.1 Linguistic Analysis

The findings are based on our proposed CLEAR pipeline. Unless otherwise stated the analysis is based on the scores of all models. Where individual models behaved differently we explicitly note this.

LLMs shorten the arguments. On the lexical level we note that models, on average, decreased text length (≈ 4.66% to 37.39% decrease). The exception to this is the Microtext corpus, where length increased (≈ 40.18% increase). It seems that the models are aiming to add details to improve the overall quality here, as the corpus consists of very short texts.

LLMs increase word length but decrease sentence length. Analysis on the lexical level further reveals that the models increase word lengths. We observe, particularly in the case of Llama 3.1, an increase in the number of 4 to 6 syllable words[8], and a decrease in shorter words.

LLMs reduce the reading ease. Larger models decreased the reading ease metrics, whereas the smaller ones increased it. This increase is not linear with the number of parameters of the models. The reason could be the increased linguistic capabilities of the larger models which make the language more complex. Manual analysis did not reveal any specific patterns that could explain this.

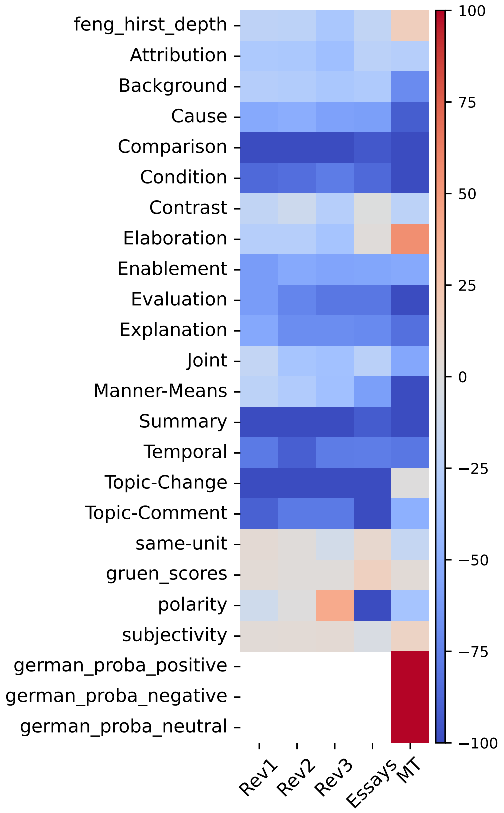


Figure 3: Changes on the semantic level for Llama 3.1. Change is measured in %.

LLMs transform the existing text. On the syntactic level the models perform operations that modify existing parts of the text (1). The models rarely add entirely new sentences or paragraphs, as well as seldom entirely delete what is there.

	Rev1	Rev2	Rev3	Essays	MT
add	0.47	0.44	0.30	0.36	0.05
copy	1.86	1.35	1.79	2.38	**2.00**
delete	0.31	0.38	0.74	0.18	0.03
fusion	1.02	0.78	0.81	0.72	0.26
merge	**5.51**	**6.76**	**7.78**	**3.11**	0.47
other	0.00	0.00	0.00	0.00	0.00

Table 1: Average number of sentence transformations performed for the Llama 3.1 model. Most common transformation per dataset in bold.

Llama 3.1 increases the number of coordinating noun phrases.[9] This is consistent across all the datasets. The Phi-3 models increase their number slightly, whereas OLMo consistently decreases it.

Llama 3.1 significantly increases the number of appositional modifiers[10]***.*** These are commonly

[8] In Figure 2 this is labeled as 'count4to6'.

[9] Labelled as 'num_coordNP' in Figure 4.

[10] Example: 'The largest model, Llama 3.1, performs best.'.

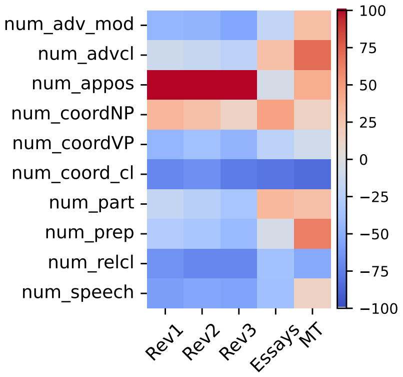


Figure 4: Changes on the syntactic level for Llama 3.1. Change is measured in %.

used to add additional details or information to other nouns or noun phrases.

LLMs consistently decrease the depth of the RST parse tree. The analysis on the semantic level (Figure 3) reveals that the rhetoric structure decreases consistently across the longer corpora. A shallow RST tree indicates that the texts are less complex and easier to understand. The Microtext corpus is the exception here.

LLMs make the tone more negative. All models perform similarly in terms of sentiment changes. On the German Microtexts the sentiment changes strongly to positive, whereas for all English texts the polarity decreases. This means the models make the texts *more negative*, but not necessarily negative *over all*. For Llama 3.1 we note an outlier for the polarity score on the Essays dataset. Without it, the average change in polarity is -11%. The value for one human-written text is almost, but not quite, zero. Table 2 shows the polarity scores of the original texts.

					MT (DE)		
Rev1	Rev2	Rev3	Essays	MT (EN)	Neg	Neutral	Pos
0.11	0.11	0.11	0.16	0.08	0.19	0.77	0.04

Table 2: Sentiment scores of the original human-written texts in the corpora. German scores are probabilities. English ranges from -1 (negative) to +1 (positive).

LLMs make the arguments more coherent and persuasive. On the pragmatic level (Figure 5) we note an increase for both persuasiveness and coherence for all models on all datasets, except for OLMo on Revision 1 of the Revisions and Microtexts datasets, where there was a small decrease for persuasion (≈ -2.4 and ≈ -1.8, respectively). Based on this increase in score we can say that overall the improvement process was a success. To confirm this, we have performed a manual analysis.

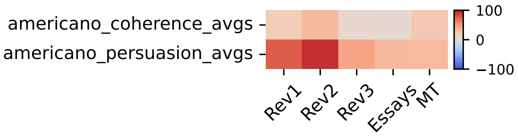


Figure 5: Changes on the pragmatic level for Llama 3.1. Change is measured in %.

5.2 Manual Analysis

The findings in this section are based on a manual analysis of the generated texts for each dataset. We use the texts generated by Llama 3.1 for the analysis.

5.2.1 Analysis of Changes

For this analysis we used 10 randomly sampled texts from each dataset. We manually compared the improved text to the original and investigated notable changes in the texts.

Llama 3.1 mimics the style of the original text. We noted that if the original text had stylistic peculiarities, the model copied them. In some texts the authors include references. The model hallucinated further references that were not present in the original argument then. None of the models we used make use of Retrieval Augmented Generation (RAG), which could cause this.

Llama 3.1 makes use of bullet points. In some cases the improved argument includes bullet points. This occurred only when the structure of the original argument lends itself well to this. Individual arguments were not well connected, such as paragraphs discussing individual claims, but not connecting to previous and following claims in neighboring paragraphs.

Llama 3.1 does not appear to check for logical quality of arguments. In the Revisions corpus there is one text that discusses that self-driving cars can get confused by GPS trackers and drive down stairs. This is not well explained in the original texts, in any revision, but Llama 3.1 did not elaborate on it. With current technologies it is doubtful that this is a relevant factor. An improved version of the argument could either elaborate on this point or outright remove it, neither of which is something the model did.

Here 'Llama 3.1' is in apposition to 'model'. Labelled as 'num_appos' in Figure 4.

Llama 3.1 refines the existing structure. Often the original argument had an implicit structure in terms of paragraphs. In many cases, especially on the Essays corpus, the authors focused on one claim per paragraph. Llama 3.1 kept this structure, and often added explicit headlines to each paragraph, to illustrate what claim the paragraph is about.

5.2.2 Analysis of Preference

For this experiment we excluded the texts used in the analysis of the changes. We randomly sampled 20 texts from each dataset. Two reviewer with a background in computer science then blindly rated the original versus the improved argument. The order of the texts was shuffled, i.e. text 1 was not always the improved argument.

On average, the improved text was preferred 79% of the time. Due to the small sample size of each dataset, we measured the agreement in percentage. Across all datasets, the reviewers agreed on 65.83% of the texts.

5.3 Length Bias

For the correlation, we use Pearson's standard correlation coefficient. Scores are based on an analysis of Llama 3.1, OLMo and both Phi-3 models for the 3-shot prompts across all datasets. Correlations are between the original length and the delta of the metrics (original vs. improved text). We aim to analyze whether models behave differently on texts of different lengths. Strong correlations imply that the models behave differently with varying input lengths. Where p-values are omitted in the text, they are < 0.001.

Lexical We note a positive correlation with average word length (≈ 0.19) and a strong negative correlation for the 1 to 3 syllable word count (≈ -0.49), sentence length (≈ -0.21) and average words per sentence (≈ -0.29). The results also show a strong correlation for the token-to-type ratio (≈ 0.46). We find no correlation in the 7 to 10 syllable word counts and only a weak one for 10 plus counts (≈ -0.07). There is a positive correlation with the improved length (≈ 0.84). The mean length decreased from ≈ 2116 to ≈ 1779 characters.

Syntactic Merge has a strong correlation (≈ 0.64) as well as delete (≈ 0.27). Both add (≈ 0.04, $p \approx 0.009$) and fusion (≈ 0.05, $p \approx 0.008$) have only a weak correlation.

Semantic We note no interesting correlations.

Pragmatic There is no correlation (≈ 0.013, $p \approx 0.02$) between the length of the argumentative texts and the persuasion scores or length and coherence (≈ -0.04, $p \approx 0.45$).

Individual model behavior All the individual models behave similarly on the syntactic level: merge has a strong positive correlation with length ($\approx 0.55 - 0.75$ for all models), as well as delete ($\approx 0.20 - 0.32$ for all models). As these are reference-based metrics, this suggests that as text length increases, so does the number of these operations. In terms of the pragmatic quality dimensions of coherence and persuasion, only Llama 3.1 and Phi-3-mini show a correlation. For Llama 3.1 we observe a correlation between length and persuasion of ≈ 0.18 while Phi-3-mini has a negative one with about ≈ 0.20.

Summary Our findings suggest that the texts become overall simplified, and shorter, but this is accomplished by using longer words. There does not appear to be a direct preference influenced by text length.

5.4 Positivity Bias

We looked at the magnitude of shifts in sentiment, specifically polarity, for the Llama 3.1 model and the few-shot approach on all datasets. We measure the strength of the sentiment shifts as follows:

$$\text{shift percentage} = \left(\frac{\Delta}{|\text{Polarity Human}|}\right) * 100 \qquad (1)$$

We find that 335 negative shifts (46.16%), 203 neutral shifts (26.40%) and 211 positive shifts (27.44%) occur. We consider a shift of above $+20\%$ positive, below -20% negative and between neutral. The mean is quite high with a value of 628.55%, but the median is negative with a value of -14.59%. The mean polarity in the original texts is $+13.18\%$ and that of the improved texts is $+11.39\%$. This indicates that while positive changes are done rarely, they are strong in magnitude when they occur. Overall, the model tends to move the improved texts towards a more neutral tone.

5.5 Argument Component Classification

We present the changes in number of argument components in Table 3. Components are identified on a sentence level. We note a large decrease in both non-argumentative components, as well

	Rev1	Rev2	Rev3	Essays	MT
MajorClaim	-0.53	-0.50	-0.41	0.26	-0.11
Claim	0.06	-0.16	-0.14	-0.42	1.13
Premise	-4.88	-6.55	-8.90	-2.54	-0.11
None	-1.65	-2.01	-3.26	-0.12	-0.89

Table 3: Changes in number of argument components

as premises. We observe an increase in sentence length, as well as an overall merging of sentences. Due to the texts becoming shorter there can be less argument components. Despite this, we observe large decreases for the non-argumentative components, which indicates that the texts become more focused. We further hypothesize that the claims and premises are merged, as suggested by the behavior on the syntactic level, which leads to the strong decrease in the number of premises.

6 Discussion

The analysis based on the scores of our text generation evaluation pipeline shows that on the lexical level overall text length decreases. We further observe an increase in 4 to 6 syllable words and a strong decrease in shorter words. On the syntactic level we note many merge and fuse operations, which means that the original text is shortened or remixed into existing sentences. Then, on the semantic level, we note a decrease in the depth of the RST parse trees. Finally, on the pragmatic level, we observe an increase in terms of coherence and persuasion, which indicates that the argument quality, in general, improved. Manual analysis supports these empirical findings. We note that the models keep the overall structure, where it exists, and do not delete or add significant chunks of text. Instead, the models *refine* and *enhance* what is already there. Notably, manual analysis revealed weak parts of certain arguments, which the model did not address or remove. These results together suggest that the models perform the improvement by focusing the texts:

- *Lexical level*: Overall text length decreases, longer words are more common, resulting in shorter sentences composed of longer words.
- *Syntactic level*: Original sentences are merged, leading to shorter texts with more focused sentences.
- *Semantic level*: Depth of the RST trees decreases, which indicates simpler texts.
- *Pragmatic level*: Argumentative quality of the texts improves, which suggests that the models' modifications are generally effective and do not compromise the integrity of the original texts.
- *Manual Analysis*: Model refines the existing text, does not significantly perform changes in terms of semantics.

In summary, it appears as though the models eliminate fluff and make the text more efficient. This is supported by our analysis of both the length and sentiment bias. To investigate the length bias we considered the token-to-type ratio as well as the lengths of the texts and sentences. The sentiment bias analyses revealed that the text shifts are towards the negative, but the original texts were positive in sentiment on average, and the improved texts are still positive, but more neutral.

7 Conclusion

By making use of the CLEAR pipeline, consisting of commonly used text generation metrics mapped to linguistic levels and performing an analysis on the individuals levels, we have found that LLMs make the texts more focused in an ArgImp setting, in the sense that (i) the texts become shorter, (ii) the average length of words increases, (iii) semantically the texts do not change. Our results suggest that the models perform well for text improvement. We note two positive factors: (i) the length of the texts decreases, but notably not in the case of the Microtexts corpus, where the input texts are already quite short, and (ii) the quality increases. We note small differences in model behavior in this task. Larger models performed better in both quality of the output texts and appear to make the texts more focused than the small models. A positivity bias could not be identified, instead the models appear to aim to make the texts more neutral, instead of shifting the tone consistently to positive or negative levels. Lastly, we could not identify a length bias. The models do not appear to prefer texts of certain lengths. We note the tendency of Llama 3.1 in particular to use longer words, which could be a form of bias. Our results suggest that this is done to increase information density without negatively impacting readability as evident by the scores on the lexical level of our analysis.

8 Limitations

Our analysis focuses on textual characteristics and linguistic qualities, while disregarding more pronounced content-based aspects, overall argument quality, and reader-focused effectiveness. In particular, we do not incorporate user studies to evaluate the perceived impact of the improvements.

In the context of Automatic Essay Scoring (AES), a wide range of essay traits is typically assessed, including content, organization, word choice, sentence fluency, conventions, prompt adherence, language, narrativity, style, and voice (Kumar et al., 2022; Do et al., 2023; Ridley et al., 2021). However, our study is limited to a narrow subset of these traits, namely text-focused linguistic qualities. Higher-order traits such as prompt adherence, content and overall organization require a more complex evaluation incorporating a detailed discourse analysis and external knowledge, which is beyond the scope of this work. By focusing on linguistic qualities, we establish a baseline for future work that may easily extend our approach to include higher-order cognitive aspects of essay quality.

Furthermore, our evaluation does not incorporate detailed argument quality assessments grounded in argumentation theory (Van Eemeren et al., 2013; Walton, 2009; Mercier and Sperber, 2011). In particular, we do not account for argument quality aspects as defined by taxonomies such as the one proposed by Wachsmuth et al. (2017), which extend beyond linguistic structure to include criteria such as logical soundness or dialectical reasonableness. A recent survey by Ivanova et al. (2024) shows that there is no consensus regarding the different quality aspects of arguments. Varying contexts and settings make use of different metrics. Due to the large number of existing argumentation datasets and settings in which argumentation occurs, it is not feasible to evaluate all possible metrics. This is further hindered by the fact that a majority of the metrics are not automated, lack publicly available models to score outputs automatically, do not have a sufficient amount of annotated data for model training available, or the datasets not being publicly available to begin with.

Finally, we rely largely on automatic scoring for the evaluation due to the extensive scale of our experiments. Our analysis involves five distinct datasets, six models, and five prompting techniques, each applied across four linguistic levels using 57 different metrics. This results in a total of $5 * 6 * 5 * 4 * 57 = 34'200$ combinations, thus making a manual evaluation for all combinations impractical. We included a manual evaluation on a small subset.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Mohammad Yeghaneh Abkenar, Weixing Wang, Hendrik Graupner, and Manfred Stede. 2024. Assessing open-source large language models on argumentation mining subtasks. *Preprint*, arXiv:2411.05639.

Adrian Akmajian, Richard A. Demers, Ann K. Farmer, and Robert M. Harnish. 2010. *Linguistics: An Introduction to Language and Communication*, 6th edition. The MIT Press, Cambridge, MA.

Fernando Alva-Manchego, Louis Martin, Antoine Bordes, Carolina Scarton, Benoît Sagot, and Lucia Specia. 2020. ASSET: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4668–4679, Online. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Ioana Buhnila, Georgeta Cislaru, and Amalia Todirascu. 2025. Chain-of-MetaWriting: Linguistic and textual analysis of how small language models write young students texts. In *Proceedings of the First Workshop on Writing Aids at the Crossroads of AI, Cognitive Science and NLP (WRAICOGS 2025)*, pages 1–15, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024a. Humans or LLMs as the judge? a study on judgement bias. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8301–8327, Miami, Florida, USA. Association for Computational Linguistics.

Guizhen Chen, Liying Cheng, Anh Tuan Luu, and Lidong Bing. 2024b. Exploring the potential of large language models in computational argumentation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2309–2330, Bangkok, Thailand. Association for Computational Linguistics.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating llms by human preference. *Preprint*, arXiv:2403.04132.

Heejin Do, Yunsu Kim, and Gary Geunbae Lee. 2023. Prompt- and trait relation-aware cross-prompt essay trait scoring. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1538–1551, Toronto, Canada. Association for Computational Linguistics.

Yao Dou, Philippe Laban, Claire Gardent, and Wei Xu. 2024. Automatic and human-AI interactive text generation (with a focus on text simplification and revision). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 5: Tutorial Abstracts)*, pages 3–4, Bangkok, Thailand. Association for Computational Linguistics.

Roxanne El Baff, Khalid Al Khatib, Milad Alshomary, Kai Konen, Benno Stein, and Henning Wachsmuth. 2024. Improving argument effectiveness across ideologies using instruction-tuned large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4604–4622, Miami, Florida, USA. Association for Computational Linguistics.

Kaveh Eskandari Miandoab and Vasanth Sarathy. 2024. “let‘s argue both sides”: Argument generation can force small models to utilize previously inaccessible reasoning capabilities. In *Proceedings of the 1st Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (CustomNLP4U)*, pages 269–283, Miami, Florida, USA. Association for Computational Linguistics.

Vanessa Wei Feng and Graeme Hirst. 2014. A linear-time bottom-up discourse parser with constraints and post-editing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 511–521, Baltimore, Maryland. Association for Computational Linguistics.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, et al. 2024. Olmo: Accelerating the science of language models. *arXiv preprint arXiv:2402.00838*.

Qingyan Guo, Rui Wang, Junliang Guo, Bei Li, Kaitao Song, Xu Tan, Guoqing Liu, Jiang Bian, and Yujiu Yang. 2024. Connecting large language models with evolutionary algorithms yields powerful prompt optimizers. In *The Twelfth International Conference on Learning Representations*.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spacy: Industrial-strength natural language processing in python. *Zenodo*.

Zhe Hu, Hou Pong Chan, Jing Li, and Yu Yin. 2025. Debate-to-write: A persona-driven multi-agent framework for diverse argument generation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4689–4703, Abu Dhabi, UAE. Association for Computational Linguistics.

Zhe Hu, Hou Pong Chan, and Yu Yin. 2024. AMERICANO: Argument generation with discourse-driven decomposition and agent interaction. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 82–102, Tokyo, Japan. Association for Computational Linguistics.

Zhiqiang Hu, Roy Ka-Wei Lee, Charu C. Aggarwal, and Aston Zhang. 2022. Text style transfer: A review and experimental evaluation. *SIGKDD Explor. Newsl.*, 24(1):14–45.

Rositsa V Ivanova, Thomas Huber, and Christina Niklaus. 2024. Let‘s discuss! quality dimensions and annotated datasets for computational argument quality assessment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20749–20779, Miami, Florida, USA. Association for Computational Linguistics.

Wei-Yu Kao and An-Zi Yen. 2024. MAGIC: Multi-argument generation with self-refinement for domain generalization in automatic fact-checking. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10891–10902, Torino, Italia. ELRA and ICCL.

Omid Kashefi, Tazin Afrin, Meghan Dale, Christopher Olshefski, Amanda Godley, Diane Litman, and Rebecca Hwa. 2022. Argrewrite v. 2: an annotated argumentative revisions corpus. *Language Resources and Evaluation*, pages 1–35.

Rahul Kumar, Sandeep Mathias, Sriparna Saha, and Pushpak Bhattacharyya. 2022. Many hands make light work: Using essay traits to automatically score essays. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1485–1495, Seattle, United States. Association for Computational Linguistics.

Zichao Li, Xin Jiang, Lifeng Shang, and Hang Li. 2018. Paraphrase generation with deep reinforcement learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3865–3878, Brussels, Belgium. Association for Computational Linguistics.

Lei Liu and Min Zhu. 2022. Bertalign: Improved word embedding-based sentence alignment for chinese–english parallel corpora of literary texts. *Digital Scholarship in the Humanities*, 38(2):621–634.

Aru Maekawa, Tsutomu Hirao, Hidetaka Kamigaito, and Manabu Okumura. 2024. Can we obtain significant success in RST discourse parsing by using large language models? In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2803–2815, St. Julian's, Malta. Association for Computational Linguistics.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

David M Markowitz, Jeffrey T Hancock, and Jeremy N Bailenson. 2024. Linguistic markers of inherently false ai communication and intentionally false human communication: Evidence from hotel reviews. *Journal of Language and Social Psychology*, 43(1):63–82.

Hugo Mercier and Dan Sperber. 2011. Why do humans reason? arguments for an argumentative theory. *Behavioral and brain sciences*, 34(2):57–74.

Nailia Mirzakhmedova, Marcel Gohsen, Chia Hao Chang, and Benno Stein. 2024. Are large language models reliable argument quality annotators? In *Robust Argumentation Machines*, pages 129–146, Cham. Springer Nature Switzerland.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

AF Oketunji, M Anas, and D Saina. 2023. Large language model (llm) bias index—llmbi. *Data & Policy*.

Alexis Palmer and Arthur Spirling. 2023. Large language models can argue in convincing ways about politics, but humans dislike ai authors: implications for governance. *Political Science*, 75(3):281–291.

Andreas Peldszus and Manfred Stede. 2015. An annotated corpus of argumentative microtexts. In *Argumentation and Reasoned Action: Proceedings of the 1st European Conference on Argumentation, Lisbon*, volume 2, pages 801–815.

Keqin Peng, Liang Ding, Yancheng Yuan, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2024. Revisiting demonstration selection strategies in in-context learning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9090–9101, Bangkok, Thailand. Association for Computational Linguistics.

Emily Reif, Daphne Ippolito, Ann Yuan, Andy Coenen, Chris Callison-Burch, and Jason Wei. 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)*, pages 837–848, Dublin, Ireland. Association for Computational Linguistics.

Robert Ridley, Liang He, Xin-yu Dai, Shujian Huang, and Jiajun Chen. 2021. Automated cross-prompt scoring of essay traits. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 13745–13753.

Parker Riley, Noah Constant, Mandy Guo, Girish Kumar, David Uthus, and Zarana Parekh. 2021. TextSETTR: Few-shot text style extraction and tunable targeted restyling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3786–3800, Online. Association for Computational Linguistics.

Horacio Saggion and Graeme Hirst. 2017. *Automatic text simplification*, volume 32. Springer.

Swarnadeep Saha, Omer Levy, Asli Celikyilmaz, Mohit Bansal, Jason Weston, and Xian Li. 2024. Branch-solve-merge improves large language model evaluation and generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8352–8370, Mexico City, Mexico. Association for Computational Linguistics.

Muhammad Tawsif Sazid and Robert E. Mercer. 2022. A unified representation and a decoupled deep learning architecture for argumentation mining of students' persuasive essays. In *Proceedings of the 9th Workshop on Argument Mining*, pages 74–83, Online and in Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Patricia Schmidtova, Saad Mahamood, Simone Balloccu, Ondrej Dusek, Albert Gatt, Dimitra Gkatzia, David M. Howcroft, Ondrej Platek, and Adarsa Sivaprasad. 2024. Automatic metrics in natural language generation: A survey of current evaluation practices. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 557–583, Tokyo, Japan. Association for Computational Linguistics.

K.A. Schriver. 1989. Evaluating text quality: the continuum from text-focused to reader-focused methods. *IEEE Transactions on Professional Communication*, 32(4):238–255.

Matthew Shardlow. 2014. A survey of automated text simplification. *International Journal of Advanced Computer Science and Applications*, 4(1):58–70.

Lei Shu, Liangchen Luo, Jayakumar Hoskere, Yun Zhu, Simon Tong, JD Chen, and Lei Meng. 2024. Rewritelm: An instruction-tuned large language-model for text rewriting. In *Proceedings of the AAAI Conference on Artificial Intelligence, 38(17), 18970-18980*.

Valerie J Shute. 2008. Focus on formative feedback. *Review of educational research*, 78(1):153–189.

Gabriella Skitalinskaya, Maximilian Spliethöver, and Henning Wachsmuth. 2023. Claim optimization in computational argumentation. In *Proceedings of the 16th International Natural Language Generation Conference*, pages 134–152, Prague, Czechia. Association for Computational Linguistics.

Christian Stab and Iryna Gurevych. 2014. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 46–56, Doha, Qatar. Association for Computational Linguistics.

Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659.

Frans H Van Eemeren, Rob Grootendorst, Ralph H Johnson, Christian Plantin, and Charles A Willard. 2013. *Fundamentals of argumentation theory: A handbook of historical backgrounds and contemporary developments*. Routledge.

Henning Wachsmuth, Gabriella Lapesa, Elena Cabrio, Anne Lauscher, Joonsuk Park, Eva Maria Vecchi, Serena Villata, and Timon Ziegenbein. 2024. Argument quality assessment in the age of instruction-following large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1519–1538, Torino, Italia. ELRA and ICCL.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017. Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187, Valencia, Spain. Association for Computational Linguistics.

Douglas Walton. 2009. Argumentation theory: A very short introduction. In *Argumentation in artificial intelligence*, pages 1–22. Springer.

Thiemo Wambsganss, Christina Niklaus, Matthias Cetto, Matthias Söllner, Siegfried Handschuh, and Jan Marco Leimeister. 2020a. AL: An adaptive learning support system for argumentation skills. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–14, New York, NY, USA. Association for Computing Machinery.

Thiemo Wambsganss, Christina Niklaus, Matthias Söllner, Siegfried Handschuh, and Jan Marco Leimeister. 2020b. A corpus for argumentative writing support in German. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 856–869, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yiran Wang, Ben He, Xuanang Chen, and Le Sun. 2025. Can LLMs clarify? investigation and enhancement of large language models on argument claim optimization. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4066–4077, Abu Dhabi, UAE. Association for Computational Linguistics.

Zhilin Wang, Alexander Bukharin, Olivier Delalleau, Daniel Egert, Gerald Shen, Jiaqi Zeng, Oleksii Kuchaiev, and Yi Dong. 2024. Helpsteer2-preference: Complementing ratings with preferences. *Preprint*, arXiv:2410.01257.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023a. Prompting large language model for machine translation: a case study. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2023b. A survey of controllable text generation using transformer-based pre-trained language models. *ACM Comput. Surv.*, 56(3).

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Jianing Zhou and Suma Bhat. 2021. Paraphrase generation: A survey of the state of the art. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5075–5086, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Pei Zhou, Jay Pujara, Xiang Ren, Xinyun Chen, Heng-Tze Cheng, Quoc V Le, Ed H Chi, Denny Zhou, Swaroop Mishra, and Huaixiu Steven Zheng. 2024. Self-discover: Large language models self-compose reasoning structures. *arXiv preprint arXiv:2402.03620*.

Wanzheng Zhu and Suma Bhat. 2020. GRUEN for evaluating linguistic quality of generated text. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 94–108, Online. Association for Computational Linguistics.

Timon Ziegenbein, Gabriella Skitalinskaya, Alireza Bayat Makou, and Henning Wachsmuth. 2024. LLM-based rewriting of inappropriate argumentation using reinforcement learning from machine feedback. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4455–4476, Bangkok, Thailand. Association for Computational Linguistics.

A Dataset Metrics

Dataset metrics can be found in Table 4.

	Rev1	Rev2	Rev3	Essays	MT (EN)	MT (DE)
Avg. Length	3128.40	3464.22	4074.56	1919.51	452.58	452.58
# number of documents	86.00	86.00	86.00	402.00	89.00	89.00
Avg. Sentence Count	25.42	28.34	32.76	16.79	4.18	4.28
Avg. Sentence Length	102.69	102.25	104.04	95.57	90.71	99.47
Avg. Words per Sentence	20.47	20.19	20.44	19.09	18.01	15.79

Table 4: Dataset metrics of our chosen datasets.

B Prompts Used

We include the prompts used here. The few-shot prompt is used for SelfDiscover as well. We otherwise follow the approach presented by Zhou et al. (2024). For Genetic Algorithm we use the following prompts as the initial population:

- Improve the following argument
- Make the following argument better
- Enhance the following argument
- Make the next argument not suck

3-shot In k-shot prompting settings the model is given k examples in the prompt that demonstrate the task that it should solve. Performance generally increases with larger k (Peng et al., 2024; Zhang et al., 2023a). We use demonstrations from the Argument Revision Corpus. We make use of the annotated alignment of the first and second revisions. Sentences for pairs of revisions are aligned and marked with the purpose. We use the first five aligned sentences that have a purpose other than ‘identical’, for three of the essays.

Branch-Solve-Merge Branch-Solve-Merge is a prompting technique proposed by Saha et al. (2024). In a first step the LLM is asked to split the problem into separate sub-problems (Branch). The sub-problems are then solved individually (Solve) and combined together into a full solution for the original problem (Merge). In our approach we ask the LLMs to come up with individual aspects that can be improved in the original argumentation (Branch). The same LLM is then prompted to improve those individual aspects (Solve) and lastly it is prompted to combine the separate generated texts into one finished argumentative text (Merge).

Self-Discover Self-Discover is a technique proposed by Zhou et al. (2024). The LLM is first prompted to select suitable reasoning modules, from a pre-defined list, that are useful for solving the task. We use the same reasoning modules that Zhou et al. (2024) describe in their work. The model is then prompted to come up with a plan in JSON format using the modules. Finally, the plan is used to prompt the model to generate a solution.

Genetic Algorithm A recent work by Guo et al. (2024) makes use of the principles of evolutionary algorithms to optimize prompts. We include an approach based on the proposed Genetic Algorithm variant. An initial prompt is used to solve the task, performance is assessed and combined with other high-performing prompts to find an optimized prompt.

Little Brother How feedback is phrased can have a large impact on how well it is received (Shute, 2008). We came up with the idea to experiment with gentle feedback. The models first solve the task in the 3-shot setting, in the role of a ‘little brother’. Next, a ‘big brother’ model, is asked to solve the same task, but provided the solution by the little brother model. The model is then asked to provide feedback to its ‘little brother’. We used Llama 3.1 as the big brother model, and the others as the solvers in the little brother role.

C Scores

The following tables show the scores of the Llama 3.1 model with the 3-shot prompting approach. We omit the other tables due to the large amount of data. Scores for all models and approaches are included in the Github repository.

D License Terms of Used Datasets

We used the Argument Annotated Essays 2.0 (Stab and Gurevych, 2017) in our research. This dataset may only be used for academic and research purposes.

The ArgRewrite V.2 (Kashefi et al., 2022) corpus is available under the GNU General Public license.

You are given an argument about the topic "{topic}". Your task is to improve it. Respond only with the improved argument wrapped in @ symbols and nothing else. Here are some examples of improvements:
Demonstration1
Demonstration2
Demonstration3

Figure 6: Few-shot prompt

You are given an argument about the topic >topic<. Your task is to improve it. In order to do so, your task is to first propose certain aspects of the argument that can be improved, and then divide the aspects into two groups such that the argument can be improved individually for all aspects in the groups. Your output should be in the format:
Group 1: <aspects here>
Group 2: <aspects here>

Figure 7: BSM Branch prompt

Improve the following argument by focussing on the specific aspects. Respond with the improved argument wrapped in @ symbols. Try to keep the length of the improved argument similar to the original one.
Argument: >task<
Aspects: >group<

Figure 8: BSM Solve prompt

Given two arguments about the topic >topic<, your task is to merge them into a single argument. Respond with the merged argument wrapped in @ symbols.

Figure 9: BSM Merge prompt

You are given two arguments. Your task is to choose the better one. Respond with @First@ if you prefer the first one, and with @Second@ if you prefer the second one.

Figure 10: Genetic Algorithm population scoring prompt

The Microtexts corpus (Peldszus and Stede, 2015) is available under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

E Computational details

We used the following models for our experiments:

- bigscience/bloomz-3b
- bigscience/bloomz-560
- allenai/OLMo-7B-0724-Instruct-hf
- microsoft/Phi-3-medium-4k-instruct (14B parameters)
- microsoft/Phi-3-mini-4k-instruct (3.8B parameters)
- nvidia/Llama-3.1-Nemotron-70B-Instruct

> Solve this task: task. Your little brother has solved this task like this previously:
> [PREVIOUS]
> {previous}
> [/PREVIOUS]
> Check if your little brother's solution is correct. If it is not, teach them where they made a mistake, and correct it. If it is correct, state the solution and explain it. Put the corrected solution into @ symbols.

Figure 11: Little Brother prompt

> You are a lecturer of the writing class. You are given the following proposition on a controversial topic. You need to carefully read the proposition and evaluate it based on the criteria:
> - Clarity
> - Relevance
> - Logical consistency
> - Validity of reasoning
>
> Now you need to assign a score for coherence on a scale of 1 to 5, where 1 is the lowest and 5 is the highest based on the Evaluation Criteria. Note, you should be very strict when giving the score.

Figure 12: AMERICANO coherence prompt

> You are a lecturer of the writing class. You are given the following proposition on a controversial topic. You need to carefully read the proposition and evaluate it based on the criteria:
> - Language and rhetoric
> - Addressing opposing viewpoints
> - Credibility
> - Overall effectiveness
>
> Now you need to assign a score for persuasion on a scale of 1 to 5, where 1 is the lowest and 5 is the highest based on the Evaluation Criteria. Note, you should be very strict when giving the score.

Figure 13: AMERICANO persuasion prompt

All models are from the HuggingFace repository.

Our texts were generated on up to 8 V100 GPUs on a DGX2 machine over the course of four weeks. Experiments were performed consecutively and did not run the full four weeks. Llama 3.1 is the only model that needed eight GPUs, the other models ran on up to four GPUs if resources were available, but can be run on two. Total GPU hours for both text generation and scoring are around ≈ 20.

F Use of AI assistants

We used ChatGPT 4o to generate the title of the paper.

G Annotation Details

Both annotators are authors of the paper and were aware that their annotations would be used as part of this paper. They are Caucasian and from Central Europe. The instructions were to analyze the changes that are present in the improved texts and to choose the argument that they consider to be better in the preference analysis.

index	Rev1	Rev2	Rev3	Essays	MT
add	46.51	44.19	30.23	36.32	5.06
copy	186.05	134.88	179.07	237.56	200.00
delete	31.40	38.37	74.42	17.91	2.81
fusion	102.33	77.91	81.40	71.89	26.40
merge	551.16	675.58	777.91	311.44	47.19
other	0.00	0.00	0.00	0.00	0.00

Table 5: BERTAlign changes

score_name	Rev1	Rev2	Rev3	Essays	MT
linguaf_avg_word_length	26.26	25.43	25.01	23.58	17.77
linguaf_char_count	-20.51	-23.48	-33.29	-0.28	47.54
linguaf_digit_count	2.53	22.66	3.43	-8.09	-7.05
linguaf_letter_count	-21.74	-24.76	-34.45	-1.06	47.46
linguaf_avg_sentence_length	5.30	8.67	7.30	16.99	52.27
linguaf_avg_words_per_sentence	-16.59	-13.43	-14.11	-5.24	30.39
lexical_ttr	35.32	35.55	40.82	25.40	1.73
linguaf_flesch_kincaid_grade	31.07	32.12	29.37	42.29	57.07
linguaf_flesch_reading_ease	-40.96	-41.29	-40.62	-43.37	-44.43
original_length	312839.53	346422.09	407455.81	191951.49	47249.44
count1to3	-18.61	-22.22	-32.49	7.86	61.73
count4to6	64.32	65.80	36.06	96.24	58.65
count7to10	0.00	-0.41	-0.40	-0.21	-0.39
count10plus	0.00	0.00	-0.39	0.00	0.00
length_change	-24.95	-27.48	-37.39	-4.66	40.18
levenshtein_levenshtein	2045.71	2265.07	2687.08	1287.39	407.94

Table 6: Lexical Level

score_name	Rev1	Rev2	Rev3	Essays	MT
add	46.51	44.19	30.23	36.32	5.06
copy	186.05	134.88	179.07	237.56	200.00
delete	31.40	38.37	74.42	17.91	2.81
fusion	102.33	77.91	81.40	71.89	26.40
merge	551.16	675.58	777.91	311.44	47.19
other	0.00	0.00	0.00	0.00	0.00
num_adv_mod	-45.75	-47.55	-56.07	-18.38	29.56
num_advcl	-12.24	-15.63	-23.12	28.10	70.23
num_appos	132.15	137.54	151.76	-6.37	39.52
num_coordNP	35.08	28.10	14.49	46.06	13.69
num_coordVP	-45.84	-38.28	-47.45	-23.03	-9.93
num_coord_cl	-72.38	-67.23	-77.94	-81.12	-84.47
num_part	-17.53	-25.56	-35.03	33.06	27.43
num_prep	-29.26	-33.67	-42.91	-6.82	62.17
num_relcl	-65.74	-72.16	-72.42	-38.11	-53.40
num_speech	-59.85	-55.08	-56.76	-39.01	14.29
improved_length	2347.79	2512.23	2550.88	1830.12	662.37
original_length	3128.40	3464.22	4074.56	1919.51	472.49

Table 7: Syntactic Level

score_name	Rev1	Rev2	Rev3	Essays	MT
feng_hirst_depth	-21.17	-22.09	-33.26	-19.39	17.15
Attribution	-31.49	-32.23	-39.57	-23.15	-26.27
Background	-27.05	-28.40	-33.39	-29.81	-70.00
Cause	-53.88	-51.09	-57.85	-59.51	-91.67
Comparison	-100.00	-100.00	-100.00	-94.12	-100.00
Condition	-86.60	-84.09	-77.82	-86.11	-100.00
Contrast	-19.13	-11.90	-25.84	-0.24	-22.45
Elaboration	-26.43	-26.56	-35.18	2.19	54.79
Enablement	-61.29	-53.89	-56.55	-55.63	-53.85
Evaluation	-60.98	-72.97	-79.81	-79.78	-100.00
Explanation	-54.65	-68.97	-69.29	-70.40	-83.33
Joint	-18.71	-34.98	-38.28	-24.17	-55.02
Manner-Means	-22.55	-29.05	-38.89	-59.72	-100.00
Summary	-100.00	-100.00	-100.00	-92.31	-100.00
Temporal	-78.33	-90.74	-77.35	-76.77	-80.00
Topic-Change	-100.00	-100.00	-100.00	-100.00	0.00
Topic-Comment	-90.22	-78.57	-78.57	-100.00	-50.00
same-unit	5.10	2.01	-8.18	7.97	-16.89
gruen_scores	4.02	2.28	1.94	15.11	3.38
polarity	-10.53	0.54	40.92	-1157.79	-35.83
subjectivity	3.60	4.50	4.90	-3.29	13.20
german_proba_positive	nan	nan	nan	nan	162.27
german_proba_negative	nan	nan	nan	nan	107.08
german_proba_neutral	nan	nan	nan	nan	146.89

Table 8: Semantic Table

score_name	Rev1	Rev2	Rev3	Essays	MT
americano_coherence_avgs	18.13	32.60	8.35	6.45	23.11
americano_persuasion_avgs	76.18	91.32	44.00	32.52	31.31

Table 9: Pragmatic Level

dataset	Rev1	Rev2	Rev3	Essays	MT
Claim	0.06	-0.16	-0.14	-0.42	1.13
MajorClaim	-0.53	-0.50	-0.41	0.26	-0.11
None	-1.65	-2.01	-3.26	-0.12	-0.89
Premise	-4.88	-6.55	-8.90	-2.54	-0.11

Table 10: Argument Mining Components