Towards CEFR-targeted Text Simplification for Question Adaptation

Luca Benedetto, Paula Buttery
ALTA Institute, Dept. Computer Science and Technology, University of Cambridge, UK
{name.surname}@cl.cam.ac.uk

Abstract

Text Simplification (TS) can adapt educational content to learners' proficiency levels. In reading comprehension questions, passage complexity directly affects the question difficulty; thus, TS could enable automatic question adaptation by generating multiple versions of a reading passage. However, despite the potential of TS and its applications in other domains, the feasibility, reliability, and robustness of TS for question adaptation remains unexplored. In this paper, we conduct the first evaluation of LLMs for CEFR targeted text simplification aimed at question adaptation. Specifically, we investigate whether LLMs can perform CEFR-targeted text simplification and how this affects question answerability. Evaluating four LLMs on two English learning datasets, we show that they can mostly perform targeted simplification with readability values correlating with reference CEFR levels, but alignment is imperfect. Crucially, the simplified texts generally preserve the information needed to for question answering, and questions associated with texts simplified at lower levels show reduced difficulty in virtual pretesting. These preliminary findings show the potential of LLMs for educational content adaptation, but practical deployment will need improved CEFR alignment.

1 Introduction

Text Simplification (TS) improves the the accessibility of content for populations with limited literacy and non-native speakers (Al-Thanyyan and Azmi, 2021) in domains such as medicine (Ong et al., 2007; Segura-Bedmar and Martínez, 2017), law (Bouayad-Agha et al., 2009), and others. In educational assessment, TS can ensure that exams evaluate subject mastery rather than language proficiency, which is particularly important in high-stakes exams where the effects of question wordings must be minimised (Yaneva et al., 2019). Also, TS can be used to adapt material to each student's skill, using the *zone of proximal development* concept (Shabani et al., 2010). This work focuses on question adaptation for reading comprehension questions, whose difficulty heavily depends on the complexity of the reading passage (Benedetto et al., 2023; Huang et al., 2018; Benedetto et al., 2020). Thus, generating multiple versions of the same passage at different levels could enable difficulty adaptation of existing questions.

Language proficiency in education is typically measured with scales such as the Common European Framework of Reference for Languages (CEFR). CEFR-targeted TS would potentially allow precise question adaptation – for instance, B2-level questions could be adapted to B1 learners by simplifying the associated passages. Notably, different question types target different proficiency dimensions and simplification may not work equally well on all of them (e.g., at different levels of Bloom's Taxonomy (Bloom et al., 1964)), and excessive simplification risks removing information needed to answer questions. As the first study on this task, we do not focus explicitly on different questions types, but study overall feasibility.

We address the following research questions. i) Can LLMs simplify reading passages to specific CEFR levels, and how do different LLMs perform? ii) How does TS affect the answerability of the associated Multiple-Choice Questions (MCQs)? iii) Are LLMs a promising avenue towards Text Simplification for question adaptation?

Experimenting with both proprietary and open-weight models, we find that most models can indeed perform zero-shot CEFR-targeted simplification. However, while readability values of the simplified texts correlate with reference values, alignment with CEFR levels remains imperfect. Most of the simplified texts preserve the information needed to answer the associated questions, and the questions associated with lower level texts show lower difficulty from virtual pretest-

Proceedings of Recent Advances in Natural Language Processing,pages 150–157
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-018

ing. Our findings suggest that LLMs are a promising avenue towards text simplification for question adaptation, although more work should focus on improved alignment with the CEFR scale. Code, output, and additional material is available at https://github.com/lucabenedetto/cefr-text-simplification-llms.

2 Related Works

This paper builds upon the recent line of work in LLM-based Text Simplification (Al-Thanyyan and Azmi, 2021; Cripwell et al., 2023; Feng et al., 2023; Jamet et al., 2024). We draw particular inspiration from Kew et al. (2023) and their approach to benchmarking LLMs on text simplification; however, we focus specifically on TS for question adaptation in language learning, rather than general TS.

Also, our work is closely related to previous research on controllable TS, in particular the work by Farajidizaji et al. (2024). However, there are two significant differences: we target specific CEFR levels rather than readability scores, and we evaluate the impact that text simplification has on the questions associated with the reading passage.

Our focus on educational assessment and question adaptation connects with previous literature on question evaluation and question difficulty estimation from text (AlKhuzaey et al., 2021). More specifically, previous research explored the relationship between the complexity of a reading passage and the difficulty of the associated questions (Benedetto et al., 2023; Huang et al., 2018). These studies serve as motivation to this paper and support our hypothesis that simplifying reading passages can systematically control question difficulty.

3 Methodology

3.1 Prompting Strategy

We experiment with four zero-shot prompts across two templates, which differ in their simplification approach: i) template *a* provides a direct instruction to perform text simplification, ii) template *b* is a persona-based prompt (Lee et al., 2023) where the LLM acts as a "skilled English teacher". Each template has two versions: with and without explicit CEFR level descriptors from the council of Europe website.[1] The prompts and the CEFR descriptors used are available in the additional material.

[1] www.coe.int/en/web/common-european-framework-reference-languages/table-1-cefr-3.3-common-reference-levels-global-scale

3.2 Evaluation Framework

For evaluation purposes, we combine standard metrics from text simplification and metrics related to our specific objective of question adaptation.

Readability Assessment. We study whether simplified texts achieve appropriate difficulty levels using eight readability indexes, following established practices in TS (Kew et al., 2023). We consider readability indexes previously used in text simplification and educational assessment literature: (Benedetto, 2023; Kew et al., 2023; Huang et al., 2018): Automated Readability Index (Senter and Smith, 1967), Coleman-Liau Index (Coleman, 1965), Dale-Chall Readability Score (Dale and Chall, 1948), Flesch Reading Ease (Flesch, 1948), Flesch-Kincaid Grade Level (Kincaid et al., 1975), Gunning FOG Index (Gunning, 1952), Linsear Write Formula (Klare, 1974), and SMOG Index (Mc Laughlin, 1969). More specifically, we compare the readability distribution of the simplified texts and of the (human curated) passages at different CEFR levels using Earth Mover's Distance[2] to quantify text simplification alignment.

Question Answerability. To assess whether text simplification preserves the information needed for answering the associated comprehension questions, we measure the *answerability* of the reading comprehension questions when using both the original and the simplified passages. We do this by prompting GPT-4o to answer the multiple-choice questions using the two versions of the reading passage and analyse changes in accuracy as indicators of information loss during text simplification.

Vocabulary Alignment. Lastly, to evaluate the lexicon used in simplified texts, we use the CEFR-J vocabulary list,[3] which provides (manually curated) lists of words that are suitable for learners of different CEFR levels. By analysing whether simplified texts use vocabulary at the appropriate levels – and comparing their distributions with the reference datasets – we can evaluate whether LLMs can adapt lexical choices to target proficiency levels.

3.3 Experimental Datasets

Cambridge English Readability Dataset (CERD). It contains 331 reading passages from Cambridge English Exams spanning A2-C2 CEFR levels (Xia et al., 2016). The dataset[4] provides the

[2] https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.wasserstein_distance.html

[3] github.com/openlanguageprofiles/olp-en-cefrj

[4] Available at: ilexir.co.uk/datasets/index.html

expert-curated CEFR level of each reading passage, thus enabling the evaluation of text simplification against this reference. Distribution across CEFR levels is fairly balanced, with 71 reading passages from level B2 (the most represented in the dataset) and 60 for level B1 (the least frequent).

Cambridge MCQs Reading Dataset (Cam MCQ). Released by Cambridge University Press & Assessment (Mullooly et al., 2023),[5] it contains 120 reading passages targeting learners at B1-C2 of the CEFR. Each reading passage is associated with one or more Multiple Choice Questions (MCQs), for a total of 795 MCQs. This dataset enables our question answerability evaluation, since it provides both the passages and the comprehension questions (also, the CEFR levels associated with each text enable the readability-based evaluation, similarly to CERD). The target CEFR levels are not uniformly distributed, with B2 being the most frequent (422 questions from 58 passages) and C2 the least frequent (62 MCQs from 9 reading passages).

3.4 Models

We experiment with four LLMs, considering both commercial and open-weights models: GPT-4o and GPT-4o-mini[6] from OpenAI (specifically using *gpt-4o-2024-08-06* and *gpt-4o-mini-2024-07-18*), Google's Gemma 7B (Gemma Team and DeepMind, 2024) and Meta's Llama 3 8B (Meta, 2024). For the open-weight models we use the weights of the instruction-tuned versions available via the HuggingFace transformers library (Wolf et al., 2020): *google/gemma-7b-it*, and *meta-llama/Meta-Llama-3-8B-Instruct*.

For all models we use zero-shot prompting with identical instruction, adapted for model-specific formatting if needed (e.g., `<|begin_of_text|>` for Llama 3 8B). We do not perform any fine-tuning, use the default temperature settings and limit the outputs to 1000 tokens, which is sufficient for the reading passages considered in this study.

4 Results and Discussion

4.1 Usability of LLM Output

Before diving into the evaluation on simplification, we measure whether the LLMs successfully follow the given instructions, and observe that all four models do indeed produce text which is usable for text simplification. GPT models occasionally fail to follow instructions (on average, 0.2% of the times for GPT-4o-mini and 0.3% for GPT-4o), while open-weight models in this setup show near perfect instruction following.

Model	Prompts a.1	a.2	b.1	b.2	μ
Gemma 7B	3.2	3.1	17.7	24.8	12.2
Llama 3 8B	42.4	36.0	71.6	60.3	52.6
GPT-4o-mini	0.2	0.3	0.5	0.6	0.4
GPT-4o	0.1	0.1	1.2	1.3	0.7

Table 1: Average length (in number of words) of the overhead text added by each LLM. Results are averaged across both datasets and all target CEFR levels.

However, the models differ significantly in verbosity:[7] Table 1 shows that Llama 8B is particularly verbose and produces texts with 52 words of overhead on average, which requires a significant amount of post-processing.[8] On the other hand, GPT models are the least verbose, generally providing at most a minimal and consistent header (*"Text:"*), and Gemma 7B is in between, generally using a very short header (*"**Simplified Text:**"*) but more verbose responses on occasions.

4.2 Readability Evaluation

Main takeaway: *Llama 3 8B and the GPT models successfully performed TS, but tend to oversimplify texts, specifically at higher CEFR levels.*

Our evaluation is based upon eight readability indexes (§3.2), and we study how the values for the simplified texts align with the readability of the reference data at different CEFR levels. We find that the eight indexes follow similar patterns, thus we show in the main body of text only the ARI (Automated Readability Index) as representative (all are available in the additional material).

As a reference, we measure the ARI of the texts in the original datasets, separately for each (expert-curated) CEFR level, and show their distribution in Figure 1. As expected, ARI values increase for higher CEFR levels and there are particularly visible steps from A2 to B1 and from B2 to C1.

Figure 2 shows the ARI distributions for the texts simplified by each model, showing on the horizon-

[5] Available upon request at englishlanguageitutoring.com/datasets/cambridge-multiple-choice-questions-reading-dataset

[6] platform.openai.com/docs/models/gpt-4o-mini

[7] We define *verbosity* as the average difference between the overall length of the responses and the length of the simplified passage (i.e., *verbosity* is the length of the overhead text). Simplified texts are extracted automatically with manually curated regular expressions (available in the additional material).

[8] It provides descriptions of what it did (e.g., "*I simplified the text by: ...*"), and to use very different wording.

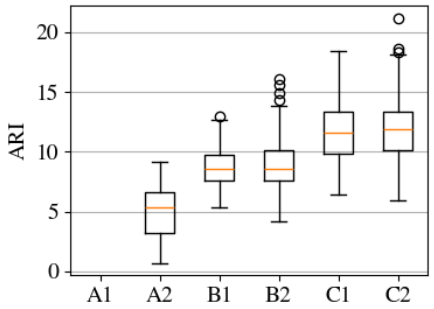

Figure 1: Distribution of the ARI for different CEFR levels in the experimental datasets.

tal axis the target level and on the vertical axis the ARI. Notably, the different prompts we experiment with do not seem to lead to significant differences. Llama 3 8B and the GPT models produce the most promising results, showing increasing ARI values for higher CEFR levels, although without the step-like behaviour visible in reference data (Figure 1). However, even though the median ARI for the texts simplified at A2 level is in line with the reference value, the values for higher CEFR levels (in particular C1) are lower than the reference ones, suggesting that the LLMs tend to over-simplify the reading passages. On the other hand, Gemma 7B has overall poor performance: it produces texts at similar readability levels regardless of the target CEFR, and there is a consistent drop between the ARI at level B2 and C1, which is incorrect.

Table 2 quantifies these observations using Earth Mover's Distance (EMD, a.k.a. Wasserstein Distance) between the ARI distributions of the simplified and reference texts; lower values indicate better alignment.[9] The table shows that Llama 3 8B mini consistently performs as one of the best models, if not the best, followed by GPT-4o-mini and GPT-4o. Gemma 7B performs particularly well for level B1 (with prompts *b.**) and B2, but very poorly for the others, thus suggesting that it mostly produces texts at B1-B2 levels regardless of the request, in line with the results shown in Figure 2. All models show larger EMD values for target C1, thus supporting the intuition that they tend to oversimplify the texts. As for the different prompts, there seem to be no significant differences in performance.

[9]Please note that the datasets do not provide the same texts at different levels of the CEFR, thus we cannot compare the simplified texts with a gold reference simplification.

Model	Prompt	EMD (ARI)			
		A2	B1	B2	C1
Gemma 7B	a.1	1.92	1.40	0.53	3.98
	a.2	1.58	1.56	0.61	4.12
	b.1	2.61	0.88	0.53	3.00
	b.2	2.42	**0.86**	0.47	3.24
	μ	*2.13*	*1.17*	*0.53*	*3.58*
Llama 3 8B	a.1	0.92	1.73	**0.27**	1.94
	a.2	**0.85**	1.75	**0.21**	1.32
	b.1	0.87	1.65	**0.28**	1.95
	b.2	**0.55**	1.55	0.39	**1.04**
	μ	*0.80*	*1.67*	*0.29*	*1.56*
GPT-4o-mini	a.1	1.21	**0.81**	0.71	1.29
	a.2	1.11	0.87	0.80	**1.05**
	b.1	1.37	**0.84**	0.55	1.49
	b.2	1.01	1.08	0.62	**1.28**
	μ	*1.11*	*0.90*	*0.67*	*1.28*
GPT-4o	a.1	1.17	1.07	0.46	2.23
	a.2	1.62	1.20	0.55	2.00
	b.1	1.18	1.00	0.46	2.16
	b.2	**0.83**	1.30	0.45	1.86
	μ	*1.2*	*1.14*	*0.48*	*2.06*

Table 2: Earth Mover's Distance (EMD) between the distribution of the ARI in the simplified texts and in the reference texts, separately for each model-prompt pair. Results are aggregated for the two datasets; in bold the three best results for each target CEFR level.

4.3 MCQ Answerability

Main takeaway: *Text simplification appears to reduce question difficulty as hypothesised, but over-simplification reduces question answerability (likely due to information removal).*

We evaluate[10] MCQ answerability by using GPT-4o to answer MCQs using both the simplified texts and the original ones, since Question Answering (QA) accuracy variations can provide insight into how text simplification affects the answerability of the items. The complete results are presented in Table 3, which shows in the first two rows the QA accuracy i) of the of the random baseline (25%), and ii) of GPT-4o when using the original texts. Notably, the accuracy drops significantly compared to the original texts (from 89-97% to 56-87%, depending on model and level); this might indicate that text simplification consistently removes information needed for some questions,[11] but might also be due to text contamination since we are using a public dataset (GPT-4o might have been trained on the original texts and seen the text/question pairs).

Focusing on the questions which are still answerable, QA accuracy decreases as target CEFR level increases (with the exception of A1). This

[10]The CERD dataset does not provide questions, hence we perform this evaluation on Cam MCQ only.

[11]Possibly due to the tendency to *over-simplify* the texts, or to text simplification not being possible for all question types.

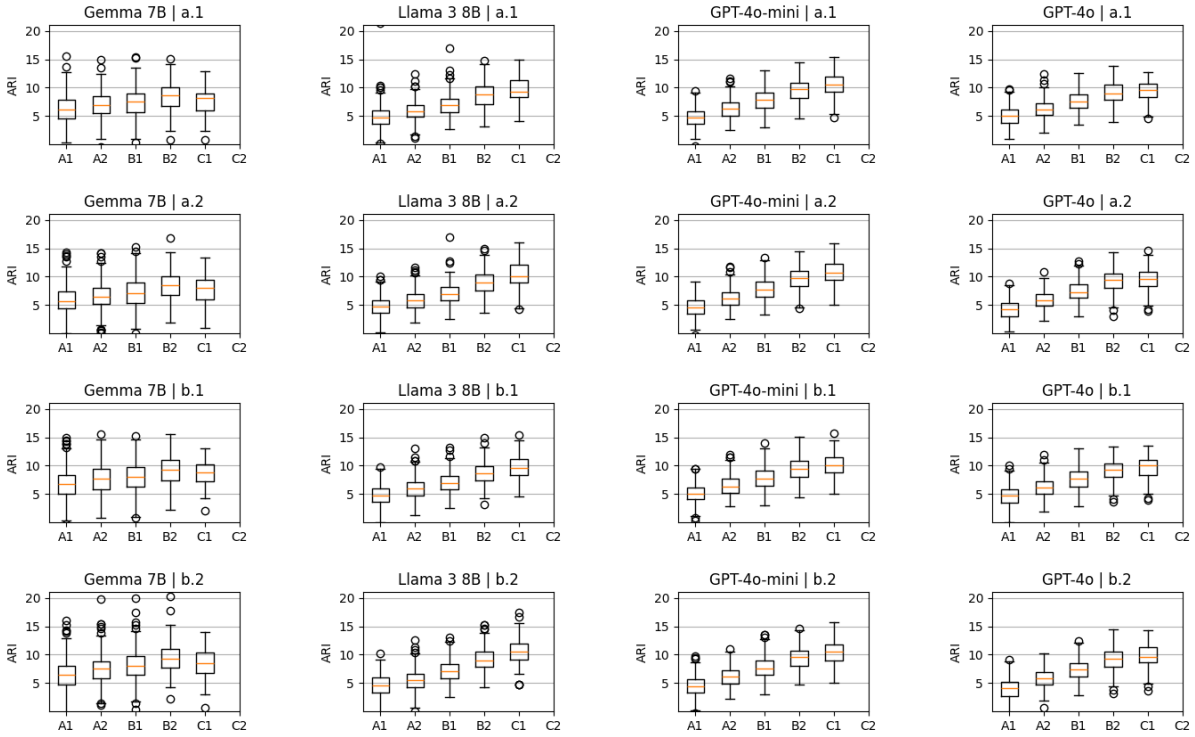

Figure 2: Boxplots showing the distribution of the ARI (Automated Readability Index) for the texts simplified with the four LLMs and the different prompts; the x-axes indicate the target CEFR level, the (shared) y-axis the ARI. Each column shows an LLM, and each row one of the prompts. The results for the two datasets are aggregated.

Model	**Pr.**	**QA Accuracy**				
		A1	A2	B1	B2	C1
Random	-	.25	.25	.25	.25	.25
Reference	-	n/a	n/a	.97	.92	.89
Gemma 7B	a.1	.68	.70	.65	.60	.57
	a.2	.66	.69	.64	.61	.58
	b.1	.72	.70	.66	.60	.56
	b.2	.69	.72	.67	.61	.58
	avg	.69	.70	.65	.60	.57
Llama 3 8B	a.1	.73	.75	.71	.62	.57
	a.2	.72	.75	.72	.61	.58
	b.1	.71	.74	.71	.62	.59
	b.2	.70	.74	.71	.60	.58
	avg	.71	.74	.71	.61	.58
GPT-4o-mini	a.1	.81	.86	.81	.65	.59
	a.2	.82	.83	.82	.65	.59
	b.1	.82	.87	.81	.64	.58
	b.2	.82	.84	.80	.64	.58
	avg	.82	.85	.81	.64	.58
GPT-4o	a.1	.81	.84	.79	.64	.61
	a.2	.76	.83	.80	.65	.61
	b.1	.79	.85	.80	.64	.58
	b.2	.77	.84	.79	.63	.57
	avg	.78	.84	.79	.64	.59

Table 3: Results of the answerability evaluation; we report the accuracy obtained by GPT-4o in answering the MCQs using the simplified texts as reading passages (separately for different target CEFR levels and overall).

consistent decrease suggests that, as hypothesised, question difficulty might be controlled with text simplification, but caution is required due to the risk of over-simplification. Indeed, the lower accuracies observed on texts simplified at higher CEFR levels are unlikely due to answer-removal from simplification (since the texts are longer), while the lower accuracy at target level A1 is likely due to over-simplification (because the simplified texts are very short). Considering different models, the text simplified with the two GPT models lead to higher QA accuracy (more often for the *mini* version), which suggests better simplification capabilities.[12]

4.4 Word-list Evaluation

Main takeaway: *Most models can adapt vocabulary to target CEFR levels, with GPT models and Llama 3 8B performing best, but underuse high-level words in high-level texts.*

Using the expert-curated CEFR-J vocabulary list, we analyse whether the models adjust their vocabulary to different CEFR target levels. Similarly to what we did for the ARI in §4.2, we construct

[12] Although it is worth noting that it might also be due to the fact that we use GPT-4o for question answer, and a different LLM should be tested to further support this claim.

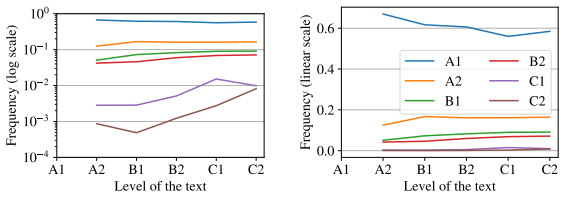

Figure 3: Frequency of words from different levels of the CEFR-J word lists in texts of different manual-curated CEFR levels (x-axis). We show the frequency in log scale (on the left) and linear scale (on the right).

a baseline by evaluating how the distribution of words from the vocabulary list varies depending the *true* CEFR level of the reading passages (as defined in the datasets), and show this in Figure 3.[13] The images show that words from lower levels are the most common across levels; the frequency of A1-level words slightly decreases for higher levels, while words associated with advanced CEFR levels get more frequent in higher levels (this is especially visible for words from levels C1 and C2).

Table 4 shows in detail word frequency distribution across levels. Gemma performs poorly, consistently overusing C1 and C2 level words regardless of target level. In contrast, GPT models and Llama 3 8B show vocabulary distributions closer to the reference values, though underusing C1-level words in C1 targeted texts, thus suggesting again over-simplification.

	Model	Target CEFR for TS A1	A2	B1	B2	C1
A1	*Reference*	-	*.67*	*.62*	*.61*	*.56*
	Gemma 7B	.62	.60	.59	.57	.60
	Llama 3 8B	.65	**.64**	.62	**.59**	.58
	GPT-4o	.63	.62	.60	.57	**.57**
	GPT-4o-mini	.64	.63	**.61**	.58	.57
A2	*Reference*	-	*.12*	*.17*	*.16*	*.16*
	Gemma 7B	.14	.15	.15	.15	.15
	Llama 3 8B	.13	**.13**	.14	.15	.16
	GPT-4o	.12	.14	.15	.16	**.16**
	GPT-4o-mini	.13	.14	**.15**	**.16**	.16
B1 ($\times 10^1$)	*Reference*	-	*.51*	*.73*	*.82*	*.90*
	Gemma 7B	.71	.76	.80	.86	.85
	Llama 3 8B	.63	.68	**.75**	.86	.94
	GPT-4o	.60	.68	.77	.86	**.92**
	GPT-4o-mini	.61	**.67**	.75	**.84**	.92
B2 ($\times 10^1$)	*Reference*	-	*.42*	*.46*	*.60*	*.69*
	Gemma 7B	.54	.57	.61	.66	.66
	Llama 3 8B	.52	.55	.58	.63	**.68**
	GPT-4o	.44	**.48**	**.53**	.58	.61
	GPT-4o-mini	.48	.51	.55	**.61**	.64
C1 ($\times 10^2$)	*Reference*	-	*.28*	*.28*	*.51*	*1.52*
	Gemma 7B	.60	.70	.84	1.14	.81
	Llama 3 8B	.40	.42	.51	.71	**.82**
	GPT-4o	.37	.39	.51	.71	.67
	GPT-4o-mini	.36	**.37**	**.45**	**.60**	.65
C2 ($\times 10^2$)	*Reference*	-	*.09*	*.05*	*.12*	*.27*
	Gemma 7B	.15	.18	.21	.34	.44
	Llama 3 8B	.06	.06	**.08**	.14	.29
	GPT-4o	.06	**.07**	.10	.15	**.26**
	GPT-4o-mini	.06	**.07**	.09	**.13**	.24

Table 4: Frequency (in %) of words from different CEFR levels (the horizontal blocks) in texts simplified at different CEFR levels (the columns). In bold are the best performing models (i.e. closest to the human-curated reference). The '$\times 10^x$' indicate that the values have been multiplied by 10^x to improve readability (e.g., .92 in block B1 indicates a frequency of 0.092%).

5 Conclusions and Future Works

We performed a first evaluation of the Text Simplification (TS) capabilities of LLMs for CEFR-targeted question adaptation in educational assessment, and studied how this impacts the answerability of reading comprehension MCQs. We found that Llama 3 8B, GPT-4o, and GPT-4o-mini can simplify texts at different readability levels and use a different vocabulary depending on the target CEFR level; however, CEFR-alignment is quite challenging, in line with previous research (Benedetto et al., 2025). TS hinders the answerability of some questions, but most can still be correctly answered by a QA model (in our case, GPT-4o). We also see a trend of decreasing QA accuracy for passages simplified at higher CEFR levels, thus supporting the possibility of question adaptation by TS. GPT models are not clearly better than Llama 3 8B, but are much less verbose, which makes post-processing easier; also, there is not a clear advantage of GPT-4o over GPT-4o-mini.

Future work will focus on improving the evaluation, working towards a framework similar to the ones proposed by Kew et al. (2023) but specific for TS for question adaptation. Additional metrics will be based on CEFR classification, and EGP tagging (O'Keeffe and Mark, 2017), to better study alignment with target CEFR levels. Also, while we are using GPT-4o for the answerability evaluation, future work should better study the correlation with a manual evaluation, possibly focusing separately on different types of questions. Lastly, future works could focus on the simplification models, working on the pedagogical alignment of LLMs (in terms of *teaching* them the CEFR), fine-tuning (large or small) language models for CEFR-targeted simplification, or leveraging in-context learning for trying to generate better simplifications.

[13] We use the frequency instead of the number of occurrences to account for the fact that higher-level texts are longer.

Acknowledgments

This paper reports on research supported by Cambridge University Press & Assessment. We thank all our ALTA colleagues for early discussions about this project, and the anonymous reviewers for their comments.

References

Suha S Al-Thanyyan and Aqil M Azmi. 2021. Automated text simplification: a survey. *ACM Computing Surveys (CSUR)*, 54(2):1–36.

Samah AlKhuzaey, Floriana Grasso, Terry R Payne, and Valentina Tamma. 2021. A systematic review of data-driven approaches to item difficulty prediction. In *International Conference on Artificial Intelligence in Education*, pages 29–41. Springer.

Luca Benedetto. 2023. A quantitative study of NLP approaches to question difficulty estimation. In *International Conference on Artificial Intelligence in Education*, pages 428–434. Springer.

Luca Benedetto, Andrea Cappelli, Roberto Turrin, and Paolo Cremonesi. 2020. Introducing a framework to assess newly created questions with Natural Language Processing. In *International Conference on Artificial Intelligence in Education*, pages 43–54. Springer.

Luca Benedetto, Paolo Cremonesi, Andrew Caines, Paula Buttery, Andrea Cappelli, Andrea Giussani, and Roberto Turrin. 2023. A survey on recent approaches to question difficulty estimation from text. *ACM Computing Surveys*, 55(9):1–37.

Luca Benedetto, Gabrielle Gaudeau, Andrew Caines, and Paula Buttery. 2025. Assessing how accurately large language models encode and apply the Common European Framework of Reference for languages. *Computers and Education: Artificial Intelligence*, 8:100353.

Benjamin Samuel Bloom, Max D Engelhart, Edward J Furst, Walker H Hill, and David R Krathwohl. 1964. *Taxonomy of educational objectives*, volume 2. Longmans, Green New York.

Nadjet Bouayad-Agha, Gerard Casamayor, Gabriela Ferraro, Simon Mille, Vanesa Vidal, and Leo Wanner. 2009. Improving the comprehension of legal documentation: the case of patent claims. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, pages 78–87.

Edmund B Coleman. 1965. On understanding prose: some determiners of its complexity. *NSF Final Report GB-2604. Washington, DC: National Science Foundation*.

Liam Cripwell, Joël Legrand, and Claire Gardent. 2023. Document-level planning for text simplification. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 993–1006, Dubrovnik, Croatia. Association for Computational Linguistics.

Edgar Dale and Jeanne S Chall. 1948. A formula for predicting readability: Instructions. *Educational research bulletin*, pages 37–54.

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.

Yutao Feng, Jipeng Qiang, Yun Li, Yunhao Yuan, and Yi Zhu. 2023. Sentence simplification via large language models. *arXiv preprint arXiv:2302.11957*.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

Gemma Team and Google DeepMind. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Robert Gunning. 1952. Technique of clear writing.

Yi-Ting Huang, Meng Chang Chen, and Yeali S Sun. 2018. Development and Evaluation of a Personalized Computer-aided Question Generation for English Learners to Improve Proficiency and Correct Mistakes. *arXiv preprint arXiv:1808.09732*.

Henri Jamet, Yash Raj Shrestha, and Michalis Vlachos. 2024. Difficulty Estimation and Simplification of French Text Using LLMs. In *International Conference on Intelligent Tutoring Systems*, pages 395–404. Springer.

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 simplification. *arXiv preprint arXiv:2310.15773*.

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. Technical report, Naval Technical Training Command Millington TN Research Branch.

George R Klare. 1974. Assessing readability. *Reading research quarterly*, pages 62–102.

Joosung Lee, Minsik Oh, and Donghun Lee. 2023. P5: Plug-and-Play Persona Prompting for Personalized Response Selection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16571–16582.

G Harry Mc Laughlin. 1969. SMOG grading-a new readability formula. *Journal of reading*, 12(8):639–646.

Meta. 2024. Introducing Meta Llama 3: The most capable openly available LLM to date.

Andrew Mullooly, Øistein Andersen, Luca Benedetto, Paula Buttery, Andrew Caines, Mark J. F. Gales, Yasin Karatay, Kate Knill, Adian Liusie, Vatsal Raina, and Shiva Taslimipoor. 2023. *The Cambridge Multiple-Choice Questions Reading Dataset*.

Ethel Ong, Jerwin Damay, Gerard Lojico, Kimberly Lu, and Dex Tarantan. 2007. Simplifying text in medical literature. *Journal of Research in Science, Computing and Engineering*, 4(1):37–47.

Anne O'Keeffe and Geraldine Mark. 2017. The English Grammar Profile of learner competence: Methodology and key findings. *International Journal of Corpus Linguistics*, 22(4):457–489.

Isabel Segura-Bedmar and Paloma Martínez. 2017. Simplifying drug package leaflets written in Spanish by using word embedding. *Journal of biomedical semantics*, 8:1–9.

RJ Senter and Edgar A Smith. 1967. Automated readability index. Technical report, Cincinnati Univ OH.

Karim Shabani, Mohamad Khatib, and Saman Ebadi. 2010. Vygotsky's zone of proximal development: Instructional implications and teachers' professional development. *English language teaching*, 3(4):237–248.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and 1 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. 2016. Text readability assessment for second language learners. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 12–22, San Diego, CA. Association for Computational Linguistics.

Victoria Yaneva, Peter Baldwin, Janet Mee, and 1 others. 2019. Predicting the difficulty of multiple choice questions in a high-stakes medical exam. In *Proceedings of the fourteenth workshop on innovative use of NLP for building educational applications*, pages 11–20.