

ExpertEase: A Multi-Agent Framework for Grade-Specific Document Simplification with Large Language Models

Anonymous EMNLP submission

Abstract

Text simplification is crucial for making texts more accessible, yet current research primarily focuses on sentence-level simplification, neglecting document-level simplification and the different reading levels of target audiences. To bridge these gaps, we introduce ExpertEase, a multi-agent framework for grade-specific document simplification using Large Language Models. ExpertEase simulates real-world text simplification by introducing expert, teacher, and student agents that cooperate on the task and rely on external tools for calibration. Experiments demonstrate that this multi-agent approach significantly enhances LLMs' ability to simplify reading materials for diverse audiences. Furthermore, we evaluate the performance of LLMs varying in size and type, and compare LLM-generated texts with human-authored ones, highlighting their potential in educational resource development and guiding future research.

1 Introduction

Text Simplification aims to make complex texts easier to understand while preserving their original meaning (Chandrasekar and Srinivas, 1997). Given the varying reading and comprehension abilities of different readers, tailoring texts to the target audience's needs is an integral part of this task (Scarton and Specia, 2018; Agrawal and Carpuat, 2023). For instance, in education, instructional materials should cater to students of various ages and cognitive levels to support learning and development.

There exists a significant gap between current text simplification research and practical applications. Existing studies mainly focuses on sentence or segment level simplification (Scarton and Specia, 2018; Maddela et al., 2021; Agrawal and Carpuat, 2023; Kew et al., 2023), with limited attention paid to document-level simplification, despite its extensive demand in real-world scenarios. Furthermore,

numerous prior studies have trained models to simplify complex samples into their simpler counterparts using monolingual parallel corpora (Zhao et al., 2018; Vu et al., 2018; Alissa and Wald, 2023), without considering the specific level of simplification. Moreover, current methods rarely incorporate user feedback into the simplification process.

In this paper, we introduce ExpertEase, a multi-agent approach to simulate real-world text simplification. As shown in Figure 1, ExpertEase comprises expert, teacher, and student agents built upon Large Language Models (LLMs). The expert agent utilizes its linguistic knowledge and example materials to generate simplified texts, while referring to readability analysis tools for calibration. The teacher and student agents provide feedback on the expert's rewritten results from different perspectives, helping to better adapt texts to users' reading levels. In the experiments, we introduced different LLMs and conducted a three-stage multi-agent study, yielding three important findings:

- The LLM-based multi-agent approach works effectively in grade-specific document simplification. The combination of expert knowledge, feedback from external tools, and input from the teacher and student agents, each contributing unique insights, significantly enhances LLMs' ability to generate texts at specified readability levels.
- In addition to the agent cooperation, collaboration between models within a single agent is also crucial, as each model has its own strengths and weaknesses. For instance, some models excel at simplification, while others are better at preserving the original meaning.
- Further analysis reveals that LLM-simplified texts exhibit good consistency in readability and linguistic features compared to human-authored ones, but employ different simplification strategies from those used by humans.

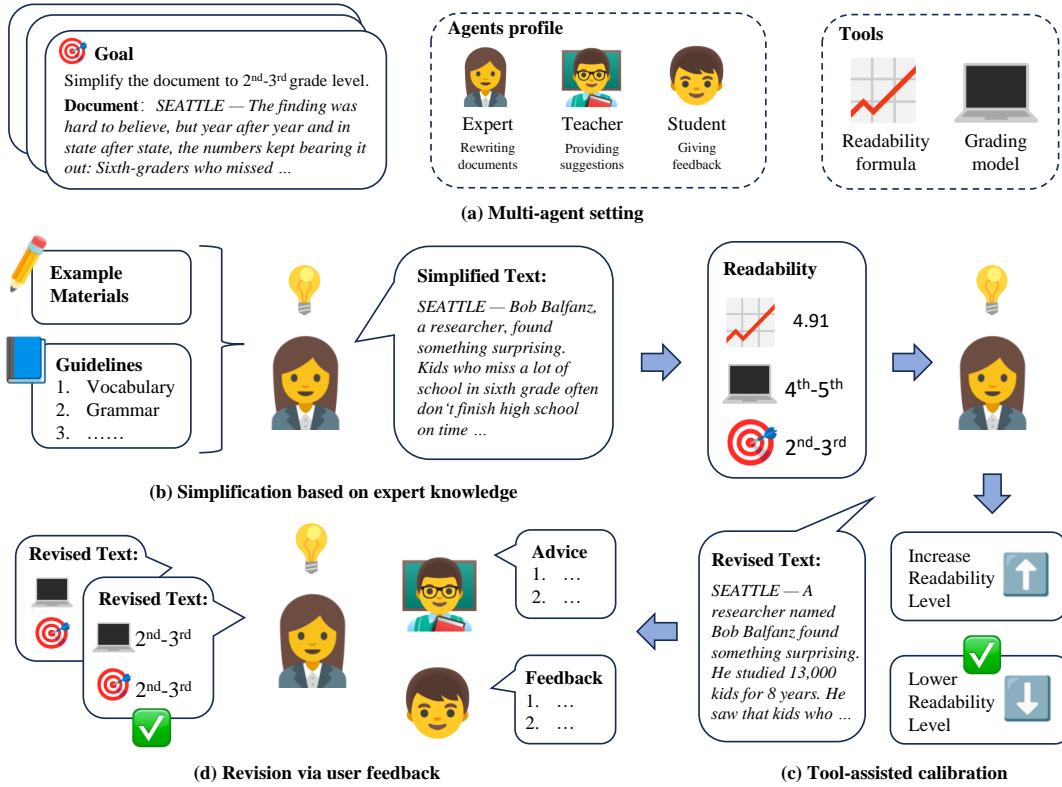


Figure 1: The proposed ExpertEase framework.

2 Related work

2.1 Text Simplification

Previous works primarily focus on the sentence level simplification, using seq-to-seq models for complex-to-simple transformation (Nisioi et al., 2017; Zhang and Lapata, 2017; Zhao et al., 2018; Vu et al., 2018; Alissa and Wald, 2023). Recent research has shifted towards controlling text difficulty. Scarton and Specia (2018) introduced the first seq-to-seq model for grade-specific simplification by annotating sequences with target audience information. Yanamoto et al. (2022) proposes a deep reinforcement learning approach for controllable simplification. In addition, Agrawal et al. (2021) and Agrawal and Carpuat (2022) suggest a non-autoregressive model for iterative input sequence editing to achieve level-controllable simplifications.

Document-level simplification has received much less attention, with the limited existing work primarily focusing on complex-to-simple transformations as well (Sun et al., 2021, 2023; Cripwell et al., 2023). Attempts at controlling readability levels in document simplification using sequence-to-sequence models have yielded suboptimal results (Alva-Manchego et al., 2019), highlighting the need for further research in this area.

Recent research suggests that LLMs show promise in text simplification, but their full potential remains untapped. Kew et al. (2023) found that the LLM’s performance in sentence-level simplification matches existing state-of-the-art baselines, with a broader range of editing operations. Additionally, Farajidizaji et al. (2024) demonstrated that zero-shot models like ChatGPT and Llama-2 can modulate text complexity, although achieving target readability remains challenging. Agrawal and Carpuat (2024) observed that prompted LLMs perform adequately but lack the accuracy of supervised systems, while Imperial and Tayyar Madabushi (2023) noted ongoing struggles in comprehending and adhering to prompts.

2.2 Multi-agents Collaboration

Recently, LLM-based multi-agent systems have achieved considerable progress in complex problem-solving and world simulation (Guo et al., 2024). For example, Du et al. (2023) employed multiple language models to propose and debate their responses and reasoning over several rounds, significantly improving mathematical and strategic reasoning performance. Xiong et al. (2023) focused on inter-consistency in commonsense reasoning tasks, observing enhanced inter-consistency through a three-stage debate framework. Tang et al.

(2023) utilized multiple LLM-based agents in collaborative medical diagnosis discussions, outperforming conventional methods across nine datasets. Inspired by these works, our framework incorporates expert, teacher, and student agents to simulate the real-world educational material development for grade-specific simplification task.

3 ExpertEase: Grade-Specific Document Simplification Framework

To effectively leverage LLMs for document-level simplification in educational settings, we propose ExpertEase, a framework that employs multi-agent collaboration to achieve precise, efficient, expert-like simplification. Figure 1 illustrates the multi-agent setting and the simplification pipeline of ExpertEase, which completes the task through three stages: (1) Simplification based on expert knowledge: the expert agent performs the initial simplification by leveraging linguistic knowledge and referring to example materials. (2) Tool-assisted calibration: LLMs receive feedback from readability tools and adjust the text. (3) Revision via user feedback: teacher and student agents provide feedback from different perspectives, enabling the expert to further refine the texts. This cooperative process among the agents and tools streamlines the simplification workflow.

3.1 Multi-Agent Setting

In real-world scenarios, experts develop learning resources and evaluate their effectiveness with target users to identify areas for improvement (Harsono, 2015). To simulate this process, we introduce expert, teacher, and student agents as shown in Figure 1(a). Furthermore, inspired by Xu et al. (2015)’s note that readability metrics can aid humans in fine-tuning simplified texts, ExpertEase introduces two readability tools to provide the expert agent with immediate feedback.

Agent profiling. The expert agent is responsible for text rewriting and adjustment. To equip LLMs with domain expertise, we construct rewriting guidelines using a data-driven approach that identifies the most critical linguistic features, as detailed in Section 3.2. Additionally, the expert agent can refer to grade-specific examples to guide its rewriting process. The student agent focuses on identifying confusing aspects of the text, while the teacher agent concentrates on detecting inappropriate content and proposing revisions.

Agent communication and feedback mechanisms. Our agent communication paradigm encourages cooperation to optimize text simplification. Each agent specializes in their designated task and interacts only with adjacent layers. The feedback mechanisms incorporate tool feedback and agent interaction. As illustrated in Figure 1(c), tool feedback assists the expert model in adjusting the text based on grade level. Agent interaction, depicted in Figure 1(d), involves the expert agent rewriting the text based on feedback received from the teacher and student agents.

3.2 Simplification based on Expert Knowledge

As depicted in Figure 1(b), the expert knowledge is derived from rewriting guidelines and example materials. To formulate the guidelines, we employed linguistic analysis tools, including TAALED (Kyle et al., 2021), TAALES (Kyle and Crossley, 2015; Kyle et al., 2018), TAASSC (Kyle, 2016), and TAACO (Crossley et al., 2016, 2019). These tools analyze texts across various grade levels, concentrating on aspects such as vocabulary (lexical diversity and sophistication), syntax (diversity and complexity of phrasal and clausal structures), and cohesion. We selected critical features exhibiting high correlation coefficients with the text grade levels, encompassing the number of words, corpus frequency-based lexical sophistication, dependents per clause and per nominal, and adjacent two-sentence overlap lemmas¹.

Furthermore, we constructed linguistically informed text simplification guidelines based on these features, as outlined in Table 1. Additionally, we introduced human-authored rewriting examples, which collectively provide guidance for LLM agent to function as an expert in text simplification.

3.3 Tool-Assisted Calibration

As previously mentioned, readability metrics can aid humans in fine-tuning simplified texts (Xu et al., 2015). Aligning with this notion, we introduce two distinct readability tools to assist the expert agent. As depicted in Figure 1(a), the first tool is the widely adopted Flesch-Kincaid Grade Level (FKGL) analyzer (Kincaid et al., 1975), which outputs a readability score based on the number of words, syllables, and sentences in a text. Recognizing that FKGL only employs surface language

¹These features were selected based on data from the Newsela corpus, while the feature selection can be applied to other datasets containing annotated text levels.

Dimensions	Guidelines
Text Length	• Remove redundant information and irrelevant details from the text, retaining the main content.
Lexical density and cohesion	• Reduce the use of function words in the text to make the meaning clearer and more concise.
Syntactic complexity	• Break down complex sentences by avoiding clauses, conjunctions, and nesting whenever possible. Reduce the use of modifiers in phrases.
Lexical complexity	• Rewrite the text using simple vocabulary , replace uncommon words with high-frequency ones , and reduce the lexical complexity and diversity in the text. • Increase readability by explaining complex concepts in the text using simple, common words .

Table 1: The linguistically informed text simplification guidelines for the expert agent. We highlighted the parts related with linguistic features in blue.

features, we introduce a second tool: a pre-trained grading model that predicts specific grade levels based on deep text representations².

Upon completion of the first round of simplification by the expert, we employ these two tools to analyze the simplified text. If the text falls outside the target grade range, the expert agent receives calibration instructions, which include the outputs of the two tools, the target grade level, and suggestions for further revision. The process is illustrated in Figure 1(c), and the detailed prompts are provided in Appendix A.7.

3.4 Revision via User Feedback

As illustrated in Figure 1(d), the final round of simplification involves rewriting based on user feedback. Specifically, the student agent identifies the words, phrases, sentences, or ideas that they find most confusing, too advanced, or inappropriate for their reading level (e.g., 4th or 5th grade). Meanwhile, the teacher agent, drawing from their teaching experience, pinpoints the items that would likely pose the greatest difficulty for typical students at a given grade level, even if some advanced students could decipher them from context. For each item, the teacher agent explains why it is challenging for that grade level and proposes revisions to make the content more age-appropriate while still conveying the core concepts. In the rewriting process, the expert agent incorporates feedback from both agents, ensuring that changes meet their needs while minimizing alterations to other parts of the text. Further details can be found in Appendix A.8.

4 Experiments

4.1 Datasets

To implement and evaluate the proposed method, we utilized three corpora in our experiments:

²See details of the grading model in Appendix A.1.

Newsela³(Xu et al., 2015), Weebit (Vajjala and Meurers, 2012), and CLEAR (Crossley et al., 2022), as listed in Table 2.

Dataset	Grade/Age	class	Texts
Newsela	grade 2-12	5	9565
Weebit	age 7-16	5	3122
CLEAR	grade 3-12	-	4724

Table 2: Dataset Statistics

Newsela consists of 1,911 news articles, each simplified at least 4 times by professional editors for children at different grade levels. We employed 10% of the Newsela articles as the test set for our grade-specific document simplification task, yielding 191 original articles, and 761⁴ human simplified references.

To construct the grading model, an essential tool in ExpertEase, we used the remaining 80% of the Newsela data for training, and 10% for validation. The Weebit corpus was also introduced into the training set⁵. The aforementioned Newsela test set and the CLEAR corpus, which contains human-assessed readability scores, are used to evaluate the grading model’s performance.

4.2 Models

We built the ExpertEase framework with various LLMs, including both commercial and open-source models of different sizes. For commercial models, we evaluated OpenAI’s GPT-3.5, GPT-4, and GPT-4o, as well as Anthropic’s Claude3-haiku and Claude3-sonnet. For open-source models, we selected Llama3-8B, Llama3-70B, Mixtral-8x7B,

³We used the January 29, 2016 version: <https://newsela.com/data/>. Each article has one original and 4 or 5 simplified versions, but 38 articles do not have 4-version. Additionally, we combined the 42 articles with 5-version into the 4-version set and excluded non-English data.

⁴3 articles lacked 4-version human references.

⁵We mapped its age groups into Newsela’s four version labels, as detailed in Appendix A.2.

and Gemma-7B. The specific settings of the models can be found in Appendix A.4.

4.3 Evaluation

We evaluated model performance across four dimensions: simplicity, readability differences, semantic consistency, and content preservation⁶.

Simplicity Commonly used metrics like SARI (Xu et al., 2015) and D-SARI (Sun et al., 2021) are not aligned with specific grade levels, while FKGL is criticized for poor robustness and limited generalizability (Tanprasert and Kauchak, 2021; Crossley et al., 2022). Therefore, we trained a grading model using Longformer-base-4096 (Beltagy et al., 2020) to more accurately measure grade-specific simplification effectiveness⁷. As presented in Table 3, compared to FKGL⁸, the Longformer-based grading model achieves much higher accuracy and F1 score on the Newsela test set, and exhibits a clearly stronger correlation coefficient with the readability rankings from the CLEAR corpus.

Grading Model	Accuracy	F1	Correlation
FKGL	43.63	38.28	0.5165
Longformer	87.88	87.78	0.6131

Table 3: The performance of Longformer model and the FKGL method on the Newsela test set and CLEAR.

Readability Difference Although FKGL does not align precisely with absolute grade levels, it effectively measures relative differences between texts. Therefore, we assess the readability differences between the model’s output and the human reference by calculating the mean and standard deviation of their FKGL score differences.

Consistency We measured the semantic consistency between the model’s output and human-written text using two metrics: (1) ROUGE-1 score (Lin, 2004), which demonstrates high alignment with human ratings in text summarization tasks (Scialom et al., 2021); and (2) text embedding similarity, computed via text-embedding-3-large, OpenAI’s top performing embedding model.

⁶We did not include fluency metrics since texts generated by LLMs inherently exhibit high levels of fluency.

⁷See Appendix A.1 for training parameters and settings.

⁸The correspondence between FKGL scores and grade levels is referenced in the Common Core Standards. See Appendix A.3 for details.

Preservation Content preservation was evaluated using the text embedding similarity (based on text-embedding-3-large) between the model’s output and the original text. As the model simplified texts to four different versions, we calculated the average similarity for each version and reported the range across version 1 to 4.

4.4 Results

Figure 2 presents the overall results of our pipeline, including the simplicity accuracy and semantic consistency of the 58 groups of results achieved by different models across the three stages. It is evident that ExpertEase effectively mirrors the process of human experts in creating grade-specific simplified texts. By integrating expert knowledge, tool assistance, and user feedback, the model’s simplicity accuracy consistently improves while maintaining high semantic consistency with human output.

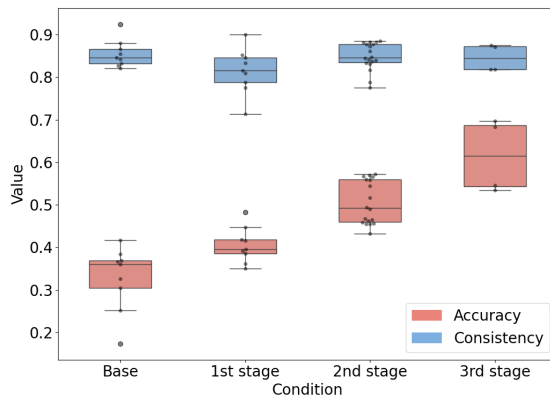


Figure 2: The overall results of the ExpertEase pipeline. Base denotes the simply prompted LLM (see Appendix A.6). The three stages correspond to the introduction of expert knowledge, tool assistance, and user feedback, as illustrated in Figure 1(b)(c)(d).

First-stage results Table 4 presents the detailed results of the first round, where different LLMs played the role of the expert agent. First, linguistically informed guidelines and example effectively enhanced grade-specific simplification accuracy and F1 scores for most models. Moreover, for smaller models, the combined use of both strategies yielded superior results. Secondly, there’s a trade-off between simplicity and retaining meaning, yet most models maintained 0.8 to 0.9 similarity in consistency and meaning preservation. Notably, GPT-4o exhibited exceptional alignment with human references, exceeding 0.9 consistency on average, with ROUGE score surpassing 40%. Claude3-haiku achieved the best overall performance, with an F1 score of 45.13% while main-

Model	Prompt	Simplicity		Readability_Diff	Consistency		Preservation
		Accuracy	F1	*FK_Diff	*Cons_Sim	ROUGE-1	*Pre_Sim
Mixtral-8x7B	base	25.1	29.5	1.49±2.14	0.8656	36.31	0.8358~0.8627
	guidelines	29.96	34.58	1.32±2.11	0.8599	34.91	0.8354~0.8518
	example	39.95	40.58	0.43±1.95	0.7926	30.17	0.7386~0.7697
	guidelines+example	41.52	43.4	0.66±1.95	0.8092	30.85	0.7592~0.7954
Gemma-7B	base	36.93	33.4	1.11±2.07	0.8206	23.41	0.7763~0.8040
	guidelines	39.42	38.66	1.45±2.01	0.8233	25.1	0.7820~0.8131
	example	35.09	29.8	-0.45±2.32	0.6965	21.54	0.6391~0.6765
	guidelines+example	41.79	39.65	1.07±2.10	0.7755	22.93	0.7397~0.7521
Llama3-8B	base	38.37	40.78	1.54±1.84	0.8274	32.57	0.7514~0.8288
	guidelines	40.08	43.6	1.51±1.73	0.8185	34.59	0.7475~0.8118
	example	48.09	48.91	1.04±1.82	0.7577	30.04	0.6467~0.7810
	guidelines+example	48.23	50.42	1.40±1.68	0.7882	31.03	0.6845~0.7877
Llama3-70B	base	41.66	40.98	0.56±1.73	0.8315	33.24	0.7594~0.8202
	guidelines	42.31	40.15	0.16±1.64	0.8184	33.7	0.7410~0.8053
	example	36.01	30.57	-0.21±1.67	0.7797	30.64	0.6918~0.7535
	guidelines+example	36.14	29	-0.53±1.69	0.7135	28.66	0.6323~0.6731
GPT-3.5	base	32.59	30.23	0.43±2.15	0.8455	28.78	0.8024~0.8400
	guidelines	32.59	28.82	-0.01±2.14	0.8435	28.33	0.7908~0.8373
	example	34.82	30.85	0.38±2.11	0.8174	24.82	0.7600~0.8172
	guidelines+example	34.95	30.4	-0.01±2.02	0.8156	24.75	0.7471~0.8181
GPT-4o	base	17.35	20.68	-0.14±1.81	0.9243	50.59	0.8964~0.9380
	guidelines	28.25	30.63	-0.39±1.77	0.9171	48.2	0.8812~0.9319
	example	35.87	37.67	-0.58±1.69	0.9111	45.3	0.8643~0.9257
	guidelines+example	38.5	38.46	-0.71±1.69	0.8997	41.89	0.8545~0.9083
GPT-4	base	30.49	32.14	0.91±2.04	0.8797	36.17	0.8266~0.8792
	guidelines	39.55	41.12	0.36±1.86	0.8759	35.44	0.8203~0.8749
	example	42.44	43.77	0.46±1.87	0.8581	32.46	0.7975~0.8610
	guidelines+example	39.55	36.51	-0.23±1.82	0.8520	31.73	0.7876~0.8560
Claude3-haiku	base	36.01	37.33	1.07±1.98	0.8424	35.32	0.7900~0.8214
	guidelines	34.82	34.9	0.89±1.96	0.8419	34.99	0.7914~0.8183
	example	44.02	44.11	0.95±1.94	0.8270	32.94	0.7725~0.8068
	guidelines+example	44.68	45.13	0.89±1.83	0.8333	33.01	0.7740~0.8119
Claude3-sonnet	base	36.66	36.01	0.54±1.72	0.8550	37.07	0.7929~0.8448
	guidelines	37.32	34.26	0.12±1.74	0.8521	36.61	0.7836~0.8432
	example	44.55	41.42	-0.11±1.64	0.8474	37.12	0.7767~0.8451
	guidelines+example	39.16	34.28	-0.43±1.64	0.8460	36.0	0.7814~0.8326

Table 4: The first-stage experimental results. *Cons_Sim: the similarity between human and model texts; *Pre_Sim: the similarity between original and model texts; *FK_Diff: the mean of FKGL \pm the standard deviation of FKGL. We have **bolded** the best results for each model.

taining 0.83 consistency with human references. Therefore, we selected GPT-4o and Claude3-haiku guidelines+example as the targets for refinement in the second round.

Second-stage results As shown in Table 5, the models significantly enhanced their performance, aided by FKGL and Longformer tools. It is worth noting that collaborative efforts among models can lead to better performance. For example, when GPT-3.5 corrected Claude3-haiku’s output and Gemma-7B corrected GPT-4o’s, accuracy significantly increased, and the models maintained good consistency and meaning preservation. Overall, GPT-3.5 outputs from Claude3-haiku achieved the highest simplification accuracy, while GPT-4 simplified on its own outputs performed best in consistency and preservation. Consequently, we selected these two sets of results for the next round, and they also represent two distinct working mechanisms: multi-model collaboration and individual model processing.

Third-stage results In the third stage, the teacher and student agents provide suggestions and feedback on the second-round rewriting results, and the expert agent then makes revisions based on this information⁹. As shown in Table 6, student and teacher feedback significantly helped the model improve simplicity accuracy and F1 scores, and the resulting readability scores were closer to those of human-rewritten texts. Meanwhile, all models maintained high semantic consistency with human rewrites and effectively preserved the original meaning. Notably, the multi-model collaboration yielded better simplicity accuracy/F1 than the GPT-4o single-model approach, while the latter exhibited greater advantages in consistency and preservation.

⁹According to the pilot study results, GPT-4o was chosen to play all three agent roles due to its strong performance in understanding and following the instructions for each role.

Stage 1	Stage 2	Simplicity		Readability_Diff	Consistency		Preservation
		Accuracy	F1	FK_Diff	Cons_Sim	Rouge F1	Pre_Sim
Claude3-haiku	None	44.68	45.13	0.89±1.83	0.8333	33.01	0.7740~0.8119
	GPT-3.5	57.16 ↑	65.03↑	2.11±2.18	0.8413	30.34	0.7793~0.8381
	Llama3-70B	56.9↑	66.06 ↑	2.30±2.47	0.7883	30.19	0.7521~0.7446
	Gemma-7B	56.64↑	65.29↑	2.71±2.56	0.8326	28.65	0.7692~0.8272
	Claude3-haiku	56.5↑	62.41↑	2.11±1.99	0.8162	31.81	0.7504~0.7941
	GPT-4	55.98↑	65.12↑	3.25±3.56	0.8393	29.43	0.7743~0.8357
	GPT-4o	55.85↑	63.73↑	1.91±1.89	0.8443	31.6	0.7833~0.8392
	Claude3-sonnet	54.4↑	63.98↑	3.26±3.36	0.8374	30.22	0.7719~0.8328
	Llama3-8B	51.64	61.09↑	2.71±2.49	0.7751	30.28	0.7378~0.7294
Mixtral-8x7B	49.41	59.19↑	2.85±2.71	0.8300	29.3	0.7791~0.8219	
GPT-4o	None	38.5	38.46	-0.71±1.69	0.8997	41.89	0.8545~0.9083
	Gemma-7B	49.01 ↑	59.56 ↑	2.15±3.24	0.8721	35.35	0.8316~0.8697
	claude3-haiku	46.78↑	57.19↑	1.69±3.03	0.8774	38.23	0.8334~0.8729
	Llama3-8B	46.52↑	57.47↑	1.81±3.10	0.8610	37.52	0.8215~0.8545
	Claude3-sonnet	46.25↑	57.0↑	2.06±3.40	0.8826	37.84	0.8383~0.8850
	GPT-4	45.86	56.94↑	2.39±3.69	0.8809	36.76	0.8384~0.8832
	GPT-4o	45.6	55.39↑	1.42±2.92	0.8849	38.98	0.8417~0.8874
	GPT-3.5	45.47	55.55↑	1.77±3.04	0.8823	37.53	0.8417~0.8819
	Llama3-70B	45.47	56.26↑	1.79±3.10	0.8465	37.63	0.8203~0.8281
	Mixtral-8x7B	43.23	54.3↑	2.03±3.20	0.8774	37.01	0.8406~0.8751

Table 5: Results from the second stage experiments: tool-assisted calibration. The best results for second stage are highlighted in **bold**, and results with an improvement exceeding 20% are indicated with ↑.

Stage 1-2	Stage 3	Simplicity		Readability_Diff	Consistency		Preservation
		Accuracy	F1	FK_Diff	Cons_Sim	Rouge F1	Pre_Sim
Haiku GPT-3.5	None	57.16	65.03	2.11±2.18	0.8413	30.34	0.7793~0.8381
	GPT-4o-teacher	69.65 ↑	74.73	1.32±1.69	0.8178	31.56	0.7764~0.8320
	GPT-4o-student	68.33	72.99	1.27±1.66	0.8174	31.59	0.7750~0.8296
GPT-4o GPT-4o	None	45.6	55.39	1.42±2.92	0.8849	38.98	0.8417~0.8874
	GPT-4o-teacher	53.35	61.4	0.66±2.18	0.8742	39.92	0.8344~0.8855
	GPT-4o-student	54.53	62.36	0.63±2.14	0.8708	39.91	0.8306~0.8805

Table 6: Results from the third stage experiments: expert revision based on user feedback. The best results for third stage are highlighted in **bold**.

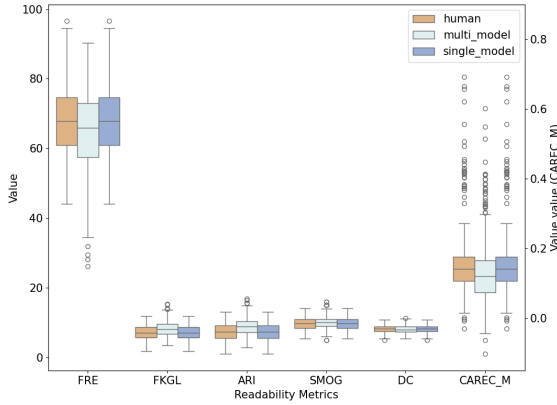


Figure 3: Readability comparison between human and model simplified texts. See metric descriptions in Appendix A.10.

5 Discussion

Experiments demonstrate that ExpertEase, leveraging multi-agent collaboration, achieves progressively better results across three stages (see Appendix A.9 for stage-wise rewriting examples). In this section, we aim to further investigate the consistency and divergence between LLM-generated and human-authored rewrites. To this end, we discuss our findings from three perspectives: readability,

linguistic features, and meaning preservation.

5.1 Readability and Linguistic Features

First, we examine the readability consistency between LLM-simplified and human-simplified texts using a series of readability metrics from the ARTE tool (Choi and Crossley, 2022). Figure 3 illustrates that LLMs closely align with human performance across various metrics, with the single-model GPT-4o approach demonstrating even stronger alignment in this aspect.

Considering that text simplicity and readability are determined by various linguistic features, we decided to conduct a comparative analysis of the texts across different dimensions, including lexical diversity, lexical sophistication, syntactic complexity, and cohesion. For each dimension, we selected three classic linguistic indices based on previous research. We found that humans and models exhibit relatively minor differences in linguistic features as well. They were consistent in lexical and syntactic sophistication. The models demonstrated slightly lower lexical diversity and marginally higher diver-

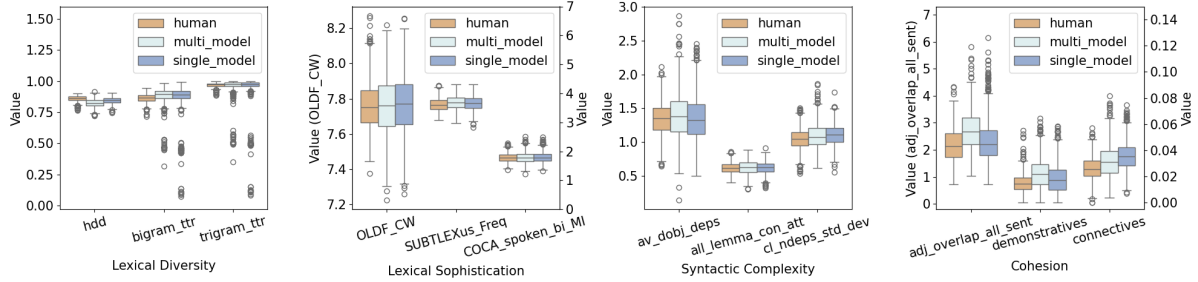


Figure 4: The linguistic features of human and model simplified texts. See feature descriptions in Appendix A.11.

Type	Example A (simplified to the 4th to 5th grade)	Example B (simplified to the 2nd to 3rd grade)
Orig.	... Last year, NASA gave contracts to SpaceX and Boeing to take astronauts to the space station. The contracts were worth \$6.8 billion. Now, it is looking for another company to work with to deliver supplies The athletic shoe and apparel maker said Thursday it will provide free design resources to schools looking to shelve Native American mascots, nicknames, imagery or symbolism. The German company also pledged to provide financial support to ensure the cost of changing is not prohibitive ...
Human	... Last year, NASA gave contracts to SpaceX and Boeing to take astronauts to the space station. The contracts were worth \$6.8 billion. Now, it is looking for another company to work with to deliver supplies Adidas will help schools design new uniforms. It will also help them to design new logos. Logos are the pictures on uniforms or signs. It costs a great deal of money to change logos and mascots. Adidas will help schools pay for it ...
Model	... NASA has given SpaceX and Boeing a lot of money to transport astronauts to the space station. Soon, NASA will choose more companies to deliver supplies, and this new contract will be very valuable They will give free help to schools that want to change their mascots, nicknames, and logos. Adidas will also help pay for the changes so it's not too expensive ...

Table 7: Examples of human and model text simplification. **Bold** denotes complex expressions; **brown** indicates simplified text; **blue** marks newly added text; **red** highlights misinformation.

433 sity in syntactic structures. In terms of cohesion,
 434 the models exhibited higher values in adjacent con-
 435 nections, pronouns, and connectives, resulting in
 436 more coherent expressions.

437 5.2 Meaning Preservation

438 Using text similarity methods, we discovered that
 439 the GPT-4o single-model approach achieves the
 440 highest meaning preservation among the models,
 441 ranging from 0.83 to 0.89. However, this falls sig-
 442 nificantly lower than the similarity between human-
 443 simplified texts and the original texts (0.88 to 0.95).
 444 We further manually analyzed the differences be-
 445 tween the two types of texts and found that the
 446 main reason lies in the fact that model-simplified
 447 texts are generally shorter. This is because humans
 448 tend to make minimal changes. As shown in Exam-
 449 ple A from Table 7, when asked to simplify the text
 450 to a 4-5 grade level, humans made no changes at
 451 all, while the model explained the complex concept
 452 of "contract" using simpler language. Addition-
 453 ally, we discovered that humans sometimes add
 454 sentences to provide supplementary explanations
 455 for the content in the text. For instance, in Example
 456 B, humans explained what LOGO is, whereas the

model directly simplified the original text.

457 6 Conclusion

458 This paper proposed ExpertEase, an effective
 459 framework for grade-specific document simplifi-
 460 cation. By integrating expert knowledge, feedback
 461 mechanisms from external tools, and collabora-
 462 tive inputs from teacher and student agents, our
 463 approach significantly enhances LLMs' ability to
 464 generate texts tailored to specified readability lev-
 465 els. Furthermore, our findings highlight the im-
 466 portance of agent-level and model-level collabo-
 467 rations in achieving superior performance. More-
 468 over, we identified distinct simplification strate-
 469 gies employed by models and humans, suggest-
 470 ing the potential for incorporating human-in-the-
 471 loop simplification pipelines when producing edu-
 472 cational resources. It is worth noting that our frame-
 473 work is universally applicable to various types
 474 of LLMs. Not only can it help achieve efficient
 475 grade-specific simplification capability, but this
 476 multi-agent framework, which incorporates experts,
 477 teachers, students, and tools, can also serve as a
 478 reference for the development of more educational
 479 applications.
 480

7 Limitations

Due to insufficient data, we did not train an end-to-end model to contrast with our framework. Additionally, we did not conduct human evaluation, as Agrawal and Carpuat (2024) did, to assess the multi-dimensional effects of simplifying text. Finally, constrained by test data availability, our experiments primarily focused on the K12 domain; however, our framework has potential applications across various other domains. Moving forward, we aim to expand our framework into more domains and conduct comprehensive evaluations to advance text simplification research.

References

Sweta Agrawal and Marine Carpuat. 2022. An imitation learning curriculum for text editing with non-autoregressive models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7550–7563.

Sweta Agrawal and Marine Carpuat. 2023. [Controlling Pre-trained Language Models for Grade-Specific Text Simplification](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12807–12819, Singapore. Association for Computational Linguistics.

Sweta Agrawal and Marine Carpuat. 2024. Do text simplification systems preserve meaning? a human evaluation via reading comprehension. *Transactions of the Association for Computational Linguistics*, 12:432–448.

Sweta Agrawal, Weijia Xu, and Marine Carpuat. 2021. A non-autoregressive edit-based approach to controllable text simplification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3757–3769.

Sarah Alissa and Michael Wald. 2023. [Text simplification using transformer and BERT](#). *Computers, Materials & Continua*, 75(2):3479–3495.

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2019. [Cross-sentence transformations in text simplification](#). In *Proceedings of the 2019 Workshop on Widening NLP*, pages 181–184, Florence, Italy. Association for Computational Linguistics.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Raman Chandrasekar and Bangalore Srinivas. 1997. Automatic induction of rules for text simplification. *Knowledge-Based Systems*, 10(3):183–190.

Joon Suh Choi and Scott A Crossley. 2022. Advances in readability research: a new readability web app for english. In *2022 International Conference on Advanced Learning Technologies (ICALT)*, pages 1–5. IEEE.

Liam Cripwell, Joël Legrand, and Claire Gardent. 2023. [Context-Aware Document Simplification](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13190–13206, Toronto, Canada. Association for Computational Linguistics.

Scott Crossley, Aron Heintz, Joon Suh Choi, Jordan Batchelor, Mehrnoush Karimi, and Agnes Malatinszky. 2022. [A large-scaled corpus for assessing text readability](#). *Behavior Research Methods*, 55(2):491–507.

Scott A Crossley, Kristopher Kyle, and Mihai Dascalu. 2019. The tool for the automatic analysis of cohesion 2.0: Integrating semantic similarity and text overlap. *Behavior research methods*, 51:14–27.

Scott A Crossley, Kristopher Kyle, and Danielle S McNamara. 2016. The tool for the automatic analysis of text cohesion (taaco): Automatic assessment of local, global, and text cohesion. *Behavior research methods*, 48:1227–1237.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multi-agent debate. *arXiv preprint arXiv:2305.14325*.

Asma Farajidizaji, Vatsal Raina, and Mark Gales. 2024. [Is it possible to modify text to a target readability level? an initial investigation using zero-shot large language models](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9325–9339, Torino, Italia. ELRA and ICCL.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xi-angliang Zhang. 2024. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*.

Y. M. Harsono. 2015. [DEVELOPING LEARNING MATERIALS FOR SPECIFIC PURPOSES](#). *TEFLIN Journal - A publication on the teaching and learning of English*, 18(2):169.

Joseph Marvin Imperial and Harish Tayyar Madabushi. 2023. [Flesch or fumble? evaluating readability standard alignment of instruction-tuned language models](#). In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 205–223, Singapore. Association for Computational Linguistics.

Tannon Kew, Alison Chi, Laura Vásquez-Rodríguez, Sweta Agrawal, Dennis Aumiller, Fernando Alva-Manchego, and Matthew Shardlow. 2023. [BLESS: Benchmarking Large Language Models on Sentence](#)

587	Simplification . In <i>Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing</i> , pages 13291–13309, Singapore. Association for Computational Linguistics.	
588		
589		
590		
591	J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.	
592		
593		
594		
595		
596	Kristopher Kyle. 2016. <i>Measuring Syntactic Development in L2 Writing: Fine Grained Indices of Syntactic Complexity and Usage-Based Indices of Syntactic Sophistication</i> . Ph.D. thesis, Georgia State University.	
597		
598		
599		
600		
601	Kristopher Kyle, Scott Crossley, and Cynthia Berger. 2018. The tool for the automatic analysis of lexical sophistication (taales): version 2.0. <i>Behavior research methods</i> , 50:1030–1046.	
602		
603		
604		
605	Kristopher Kyle and Scott A Crossley. 2015. Automatically assessing lexical sophistication: Indices, tools, findings, and application. <i>Tesol Quarterly</i> , 49(4):757–786.	
606		
607		
608		
609	Kristopher Kyle, Scott A Crossley, and Scott Jarvis. 2021. Assessing the validity of lexical diversity indices using direct judgements. <i>Language Assessment Quarterly</i> , 18(2):154–170.	
610		
611		
612		
613	Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries . In <i>Text Summarization Branches Out</i> , pages 74–81, Barcelona, Spain. Association for Computational Linguistics.	
614		
615		
616		
617	Mounica Maddela, Fernando Alva-Manchego, and Wei Xu. 2021. Controllable Text Simplification with Explicit Paraphrasing . In <i>Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies</i> , pages 3536–3553, Online. Association for Computational Linguistics.	
618		
619		
620		
621		
622		
623		
624	Sergiu Nisioi, Sanja Štajner, Simone Paolo Ponzetto, and Liviu P. Dinu. 2017. Exploring Neural Text Simplification Models . In <i>Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)</i> , pages 85–91, Vancouver, Canada. Association for Computational Linguistics.	
625		
626		
627		
628		
629		
630		
631	Carolina Scarton and Lucia Specia. 2018. Learning Simplifications for Specific Target Audiences . In <i>Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)</i> , pages 712–718, Melbourne, Australia. Association for Computational Linguistics.	
632		
633		
634		
635		
636		
637	Thomas Scialom, Louis Martin, Jacopo Staiano, Eric Villemonte de La Clergerie, and Benoît Sagot. 2021. Rethinking automatic evaluation in sentence simplification. <i>arXiv preprint arXiv:2104.07560</i> .	
638		
639		
640		
	Renliang Sun, Hanqi Jin, and Xiaojun Wan. 2021. Document-Level Text Simplification: Dataset, Criteria and Baseline . In <i>Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing</i> , pages 7997–8013, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.	641
		642
		643
		644
		645
		646
		647
	Renliang Sun, Wei Xu, and Xiaojun Wan. 2023. Teaching the Pre-trained Model to Generate Simple Texts for Text Simplification . In <i>Findings of the Association for Computational Linguistics: ACL 2023</i> , pages 9345–9355, Toronto, Canada. Association for Computational Linguistics.	648
		649
		650
		651
		652
		653
	Xiangru Tang, Anni Zou, Zhuosheng Zhang, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gestein. 2023. Medagents: Large language models as collaborators for zero-shot medical reasoning. <i>arXiv preprint arXiv:2311.10537</i> .	654
		655
		656
		657
		658
	Teerapaun Tanprasert and David Kauchak. 2021. Flesch-Kincaid is Not a Text Simplification Evaluation Metric . In <i>Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)</i> , pages 1–14, Online. Association for Computational Linguistics.	659
		660
		661
		662
		663
		664
	Sowmya Vajjala and Detmar Meurers. 2012. On Improving the Accuracy of Readability Classification using Insights from Second Language Acquisition. In <i>Proceedings of the Seventh Workshop on Building Educational Applications Using NLP</i> , pages 163–173, Montréal, Canada. Association for Computational Linguistics.	665
		666
		667
		668
		669
		670
		671
	Tu Vu, Baotian Hu, Tsendsuren Munkhdalai, and Hong Yu. 2018. Sentence Simplification with Memory-Augmented Neural Networks . In <i>Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)</i> , pages 79–85, New Orleans, Louisiana. Association for Computational Linguistics.	672
		673
		674
		675
		676
		677
		678
		679
	Kai Xiong, Xiao Ding, Yixin Cao, Ting Liu, and Bing Qin. 2023. Examining inter-consistency of large language models collaboration: An in-depth analysis via debate. In <i>Findings of the Association for Computational Linguistics: EMNLP 2023</i> , pages 7572–7590.	680
		681
		682
		683
		684
	Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in Current Text Simplification Research: New Data Can Help . <i>Transactions of the Association for Computational Linguistics</i> , 3:283–297.	685
		686
		687
		688
		689
	Daiki Yanamoto, Tomoki Ikawa, Tomoyuki Kajiwar, Takashi Ninomiya, Satoru Uchida, and Yuki Arase. 2022. Controllable Text Simplification with Deep Reinforcement Learning. In <i>Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)</i> , pages 398–404,	690
		691
		692
		693
		694
		695
		696
		697

698 Online only. Association for Computational Linguistics.
699

700 Xingxing Zhang and Mirella Lapata. 2017. [Sentence Simplification with Deep Reinforcement Learning](#).
701 In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages
702 584–594, Copenhagen, Denmark. Association for
703 Computational Linguistics.
704
705

706 Sanqiang Zhao, Rui Meng, Daqing He, Andi Saptono,
707 and Bambang Parmanto. 2018. [Integrating Transformer and Paraphrase Rules for Sentence Simplification](#).
708 In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*,
709 pages 3164–3173, Brussels, Belgium. Association
710 for Computational Linguistics.
711
712

713 A Appendix

714 A.1 Grading Model Training and Prediction

715 In this study, we utilize the Longformer-base-4096
716 model to investigate our research hypotheses. The
717 model is trained with an attention window of 512
718 tokens and a maximum sequence length of 2048
719 tokens. The training process employs a learning
720 rate of 1e-05, with weight decay set to 0.01. We
721 utilize mixed precision training (FP16) to enhance
722 computational efficiency.

723 The training and evaluation proceed over 10
724 epochs, with a per-device batch size of 3 for both
725 phases. To ensure thorough evaluation, we adopt
726 an evaluation strategy based on steps, assessing per-
727 formance every 50 steps. Correspondingly, model
728 checkpoints are saved every 50 steps, with a maxi-
729 mum of 2 checkpoints retained to manage storage
730 constraints.

731 Early stopping is implemented with a patience
732 of 8, using the evaluation loss as the metric for
733 determining the best model. The training regimen
734 includes a warmup phase comprising 50 steps. A
735 fixed seed value of 4 ensures reproducibility of the
736 results.

737 Logging is configured to capture metrics every
738 50 steps, aligning with our evaluation and saving
739 intervals.

740 The training data comprises corpora from
741 Newsela and Weebit. Specifically, the training set
742 contains 10,773 entries sourced from both Newsela
743 and Weebit, while the test set includes 957 entries
744 exclusively from Newsela.

745 During the prediction phase, the model gener-
746 ates prediction values, which may be continuous
747 real numbers. To convert these continuous pre-
748 dictions into discrete class labels, the following

749 steps are taken: (1) Rounding: The prediction val-
750 ues are rounded to the nearest integers. This step
751 transforms the continuous prediction values into
752 discrete integer values. (2) Clipping: To ensure
753 the prediction values fall within the predefined cat-
754 egory range (i.e., [0, 4]), the rounded predictions
755 are clipped. Specifically, any prediction less than
756 0 is set to 0, and any prediction greater than 4 is
757 set to 4. These processing steps ensure that the fi-
758 nal predicted labels fall within the range of 0 to 4,
759 conforming to the predefined class label range.

760 A.2 Tagging System Alignment

761 Newsela’s Design Referenced the Common Core
762 Standards, with each version corresponding to spe-
763 cific grade levels in the Common Core framework.
764 Therefore, the prompts in experiments translate
765 Newsela’s tagging system into corresponding age
766 groups to facilitate comprehension by large lan-
767 guage models. Additionally, when training the
768 grading model, we mapped the age groups of the
769 Weebits corpus to Newsela’s 0-4 version labeling
770 system. The following tabel 8 shows the correspon-
771 dence between the data tagging systems.

Common Core	Newsela	Weebit
Grades 2-3	4	Ages 7-9
Grades 4-5	3	Ages 9-10
Grades 6-8	2	Ages 11-14
Grades 9-10	1	Ages 14-16
Grades 11-CCR	0	None

Table 8: The correspondence between the data tagging systems

772 A.3 The correspondence between FKGL score 773 ranges and Common Core bands

774 In this research, we employ the FKGL score ranges
775 provided for each grade band as outlined in the
776 Common Core Standards to predict the grade level
777 of texts within the Newsela test set. The correspon-
778 dence between FKGL score ranges and Common
779 Core bands is illustrated in Table 9. Each text in
780 the test set is allocated to the grade level whose
781 FKGL median is nearest to the FKGL score of the
782 text. Further details can be found in the Common
783 Core Standards Appendix A: [New Research on Text Complexity](#).
784

785 A.4 Large Language Models and Experiment 786 Setting

787 For commercial models, we evaluated OpenAI’s
788 GPT series: gpt-3.5-turbo-0125(GPT-3.5), gpt-4-
789 0125-preview (GPT-4), and gpt-4o-2024-05-13

Common Core	FKGL
Grades 2-3	1.98-5.34
Grades 4-5	4.51-7.73
Grades 6-8	6.51-10.34
Grades 9-10	8.32-12.12
Grades 11-CCR	10.34-14.2

Table 9: The correspondence between FKGL score ranges and Common Core bands

(GPT-4o), as well as Anthropic’s Claude series: claude-3-haiku-20240307 (claude3-haiku) and claude-3-sonnet-20240229 (claude3-sonnet). Regarding open-source models, we selected Meta-Llama-3-8B-Instruct (Llama3-8b), Meta-Llama-3-70B-Instruct (Llama3-70b), Mixtral-8x7B-Instruct-v0.1 (Mixtral-8x7b), and gemma-1.1-7b-it (Gemma-7b).

All models were prompted between May 22, 2024, and June 14, 2024. To ensure experiment reproducibility, the maximum response length for all models was set to 4000, and the decoding temperature was set to 0. For the OpenAI models, the random seed parameter was set to 34.

A.5 Sample in Prompt

Segments from the Newsela corpus identified as "10dollarbill-woman" were utilized as the sample text in the prompts. This passage was not included in LLMs test set. In all prompts requiring examples, examples were provided based on the current version of the text and the target level version. For instance, if the current text version is 0 and the target level is at version 3, the provided example would be the corresponding segments from version 0 and version 3 of the article. Table 10 is the specific content of the segments employed in this study.

A.6 Prompts for Initial Expert Simplification

In the first phase, in addition to the two prompts shown below, we also utilized two additional prompt variants: Example and Guidelines. The Example prompt is derived from the Example-Guidelines prompt by removing the guidelines section, while the Guidelines prompt is derived from the Example-Guidelines prompt by removing the examples section.

version	text
0	An abolitionist. The longest-serving first lady. The Labor secretary through the Great Depression. The founder of the Girl Scouts. These are some of the candidates to be the first woman on U.S. currency notes in more than a century. Treasury Secretary Jacob J. Lew announced the plans this week, saying the all-male lineup on American money has gone on long enough. "We will right that wrong, and when the new, redesigned \$10 note is released, it will bear the portrait of a woman," he said at the National Archives in Washington.
1	The all-male lineup on American money has gone on long enough, Treasury Secretary Jacob J. Lew said. "We will right that wrong, and when the new, redesigned \$10 note is released, it will bear the portrait of a woman," he said in Washington, D.C., recently.
2	It’s time for a woman to be honored on American money, U.S. Treasury Secretary Jacob J. Lew said. The all-male lineup has gone on long enough. "We will right that wrong," Lew said. "And when the new, redesigned \$10 note is released, it will bear the portrait of a woman."
3	It is time that a woman be on American money, the head of the U.S. Treasury Department said. "We will right that wrong," Treasury Secretary Jacob J. Lew promised.
4	Pictures of men are on all American paper money. It is time for a woman’s face, a top government worker said. "We will right that wrong," Jacob J. Lew said. He is the head of Treasury Department. It prints money for the American government.

Table 10: examples in prompts

Base

System: You are an helpful assistant.

User: Rewrite the following text into a simpler version that a {9th to 10th} grade student could easily understand. Use simple words and short sentences while preserving the main ideas as much as possible.

Complex: {text}

Simple:

Example-Guidelines

System: You are an helpful assistant.

User: Rewrite the following text into a simpler version that a {9th to 10th grade} student could easily understand. Use simple words and short sentences while preserving the main ideas as much as possible.

You can refer to the following guidelines to modify complex text to a reading level suitable for {9th to 10th grade}:

1. Remove redundant information and irrelevant details from the text, retaining the main content.
2. Reduce the use of function words in the text to make the meaning clearer and more concise.
3. Break down complex sentences by avoiding clauses, conjunctions, and nesting whenever possible. Reduce the use of modifiers in phrases.
4. Rewrite the text using simple vocabulary, replace uncommon words with high-frequency ones, and reduce the lexical complexity and diversity in the text.
5. Increase readability by explaining complex concepts in the text using simple, common words.

Here is an example of rewriting complex text for a {9th to 10th grade} reading level:

Complex: {sample text}

Simple: {sample text}

Complex: {test data}

Simple:

A.7 Prompts for Tool-assisted Calibration

Lower Grade Level

System: You are an helpful assistant.

User: Rewrite the following text, preserving the original meaning but simplifying the language to make it appropriate for {2nd to 3rd grade} students. The current text has a Flesch-Kincaid Grade Level (FKGL) score of {6.15}, indicating it is suitable for {4th to 5th grade}. Lower the FKGL score by:

- Shortening sentences
- Replacing complex vocabulary with simpler synonyms
- Clarifying any confusing or abstract concepts

Your rewritten text should be accessible to 2nd to 3rd grade students while preserving the main ideas.

Source Text: {1st-stage output text}

Rewritten Text:

Increase Grade Level

System: You are an helpful assistant.

User: Rewrite the following text to be suitable for {6th to 8th grade} students while preserving its original meaning. The current text has a Flesch-Kincaid Grade Level (FKGL) score of {8.01}, appropriate for {4th to 5th grade} students. Increase the FKGL score by:

- Combine some short sentences into longer ones, but avoid making them overly complex.
- Incorporate some more advanced vocabulary that is appropriate for {6th to 8th} grade level
- Maintain the text's coherence, logical flow and original meaning.

Focus on hitting the target grade level without unnecessarily overcomplicating the sentence structures or word choices. The rewritten text should be readable and understandable for average students in {6th to 8th} grade, similar to the following reading sample:

{sample text}

Source Text: {1st-stage output text}

Rewritten Text:

A.8 Prompts for Revision via User Feedback

Student Agent

System: you are a typical {4th or 5th grade} student.

User: Please carefully read the following text that was written for students at your grade level.

As you read, make a list of the top 3-5 words, phrases, sentences or ideas that you find most confusing, too advanced, or inappropriate for a {4th or 5th grade} audience. Remember, you don't need to list everything

you don't fully understand - try to use the surrounding context to figure out the meaning if you can. Only list the parts that are very unclear or seem inappropriate for your grade level.

Text: {2nd-stage output text}

List:

Teacher Agent

System: you are a highly experienced middle school teacher who has taught {middle school} reading for many years.

User: Please carefully review the following reading material that experts have written for {middle school} students.

Identify the top 3-5 words, phrases, sentences, or ideas that would likely be most confusing, too advanced, or inappropriate for typical {middle school}, even if some advanced students could figure them out from context. For each problematic item, briefly explain why it's challenging for this grade level and suggest how to revise it to be more age-appropriate while still conveying the core concepts. Consider factors like vocabulary, sentence structure, abstractness of ideas, and developmental fit of examples used. Only list the issues and suggestions. The goal is to ensure the material is accessible and engaging for the vast majority of middle school students.

Text: {2nd-stage output text}

Suggestions:

A.9 The example of model outputs in all stage

original text

Target grade: Grades 4-5

The grade of following text: Grades 11 and above

Text: Last year, NASA gave contracts to SpaceX and Boeing to take astronauts to the space station. The contracts were worth \$6.8 billion. Now, it is looking for another company to work with to deliver supplies.

human simplification

Text: Last year, NASA gave contracts to SpaceX and Boeing to take astronauts to the space station. The contracts were worth \$6.8 billion. Now, it is looking for another company to work with to deliver supplies.

The following outlines the process in which the model gradually refined its output to precisely control reading levels within grades 4-5 over three stages:

1st stage simplification

Target grade: Grades 4-5

The grade of following text: Grades 2-3

Text: NASA gave SpaceX and Boeing a lot of money to take astronauts to the space station. Soon, NASA will choose more companies to take supplies. This new contract will be worth a lot of money too.

2nd stage simplification

Target grade: Grades 4-5

The grade of following text: Grades 6-8

Text: NASA has provided SpaceX and Boeing with significant funding to transport astronauts to the space station. Soon, NASA will select more companies to deliver supplies, and this new contract will be very valuable.

3rd stage simplification

Teacher suggestions:

****Phrase**:** "NASA has provided SpaceX and Boeing with significant funding to transport astronauts to the space station."

- ****Explanation**:** The term "significant funding" might be too advanced.

- ****Revision**:** "NASA has given SpaceX and Boeing a lot of money to take astronauts to the space station." – This is more direct and easier to understand.

Target grade: Grades 4-5

The grade of following text: Grades 4-5

Expert revision: NASA has given SpaceX and Boeing a lot of money to transport astro-

nauts to the space station. Soon, NASA will choose more companies to deliver supplies, and this new contract will be very valuable.

A.10 Readability Indices

Our readability Indices are obtained from ARTE, which automatically calculates a variety of readability formulas for texts. The details of each metric are provided in Table 13. The t-test results for the readability scores of human-generated text and model-generated text are presented in Table 11. For more information, please refer to <https://www.linguisticanalysistools.org/arte.html>.

Column	*T-Statistic_single	*T-Statistic_multi
FRE	-3.65***	-9.97***
FKGL	8.72***	21.65***
ARI	11.78***	24.23***
SMOG	2.27**	10.98***
DC	-2.89**	-3.18**
CAREC_M	-13.66***	-11.08***

Table 11: The t-test results comparing the readability scores between human-generated texts and model-generated texts.*T-Statistic_multi: Represents the T-test result combining multiple models.*TStatistic_single: Indicates the T-test result for a single model. $p < 0.05$ is marked with *, $p < 0.01$ with **, and $p < 0.001$ with ***.

A.11 Linguistic Indices

The indices we used are derived from TAALED, TAALES, TAASSC, and TAACO. These tools are respectively utilized for analyzing lexical diversity, lexical sophistication, syntactic complexity, and cohesion. The t-test results for human-generated text and model-generated text are shown in Table 12. The explanations of the Indices are provided in Table 14, and further details can be found at <https://www.linguisticanalysistools.org/>.

Aspect	Indices	T-Statistic_single	T-Statistic_multi
Lexical Diversity	hdd	-17.40***	-35.78***
	bigram_lemma_ttr	1.17	7.56***
	trigram_lemma_ttr	-4.95***	-3.67***
Lexical Sophistication	SUBTLEXus_Freq_CW_Log	6.85***	8.75***
	OLDF_CW	1.487	-0.734
	COCA_spoken_bi_MI	0.49	0.39
Syntactic Complexity	av_dobj_deps	0.99	3.58***
	all_lemma_con_attested	0.59	1.48
	cl_ndeps_std_dev	8.92***	6.65***
Cohesion	adjacent_overlap_all_sent_div_seg	4.93***	17.53***
	all_demonstratives	9.14***	18.85***
	basic_connectives	22.23***	11.69***

Table 12: The t-test results comparing linguistic metrics between model-generated and human-generated texts. $p < 0.05$ is marked with *, $p < 0.01$ with **, and $p < 0.001$ with ***.

Index name	Formula Name	References
FRE	Flesch Reading Ease Formula	Flesch, R. (1948). A new readability yardstick. <i>Journal of applied psychology</i> , 32(3), 221.
FKR	Flesch Kincaid Grade Level Formula	Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., & Chissom, B. S. (1975). Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.
ARI	Automated Readability Index	Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., & Chissom, B. S. (1975). Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.
SMOG	SMOG Grading	Mc Laughlin, G. H. (1969). SMOG grading-a new readability formula. <i>Journal of reading</i> , 12(8), 639-646.
DC	New Dale-Chall Readability Formula	Chall, J. S., & Dale, E. (1995). <i>Readability revisited: The new Dale-Chall readability formula</i> . Brookline Books.
CAREC_M	Crowdsourced algorithm of reading comprehension modified	Crossley, S. A., Skalicky, S., & Dascalu, M. (2019). Moving beyond classic readability formulas: new methods and new models. <i>Journal of Research in Reading</i> , 42(3-4), 541-561.

Table 13: The details of readability indices

Aspect	Index Name	In Text Name	Explanation
Lexical Diversity	hdd	Hypergeometric Distribution D (HDD)	Assessing the likelihood of unique words in a sample using hypergeometric distribution.
	bigram_ttr	Bigram Lemma TTR	Number of unique bigram lemmas (types) divided by the number of total bigram lemmas (tokens)
	trigram_ttr	Trigram Lemma TTR	Number of unique trigram lemmas (types) divided by the number of total trigram lemmas (tokens)
Lexical Sophistication	COCA_spoken_bi_MI	COCA Spoken Bigram Association Strength (MI)	Mean Mutual Information Score
	OLDF_CW	Average log HAL frequency of closest orthographic neighbors CW	Mean log HAL frequency of a word's 20 closest neighbors; neighbors determined using Levenshtein orthographic distance
	SUBTLEXus_Freq_CW_Log	SUBTLEXus Frequency CW Logarithm	Mean Frequency Score
Syntactic Complexity	cl_ndeps_std_dev	Clause Dependents Standard Deviation	Dependents per clause (standard deviation)
	all_lemma_con_attested	All Lemma Construction Attested Percentage	percentage of lemma construction combinations in text that are in reference corpus - all
	av_dobj_deps	Average Dependents per Direct Object	Dependents per direct object
Cohesion	adjacent_overlap_all_sent	Adjacent Sentence Overlap All Lemmas (sentence normed)	Number of lemma types that occur at least once in the next sentence
	all_demonstratives	Demonstratives	Number of demonstratives
	basic_connectives	Basic Connectives	Number of basic connectives

Table 14: The explanations of the linguistic indices