Linguistic Transformations in Argument Improvement: Analyzing Large Language Models' Rewriting Strategies

Anonymous ACL submission

Abstract

Text rewriting is a task that is related to, but different from, general text generation. While LLMs have been extensively studied on general text generation tasks, there is less research on text rewriting, and particularly on the behavior of models on this task. In this paper we analyze what changes LLMs in a text rewriting setting. We focus specifically on argumentative texts and their improvement, a task named Argument Improvement (ArgImp). We present an evaluation pipeline consisting of metrics on four linguistic levels. This pipeline is used to score improved arguments on diverse corpora 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 this task by shortening the vocabulary while simultaneously increasing average word length and merging sentences. Overall we note an increase in the persuasion and coherence dimensions. Our findings were made possible by splitting the analysis on the four linguistic levels in our evaluation pipeline.

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 other things. It can be seen as a form of controllable text generation (Zhang et al., 2023b), where a given text is modified based on specific user requirements, such as improving its 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 argument or argumentative text, respectively, 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). However, the task of ArgImp remains largely unexamined.

This work aims to bridge this gap by investigating the linguistic transformations performed by LLMs when prompted to improve a given argumentative text. Specifically, we analyze how these models alter argumentative texts at four distinct linguistic levels: We examine word choice *(lexical level)*, sentence structure *(syntactic level)*, meaning shifts *(semantic level)*, and rhetorical effectiveness *(pragmatic level)*. 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. We include an investigation into known biases in this setting.

To tackle this problem, we have created an evaluation pipeline consisting of 57 metrics used in natural language generation (NLG). 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

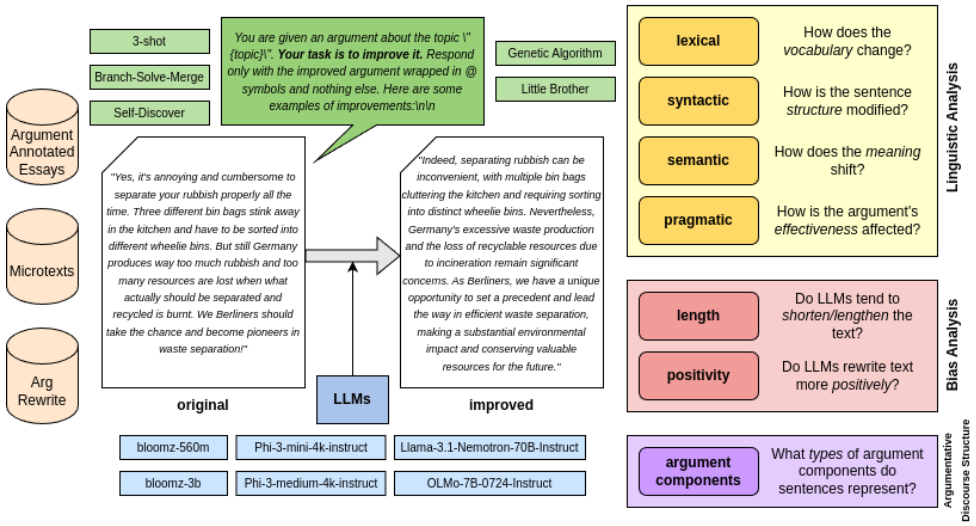

Figure 1: Overview of our experimental setup for the task of ArgImp. We evaluate the quality of argumentative texts rewritten by LLMs prompted for improvement. We apply six models across five datasets (each revision of the ArgRewrite corpus is treated as a distinct dataset). The evaluation spans four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

make exactly when used in an ArgImp setting. 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.

Our contributions are as follows: (i) a comprehensive pipeline for evaluating the output quality of text rewriting tasks, consisting of 57 different metrics[1]; (ii) an analysis of LLM behavior on four different linguistic levels for the task of ArgImp; and (iii) an investigation of LLM biases in an ArgImp setting.

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, Zhang et al. (2025) 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.

There is a growing line of research that focuses on developing argumentative writing support tools that provide users with feedback on the quality of their argumentative texts with the objective of guiding them in generating high-quality persuasive texts (Sinikallio et al., 2025). For instance, AL, an adaptive learning support system for argumentation skills, offers formative feedback through in-text highlighting of argumentative components, qualitative scores and graph-based visualizations of argument discourse structures (Wambsganss et al., 2020a). More recently, Gubelmann et al. (2024) introduced Artist, a framework that integrates LLM-based improvement suggestions. They conducted a user study with students to evaluate the effectiveness of such feedback. The results indicate that the students generally find the feedback provided by the LLMs to be helpful and of high quality. These frameworks primarily adopt a reader-

[1]The code is available at *anonymized for review*.

oriented perspective, as their goal is to provide feedback that helps students develop the skills to refine their texts manually. Our work, in contrast, takes a text-centric approach, focusing on the linguistic quality of LLM-generated improvements in argumentative texts.

3 Argument Improvement with LLMs

We aim to evaluate the quality of argumentative texts rewritten by LLMs prompted for improvement. 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 from one another in an ArgImp setting?

3.1 Model Selection

We aim to provide a broad overview over LLM behavior. For that reason we selected multiple models of different families, and varying sizes. We considered adaption rate of the models in our selection process. The models we used for our experiments are 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)[2].

3.2 Datasets

Our aim is to present results on a diverse set of datasets representing different argumentative settings. We use the well-known Argument Annotated Essays 2.0 corpus by Stab and Gurevych (2017). The texts in this corpus are student-generated essays. We further include the Microtexts corpus (Peldszus and Stede, 2015). In contrast to the essays, the texts in this corpus consist of very short argumentative texts. These texts are closer to how argumentation occurs in informal settings. The corpus consists of both English and German texts, which allows us to show results in two different languages. Lastly we make use of the ArgRewrite V.2 corpus by Kashefi et al. (2022). This corpus consists of three revisions of argumentative essays. Students wrote the initial version, received feedback to revise their texts to produce a second version, and lastly refined their texts further in different settings. We treat each individual set of revisions (original/revision 1, revision 2 and revision 3) as separate datasets to analyze model behavior across different versions of the original texts.

3.3 Prompting Techniques

Prompts have a large impact on the output of LLMs (Cheng et al., 2024; Long et al., 2024). There is currently, to our knowledge, no research investigating how to efficiently prompt LLMs in an ArgImp setting. Our selection is as follows.

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

[2]All models are from the HuggingFace repository.

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.

4 Evaluation

4.1 Linguistic Analysis

We employ a wide range of NLG evaluation metrics. 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. While most scores are related to the form and structure of the texts, we also include metrics that are account for the meaning.

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 We expect the models to change the structure of the argumentative texts. To investigate these modifications, we analyze the syntax of the sentences using dependency parse tags generated by spaCy (Honnibal et al., 2020). 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. In our work, we applied this technique to align sentences from the original texts with their corresponding improved versions. This allowed 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).

Semantic Analysis To capture changes in meaning between the original and improved texts, we include a sentiment classifier, GRUEN score metrics (Zhu and Bhat, 2020), and a discourse analysis using an RST parser. In that way, we aim to capture both changes in the general tone and more nuanced shifts in meaning resulting from the models' improvements.

Pragmatic Analysis We adopt the approach by Hu et al. (2024) to evaluate the texts' persuasiveness and coherence as key aspects of pragmatics. 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, particularly in an evaluation setting, have both a length[3] (Chen et al., 2024a; Zheng et al., 2023) and a positivity bias (Palmer and Spirling, 2023; Buhnila et al., 2025; Markowitz et al., 2024). It is unclear whether this can be observed in a rewriting setting such as ArgImp as well. We investigate this 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 analyze the argument discourse structure of the texts by comparing the original and improved versions in terms of their argument components. More specifically, we classify each sentence into one of the following four types of argument components:

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

claim, premise, major claim, or none. For the English datasets, we make use of an implementation of the best-performing approach proposed in Stab and Gurevych (2014b), which is based on an SVM classifier trained on the argument annotated essay corpus (Stab and Gurevych, 2014a) and achieving an accuracy of 0.77. For the German Microtext dataset, we apply the same classification method, trained on the corpus introduced by Wambsganss et al. (2020b), achieving an accuracy of 0.65 as reported by Wambsganss et al. (2020a). We then compare the distribution of argument component types between the original and revised texts to assess structural modifications.

5 Results

Due to the large number of possible combinations[4] we focus our in-depth analysis. The performance of Llama 3.1 is expected to be highest, both due to its comparatively high performance on various benchmarks (Chiang et al., 2024) and its parameter size. As few-shot prompting is the most widely used of our approaches we use the combination of both Llama 3.1 as well as the few-shot prompting approach for a deeper analysis. We include an analysis of the remaining results in a more general form due to the sheer size of the experimental setup. Detailed scores in tabular form can be found in Appendix B. The heatmaps used in this section are based on the scores of Llama 3.1 and the 3-shot prompting approach. Blue indicates a decrease, red an increase. All scores indicate a percentage change relative to the scores of the original human-written texts. The heatmaps scale from -200 to +200. We describe outliers in the analysis of each level. Both Bloomz models generated very short texts that are not full argumentative texts. We omit them from the analysis for this reason.

5.1 Lexical Analysis

Figure 2 provides an overview of the scores on the lexical level. Levenshtein edit distances are included in Table 1. We note that Llama 3.1 shortened the texts on all datasets but Microtexts, where length increased on average by about 40%. This behavior is consistent with OLMo and the two Phi-3 models. The models generally increased the average word lengths but made sentences shorter. We observe that the larger models decreased the

[4]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 $6 * 5 * 5 * 4 = 600$.

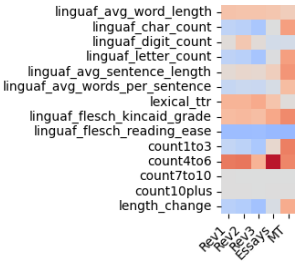

Figure 2: Changes on the lexical level

reading ease metrics, whereas the smaller ones increased it. This increase is not linear with the number of parameters of the models however. Llama 3.1 in particular shows a strong increase in the number of 4 to 6 syllable words and decrease in words with less syllables. This could be related to the length bias as discussed in Section 5.5. It is possible that the larger models inherently prefer longer words, which could be connected to their autoregressive training and general generation mechanics. Tokens in LLMs do not correspond with syllables, but it holds that in text the majority of characters are not whitespace. The lack of whitespace in the training data could lead to the models generating lengthier words, which in turn have more syllables. This fits with the observed increase in average sentence length as well as the decrease of the average number of words per sentence, particularly in the case of the Llama 3.1 model.

	Rev1	Rev2	Rev3	Essays	MT
levenshtein	2045.71	2265.07	2687.08	1287.39	407.94

Table 1: Changes for Levenshtein distance metric

5.2 Syntactic Analysis

As the classification of the sentence transformations operations is reference-based, we provide an overview of the changes in Table 2. The values in Figure 3 are reference-free and show the percentage increase/decrease in metrics. Llama 3.1 deleted parts of the text rarely, but also added new ones rarely. Instead it opted to modify the sentences in some way, with merge being its most popular type of transformation operation. This is in line with the findings of the lexical analysis in the previous section, that the texts are generally shortened. This hints towards the model making an attempt to make the text more focused by cutting out existing parts, but not deleting large sections of the text. If the model deleted entire sentences

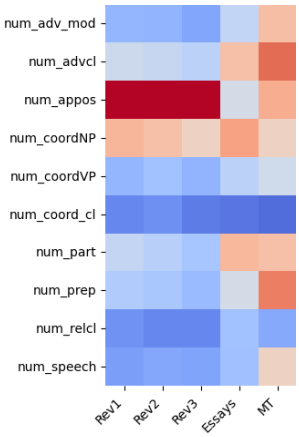

Figure 3: Changes on the syntactic level

or paragraphs then the delete value would be high, but it is its least used action on average. OLMo on the other hand was more aggressive. Its delete score is high ($\approx 273 - 357$) with the exception of the Microtexts corpus. Just like Llama 3.1 it often copied existing chunks. Its most frequent action, outside of copy, is merge, which is in line with the Llama 3.1 model. The two Phi-3 models behave similarly to OLMo. Both Phi-3 models delete more than Llama 3.1, with the the medium model being more moderate. Figure 3 shows the percentage changes of counts of selected dependency parse tags in the improved texts relative to the original human-written ones. We note that Llama 3.1 is the only model that increases the number of coordinating noun phrases, for all datasets, whereas OLMo commonly decreases them, or only increase them marginally on some of the datasets (Phi-3 models). It also quite significantly increases the number of appositional modifiers ('The largest model, Llama 3.1, performs best.', here 'Llama 3.1 is in apposition to 'model'). This hints towards the model making modifications that aim to make the text more understandable. This is at first contrary to the previous findings that the texts are shortened. We discuss the implications of this further in Section 6.

	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

Table 2: Types of sentence transformations

5.3 Semantic Analysis

We provide an overview of the percentage changes in the scores on this level in Figure 4. The most notable finding is that all models consistently decreased the depth of the RST parse tree on all datasets, but increased it for the Microtexts dataset. The Microtexts are all very short arguments, and it appears as though the models consider them, or at least the overall rhetorical structure, to be too short. We refer back to Section 5.1 where we find that the models shorten all texts, with the exception of the Microtexts, and 5.2, where we found that often the models merge or split the original sentences in some way. We discuss this further in Section 6. For Llama 3.1 we note an outlier for the polarity score on the Essays dataset. Without it, the average change is -11%. The value for one human-written text is almost, but not quite, zero. Overall the models perform very similarly in terms of sentiment changes. On the German Microtexts there is a large increase in sentiment, whereas for all English texts the polarity is decreased. This means the models make the texts *more negative*, but not necessarily negative *over all*. We also note an increase in terms of subjectivity. We also include GRUEN score in our analysis, which has an increase across all datasets and models, but has the strongest increase on the Essays dataset. Similar to the other levels, the changes on the other metrics are largest on the Microtexts corpus texts. We discuss this further in Section 6.

5.4 Pragmatic Analysis

For the persuasion and coherence scores we note an increase 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). Interestingly the increase in score is largest for Revision 2 for both Llama 3.1 and OLMo. Humans wrote the original text (Revision 1) and then improved that using expert feedback to produce Revision 2. In our setting the models were asked to improve Revision 1, Revision 2 and Revision 3 separately, without this feedback. Revision 1 can be expected to be comparatively unrefined, relative to the other revisions, and as such has the most room for improvement, and Revision 3 the least. We expected the scores to decrease as revisions increase, as the texts improve with increasing revision as well. As for both dimensions, coherence

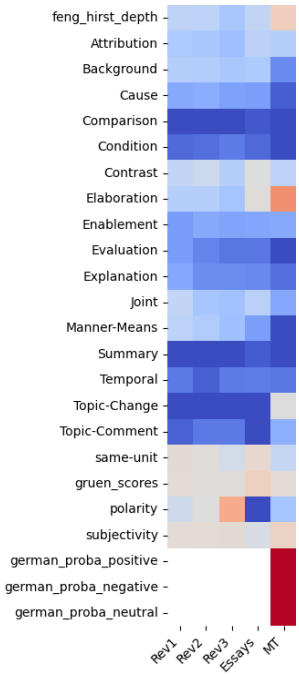

Figure 4: Changes on the semantic level

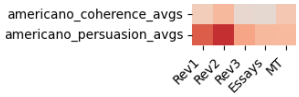

Figure 5: Changes on the pragmatic level

and persuasion, there is a positive increase, we can say that overall the improvement process was a success.

5.5 Length Bias

We describe findings on each level. For the correlation, we use Pearson's standard correlation coefficient. We aim to analyze whether models behave differently on texts of different lengths, by means of using correlations.

Lexical We note a correlation between the length and the average word length (≈ 0.3) and a strong negative correlation for the 1 to 3 syllable word count (≈ -0.55), sentence length (≈ -0.40) and average words per sentence (≈ -0.45). The results also show a strong correlation for the token-to-type ratio (≈ 0.58) and a negative correlation for the Flesch-Kincaid grade (≈ -0.25).

Syntactic There is a strong correlation for fusion (≈ 0.57), a weak correlation for add and copy and a weak negative correlation for delete and merge (≈ -0.17 and ≈ -0.23).

Semantic We note no interesting correlations.

Pragmatic There is a weak correlation (≈ 0.13) between the length of the argumentative texts and the persuasion scores. There is no correlation between length and coherence.

Behavior for the other prompts For the other prompts we note largely the same behavior as for the few-shot approach. The overall values differ, but the general trends are the same: the models shorten the texts and sentences, but increase particularly the number of 4 to 6 syllable words. In terms of argument quality we note interesting differences on the pragmatic level (scores are Coherence/Persuasion, for Llama 3.1): few-shot has +18/+55%, Genetic Algorithm +45/+59%, SelfDiscover +40/+74%, Branch-Solve-Merge +58/+84% and Little Brother the largest increase in persuasion, with scores of +34/+101%. For Branch-Solve-Merge we note that OLMo performs a lot of merge operations, the overall changes it makes seem to be less pronounced in this case than in the case of few-shot prompting.

Summary Outside of a weak correlation on the persuasion score we do not find any indications that the models prefer texts of a certain length when rating scores. We find a strong negative correlation for the number of 1 to 3 syllable words, as well as words per sentence and sentence length, which is similar to the findings discussed in Section 5.2. The models decrease the overall length of the texts, but do so by increasing the length of the words. The Flesch-Kincaid reading grade score decreases with both average sentence length and average syllables per word, so the correlation there follows from the discussed behavior. The token-to-type ratio also has a strong correlation as previously discussed. This supports our hypothesis that the texts become shorter, as the words become longer: longer words are less likely to be re-used, thus increasing the types present, and a shorter text has less tokens. Both are factors leading to a higher ratio.

5.6 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:

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

Using this formula, we find that 335 negative shifts (46.16%), 203 neutral shifts (26.40%) and 211 pos-

	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 values of argument components

itive 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 model is $+11.39\%$. This indicates that while positive changes are done rarely, they are strong in magnitude when they occur. The model appears to move the improved texts towards a more neutral sentiment.

5.7 Argument Component Classification

We present the changes in values 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 as premises. As discussed in previous sections we observe an increase in sentence length, as well as an overall merging of sentences. Due to the texts becoming shorter on average 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 premises.

6 Discussion

Our analysis indicates that the models aim to shorten the overall texts in the ArgImp setting. Results on the lexical level show that overall text length decreases, as well as 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 actions, 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. 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. These indicate shorter sentences, with longer words.
- *Syntactic level*: Original sentences are merged. This again suggests both shorter texts, as well as more focused sentences.
- *Semantic level*: Depth of the RST trees decreases.
- *Pragmatic level*: Argumentative quality increases, which suggests the changes made are effective, the models did not destroy the texts.

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 categorizing commonly used text generation metrics into linguistic levels and performing an analysis on the individual linguistic levels we have found that LLMs make the texts more focused in an ArgImp setting. Additionally, our results suggest that the improvement process is a task that LLMs can perform well. 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. The larger models performed better in both quality of the 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 in terms of quality assessment by the LLMs. 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 make the texts more focused and increase information density, while not having an adverse affect on 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 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 manual evaluation impractical.

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.

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.

Jiale Cheng, Xiao Liu, Kehan Zheng, Pei Ke, Hongning Wang, Yuxiao Dong, Jie Tang, and Minlie Huang. 2024. Black-box prompt optimization: Aligning large language models without model training. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3201–3219, 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.

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.

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*.

Reto Gubelmann, Michael Burkhard, Rositsa V. Ivanova, Christina Niklaus, Bernhard Bermeitinger, and Siegfried Handschuh. 2024. Exploring the usefulness of open and proprietary llms in argumentative writing support. In *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky*, pages 175–182, Cham. Springer Nature Switzerland.

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, 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.

Do Long, Yiran Zhao, Hannah Brown, Yuxi Xie, James Zhao, Nancy Chen, Kenji Kawaguchi, Michael Shieh, and Junxian He. 2024. Prompt optimization via adversarial in-context learning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7308–7327, Bangkok, Thailand. Association for Computational Linguistics.

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.

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.

Laura Sinikallio, Lili Aunimo, and Tomi Männistö. 2025. Systematic review on the current state of computer-supported argumentation learning systems. *Information and Software Technology*, 178:107598.

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. 2014a. Annotating argument components and relations in persuasive essays. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1501–1510, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Christian Stab and Iryna Gurevych. 2014b. 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).

Kexun Zhang, Jane Dwivedi-Yu, Zhaojiang Lin, Yuning Mao, William Yang Wang, Lei Li, and Yi-Chia Wang. 2025. Extrapolating to unknown opinions using LLMs. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7819–7830, Abu Dhabi, UAE. Association for Computational Linguistics.

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 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

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

B 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.

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

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 4: 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 5: 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 6: 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 7: 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 8: 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 9: Argument Mining Components

C 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.

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

D 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

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.

E Use of AI assistants

We used ChatGPT to generate the title of the paper.