Quantifying the Risks of LLM- and Tool-assisted Rephrasing to Linguistic Diversity

Mengying Wang
University of Konstanz
mengying.w.wang@gmail.com

Andreas Spitz
University of Konstanz
andreas.spitz@uni.kn

Abstract

Writing assistants and large language models see widespread use in the creation of text content. While their effectiveness for individual users has been evaluated in the literature, little is known about their proclivity to change language or reduce its richness when adopted by a large user base. In this paper, we take a first step towards quantifying this risk by measuring the semantic and vocabulary change enacted by the use of rephrasing tools on a multi-domain corpus of human-generated text.

1 Introduction

Writing assistants such as Grammarly or Quillbot are widely used in writing tasks by native and non-native speakers alike. To aid the user in the composition of text, writing assistants (WATs) provide sophisticated and comprehensive functions, such as grammar correction, spell-checking, and specialized rephrasing and embellishment. More recently, large language models (LLMs) are increasingly being integrated into writing assistants (Fok and Weld, 2023), including Grammarly and Google's Smart Compose (Chen et al., 2019). These advanced WATs offer substantial assistance in the writing process (Roe et al., 2023), but also introduce challenges such as hallucinations or inconsistent content and style (Ariyaratne et al., 2023; Kacena et al., 2024; Gero et al., 2022). After the introduction of ChatGPT, language models can also be – and are – used as writing assistants directly due to their chat functionality and ease of use.

Despite their widespread use, however, there is a notable lack of systematic *quantitative* investigations into how the use of writing assistants alters the produced text. Existing studies tend to focus on the effectiveness and accuracy of WATs (Gayed et al., 2022), but discount the downstream impact that a pervasive adoption of such systems might have on the style and diversity of the language we use in written day-to-day communication.

Contributions. In this paper, we address this research gap with a quantitative analysis of the effect that the use of (semi)automated rephrasing tools has on the produced text. We experiment with four traditional WATs and six LLMs to determine the effect of tool-assisted rephrasing from the perspective of corpus-level text diversity. Using token-level and vector-level metrics, we provide a comprehensive overview of the ways in which these tools modify text – and where worries about linguistic diversity may or may not be warranted.

2 Related Work

Academic work on writing assistants is relatively scarce, and focuses predominantly on measuring the effectiveness of WATs as a tool for improving writing efficiency, the accuracy of grammar, or for spell checking (Gayed et al., 2022), and we are unaware of any studies of language diversity as a result of writing assistant usage. Furthermore, analyses of WATs with respect to content are predominantly qualitative or based on manual evaluation (Ebadi et al., 2023), which introduces the potential for subjective biases. Prior research tends to examine individual tools, such as Grammarly (Ebadi et al., 2023) or Quillbot (Amyatun and Kholis, 2023) rather than providing a comparative analysis, thereby limiting generalizability. With regard to language models, prior work more commonly focuses on the collaborative writing effectiveness between LLMs and humans, rather than considering textual change (Lee et al., 2022), or on evaluating grammar and style, rather than corpus-level diversity (Reinhart et al., 2025). While Martínez et al. (2024) investigate the risk of diversity reduction as a result of language model usage, they only provide a case study on the example of ChatGPT. Otherwise, prior investigations have left the dimension of linguistic diversity largely unexplored to focus on quality (Aydin et al., 2025),

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 22561–22574
November 4-9, 2025

while their application scenarios (often limited to academic writing tasks) are narrower than the cross-domain texts we consider here.

In contrast to the above works, we conduct a quantitative evaluation that systematically compares both traditional writing assistants and multiple LLM-based tools across texts from diverse domains. Our analysis considers not only surface-level measures but also semantic-level metrics, with a particular focus on linguistic diversity.

3 Data

We use texts from a variety of domains that we rephrase with the help of writing assistants and language models, as described in the following.

3.1 Corpus Compilation

To compile the corpus, we consider English texts, predominantly written prior to 2010 to exclude those that were created with support by writing assistants or language models. Each individual text has paragraph length, and is selected to be coherent, contiguous, with content relevant to the domain, and not contain non-standard characters that may interfere during rephrasing. To investigate the impact of text type, we compile the corpus from 8 different domains (literature, academic papers, encyclopedic texts, instruction manuals, news, social media posts, interview transcripts, and speeches), resulting in a total of 819 texts.

For further details, see Appendix A. The data is available in our code repository[1].

3.2 Rephrasing

Using this corpus as input, we then utilized writing assistants and LLMs to rephrase the texts.

Writing assistant tools (WATs). As WATs, we consider the popular tools Grammarly, Quillbot, and Wordtune, and also include Rephrase as a lesser-known tool, all four of which provide rephrasing functionality. For better performance, we obtained paid membership subscription for all tools. Processing of the input texts was then done manually by one of the coauthors, rephrasing each paragraph independently in the WATs interface. To ensure consistent rephrasing, we adopted the rule-set that (1) all rephrasing suggestions had to be accepted, and (2) in case of multiple rephrasing suggestions, the first (highest ranked) option had to be used. For further details on the tools and rephrasing process, see Appendix B.

Language model rephrasing (LLMs). Since language models are increasingly used for text generation, we also employ five LLMs to rephrase the texts via zero-shot prompting. We consider three commercial models with GPT-4o mini (OpenAI, 2023), Gemini-2.5 Flash (Comanici et al., 2025) and DeepSeek R-1 (DeepSeek-AI, 2025), as well as the three open-weight models Aya-23 (Aryabumi et al., 2024), LLaMa 3 (8b) (Dubey et al., 2024), and Qwen 2.5 (7b) (Yang et al., 2024).

For all LLMs, we experiment with five different prompt templates, including one chain-of-thought prompt (see Appendix C for details) and generate one rephrase for each text per prompt per LLM. We manually check all output for LLM refusal and corruption, and discard unusable rephrased texts (see Appendix G). Model settings and hyperparameters are listed in Appendix D. An example of a rephrased paragraph for all tools can be found in Appendix I.

4 Experimental Setup

To measure changes to the texts as a result of rephrasing, we consider word-based measures and embedding-based measures, both computed at the paragraph level. For definitions, see Appendix E.

4.1 Word-based Measures

As word-based metrics, we consider changes at the sentence and the vocabulary level.

Paragraph length. As the most straightforward measure, we use the percentual change in the length of paragraphs, measured in the number of words.

Jaccard similarity. To measure the vocabulary overlap before and after rephrasing, we use the Jaccard similarity (Niwattanakul et al., 2013).

Levenshtein distance. We also consider the normalized Levenshtein distance (Yujian and Bo, 2007) between texts as a more fine-grained alternative to the Jaccard similarity.

Vocabulary Size. As a corpus-level metric, we define the vocabulary size as the number of unique words in a set of documents (e.g., all paragraphs rephrased by a specific tool).

4.2 Vector-based Measures

To measure vector-based changes, we consider the semantic similarity between paragraphs, as well as changes in the size of the cone containing the texts in latent embedding space.

[1] https://github.com/Mengying-W/Writing-Assistant-Tools

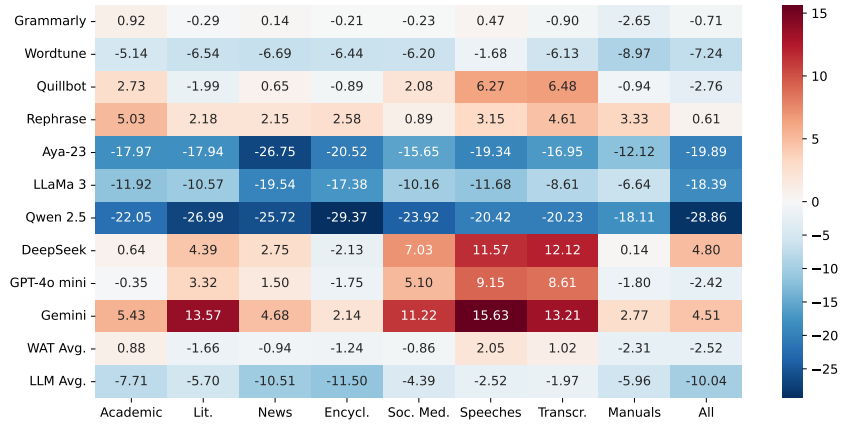

Figure 1: Percentual changes in total vocabulary size between all input and rephrased texts of a given domain.

Semantic similarity. To assess semantic changes incurred during rephrasing, we create paragraph embeddings of texts before and after rephrasing, and compute the cosine similarity. We use two models to create embeddings: sentence-BERT (Reimers and Gurevych, 2019) and ALBERT (Lan et al., 2020), with default hyperparameters. For ALBERT, we take the embedding of the `[CLS]` token as sentence representation (Choi et al., 2021), while for Sentence-BERT, which is specifically designed for sentence-level representations, we directly use the provided sentence embeddings.

Conicity. As a metric for assessing the dispersion of a set of vectors, conicity can be applied to measure the spread of token vectors in the latent space of a language model (Chandrahas et al., 2018). Intuitively, if one were to construct the smallest cone that contains all embedding vectors and has its apex at the origin, a larger conicity for a set of vectors denotes a lower spread. A larger conicity value is thus correlated with lower semantic variation in the text. We obtain token-level embeddings from two different language models to control for possible model bias, namely BERT-Large uncased (Devlin et al., 2019), and GPT-2 XL (Radford et al., 2019). We use the HuggingFace Transformers implementations with default hyperparameters and extract as embeddings one vector representation per token in the input sequence from the final hidden layer.

5 Results

In the following, we discuss text changes due to rephrasing by writing assistants (WATs) and language models (LLMs) from the perspective of word-based measures and vector-based measures.

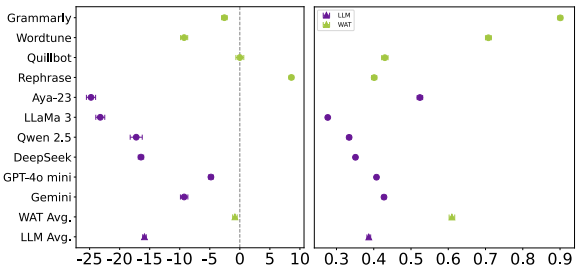

Figure 2: Percentual difference in text length after rephrasing (left) and Jaccard similarity between original and rephrased texts (right) for WATs (green) and LLMs (purple), broken down by tool. Error bars denote 99% confidence intervals.

5.1 Word-based Measures

When considering the variation in paragraph length due to rephrasing (see Figure 2, left), we find that, on average, WATs tend to not change the length of texts substantially: while Wordtune shortens the texts by 9.3%, Rephrase extends them by 8.6%, and the other tools enact little change. In contrast, LLMs consistently shorten the texts: Aya shortens most drastically by 24.8%, both GPT and Gemini shorten text to a comparable degree as WATs, with GPT producing the least shortened output among LLMs. These findings are consistent across different text domains (see Figure 3, left), with all tools shortening texts regardless of domain, with the sole exception of encyclopedia texts that are extended by WATs (in particular: Rephrase). Here as well, WATs shorten the texts to a significantly lesser degree than LLMs.

Considering the corpus-level vocabulary size (see Figure 1), we find that WATs slightly decrease the vocabulary size, with the exception of Rephrase, which slightly increases it. With the exception of Gemini and DeepSeek, LLMs also consistently

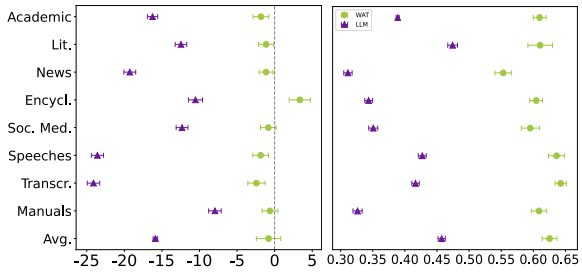

Figure 3: Percentual difference in text length after rephrasing (left) and Jaccard similarity between original and rephrased texts (right) for WATs (green) and LLMs (purple), broken down by domain. Error bars denote 99% confidence intervals.

decrease the vocabulary size – with extreme reductions of up to 29% by the open-weight LLMs. In contrast, among the closed-weight commercial models, both Gemini and DeepSeek increase the vocabulary size overall, and up to 15% on individual domains. The behavior of tools is relatively consistent across domains, although there are variations, in particular for speeches and transcripts. Particularly interesting is the fact that commercial LLMs increase the vocabulary size within some domains much more than they do on the corpus level, indicating a cross-domain linguistic shift towards the LLMs' own inherent vocabulary.

With regard to vocabulary overlap (see Figure 2, right), we find the changes resulting from almost all tools to be significant. Only Grammarly (Jaccard score of 0.90) and Wordtune (Jaccard score of 0.71) retain a relatively strong overlap with the original texts, while all other tools have Jaccard scores below 0.6. Aya is the only LLM that induces less change than two of the WATS. On average, the change in vocabulary is more pronounced for LLMs than it is for writing assistants. We again find these results to be consistent across the different domains (see Figure 3, right).

The results we obtain when using Levenshtein distance are strongly correlated with the Jaccard scores (Pearson correlation $\rho = 0.81$), so we omit the results here and include details in Appendix H.

5.2 Vector-based Measures

When considering semantic similarity between original and rephrased texts on the basis of vector embeddings (see Table 1), we find a consistently stronger deviation from the original text for LLMs than for WATs, with the exception of Rephrase. Among LLMs, semantic divergence is stronger

Type	Tool	SBERT	ALBERT
WATs	Grammarly	0.9873	0.9950
	Wordtune	0.9539	0.9924
	Quillbot	0.9382	0.9883
	Rephrase	0.8921	0.9641
	WAT avg.	0.9429	0.9850
LLMs	Aya-23	0.8979	0.9806
	LLaMa 3	0.8727	0.9731
	Qwen 2.5	0.8863	0.9765
	DeepSeek	0.9246	0.9795
	GPT-4o mini	0.9416	0.9848
	Gemini	0.9132	0.9780
	LLM avg.	0.9059	0.9787

Table 1: Cosine similarity between paragraph embeddings of original and rephrased texts.

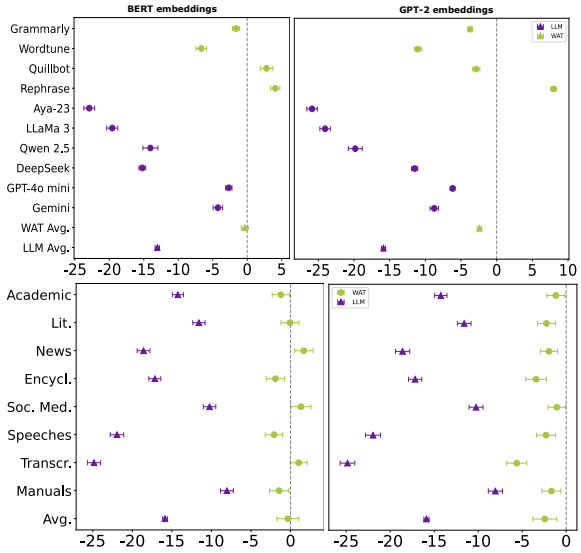

Figure 4: Percentual changes in conicity after rephrasing with WATs (green) and LLMs (purple), using BERT embeddings (left) and GPT-2 embeddings (right). Error bars denote 99% confidence intervals.

among open-weight models than for commercial models when using SBERT embeddings. However, we also find that these differences are potentially model-specific, as ALBERT embeddings consistently indicate very minor semantic changes.

These results hold when considering the domains of texts (see Table 4 in Appendix F), where we observe no stark outliers by semantic divergence for WATs. In contrast, rephrasing with LLMs leads to stronger divergence overall, in particular for literary text, social media posts, interview transcripts, and speeches.

To measure the dispersion of rephrased texts in the embedding space, we consider the changes in conicity (see Figure 4, top), which indicates that, on average, LLMs cause an increase in the semantic spread of the texts (i.e., LLM outputs are on average less similar, independent of the input), while

WATs cause close to no change. When using BERT embeddings, this difference is mostly caused by the open-weight models, while GPT-4o mini and Gemini perform similar to WATs. For GPT-embeddings, DeepSeek is also similar to WATs. Otherwise, these observations are consistent for both embedding models. Overall, the variation in conicity scores is higher when using LLMs for rephrasing than WATs. The results are also consistent when considering the text domain (see Figure 4, bottom), where BERT-based embeddings cluster around zero for WATs, while they are slightly negative when using GPT-embeddings. In contrast, conicity consistently decreases for LLMs for both embeddings. An interesting difference occurs for transcripts, which show a strong decrease in conicity for WATs when using GPT-embeddings instead of BERT-embeddings.

For the LLMs, we also show a breakdown of changes based on the used prompt templates in Appendix F, where we find prompts to have differing performance depending on the LLM, but no evidence of generally high prompt sensitivity.

6 Discussion and Conclusion

Our findings indicate three major take-aways.

6.1 LLMs vs. Writing Assistants

Based on the semantic similarity between input and rephrased texts, LLMs should not be considered a drop-in replacement for writing assistants. In our experiments, LLMs consistently enact a significantly stronger change in semantics and vocabulary than writing assistants, providing reason to caution against their indiscriminate use, in particular without considering the domain. However, future work should investigate whether this impact can be mitigated through style-sensitive prompting strategies.

6.2 LLMs Tend to Summarize

In the comparison between WATs and LLMs, we find that the latter are more strongly inclined to reduce the length of the text, despite the token generation constraint being set well above the length of the input texts. We conjecture that this may be the result of our use of neutral rephrasing prompts in combination with summarization being a likely inclusion during instruction tuning of the models.

6.3 Reduction of Linguistic Diversity

Our most drastic finding is the reduction in vocabulary, which is slight but significant for WATs, yet far more pronounced for LLMs. In particular the distribution of strong intra-domain changes in combination with lesser changes at the corpus level suggests that the vocabulary is actively shifted towards the tools' internal default vocabulary. This raises concerns of a vocabulary shrinkage and resulting loss in linguistic diversity as a result of WAT and LLM use in text composition. However, similar to Martínez et al. (2024), we find that this change is dependent on the model and thereby appears avoidable if suitable design decisions or usage patterns are encouraged, indicating that further research is necessary to prevent an incidental yet avoidable loss of linguistic diversity.

7 Limitations

As a first step in quantifying the impact of using writing assistants and LLMs for rephrasing, our experiments reveal some limitations that should be addressed in future work.

7.1 Prompting of LLMs and Style

While we experimented with prompting variations for zero-shot rephrasing with LLMs, we exclusively focused on plain prompts and avoided text style requests such as simplification or domain-specific adaptation. Intuitively, one would expect differences in vocabulary changes when more specific prompts are used, for better or worse.

7.2 Corpus Limitations

Although we included a wide range of domains in our corpus, it is far from comprehensive. Future work should expand upon this selection of domains and further investigate domain dependence of linguistic diversity reduction. Similarly, our results are restricted to English, and further languages should be considered. Furthermore, rephrasing of texts by domain experts may lead to differences in WAT-assisted rephrasing.

7.3 Experimental Scale

Due to our focus on writing assistants that required semi-manual rephrasing of texts, the size of our corpus is limited to what could feasibly be processed, as it already required days of manual labor-intensive rephrasing. However, our findings indicate that a larger corpus may be necessary to fully quantify the impact of vocabulary shift as a result of exclusively using LLMs vs. settings with a human in the loop. Since there are no constraints to the size of the data for LLM-based rephrasing beyond

available compute, future work should investigate this phenomenon with a focus on LLMs on a much larger corpus.

7.4 Qualitative Linguistic Analysis

In our exploration, we focus on quantifying the presence of change, but do not investigate the styles of change. A proper, linguistically motivated qualitative evaluation and comparison of LLM and WAT outputs for linguistic change would be of interest, but is both outside the scope of this contribution, and outside our area of expertise. To facility such future work, we make available our entire rephrased dataset.

7.5 Assumption of WAT Competence

To maintain comparability across different tools, we assumed that WATs would only suggest changes when necessary. This pragmatic choice reflects realistic usage scenarios, particularly for non-native speakers or users with lower proficiency. At the same time, it distracts from potential differences in how proactive individual WATs may be in presenting suggestions, which could influence the extent of observed vocabulary change.

7.6 Vocabulary Loss for LLMs

The reduction of lexical diversity that we observe for LLM-based rephrasing could potentially be mitigated through combining prompt engineering or through constrained generation (for example through explicit lexical constraints), which should be investigated in future work.

AI Statement

Language model-based AI tools (ChatGPT) were used as coding assistants in the implementation and as writing assistants in creating a draft of the manuscript. The final version of the manuscript was re-written without AI input.

References

Riza Laras Amyatun and Adhan Kholis. 2023. Can artificial intelligence (AI) like QuillBot ai assist students' writing skills? assisting learning to write texts using ai. *ELE Reviews: English Language Education Reviews*, 3(2).

Sisith Ariyaratne, Karthikeyan P Iyengar, Neha Nischal, Naparla Chitti Babu, and Rajesh Botchu. 2023. A comparison of chatgpt-generated articles with human-written articles. *Skeletal Radiology*, 52(9).

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, Kelly Marchisio, Max Bartolo, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Aidan N. Gomez, Phil Blunsom, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. 2024. Aya 23: Open weight releases to further multilingual progress. *CoRR*, abs/2405.15032.

Omer Aydin, Enis Karaarslan, Fatih Safa Erenay, and Nebojsa Bacanin. 2025. Generative AI in academic writing: A comparison of deepseek, qwen, chatgpt, gemini, llama, mistral, and gemma. *arXiv preprint arXiv:2503.04765*.

Chandrahas, Aditya Sharma, and Partha P. Talukdar. 2018. Towards understanding the geometry of knowledge graph embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL*.

Mia Xu Chen, Benjamin N. Lee, Gagan Bansal, Yuan Cao, Shuyuan Zhang, Justin Lu, Jackie Tsay, Yinan Wang, Andrew M. Dai, Zhifeng Chen, Timothy Sohn, and Yonghui Wu. 2019. Gmail smart compose: Real-time assisted writing. In *International Conference on Knowledge Discovery & Data Mining, KDD*.

Hyunjin Choi, Judong Kim, Seongho Joe, and Youngjune Gwon. 2021. Evaluation of BERT and ALBERT sentence embedding performance on downstream NLP tasks. In *25th International conference on pattern recognition, ICPR*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, and et al. 2024. The Llama 3 herd of models. *CoRR*, abs/2407.21783.

Saman Ebadi, Mina Gholami, and Shokoufeh Vakili. 2023. Investigating the effects of using grammarly in efl writing: The case of articles. *Computers in the Schools*, 40(1).

Raymond Fok and Daniel S Weld. 2023. What can't large language models do? the future of ai-assisted academic writing. In *In2Writing Workshop at CHI*.

John Maurice Gayed, May Kristine Jonson Carlon, Angelu Mari Oriola, and Jeffrey S. Cross. 2022. Exploring an AI-based writing assistant's impact on english language learners. *Comput. Educ. Artif. Intell.*, 3.

Katy Ilonka Gero, Vivian Liu, and Lydia B. Chilton. 2022. Sparks: Inspiration for science writing using language models. In *Designing Interactive Systems Conference, DIS*.

Melissa A Kacena, Lilian I Plotkin, and Jill C Fehrenbacher. 2024. The use of artificial intelligence in writing scientific review articles. *Current Osteoporosis Reports*, 22(1).

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *International Conference on Learning Representations, ICLR*.

Mina Lee, Percy Liang, and Qian Yang. 2022. Coauthor: Designing a human-ai collaborative writing dataset for exploring language model capabilities. In *Proceedings of the 2022 CHI conference on human factors in computing systems*.

Gonzalo Martínez, José Alberto Hernández, Javier Conde, Pedro Reviriego, and Elena Merino-Gómez. 2024. Beware of words: Evaluating the lexical diversity of conversational llms using chatgpt as case study. *ACM Trans. Intell. Syst. Technol.*

Suphakit Niwattanakul, Jatsada Singthongchai, Ekkachai Naenudorn, and Supachanun Wanapu. 2013. Using of jaccard coefficient for keywords similarity. In *International Multiconference of Engineers and Computer Scientists*.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP*.

Alex Reinhart, Ben Markey, Michael Laudenbach, Kachatad Pantusen, Ronald Yurko, Gordon Weinberg, and David West Brown. 2025. Do LLMs write like humans? variation in grammatical and rhetorical styles. *Proceedings of the National Academy of Sciences*, 122(8):e2422455122.

Jasper Roe, Willy A Renandya, and George M Jacobs. 2023. A review of AI-powered writing tools and their implications for academic integrity in the language classroom. *Journal of English and Applied Linguistics*, 2(1):3.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, and et al. 2024. Qwen2.5 technical report. *CoRR*, abs/2412.15115.

Li Yujian and Liu Bo. 2007. A normalized Levenshtein distance metric. *IEEE transactions on pattern analysis and machine intelligence*, 29(6).

A Text Collection

A.1 Data Sources

Literature. We randomly select 5 paragraphs from 20 novels written in the 1950s, which are divided between science fiction and romance, and cover authors from the U.S. and U.K. equally.

Academic papers. For the years 2000, 2010, and 2020, we randomly select 7 papers published in NLP from the ACL Anthology[2] and chose 5 paragraphs from each paper at random.

Encyclopedia texts are extracted randomly from articles in Wikipedia[3] and the Encyclopedia Britannica[4] in equal amounts, using 17 keywords each from politics, climate change, and technology that are randomly generated by ChatGPT.

Instruction Manual texts cover 20 instruction manuals for electronic and non-electronic products, with 5 paragraphs per manual that we downloaded from Manualsrepo[5].

News. To include news articles, we consider the topics of politics, climate change, and technology. For each, we randomly select 7 articles released between 2011 and 2013 by CNN[6] and BBC[7] and extract 5 paragraphs per article.

Social Media texts also cover politics, climate change, and technology. For each topic, we chose 17 posts from Instagram and Reddit, using the same randomly generated search terms as for encyclopedias.

Speeches. To cover politics, we consider transcripts of State of the Union addresses for 7 U.S. presidents. For the topics climate and technology, we use transcripts of 7 TED talks[8] each. Per transcript, we extract 5 paragraphs at random.

Interview Transcripts were collected from celebrity interviews on Collider[9], with half the interview transcripts stemming from native English speakers, and half from non-native speakers.

[2] https://aclanthology.org
[3] https://www.wikipedia.org
[4] https://www.britannica.com
[5] https://manualsrepo.com
[6] https://edition.cnn.com/sitemap.html
[7] https://www.bbc.com/news
[8] https://www.ted.com
[9] https://collider.com

A.2 Selection Criteria

To minimize bias in the data, we defined a set of global criteria applying to all domains, as well as some domain-specific selection and randomization criteria. Global criteria include:

- Ensuring that the length of the paragraphs is roughly similar.
- Ensuring that no special characters are contained in the texts.
- Ensuring that no names appear in the text.
- Ensuring that the text is contiguous and relevant to the domain.
- Restricting the text to English content.

As domain-specific selection criteria, we also consider the following.

Literature. Paragraphs are sampled randomly from the entire book, such that each text has roughly the same length and contains ten sentences. If necessary, adjacent paragraphs are merged to create a text of sufficient length. The inclusion of fictional names is avoided.

NLP Papers. We exclude the abstract of papers from the selection. We avoid the inclusion of formulae or mathematical characters. Texts are selected to be roughly ten sentences long and from contiguous sequences of text in the paper.

Encyclopedias. We avoid text with non-standard symbols or characters. All text samples possess similar lengths and numbers of sentences.

Instruction Manuals. Several neighboring paragraphs from the same section are selected to create each text sample. The number of sentences and the length are kept comparable.We avoid sections that strongly rely on numbered instructions indicating steps or sequences.

News. Texts are chosen to contain content relevant to the selected topic. Each text is a contiguous segment from the article with a consistent length and amount of sentences.

Social Media. Each paragraph is taken from a complete comment or discussion. Paragraphs including non-English phrases are avoided. Comparable text lengths and identical number of sentences is ensured, with no non-standard symbols.

Speeches. Texts are chosen to represent a consistent response to the same issue. Annotations such as applause and laughing from the audience are removed manually. We ensure similar text length and number of sentences, with no non-standard symbols.

Interview Transcripts. We select coherent responses or a question with a subsequent response to constitute each text.

Domain	Politics	Climate	Technology
News	35	35	35
Encycl.	34	34	34
Soc. Med.	34	34	34
Speeches	35	35	35
	2020	2010	2000
Academic	35	35	35
	Romance	SciFi	
Lit.	50	50	
	Entertainment	Tech	
Transcr.	50	50	
	Electr.	Mech.	
Manuals	50	50	

Table 2: Counts of texts by domain. Each domain contributes roughly 100 texts to the corpus.

A.3 Data Composition

The final corpus consists of 819 paragraph-length texts. For an overview, see Table 2.

B Rephrasing Criteria

For rephrasing texts with writing assistants, we follow a set of general criteria. All writing assistants are used in their default settings. Each paragraph is rephrased separately to maintain the original meaning and structure, without reformatting citations or lists. The length of paragraphs is designed to allow us rephrasing without having to split texts.

Additionally, we use specific settings for each of the assistants to make their performance as close as possible, despite differing features.

- **Grammarly**[10] All services are utilized. Rephrasing done sequentially from the start to the end of the text. The first option is always chosen, and no reformatting of citations or lists performed. Paragraphs are not split, and rephrasing continues until no further suggestions are made. Extra attention is given to speech and interview texts to avoid the inclusion of recurring, alternating patterns that is sometimes suggested by Grammarly.

[10] https://app.grammarly.com/

ID	Prompt
P1	Rewrite the following paragraph: Paragraph: ⟨input⟩ Rewritten version:
P2	How would you rephrase this paragraph while preserving its original meaning? Paragraph: ⟨input⟩ Rephrased version:
P3	Rephrase the following paragraph without changing the main content: Paragraph: ⟨input⟩ Rephrased version:
P4	Rephrase the following paragraph while preserving its meaning. Follow these steps: 1): Split the paragraph into individual sentences. 2): Rephrase each sentence naturally while keeping the overall flow. 3): Combine the rephrased sentences into a coherent paragraph. Paragraph: ⟨input⟩ Rephrased version:
P5	Imagine you are an advanced language model capable of rephrasing text while preserving its original meaning. If this were your paragraph, how would you naturally rephrase it? Paragraph: ⟨input⟩ Your rephrased version:

Table 3: Prompts used for rewriting with Aya-23, LLaMa 3, Qwen 2.5, DeepSeek, Gemini 2.5 Flash and GPT-4o mini.

- **Wordtune**[11] All services are used, and rephrasing is done by paragraph, not sentence by sentence. The first suggestion is always selected, and rephrasing continues until no further suggestions are provided. Original content is kept for sentences that are too long for the tool to to offer advice.
- **Quillbot**[12] All services are used, and the first generated content is accepted without any manual modifications.
- **Rephrase**[13] All services are used, and the first generated content is accepted without any manual modifications.

C Prompts for Rephrasing

Since the output of LLMs is likely to vary based on the used prompt template, we experimented with five different prompts for rephrasing with increasing complexity, shown in Table 3. In particular, P4 uses a chain-of-thought approach, while P5 appeals to the LLM's agency.

D Model Parameter Settings

In our experiments with LLMs for rewriting, we use the following hyperparameter settings.

Qwen. We used the Qwen-7B-Chat model with sampling enabled (`do_sample = True`), a temperature of `0.7`, and nucleus sampling with `top_p = 0.9`. This configuration encourages diversity while maintaining output relevance.

Aya-23. We used the Aya-23 8B model for text generation with sampling enabled (`do_sample = True`), a temperature of `0.7`, and nucleus sampling with `top_p = 0.9`. The maximum number of new tokens was limited to `200`, and generation was terminated upon reaching either the `eos_token_id` or the maximum token budget. The model operated in 16-bit precision on a single GPU.

LLaMA3. We used the Meta-LLaMA3-8B-Instruct model with sampling enabled (`do_sample = True`), a temperature of `0.7`, and nucleus sampling (`top_p = 0.9`). To discourage repetitive outputs, we applied a `repetition_penalty` of 1.1. The model was loaded in 16-bit precision.

GPT-4o mini. We accessed GPT-4o mini via the OpenAI API, using a temperature of `1.0` to promote response diversity. For each prompt, generation was conducted in a chat format with a single user message, and responses were collected without any further sampling or decoding configuration.

DeepSeek. We accessed the DeepSeek-Chat model via the OpenAI-compatible API, using a temperature of `0.7` and nucleus sampling with `top_p = 0.9`. Prompts were provided in a chat-based format, with each interaction consisting of a single user message. No additional decoding constraints (e.g., top-k, repetition penalty) were applied.

Gemini 2.5 Flash. We accessed the Gemini-2.5-Flash model via the Google API. Generations were produced using the default decoding configuration, without additional constraints such as nucleus or top-k sampling. Prompts were provided as single user messages, and responses were collected in text format.

All open-weight models were run on an NVIDIA A-40 GPU. Across all six models and five prompts, we generated $819 \cdot 6 \cdot 5 = 24{,}570$ rephrased outputs with the language models ($4{,}095$ per LLM).

E Metric Definitions

In the following, we provide further details on the metrics used in the experiments.

[11] https://www.wordtune.com/rewrite
[12] https://quillbot.com
[13] https://www.rephrase.info

	Tool	Academic	Lit.	News	Encycl.	Soc. Med.	Speeches	Transcr.	Manuals
SBERT	Grammarly	0.9866	0.9863	0.9943	0.9935	0.9871	0.9857	0.9794	0.9855
	Wordtune	0.9783	0.9497	0.9572	0.9647	0.9436	0.9392	0.9368	0.9610
	Quillbot	0.9672	0.9116	0.9518	0.9532	0.9314	0.9217	0.9185	0.9485
	Rephrase	0.8563	0.9091	0.9252	0.9047	0.8963	0.8803	0.8624	0.9028
	WAT avg.	0.9471	0.9392	0.9571	0.9540	0.9396	0.9317	0.9243	0.9495
	Aya-23	0.9428	0.8884	0.9095	0.9245	0.8753	0.8665	0.8554	0.9197
	LLaMa 3	0.9161	0.8531	0.9062	0.9069	0.8375	0.8391	0.8163	0.9020
	Qwen 2.5	0.9343	0.8580	0.9150	0.9311	0.8641	0.8427	0.8159	0.9271
	DeepSeek	0.9527	0.9233	0.9482	0.9488	0.9003	0.8916	0.8895	0.9414
	GPT-4o mini	0.9640	0.9293	0.9618	0.9679	0.9245	0.9111	0.9110	0.9416
	Gemini	0.9501	0.9176	0.9397	0.9550	0.8818	0.8806	0.8390	0.9433
	LLM avg.	0.9433	0.8950	0.9301	0.9391	0.8806	0.8719	0.8542	0.9327

Table 4: Cosine similarity between paragraph embeddings of original and rephrased texts, grouped by domain.

E.1 Paragraph length

For texts before rephrasing (B) and after rephrasing (A), we compute

$$length(A, B) = \frac{tokens(A) - tokens(B)}{tokens(B)} \quad (1)$$

with negative values denoting that the number of tokens decreased due to rephrasing. For tokenization, we use NLTK.

E.2 Jaccard similarity

To measure the vocabulary overlap between paragraphs before and after rephrasing, the Jaccard similarity treats both texts as sets of words and computes the overlap as:

$$Jaccard(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (2)$$

with a value of 1 denoting a perfect match in the vocabulary. For preprocessing, we use NLTK to tokenize, convert letters to lowercase, remove punctuation marks, and eliminate stopwords.

E.3 Semantic similarity

To assess the semantic changes incurred during rephrasing, we create embeddings se of paragraphs before (B) and after (A) rephrasing, and compute the semantic similarity as

$$sesim(A, B) = cosine(se(A), se(B)) \quad (3)$$

E.4 Conicity

We use the definition of conicity by Chandrahas et al. (2018). It is based on the concept of alignment to the mean (ATM), which is defined for a vector $v \in V$ in a set of vectors as

$$ATM(v, V) = cosine(v, \frac{1}{|V|} \sum_{x \in V} x) \quad (4)$$

The conicity is then defined as the average ATM over all vectors in V.

$$conicity(V) = \frac{1}{|V|} \sum_{v \in V} ATM(v, V) \quad (5)$$

F Additional Results

F.1 Semantic Similarity by Domain

In Table 4, we show cosine similarities between paragraph embeddings before and after rephrasing, grouped by the domain of the text.

F.2 Text Change by Prompt

To investigate the impact of using different prompts when rephrasing with LLMs, we show the four metrics broken down by prompt in Figure 5.

With regard to **Jaccard similarity**, we find the commercial models to be stable to prompt variations, with the sole exception of chain-of-though prompting for DeepSeek, which deviates slightly. Among the open-weight models, LLaMa and Qwen are also relatively stable, yet Aya-23 shows a strong sensitivity to changes in the prompt. Overall, variations in the prompt template seem to have a very limited effect on vocabulary changes as a result of LLM-based rephrasing.

Considering paragraph **length**, we observe a similar yet more pronounced picture. The commercial models are relatively robust to prompt variations, while the length of rephrased texts of all three open-weight models varies strongly, indicating a much lower consistency.

For dispersion measures via **conicity**, we find an identical effect, independent of the use embedding model: commercial models are stable under prompt template variation, while the conicity of open-weight models fluctuates strongly.

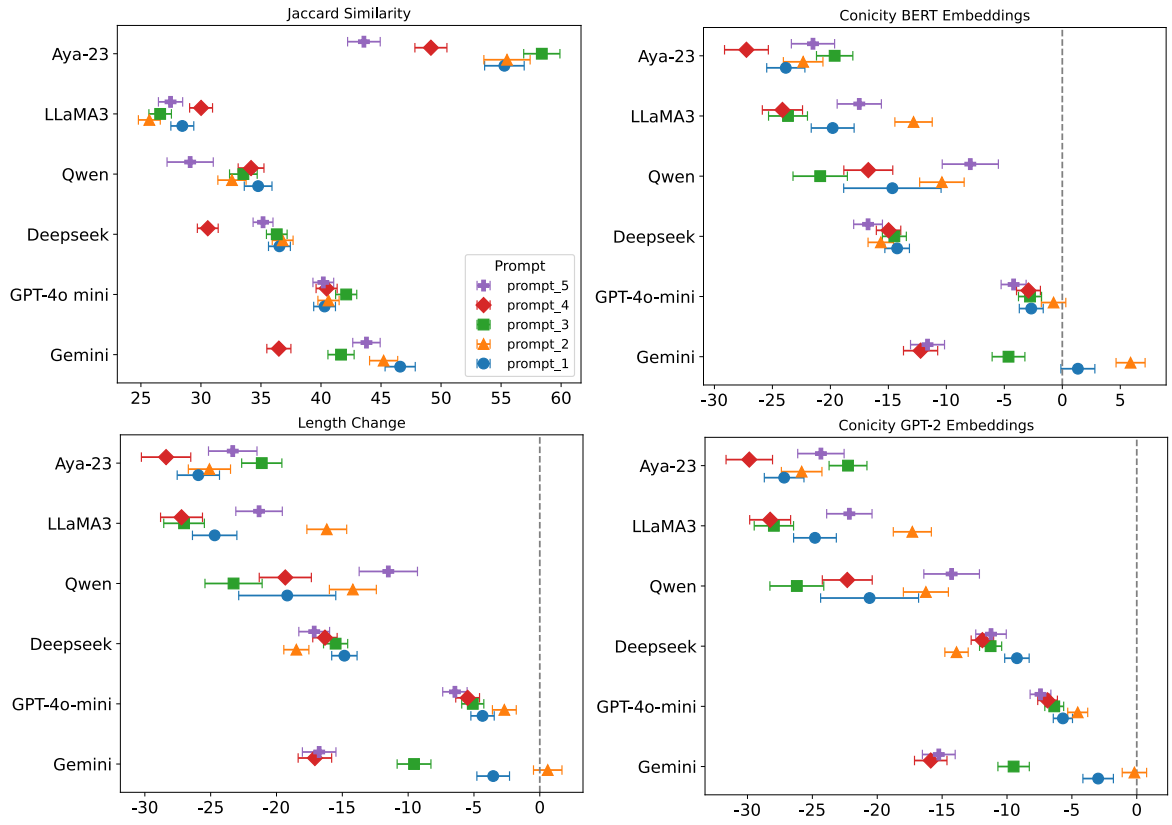

Figure 5: Text change as a result of LLM-based rephrasing, broken down by prompt template. Error bars denote 99% confidence intervals.

Model	P1	P2	P3	P4	P5
Qwen 2.5	27	25	25	26	547
LLaMa 3	1	1	5	1	163
Aya-23	1	0	0	1	7
DeepSeek	0	0	0	12	0
GPT-4o mini	0	0	1	0	1
Gemini 2.5	0	0	0	0	0

Table 5: Number of empty and invalid rephrased LLM outputs among the 819 input paragraphs, broken down by the used prompts 1 to 5.

With regard to the prompt type, we find no observable trend in which one prompt is more or less likely to induce strong text change. In our main experiments, we therefore include all five prompt variations to obtain a broader representation of possible user interactions.

G Manual Output Evaluation

To assess the quality of the LLM-generated rephrased texts, we manually checked them for quality and discarded empty results as well as those containing LLM refusal (e.g., `I am sorry but cannot help you with this`).

The results are shown in Table 5, which summarizes the number of empty or invalid outputs generated by each LLM across the five prompt variants. Overall, we find the performance of all models to be suitable for all prompts, with the exception of Qwen, which also produced non-English responses in a few cases (no other model suffered from this issue). The values we report in Table 5 for Qwen correspond to the total number of unusable responses per prompt, which include non-English responses that occurred 26 times for P1, 25 for P2, 25 for P3, 26 for P4, and 24 for P5. We also note that Qwen failed to properly respond to P5 for 66.8% of inputs while LLaMa failed on 19.9% of input for this prompt template, indicating a stark incompatibility with this promoting style, while other models showed no such issue. Notably, Gemini-2.5 produced consistent outputs across all prompts, without empty results or refusals.

H Levenshtein Distance

To measure the vocabulary-level text change, we also considered Levenshtein distance as a character-level metric, in contrast to the word-level Jaccard

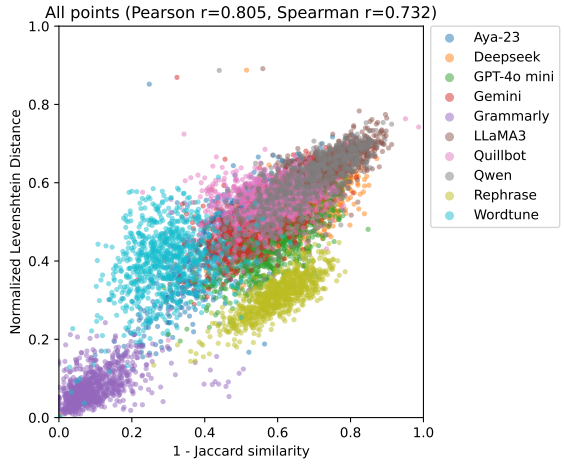

Figure 6: Scatter plot of Levenshtein distance vs. Jaccard similarity for all text pairs.

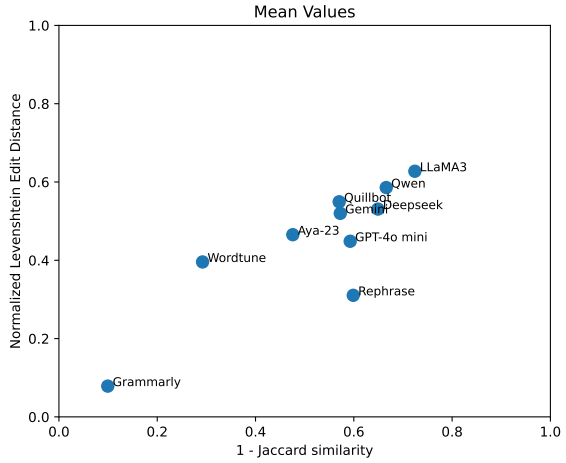

Figure 7: Average Levenshtein distance vs. Jaccard similarity, aggregated by rephrasing tool.

similarity. However, as shown in Figure 6, we find that they are strongly correlated, so we focus on the more intuitive word-level Jaccard similarity in our main evaluation. Nevertheless, the distribution of scores by tool provides an visual representation of the overall level of change that is enacted by the tools. In Figure 7, we therefore also show the averages for each LLM and WAT, which again highlights the much lower change that is induced by WATs when compared to LLMs.

I Rephrasing Examples

To provide an intuition for the data set that we use in our analyses, we provide an example paragraph alongside rephrased versions generated with writing assistant tools in Table 6 and with LLMs using Prompt 2 in Table 7.

Tool	Text
Original	In the afternoon came occupational therapy. The TV screen in each cell illuminated and the patient thrust his hands into the shadow frame of the screen. He saw three-dimensionally and he felt the broadcast objects and tools. He cut hospital uniforms, sewed them, manufactured kitchen utensils, and prepared foods. Although actually he touched nothing, his motions were transmitted to the shops where the work was accomplished by remote control. After one short hour of this relief came the darkness and silence again. But every so often... once or twice a week (or perhaps once or twice a year) came the muffled thud of a distant explosion. The concussions were startling enough to distract Foyle from the furnace of vengeance that he stoked all through the silences. He whispered questions to the invisible figures around him in Sanitation.
Grammarly	In the afternoon came occupational therapy. The TV screen in each cell illuminated, and the patient thrust his hands into the shadow frame of the screen. He saw three-dimensionally, and he felt the broadcast objects and tools. He cut hospital uniforms, sewed them, manufactured kitchen utensils, and prepared food. Although he touched nothing, his motions were transmitted to the shops where the work was accomplished by remote control. After one short hour of this relief came the darkness and silence again. But the muffled thud of a distant explosion came every so often... once or twice a week (or perhaps once or twice a year). The concussions were startling enough to distract Foyle from the furnace of vengeance that he stoked all through the silences. He whispered questions to the invisible figures around him in Sanitation.
Wordtune	Occupational therapy followed in the afternoon. A TV screen illuminated in each cell, and the patient thrust his hands into the shadow frame. The broadcast objects and tools were felt by him in three dimensions. He cut and sewed hospital uniforms, manufactured kitchen utensils, and prepared meals. Even though he did not touch anything, his motions were transmitted to the shops by remote control. One short hour later, darkness and silence returned. Occasionally... (perhaps once or twice a year) I heard the muffled thud of a distant explosion. Through the silences, Foyle stoked the furnace of vengeance despite the concussions that startled him. In Sanitation, he whispered questions to invisible figures.
Quillbot	The second session of occupational therapy began. Each cell's TV screen came to life, and the patient shoved his palms into the screen's shadow frame. He felt the broadcast tools and items as well as saw them in three dimensions. He made cooking utensils, prepared meals, and sewn and cut hospital uniforms. Even though he didn't actually touch anything, his movements were sent to the shops where the job was being done remotely. This relief lasted for only an hour before the darkness and stillness returned. But every now and then, perhaps once or twice a week (or perhaps once or twice a year), a faraway explosion's muffled thud could be heard. The concussions startled Foyle enough to divert him from the raging fire of revenge he had been stoking throughout the silences. He asked queries inaudibly to the shadowy people in Sanitation all around him.
Rephrase	In the afternoon got here occupational therapy. The TV display screen in every mobileular illuminated and the affected person thrust his arms into the shadow body of the display screen. He noticed three-dimensionally and he felt the published gadgets and tools. He reduce health center uniforms, sewed them, synthetic kitchen utensils, and organized foods. Although simply he touched nothing, his motions had been transmitted to the stores in which the paintings become done through faraway control. After one brief hour of this comfort got here the darkness and silence again. But each so often... a couple of times a week (or possibly a couple of times a year) got here the muffled thud of a far off explosion. The concussions had been startling sufficient to distract Foyle from the furnace of vengeance that he stoked at some stage in the silences. He whispered inquiries to the invisible figures round him in Sanitation.

Table 6: Examples of rephrased versions using WATs of an original paragraph from the 1956 novel "The Stars My Destination" by Alfred Bester from the literature set in our data.

Tool	Text
Original	In the afternoon came occupational therapy. The TV screen in each cell illuminated and the patient thrust his hands into the shadow frame of the screen. He saw three-dimensionally and he felt the broadcast objects and tools. He cut hospital uniforms, sewed them, manufactured kitchen utensils, and prepared foods. Although actually he touched nothing, his motions were transmitted to the shops where the work was accomplished by remote control. After one short hour of this relief came the darkness and silence again. But every so often... once or twice a week (or perhaps once or twice a year) came the muffled thud of a distant explosion. The concussions were startling enough to distract Foyle from the furnace of vengeance that he stoked all through the silences. He whispered questions to the invisible figures around him in Sanitation.
Aya-23	The TV screen in each cell lit up in the afternoon and the patient thrust his hands into the shadow frame of the screen. He saw three-dimensionally and felt the broadcast objects and tools. He cut hospital uniforms, sewed them, manufactured kitchen utensils, and prepared foods. Although he actually touched nothing, his motions were transmitted to the shops where the work was accomplished by remote control. After one short hour of this relief came the darkness and silence again. But every so often... once or twice a week (or perhaps once or twice a year) came the muffled thud of a distant explosion. The concussions were startling enough to distract Foyle from the furnace of vengeance that he stoked all through the silences. He whispered questions to the invisible figures around him in Sanitation.
LLaMa 3	Later that day, occupational therapy sessions took place. In each cell, patients used special screens that projected 3D images and allowed them to interact with virtual objects. They could perform various tasks like sewing, cooking, and crafting, even though they weren't physically touching anything. Their movements were transmitted remotely to actual workshops, where the tasks were completed. This respite lasted for about an hour before the darkness and silence returned. Occasionally, however, a faint rumbling sound would interrupt the stillness, jolting Foyle out of his reverie. These sudden explosions served as a distraction from his burning desire for revenge, which he nurtured during the quiet periods. As he sat in the darkness, he would whisper questions to the unseen figures surrounding him, who were presumably connected to the Sanitation department.
Qwen 2.5	Occupational therapy took place in the afternoon, with the TV screens in each cell lighting up and patients placing their hands in the shadows of the screen. They experienced the objects and tools being broadcast in 3D, cutting hospital uniforms, sewing them, making kitchen utensils, and preparing food. Despite not physically touching anything, their actions were remotely controlled to complete tasks in shops. After a brief respite, the darkness and silence returned, but occasionally loud explosions could be heard, which were enough to divert Foyle's attention from his burning desire for revenge. He spoke to unseen figures around him in Sanitation during these moments.
DeepSeek	In the afternoon, occupational therapy began. The TV screens in the cells lit up, and the patient placed his hands into the shadow frame of the display. He perceived the broadcast objects and tools in three dimensions, feeling them as if they were real. He cut and stitched hospital uniforms, crafted kitchenware, and prepared meals. Though he never physically touched anything, his movements were relayed to workshops where the tasks were completed remotely. After just an hour of this brief respite, darkness and silence returned. Yet occasionally, once or twice a week, or maybe even once or twice a year, a distant explosion would resonate with a dull thud. The shockwaves were jarring enough to momentarily pull Foyle's focus away from the burning vengeance he nurtured during the long silences. He murmured questions to the unseen figures around him in Sanitation.
GPT-4o mini	In the afternoon, occupational therapy began. The TV screens in each room lit up, and the patient reached his hands into the dark frame of the screen. He experienced a three-dimensional view and interacted with the objects and tools being broadcasted. He cut and sewed hospital uniforms, created kitchen utensils, and prepared meals. Although he didn't physically touch anything, his actions were transmitted to the workshops where the tasks were completed remotely. After just an hour of this distraction, darkness and silence returned. However, occasionally, once or twice a week or perhaps even once or twice a year, the distant sound of an explosion would echo through the air. These concussions were sufficient to pull Foyle's attention away from the simmering desire for vengeance that he nurtured during the quiet moments. He whispered questions to the unseen figures around him in Sanitation.
Gemini 2.5	Afternoons were dedicated to occupational therapy. Each cell's TV screen would light up, and the patient would insert his hands into its shadow frame, enabling him to see and feel broadcast objects and tools three-dimensionally. He engaged in tasks like cutting and sewing hospital uniforms, crafting kitchen utensils, and preparing meals. Despite the immersive experience, he made no physical contact; his movements were transmitted to workshops where the work was completed remotely. After just one hour of this virtual respite, darkness and silence would return. Occasionally, however—perhaps a couple of times a week or even yearly—a muffled distant explosion would occur. The resulting concussions were jarring enough to pull Foyle from the intense vengeful thoughts he nurtured during the quiet hours, prompting him to whisper questions to the unseen presences he perceived in Sanitation.

Table 7: Examples of rephrased versions using LLMs with prompt template 2 of an original paragraph from the 1956 novel "The Stars My Destination" by Alfred Bester from the literature set in our data.