Inclusive Easy-to-Read Text Generation for Individuals with Cognitive Impairments

Paper #0534

Abstract. Ensuring accessibility for individuals with cognitive impairments is essential for autonomy, self-determination, and full citizenship. However, manual Easy-to-Read (ETR) text adaptations are slow, costly, and difficult to scale, limiting access to crucial information in healthcare, education, and civic life. AI-driven ETR generation offers a scalable solution but faces key challenges, including dataset scarcity, domain adaptation, and balancing frugal learning with Large Language Models (LLMs). In this paper, we introduce ETR-fr, the first dataset for ETR text generation fully compliant with European ETR guidelines. We implement parameter-efficient fine-tuning on PLMs and LLMs to establish strong generative baselines. And, to ensure high-quality accessible outputs, we propose a rigorous evaluation framework, combining automated metrics with manual assessment based on a 36-question evaluation form following the European guidelines. Overall results show that PLMs perform on par with LLMs and effectively adapt to out-of-domain texts.

1 Introduction

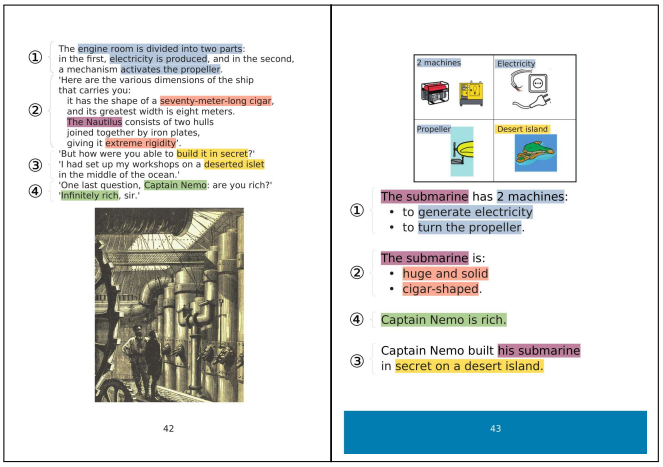


Figure 1. Extract of the Easy-to-Read book *Twenty Thousand Leagues Under the Seas* by Jules Verne from François Baudez Publishing. The original document is in French, but we translated it into English to ease comprehension. **Left page** is the original text with an illustration. **Right page** is the ETR transcription with the main information plus its captioned *vignettes*. We have highlighted and numbered the paragraphs to show the matches between the original and the ETR versions.

In line with global initiatives like the United Nations Sustainable Development Goals[1] and the Leave No One Behind Principle[2], ensuring accessibility for individuals with cognitive impairments is essential to foster autonomy, self-determination, and full citizenship. Individuals with intellectual disabilities deserve equal rights to participate in society, make informed choices, and engage fully in their communities. However, they continue to face significant obstacles, especially in accessing written information, which is essential for healthcare, education, employment, and civic engagement. Mental health disorders and intellectual disabilities affect millions worldwide, with an estimated 1.3% of the global population experiencing significant cognitive challenges [33]. In Europe alone, 4.2 million individuals are affected, while in France, between 650,000 and 700,000 people live with intellectual disabilities, including thousands of children born each year with conditions that impact their ability to comprehend written materials [11].

Easy-to-Read (ETR) is a well-established method for simplifying complex documents, ensuring that people with cognitive impairments can understand and use key information autonomously [37]. European organizations and institutions, including France's National Solidarity Fund for Autonomy[3], are increasingly producing simplified materials, indicating growing recognition of its value in improving accessibility for diverse populations. However, the current manual adaptation process is slow, costly, and subject to strict certification requirements, making it difficult to scale [6].

Developing effective AI-driven accessibility tools comes with several challenges. One major obstacle is the construction of high-quality datasets, ensuring that AI models learn to generate clear and meaningful adapted texts. Additionally, a balance must be struck between frugal learning approaches which enable low-resource, efficient adaptation and large language model (LLM) based techniques, which leverage extensive linguistic knowledge for high-quality text simplification. Open-source development ensures transparency and collaboration while empowering individuals to customize solutions and fully participate as equal citizens.

Generating high-quality ETR texts is challenging due to the need for linguistic simplification and strict adherence to accessibility guidelines. To address these challenges, we introduce ETR-fr, the first dataset specifically designed for ETR text generation tailored to cognitively disabled users. This dataset comprises 523 aligned text pairs and fully complies with European ETR guidelines. We develop robust generative models using parameter-efficient fine-tuning strategies, such as prefix-tuning [23] and Low-Rank Adaptation (LoRA) [13] applied to frugal backbones like mBART [27] and mBARThez [17], as well as large language models like Mistral-7B [16] and Llama-2-7B [39]. On the other hand, to ensure the highest quality in generating accessible texts, rigorous evaluation is essential. The different generative models undergo intrinsic evaluation using a comprehensive set of metrics derived from text simplification and text summarization. However, given the critical need for clarity, coherence, and accessibility in this context, manual evaluation plays a central role. Our main contributions are summarized as follows:

- Introduction of ETR-fr, the first dataset fully compliant with European ETR guidelines, designed for ETR text generation tailored to cognitively disabled users.

[1] https://sdgs.un.org/goals
[2] https://unsdg.un.org/2030-agenda/universal-values/leave-no-one-behind
[3] https://www.cnsa.fr/

- Implementation of parameter-efficient fine-tuning strategies, such as prefix-tuning and LoRA, applied to frugal backbones and LLMs.
- Comprehensive evaluation framework using intrinsic metrics from text simplification and summarization, complemented by a 36-question manual assessment based on European ETR guidelines.
- Investigation of model's ability to generalize ETR text generation from our ETR-fr to sources beyond its training data, showing that it can effectively handle a wide range of content, including politically focused materials.

2 Easy-to-Read Framework

Creating accessible texts for individuals with cognitive disabilities follows the Easy-to-Read framework, which adapts content to align with the European Easy-to-Read guidelines [37] (see example in Figure 1). The key principles are outlined as follows.

Clear and simple language: Use everyday vocabulary, avoiding technical jargon. Sentences should be short, direct, and in active voice to specify who is performing an action. Each sentence should convey only one idea, and consistent terminology should be used throughout the text.

Examples and analogies: Provide concrete examples and relatable analogies to explain abstract or complex ideas, linking them to familiar situations for better comprehension.

Structure and organization: Arrange content into clearly defined sections with descriptive headings and subheadings. Information should follow a logical sequence, grouping related concepts while avoiding lengthy paragraphs. Important points should be highlighted using lists where appropriate.

Accessible content: Begin with a summary outlining key points in simple terms. If technical terms are necessary, introduce clear definitions. For complex concepts or procedures, explain each step systematically with concrete examples.

Visuals and illustrations: Incorporate relevant images, charts, or diagrams to reinforce key messages. Visuals should be simple, directly connected to the text, and include concise explanatory captions.

Following the ETR guidelines, ensuring the validity of ETR content requires approval from both experts and the target audience. The manual ETR transcription process involves summarizing content and simplifying it through an iterative collaboration between human experts and individuals with cognitive impairments. This co-creation process is essential for obtaining the official European ETR label[4].

3 Related Work

Automating ETR generation could significantly streamline document creation and bridge the digital divide. However, research in this area remains scarce to the exception of very few studies mainly endeavored in Europe [4, 32]. In contrast, related fields such as text simplification [1, 24] and text summarization [43] have been widely studied.

Within the natural language processing field, various studies and tools have been developed to support individuals with cognitive disabilities by enhancing augmentative communication methods [31, 34], with dialogue agents being a widely explored solution [14]

Within the context of inclusive text generation, [10] introduced an email-writing interface based on the LaMDA LLM, offering features such as summarization, subject line generation, and text revision. However, human evaluation indicates that current LLMs still fall short in accuracy and quality for dyslexic users, highlighting the need for further research. In French, the Hector system [38] integrates word embeddings with rule-based methods for dyslexia-friendly text adaptation. While syntactic transformations improve readability, results show a decline in performance at the discourse and lexical levels.

With respect to the specific domain of ETR generation, a Finnish study [7] created a dataset aligning news articles with their Easy Finnish (*selkosuomi*) versions through automatic alignment. However, the authors acknowledge potential inaccuracies in text pairing and note that Easy Finnish does not strictly adhere to ETR guidelines. Additionally, they introduce baseline models for ETR sentence generation using fine-tuned mBART and FinGPT [29]. Similarly, the ClearText project [8] aims to develop the ClearSim corpus for simplifying Spanish public administrative texts. The current public version[5] contains three ETR document pairs with 201 misaligned pages, limiting its suitability for learning purposes. In particular, the project fully fine-tunes a Spanish T5 model, with plans to expand to 18,000 texts—15,000 generated by ChatGPT and 3,000 transcribed by experts. More recently, [32] introduced an automatically aligned Spanish ETR corpus alongside a fine-tuned Llama-2-7B model. An expert-led evaluation highlights progress in accessibility and underscores ongoing challenges in producing high-quality, guideline-compliant document-level generation. This study highlights the challenges of cross-lingual transfer, demonstrating that the translate-simplify-retranslate strategy often leads to incorrect or untranslated outputs.

Although these initiatives reflect a growing interest in ETR generation, they highlight the absence of high-quality resources that fully adhere to European ETR guidelines. To address this gap, we introduce ETR-fr, the first expert-transcribed ETR dataset specifically designed for users with cognitive disabilities.

4 ETR-fr Dataset

While there are datasets for text simplification and text summarization [9, 12, 17, 26], there is still a lack of high-quality document-aligned corpora for ETR generation in the general case and for the French language in specific.

To address this gap, the ETR-fr dataset is derived from children's books in the Easy-to-Read-and-Understand[6] collection by François Baudez Publishing[7]. This collection, designed for readers with cognitive impairments, consists of eleven books transcribed into ETR following European guidelines. Each book presents the original text on the left page and its ETR transcription on the right, as illustrated in Figure 1. From the eleven books, we extracted 523 aligned page pairs $(\text{source}, \text{target})$, where the source represents the original text and the target its ETR transcription, forming the ETR-fr dataset. Table 1 provides the key ETR-fr dataset attributes including readability indices KMRE [18] and LIX [2], compression ratios, and novelty percentages. On average, ETR-fr achieves a 50.05% compression rate, reducing token count by 56.61 and sentence count by 2.17. The average novelty rate [35] is 53.80%, reflecting the proportion of newly introduced unigrams in target texts. Readability improves by 7.51 points

[4] https://www.inclusion-europe.eu/wp-content/uploads/2021/02/How-to-use-ETR-logo..pdf

[5] https://github.com/gplsi/corpus-cleartext-cas-v1.0/tree/main

[6] Known in French as *Facile à Lire et à Comprendre*.

[7] http://www.yvelinedition.fr/Facile-a-lire

Table 1. Statistics between ETR-fr, OrangeSum, Alector, Finnish-Easy and ClearSim datasets. Compression and novelty ratios are not given for ClearSim as the publicly available version is not aligned. The LIX readability index is used instead of KMRE for Finnish-Easy and ClearSim as it is language-independent. Results are given on average with corresponding standard deviation over documents.

		French			Finnish and Spanish	
		ETR-fr (ours)	**Alector**	**OrangeSum**	**Finnish-Easy**	**ClearSim**
Dataset size		523	79	24,401	1587	207
Vocabulary size	**source**	4547	3129	80,295	98,833	6067
	target	1765	2538	23,092	18,934	2952
Num. of words	**source**	$102.76_{\pm 42.84}$	$306.48_{\pm 90.83}$	$375.98_{\pm 183.34}$	$348.47_{\pm 266.71}$	$429.13_{\pm 225.28}$
	target	$46.15_{\pm 16.73}$	$285.63_{\pm 85.34}$	$34.00_{\pm 12.17}$	$55.00_{\pm 16.61}$	$147.78_{\pm 59.54}$
Num. of sentences	**source**	$9.30_{\pm 5.12}$	$20.56_{\pm 8.95}$	$17.15_{\pm 8.85}$	$30.82_{\pm 24.05}$	$23.00_{\pm 12.77}$
	target	$7.13_{\pm 3.85}$	$22.72_{\pm 9.79}$	$1.86_{\pm 0.94}$	$6.97_{\pm 2.13}$	$11.88_{\pm 5.44}$
Sentence length	**source**	$12.57_{\pm 5.63}$	$16.82_{\pm 6.14}$	$22.77_{\pm 5.99}$	$11.29_{\pm 1.83}$	$20.13_{\pm 9.21}$
	target	$7.89_{\pm 4.55}$	$13.87_{\pm 4.08}$	$21.68_{\pm 10.82}$	$8.04_{\pm 1.55}$	$13.04_{\pm 6.61}$
KMRE $\uparrow$	**source**	$91.43_{\pm 9.41}$	$88.56_{\pm 8.23}$	$69.80_{\pm 9.47}$	–	–
	target	$98.94_{\pm 10.60}$	$95.25_{\pm 7.15}$	$68.32_{\pm 16.07}$	–	–
LIX $\downarrow$	**source**	$33.59_{\pm 8.72}$	$39.06_{\pm 9.44}$	$49.95_{\pm 7.90}$	$67.44_{\pm 5.82}$	$59.12_{\pm 8.89}$
	target	$26.89_{\pm 9.68}$	$34.19_{\pm 8.27}$	$50.39_{\pm 13.43}$	$58.12_{\pm 8.47}$	$45.30_{\pm 10.24}$
Comp. ratio (%)		$50.05_{\pm 20.55}$	$6.84_{\pm 4.47}$	$89.16_{\pm 6.34}$	$75.40_{\pm 21.71}$	–
Novelty (%)		$53.80_{\pm 16.14}$	$17.84_{\pm 8.72}$	$38.24_{\pm 19.71}$	$54.74_{\pm 16.55}$	–

from source to ETR output. For training purposes, ETR-fr is split into three fixed subsets. The test set comprises two books selected to maximize diversity in text length, word count, sentence structure, compression, novelty, and readability. The remaining nine books are divided into training and validation sets via a stratified split. Table 2 outlines these partitions.

As ETR generation involves both text simplification and summarization, we explore its relationship with these tasks by comparing key attributes of ETR-fr with Alector [9] and OrangeSum [17], two French-language datasets respectively built for text simplification and summarization. As shown in Table 1, ETR-fr balances the features of OrangeSum, which exhibits a high compression rate (89.16%), and Alector, which has a lower reduction rate (6.84%). Unlike OrangeSum, where target texts have higher KMRE scores (indicating reduced readability), ETR-fr and Alector improve readability, with a 7.51 and 6.69 point increase, respectively. For novelty, ETR-fr (53.80%) introduces more new content than OrangeSum (38.24%) and Alector (17.84%), highlighting its distinct approach.

Table 1 further compares ETR-fr with Easy-Finnish [7] and ClearSim [8] datasets. Unlike ETR-fr, which is explicitly tailored for cognitively impaired readers, these datasets target a broader audience, focusing on news articles (Easy-Finnish) and administrative texts (ClearSim). Easy-Finnish demonstrates a higher compression rate (75.40%), akin to OrangeSum. However, both Easy-Finnish and ClearSim exhibit lower accessibility, with significantly higher LIX readability scores: +33.85 and +25.53 points for source texts and +31.23 and +18.41 points for targets, respectively. Notably, Easy-Finnish shares ETR's high novelty ratio (54%), while ClearSim lacks compression and novelty metrics due to misalignment. Overall, ETR-fr prioritizes high readability and novelty while maintaining moderate text compression, making it well-suited for users with cognitive disabilities.

5 ETR Generation and Evaluation

To evaluate generation models on ETR-fr and establish baseline performance, we design a learning benchmark that involves parameter-efficient fine-tuning of frugal pre-trained language models (PLMs) and LLMs. Our approach also incorporates a two-step pipeline combining text simplification and summarization, mimicking human-expert strategy.

5.1 Expert-Centric Configuration

ETR transcription is traditionally a two-step process, where experts first summarize the source text before simplifying it. To emulate this methodology, we propose an expert-centric pipeline inspired by [3], which sequentially applies a document-level summarization model followed by a sentence simplification model. For summarization, we use the mBARThez encoder-decoder model trained on OrangeSum [17] and the summarized text is then processed by the MUSS model [30], which applies sentence-level simplification using default control tokens to generate the final ETR transcription.

5.2 Parameter-efficient Fine-tuning

To conduct ETR generation, we also investigate frugal parameter-efficient fine-tuning (PEFT) of sequence-to-sequence models, which are widely employed in the context of abstractive summarization and text simplification, such as mBART [27] and mBARThez [17]. Additionally, we explore the performance of LLMs, namely Mistral-7B [16] and Llama-2-7B [39] under PEFT.

With the growing sophistication of PLMs and LLMs, reducing computational costs while maintaining performance has become a priority. This has led to the development of PEFT strategies, such as prefix-tuning [23] and low-rank adaptation tuning [13]. These methods enable fine-tuning of only a small subset of parameters while keeping most model weights frozen, thereby minimizing the risk of catastrophic forgetting [41].

Table 2. Training, validation, and test splits of ETR-fr. Results are given on average with corresponding standard deviation over documents.

	Train		Validation		Test	
	source	**target**	**source**	**target**	**source**	**target**
Num. of texts	399		71		53	
Num. of words	$99.70_{\pm 39.25}$	$46.50_{\pm 16.80}$	$100.76_{\pm 48.12}$	$48.59_{\pm 17.20}$	$128.47_{\pm 52.54}$	$40.26_{\pm 14.38}$
Num. of sentences	$8.92_{\pm 4.73}$	$7.48_{\pm 3.42}$	$9.03_{\pm 5.21}$	$7.77_{\pm 3.91}$	$12.51_{\pm 6.60}$	$10.34_{\pm 3.81}$
Sentence length	$12.57_{\pm 4.53}$	$6.92_{\pm 2.91}$	$13.59_{\pm 10.53}$	$6.90_{\pm 2.30}$	$11.16_{\pm 2.86}$	$3.97_{\pm 0.88}$
KMRE ↑	$91.03_{\pm 8.67}$	$99.71_{\pm 9.43}$	$89.50_{\pm 13.49}$	$100.59_{\pm 10.30}$	$97.02_{\pm 5.48}$	$103.67_{\pm 10.71}$
Compression (%)	$49.04_{\pm 20.12}$		$44.47_{\pm 22.10}$		$65.19_{\pm 14.18}$	
Novelty (%)	$53.79_{\pm 16.32}$		$52.96_{\pm 16.24}$		$55.01_{\pm 14.80}$	

Table 3. Performance metrics for expert-centric and fine-tuned models on the ETR-fr test set (FT stands for full fine-tuning and PT for prefix-tuning). BARThez* refers to BARThez PLM finetuned on OrangeSum [17] dataset. Results are reported as average with standard deviation over 5 runs except for the pipelines. The best scores are highlighted in bold, except for novelty and compression rate, where scores closest to the values reported for the test split in Table 2 are emphasized in bold.

		ROUGE-1	**ROUGE-2**	**ROUGE-L**	**BERT-F_1**	**SARI**	**KMRE**	**Comp. ratio**	**Novelty**
Expert-centric									
BARThez*		22.85	5.30	15.28	67.54	36.87	95.26	73.38	30.17
MUSS		28.11	8.87	18.54	70.92	36.48	98.03	6.62	15.00
BARThez* +MUSS		22.42	4.48	14.64	67.58	36.70	96.70	75.61	36.51
MUSS+BARThez*		20.15	5.36	13.58	66.85	37.56	93.74	75.62	37.48
Fine-Tuning									
Mistral-7B	PT	$23.78_{\pm 12.03}$	$8.33_{\pm 4.70}$	$16.90_{\pm 8.20}$	$64.44_{\pm 15.14}$	$38.21_{\pm 1.36}$	$98.99_{\pm 0.80}$	$30.88_{\pm 18.92}$	$6.20_{\pm 5.18}$
	LoRA	$30.53_{\pm 0.52}$	$11.75_{\pm 0.58}$	$\mathbf{23.10}_{\pm 0.54}$	$72.51_{\pm 0.23}$	$\mathbf{42.27}_{\pm 0.70}$	$102.84_{\pm 0.35}$	$39.87_{\pm 3.53}$	$20.17_{\pm 1.30}$
Llama-2-7B	PT	$26.52_{\pm 1.82}$	$10.00_{\pm 0.96}$	$19.97_{\pm 1.17}$	$69.69_{\pm 0.80}$	$41.18_{\pm 0.58}$	$101.90_{\pm 1.08}$	$32.45_{\pm 2.33}$	$18.82_{\pm 2.16}$
	LoRA	$26.70_{\pm 1.07}$	$10.11_{\pm 0.50}$	$20.53_{\pm 0.76}$	$69.79_{\pm 0.54}$	$41.18_{\pm 0.34}$	$102.31_{\pm 0.52}$	$40.01_{\pm 4.08}$	$\mathbf{40.01}_{\pm 4.08}$
mBART	FT	$24.07_{\pm 0.07}$	$6.57_{\pm 0.01}$	$16.41_{\pm 0.03}$	$68.66_{\pm 0.00}$	$35.57_{\pm 0.00}$	$97.21_{\pm 0.00}$	$56.10_{\pm 0.00}$	$1.68_{\pm 0.00}$
	PT	$29.22_{\pm 0.47}$	$8.96_{\pm 0.80}$	$20.46_{\pm 0.70}$	$72.48_{\pm 0.31}$	$41.01_{\pm 0.26}$	$103.88_{\pm 1.29}$	$56.95_{\pm 3.16}$	$27.35_{\pm 4.86}$
	LoRA	$29.60_{\pm 1.01}$	$10.22_{\pm 0.79}$	$21.44_{\pm 0.66}$	$72.38_{\pm 0.96}$	$41.18_{\pm 0.50}$	$103.94_{\pm 1.35}$	$\mathbf{61.34}_{\pm 1.77}$	$19.40_{\pm 4.61}$
mBARThez	FT	$16.47_{\pm 0.01}$	$5.28_{\pm 0.02}$	$13.08_{\pm 0.05}$	$65.96_{\pm 0.00}$	$34.7_{\pm 0.00}$	$96.95_{\pm 0.00}$	$76.12_{\pm 0.00}$	$11.02_{\pm 0.00}$
	PT	$32.46_{\pm 0.74}$	$11.36_{\pm 0.38}$	$22.62_{\pm 0.60}$	$73.57_{\pm 0.18}$	$41.79_{\pm 0.77}$	$104.17_{\pm 0.19}$	$59.61_{\pm 1.52}$	$20.26_{\pm 2.39}$
	LoRA	$\mathbf{32.88}_{\pm 0.29}$	$\mathbf{11.81}_{\pm 0.31}$	$\mathbf{23.10}_{\pm 0.29}$	$\mathbf{73.73}_{\pm 0.14}$	$41.48_{\pm 0.34}$	$\mathbf{104.21}_{\pm 0.20}$	$56.52_{\pm 0.80}$	$16.89_{\pm 1.40}$

In particular, prefix-tuning modifies the attention output in each Transformer [40] layer. Let d represent the hidden state dimension and L the number of layers. For the l-th layer, the query, key, and value matrices are $Q_l \in \mathbb{R}^{N\times d}$ and $K_l, V_l \in \mathbb{R}^{M\times d}$, where N is the number of query tokens and M the number of key/value tokens. For each attention type (encoder self-attention, decoder cross-attention, decoder self-attention), a distinct prefix of key-value pairs is learned, $P = \{P_1, \ldots, P_L\}$, where $P_l \in \mathbb{R}^{\rho\times 2d}$ and ρ is the prefix length. For the l-th layer, K_l and V_l are augmented as in equation 1 where $K_l', V_l' \in \mathbb{R}^{(\rho+M)\times d}$.

$$K_l' = [P_{l,K}; K_l]\ , V_l' = [P_{l,V}; V_l] \tag{1}$$

Additionally, prefix optimization is stabilized by increasing the number of trainable parameters. This is done by introducing a distinct two-layer feed-forward network with an intermediate dimension of k, dedicated to re-parameterize the prefix for each attention type. The full prefix, parameterized by θ, is $P\theta = \{P^E, P^{Dc}, P^{Ds}\} \in \mathbb{R}^{\rho\times 6dL}$.

$$h = W_0 x + \frac{\alpha}{r} BAx \tag{2}$$

Low-rank adaptation fine-tuning (LoRA) provides an alternative by approximating full fine-tuning through a low-rank decomposition of weight matrices in the model. Specifically, it decomposes a weight matrix $W_0 \in \mathbb{R}^{d\times k}$ into two smaller matrices, $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times k}$, where $r \ll \min(d, k)$. This low-rank approximation is further scaled by a factor α to modulate the update contribution, ensuring that the changes remain low-rank while maintaining the backbone's integrity as defined in Equation 2.

LoRA can be applied to each linear layer in the Transformer architecture, such as W_Q, W_K, W_V, W_O matrices projections in the attention layers.

5.3 Evaluation Metrics

Since no dedicated evaluation metrics exist for ETR generation, we propose assessing it using standard summarization and text simplification metrics. For summarization, we report F1-scores for ROUGE-1, ROUGE-2, and ROUGE-L [25], along with BERTScore [44]. For

simplification, we include SARI [42], Kandel-Moles Readability Estimate (KMRE) [18], and novelty ratio for unigrams [17]. BLEU is excluded, as it is unsuitable for text simplification [42].

5.4 Experimental Setup

All PLMs are trained for 30 epochs, while LLMs are trained for 5 epochs, using the AdamW optimizer [28] with the following parameters: $\epsilon = 10^{-9}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a weight decay of $\lambda = 0.01$. A linear learning rate scheduler with a 10% warm-up ratio is employed. The training batch size is fixed at 8, with no gradient accumulation. The learning rate is chosen from the set $\{1 \cdot 10^{-5}, 2 \cdot 10^{-5}, 5 \cdot 10^{-5}, 1 \cdot 10^{-4}\}$, and hyperparameter tuning for prefix-tuning and LoRA is performed to maximize the harmonic mean of SARI, ROUGE-L, and BERTScore. Each best model is selected following hyperparameters search policy using grid search. In particular for prefix-tuning, we explore prefix length $\rho \in \{10, 50, 150, 250, 500\}$ and parametrization Multilayer Perceptron hidden layer $h \in \{256, 512, 1024, 2048\}$. For LoRA, we explore $r \in \{8, 16, 32, 64, 128\}$, $dropout \in \{0.0, 0.05, 0.1\}$, and which matrices to adapt for the self-attention and cross-attention layers $attn\_matrices \in \{W_Q, W_K, W_V, W_O, W_{QK}, W_{QV}, W_{KV}, W_{QKVO}\}$. Moreover, we choose $\alpha = r$ to keep a 1:1 ratio so as not to overpower the backbone [21].

For evaluation, generation performance results are averaged over five runs, distinguishing our approach from most text generation studies that typically report results from a single run or fixed seed [23, 32]. The expert-centric model is the only one evaluated in a zero-shot setting.

6 Quantitative and Qualitative Results

To rigorously evaluate the various ETR generation models, we propose a dual approach: a quantitative evaluation using both in-domain and out-of-domain test sets, and a qualitative assessment through manual evaluation by linguist-experts, based on 36 questions from the European ETR guidelines.

6.1 In-Domain Quantitative Results

Table 3 presents the evaluation metrics for all ETR generation models on the ETR-fr test set. Among the expert-centric pipelines, the MUSS model achieves the highest ROUGE-1 (28.11) and ROUGE-2 (8.87) scores but has the lowest compression ratio (6.62) and novelty (15.00), indicating a more conservative summarization approach. The model alone demonstrates moderate performance with a ROUGE-1 score of 22.85, while expert-centric combinations strike a balance between novelty and compression.

For fine-tuned models, PEFT methods outperform full fine-tuning, aligning with the findings of [41]. Mistral-7B with LoRA achieves strong results, with ROUGE-L (23.10), SARI (42.27), and novelty (20.17). Llama-2-7B, in both prefix-tuning and LoRA configurations, delivers competitive performance, with ROUGE-L scores of 19.97 and 20.53, respectively. Notably, Llama-2-7B with LoRA achieves the highest novelty score (40.01).

Among the fine-tuned models, mBART with LoRA exhibits the best compression ratio (61.34) (closest to the test split compression ratio), while maintaining strong ROUGE-1 (29.60) and ROUGE-2 (10.22) scores. The frugal mBARThez with LoRA achieves the best overall performance, with the highest ROUGE-1 (32.88), ROUGE-2 (11.81), ROUGE-L (23.10), BERTScore (73.73), and KMRE (104.21). Interestingly, prefix-tuning delivers results comparable to LoRA across both PLMs and LLMs.

6.2 Out-of-Domain Quantitative Results

The participation of persons with disabilities in political and public life is guaranteed by the United Nations Convention on the Rights of Persons with Disabilities, ratified by France. Since 2021, candidates for the French presidential election, as well as candidates for the legislative and regional elections, must now submit an ETR version of their electoral programme. In order to evaluate the robustness of the ETR models and their abilities to generalize across diverse and seemingly unrelated domains, we propose to test the different ETR models on a test set exclusively focused on political elections. Note that none of these text genre is including in the training phase as we only focus on ETR versions of children's books. Political texts were not included in the training process, as their alignment with European ETR guidelines is not guaranteed. However, they serve as a crucial test configuration.

For this purpose, the out-of-domain test set consists of 33 randomly selected ETR-aligned paragraphs from the 2022 French presidential election programs[8]. Table 5 provides details about the characteristics of the political test set (ETR-politic). When compared to the ETR-fr test set, the ETR-politic dataset shows several notable differences. The ETR-fr test set includes more texts (53 vs. 33) and its source texts are longer in both word count (128.47 vs. 96.27) and sentence count (12.51 vs. 6.42). However, its target texts are significantly shorter, averaging 40.26 words compared to 62.85 words in ETR-politic. The source and target texts in ETR-fr are simpler, as indicated by higher KMRE (97.02 vs. 75.03 for the source, 103.67 vs. 88.12 for the target). Additionally, the ETR-fr has a higher compression ratio (65.19 vs. 29.17) and lower novelty (55.01 vs. 63.78) compared to ETR-politic. In summary, the ETR-fr test set contains longer, simpler source texts and more concise target texts, whereas the ETR-politic test set introduces more novel content.

Table 4 illustrates the performance of fine-tuned models on ETR-fr when evaluated on ETR-politic. Similarly to results in §6.1, mBARThez achieves the highest scores across most metrics, particularly with the LoRA configuration. It records the top ROUGE-1 (38.12), ROUGE-2 (14.73), and ROUGE-L (28.11), along with the highest BERTScore (71.31) and a strong SARI score (40.35). Overall, LoRA emerges as the superior fine-tuning strategy, consistently yielding higher performance across all models compared to prefix-tuning. Additionally, the lower standard deviations associated with LoRA, especially for Mistral-7B and mBARThez, underline their stability. However, the analysis reveals that LLMs exhibit a negative compression rate, indicating challenges in replicating summarization behavior effectively.

6.3 Manual Qualitative Results

Manual evaluation is essential for assessing the quality of ETR text production and compliance with European ETR guidelines. These guidelines consist of 57 questions categorized by topic and weighted by importance, forming a comprehensive framework for evaluating clarity, simplicity, and accessibility[9]. By following these standards,

[8] https://www.cnccep.fr/candidats.html

[9] https://www.unapei.org/wp-content/uploads/2020/01/liste_verification-falc-score_v2020-01-14-1.xlsx

Table 4. Performance metrics, for fine-tuned models on ETR-fr, tested on the ETR-politic test set. Results are reported as average with standard deviation over 5 runs. The best scores are highlighted in bold.

		ROUGE-1	ROUGE-2	ROUGE-L	BERT-F_1	SARI	KMRE	Comp. ratio	Novelty
Mistral-7B	PT	$22.56_{\pm 11.68}$	$7.92_{\pm 4.56}$	$16.95_{\pm 8.45}$	$63.29_{\pm 9.14}$	$36.71_{\pm 1.22}$	$80.34_{\pm 3.89}$	$-9.53_{\pm 18.26}$	$12.77_{\pm 9.08}$
	LoRA	$33.16_{\pm 1.34}$	$12.04_{\pm 0.84}$	$25.0_{\pm 0.92}$	$69.45_{\pm 0.53}$	$39.39_{\pm 0.4}$	$79.66_{\pm 0.39}$	$7.9_{\pm 4.6}$	$15.33_{\pm 1.98}$
Llama-2-7B	PT	$24.64_{\pm 3.04}$	$8.9_{\pm 1.42}$	$19.44_{\pm 2.03}$	$65.35_{\pm 1.46}$	$37.74_{\pm 2.17}$	$81.89_{\pm 1.01}$	$-20.17_{\pm 19.57}$	$22.54_{\pm 3.44}$
	LoRA	$27.79_{\pm 0.75}$	$11.03_{\pm 0.18}$	$21.24_{\pm 0.35}$	$66.83_{\pm 0.37}$	$39.14_{\pm 0.15}$	$73.49_{\pm 0.98}$	$-9.22_{\pm 4.44}$	$15.41_{\pm 0.94}$
mBART	PT	$28.58_{\pm 0.79}$	$9.72_{\pm 1.42}$	$21.2_{\pm 1.6}$	$67.94_{\pm 0.49}$	$\mathbf{40.42}_{\pm 0.77}$	$\mathbf{86.98}_{\pm 1.73}$	$46.24_{\pm 3.13}$	$39.03_{\pm 6.68}$
	LoRA	$31.72_{\pm 1.57}$	$10.61_{\pm 1.05}$	$24.07_{\pm 0.95}$	$69.05_{\pm 1.25}$	$39.78_{\pm 0.81}$	$85.82_{\pm 1.61}$	$41.92_{\pm 2.06}$	$\mathbf{34.31}_{\pm 2.34}$
mBARThez	PT	$36.79_{\pm 0.68}$	$14.43_{\pm 0.72}$	$26.95_{\pm 0.65}$	$71.11_{\pm 0.35}$	$39.23_{\pm 0.6}$	$81.92_{\pm 0.8}$	$37.86_{\pm 2.43}$	$12.58_{\pm 3.57}$
	LoRA	$\mathbf{38.12}_{\pm 0.32}$	$\mathbf{14.73}_{\pm 0.67}$	$\mathbf{28.11}_{\pm 0.4}$	$\mathbf{71.31}_{\pm 0.32}$	$40.35_{\pm 0.37}$	$81.58_{\pm 0.5}$	$\mathbf{35.37}_{\pm 1.3}$	$16.74_{\pm 2.2}$

Table 5. Statistics of the political test dataset. Results are given on average with corresponding standard deviation over documents.

	ETR-politic Test Set	
	source	target
Num. of texts	33	
Num. of words	$96.27_{\pm 56.34}$	$62.85_{\pm 30.04}$
Num. of sentences	$6.42_{\pm 3.17}$	$6.09_{\pm 2.87}$
Sentence length	$15.68_{\pm 6.32}$	$11.47_{\pm 7.21}$
KMRE ↑	$75.03_{\pm 11.15}$	$88.12_{\pm 11.34}$
Compression ratio (%)	$29.17_{\pm 22.48}$	
Novelty (%)	$63.78_{\pm 13.85}$	

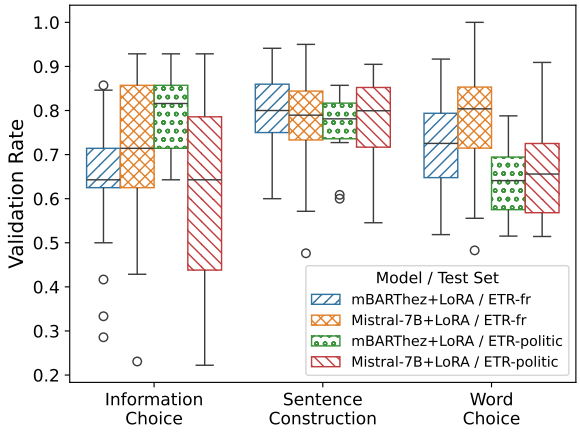
the evaluation process ensures linguistic accuracy while also verifying that the texts meet cognitive requirements, making them understandable, engaging, and suitable for the target audience.

To validate our approach, we conduct a human evaluation with three linguist-experts[10] across the ETR-fr and ETR-politic test sets. The assessment begins by focusing on the most critical criteria from the ETR guidelines checklist, including Information Choices (IC), Sentence Construction (SC), Word Choice (WC), and Illustrations[11], and consisting of 28 individual questions. Additionally, we evaluate general criteria commonly used in automatic text generation, such as Fluency, Grammar/Spelling, Relevance, Textual Coherence, and Overall Perceived Quality, gathered in extra 8 individual questions. ETR criteria are assessed using a binary scale (respected, not respected), while human judgments are rated on a 5-point Likert scale (0–4).

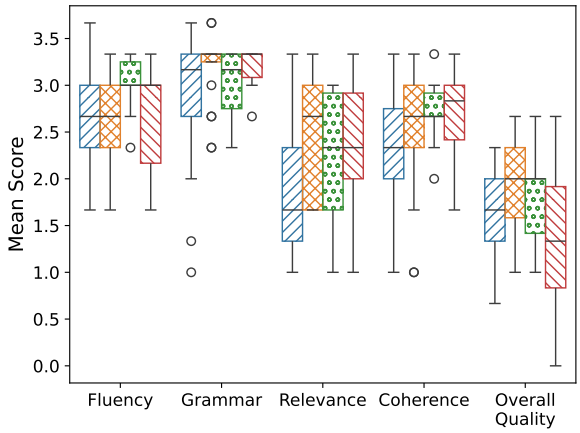
For each model, annotators were assigned to evaluate 20 texts from ETR-fr and 10 from ETR-politic randomly sampled. All annotators assessed the same set of texts, ensuring consistency in the evaluation process across models and datasets. The averaged inter-annotator agreement over the 36 criteria is $\alpha = 0.07$[12] [20].

Figure 2(a) presents the results of the ETR guidelines-based evaluation for the two best competing models: mBARThez+LoRA and Mistral-7B+LoRA. Unlike the automatic evaluation, the manual assessment shows that Mistral-7B+LoRA achieves the highest scores for IC and WC, while mBARThez+LoRA excels in SC on the ETR-fr test set. Interestingly, the trend is almost reversed on ETR-politic, where mBARThez+LoRA scores highest for IC and performs comparably to Mistral-7B+LoRA for WC and SC. Additionally, for both test sets, the frugal model exhibits the lowest dispersion score, indicating greater stability in generation.

Figure 2(b) presents the manual evaluation results for text generation quality and accuracy. Similar to the ETR-based assessment, Mistral-7B+LoRA achieves highest scores for most criteria on the ETR-fr test set, though mBARThez+LoRA performs equally well in



(a) Validation Rate of ETR Criteria



(b) Quality Score of Generation Quality Criteria

Figure 2. Manual evaluation comparisons. **(a)** Assessments from 28 ETR guidelines questions grouped into three categories. **(b)** Assessments from 8 text generation questions grouped into five categories.

[10] The linguist-experts are second-year Master's students in Language Studies. They underwent dedicated training sessions to prepare for the evaluation task. Additionally, they were unaware of the model development to ensure unbiased assessments.

[11] Results for Illustrations are not presented, as this criterion was not applicable to most of the evaluated texts.

[12] It reaches 0.20 for a binarized aggregated scores.

Fluency. However, the trend shifts significantly in the out-of-domain setting, where mBARThez+LoRA emerges as the top-performing model for Overall Perceived Quality and Fluency.

In summary, Mistral-7B+LoRA appears to overfit on ETR-fr, while mBARThez+LoRA demonstrates better generalization for ETR generation, achieving highest results on ETR-politic while maintaining strong performance on ETR-fr.

7 Limitations and Perspectives

The automatic evaluation of text generation models remains an open issue [15]. We argue that specific metrics should be developed for ETR generation, considering aspects such as novelty ratio, repetition, and coherence. Indeed, evaluation metrics for summarization and text simplification do not capture all characteristics of ETR generation, even when combined into a unique score as used in this work.

Given the limitations of automatic evaluation, manual evaluation has been performed as it is known to be a good indicator of generation capacities [22]. However, it still suffers from key drawbacks [19]. We hypothesize that the low inter-annotator agreement score presented in §6.3 is due to the fact that the ETR criteria are relatively abstract [5], which may lead to increased subjectivity in their interpretation. This variability could be mitigated through better formalization of the criteria or a comprehensive annotator training with disabled users.

While our dataset is limited in size, cross-lingual transfer remains particularly challenging due to the lack of data in other languages, especially in English. Additionally, [32] demonstrates that the translate-simplify-retranslate strategy is ineffective for ETR, often resulting in incorrect outputs. Using data from other languages also necessitates a rigorous, manual translation process involving native speakers to ensure accessibility, which restricts scalability. Although developing a multilingual model could alleviate this issue, it would still require a large-scale protocol for manual ETR transcription to create reliable resources in English.

Reinforcement learning from human feedback (RLHF) [36] could further refine ETR generation by aligning model outputs with user preferences. Collecting high-quality preference data from both expert writers and cognitively disabled users is essential to train reward models that guide language models optimization. This process would involve curated annotation tasks where users rank generated texts based on clarity, accessibility, and engagement. Expanding RLHF data collection across languages and cognitive conditions would ensure that models generate texts that are both contextually appropriate and widely usable. Moreover, this process could be a step toward automating the acquisition of the European ETR label.

8 Conclusion

This paper addresses ETR text generation for cognitively impaired individuals, aiming to enhance their self-determination and autonomy by bridging the digital divide. To support this objective, we introduced the ETR-fr dataset, a set of 523 pairs of ETR-aligned texts, and conducted an extensive empirical study using multilingual PLMs and LLMs. Our findings show that ETR generation differs significantly from traditional text simplification and summarization tasks, requiring a focused approach on cognitive accessibility. Remarkably, the small mBARThez model, combined with LoRA tuning, performs on par with larger LLMs, achieving the best results in ROUGE and BERT scores, as well as highly competitive indicators for simplification assessment, across both in-domain and out-of-domain settings. The manual evaluation conducted by three linguist-experts also highlights that the LLM-based approach tends to overfit to the main task, whereas the frugal approach exhibits better generalization, achieving the highest results on the political domain test set while maintaining strong performance on the original task.

References

[1] S. Asthana, H. Rashkin, E. Clark, F. Huot, and M. Lapata. Evaluating llms for targeted concept simplification for domain-specific texts. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6208–6226. Association for Computational Linguistics, 2024. URL https://aclanthology.org/2024.emnlp-main.357.

[2] C. H. Björnsson. Readability of Newspapers in 11 Languages. *Reading Research Quarterly*, 18(4):480–497, 1983. ISSN 0034-0553. doi: 10.2307/747382. URL https://www.jstor.org/stable/747382.

[3] S. Blinova, X. Zhou, M. Jaggi, C. Eickhoff, and S. A. Bahrainian. SIMSUM: Document-level Text Simplification via Simultaneous Summarization. In *61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 9927–9944. Association for Computational Linguistics, July 2023. doi: 10.18653/v1/2023.acl-long.552.

[4] J. Calleja, T. Etchegoyhen, and A. D. P. Martínez. Automating easy read text segmentation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 11876–11894, 2024. URL https://aclanthology.org/2024.findings-emnlp.694.

[5] E. Canut, J. Delahaie, and M. Husianycia. Vous avez dit FALC ? pour une adaptation linguistique des textes destinés aux migrants nouvellement arrivés. *Langage et Société*, N° 171(3):171–201, 2020.

[6] N. Chehab, H. Holken, and M. Malgrange. Simples - etude recueil des besoins falc. Technical report, SYSTRAN and EPNAK and EPHE and CHArt-LUTIN, 2019. URL http://51.91.138.70/simples/docs/SIMPLES_Etude_Recueil_desBesoins_FALC_HC.pdf.

[7] A. Dmitrieva and J. Tiedemann. Towards Automatic Finnish Text Simplification. In *Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context at LREC-COLING*, pages 39–50, 2024. URL https://aclanthology.org/2024.determit-1.4.

[8] I. Espinosa-Zaragoza, J. Abreu-Salas, P. Moreda, and M. Palomar. Automatic Text Simplification for People with Cognitive Disabilities: Resource Creation within the ClearText Project. In *2nd Workshop on Text Simplification, Accessibility and Readability*, pages 68–77, 2023. URL https://aclanthology.org/2023.tsar-1.7.

[9] N. Gala, A. Tack, L. Javourey-Drevet, T. François, and J. C. Ziegler. Alector: A Parallel Corpus of Simplified French Texts with Alignments of Misreadings by Poor and Dyslexic Readers. In *12th Language Resources and Evaluation Conference (LREC)*, pages 1353–1361, 2020. URL https://aclanthology.org/2020.lrec-1.169.

[10] S. M. Goodman, E. Buehler, P. Clary, A. Coenen, A. Donsbach, and al. LaMPost: Design and Evaluation of an AI-assisted Email Writing Prototype for Adults with Dyslexia. In *24th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS)*, pages 1–18, 2022. ISBN 978-1-4503-9258-7. doi: 10.1145/3517428.3544819.

[11] A. Gustavsson, M. Svensson, F. Jacobi, C. Allgulander, J. Alonso, and al. Cost of disorders of the brain in europe 2010. *European Neuropsychopharmacology*, 21(10):718–779, 2011. ISSN 0924-977X. doi: https://doi.org/10.1016/j.euroneuro.2011.08.008.

[12] R. Hauser, J. Vamvas, S. Ebling, and M. Volk. A multilingual simplified language news corpus. In *2nd Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI) within the 13th Language Resources and Evaluation Conference (LREC)*, pages 25–30, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.readi-1.4/.

[13] E. J. Hu, y. shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. LoRA: Low-rank adaptation of large language models. In *International conference on learning representations (ICLR)*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

[14] S. M. Huq, R. Maskeliūnas, and R. Damaševičius. Dialogue agents for artificial intelligence-based conversational systems for cognitively disabled: a systematic review. *Disability and Rehabilitation: Assistive Technology*, 19(3):1059–1078, 2024. ISSN 1748-3107. doi: 10.1080/17483107.2022.2146768.

[15] H. Jamet, Y. R. Shrestha, and M. Vlachos. Difficulty Estimation and Simplification of French Text Using LLMs. In A. Sifaleras and F. Lin, editors, *Generative Intelligence and Intelligent Tutoring Systems*, pages 395–404, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-63028-6. doi: 10.1007/978-3-031-63028-6_34.

[16] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, and al. Mistral 7B, 2023. arXiv:2310.06825 [cs].

[17] M. Kamal Eddine, A. Tixier, and M. Vazirgiannis. BARThez: a Skilled Pretrained French Sequence-to-Sequence Model. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9369–9390, 2021. doi: 10.18653/v1/2021.emnlp-main.740.

[18] L. Kandel and A. Moles. Application de l'indice de Flesch à la langue francaise. *Cahiers Etudes de Radio-Télévision*, 19, 1958.

[19] D. Khashabi, G. Stanovsky, J. Bragg, N. Lourie, J. Kasai, and al. GENIE: toward reproducible and standardized human evaluation for text generation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 11444–11458, 2022. doi: 10.18653/V1/2022.EMNLP-MAIN.787.

[20] K. Krippendorff. Reliability in content analysis. *Hum. Commun. Res.*, 30(3):411–433, 2004.

[21] A. N. Lee, C. J. Hunter, and N. Ruiz. Platypus: Quick, Cheap, and Powerful Refinement of LLMs, 2023. URL https://arxiv.org/abs/2308.07317v2.

[22] G. Leroy, D. Kauchak, D. Haeger, and D. Spegman. Evaluation of an online text simplification editor using manual and automated metrics for perceived and actual text difficulty. *JAMIA Open*, 5(2):ooac044, July 2022. ISSN 2574-2531. doi: 10.1093/jamiaopen/ooac044. URL https://doi.org/10.1093/jamiaopen/ooac044.

[23] X. L. Li and P. Liang. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In *59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL/IJCNLP)*, pages 4582–4597, 2021. doi: 10.18653/v1/2021.acl-long.353.

[24] Z. Li, S. Belkadi, and N. Micheletti. Investigating Large Language Models and Control Mechanisms to Improve Text Readability of Biomedical Abstracts. In *12th International Conference on Healthcare Informatics (ICHI)*, pages 265–274. IEEE Computer Society, 2024. doi: 10.1109/ICHI61247.2024.00042.

[25] C.-Y. Lin. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81. Association for Computational Linguistics, 2004. URL https://aclanthology.org/W04-1013.

[26] P. J. Liu, M. Saleh, and E. Pot. Generating wikipedia by summarizing long sequences. In *6th International Conference on Learning Representations (ICLR)*, 2018. URL https://openreview.net/forum?id=Hyg0vbWC-.

[27] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer. Multilingual Denoising Pre-training for Neural Machine Translation. *Transactions of the Association for Computational Linguistics (TACL)*, 8:726–742, 2020. doi: 10.1162/tacl_a_00343.

[28] I. Loshchilov and F. Hutter. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations (ICLR)*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

[29] R. Luukkonen, V. Komulainen, J. Luoma, A. Eskelinen, J. Kanerva, and al. FinGPT: Large Generative Models for a Small Language. In H. Bouamor, J. Pino, and K. Bali, editors, *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2710–2726, 2023. doi: 10.18653/v1/2023.emnlp-main.164.

[30] L. Martin, A. Fan, E. de la Clergerie, A. Bordes, and B. Sagot. MUSS: Multilingual Unsupervised Sentence Simplification by Mining Paraphrases. In *13th Language Resources and Evaluation Conference (LREC)*, pages 1651–1664, 2022. URL https://aclanthology.org/2022.lrec-1.176.

[31] L. J. Martin and M. Nagalakshmi. Bridging the Social & Technical Divide in Augmentative and Alternative Communication (AAC) Applications for Autistic Adults. 2024. doi: 10.48550/ARXIV.2404.17730. URL https://arxiv.org/abs/2404.17730. Publisher: arXiv.

[32] P. Martínez, A. Ramos, and L. Moreno. Exploring large language models to generate Easy to Read content. *Frontiers in Computer Science*, 6, 2024. ISSN 2624-9898. doi: 10.3389/fcomp.2024.1394705.

[33] P. K. Maulik, M. N. Mascarenhas, C. D. Mathers, T. Dua, and S. Saxena. Prevalence of intellectual disability: A meta-analysis of population-based studies. *Research in Developmental Disabilities*, 32(2):419–436, 2011. ISSN 0891-4222. doi: https://doi.org/10.1016/j.ridd.2010.12.018.

[34] T. Murillo-Morales, P. Heumader, and K. Miesenberger. Automatic Assistance to Cognitive Disabled Web Users via Reinforcement Learning on the Browser. In *Computers Helping People with Special Needs*, pages 61–72, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58805-2. doi: 10.1007/978-3-030-58805-2_8.

[35] S. Narayan, S. B. Cohen, and M. Lapata. Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1797–1807, 2018. doi: 10.18653/v1/D18-1206.

[36] L. Ouyang, J. Wu, X. Jiang, and D. Almeida. Training language models to follow instructions with human feedback, Mar. 2022. URL http://arxiv.org/abs/2203.02155. arXiv:2203.02155 [cs].

[37] Pathways. Information for all: European standards for making information easy to read and understand, 2021. URL https://www.inclusion-europe.eu/easy-to-read-standards-guidelines/.

[38] A. Todirascu, R. Wilkens, E. Rolin, T. François, D. Bernhard, and N. Gala. HECTOR: A Hybrid TExt SimplifiCation TOol for Raw Texts in French. In *13th Language Resources and Evaluation Conference (LREC)*, pages 4620–4630, 2022. URL https://aclanthology.org/2022.lrec-1.493.

[39] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, and al. Llama 2: Open Foundation and Fine-Tuned Chat Models, 2023. arXiv:2307.09288 [cs].

[40] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

[41] T. Vu, A. Barua, B. Lester, D. Cer, M. Iyyer, and N. Constant. Overcoming Catastrophic Forgetting in Zero-Shot Cross-Lingual Generation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9279–9300, 2022. doi: 10.18653/v1/2022.emnlp-main.630.

[42] W. Xu, C. Napoles, E. Pavlick, Q. Chen, and C. Callison-Burch. Optimizing Statistical Machine Translation for Text Simplification. *Transactions of the Association for Computational Linguistics (TACL)*, 4:401–415, 2016. doi: 10.1162/tacl_a_00107.

[43] P. Zakkas, S. Verberne, and J. Zavrel. Sumblogger: Abstractive summarization of large collections of scientific articles. In *Advances in Information Retrieval*, pages 371–386, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-56027-9.

[44] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations (ICLR)*, 2020. URL https://openreview.net/forum?id=SkeHuCVFDr.

We thank the reviewers for their thoughtful feedback and positive remarks about the solid literature review (R1), novelty of the task (R3), the societal impact towards social good (R2,R3), the sound methodology with an in-depth analysis of the results (R1,R2) and the resource contribution (R1,R2,R3). Below, we summarize key contributions and clarifications from our rebuttal, including new results for hybrid variants (R3). We will correct the typographical errors in equations and clarify indicated points.

R1/Dataset Scale Requirements: Despite promising results with ETR-fr, the ideal dataset should contain as much text as possible, but more importantly, it should be diverse and cover multiple domains, genres, and languages to meet the needs of cognitively impaired readers. It should also adhere to ETR guidelines and maintain alignment between source and simplified text. However, creating such datasets remains challenging due to the manual, collaborative, with cognitively impaired, nature of ETR-compliant content creation. As discussed in the paper (lignes), the ClearText project[8] attempted to build a large Spanish ETR corpus but publicly released only a few aligned pairs and the project has stopped.

R1/Experiment with Cognitively Impaired Participants: We are organizing evaluation workshops with ETR experts and partner organizations specializing in inclusive transcription. We developed an accessible web application that enables users to validate ETR outputs and provide structured feedback. This feedback is gathered using inclusive co-evaluation methods based on ETR practices. A non-impaired transcription partner asks open-ended questions to the target audience to assess their understanding of the text. These sessions assess the alignment of generated texts with ETR principles and support dataset expansion through annotation of both successful and problematic outputs. This data can also inform Reinforcement Learning from Human Feedback to improve personalization and ac-

cessibility. This evaluation was not conducted for this paper due to logistical and financial constraints.

R2/PLMs, LLMs, and Frugal Backbones: Our terminology aligns with standard NLP usage. Pre-trained Language Models (PLMs) are general-purpose models trained on large corpora, typically with fewer than one billion parameters. Examples include BART, and T5. Large Language Models (LLMs), such as Mistral, and Llama, are larger generative models designed for open-ended tasks. LLMs can be considered a subset of PLMs, see: https://arxiv.org/pdf/2406.11289. We use the term "frugal backbones" to describe PLMs with significantly fewer parameters and reduced computational demands than LLMs. This reflects the realities of organizations that serve people with cognitive impairments, many of which have limited funding and infrastructure.

R2/KMRE-LIX: These two methods are language-independent and consider sentence length and word complexity (length, number of syllables). KMRE[18] is a French-specific metric based on adaptations of the weights of Flesch-Kincaid Reading Ease formula and LIX[2] is a language independent metric. We use them to quantify the readability improvements between original and ETR-versions (see §4,tab.1). We plan to include the exact formulas if space permits in the final version.

R2,R3/Expert-Centric Pipeline: Our expert-centric pipeline is a baseline inspired by observed manual transcription practices (§5.1), introduced due to the absence of existing ETR benchmarks. None of its components were fine-tuned on ETR-fr. This allows us to evaluate how task-specific models, such as BARThez (summarization) and MUSS (simplification), perform on ETR without adaptation. If such pipeline had outperformed ETR-specific fine-tuned models, they could have offered a cost-efficient alternative for accessibility contexts. The reviewer mention a very interesting point. Indeed we investigated hybrid multitask learning strategies that address summarization, simplification, and ETR rewriting together. In particular, we implemented the MTL-LoRA (https://arxiv.org/abs/2410.09437) with the three tasks. But results are below than monotask variant:

|Models |ROUGE-1|ROUGE-2|ROUGE-L|BERTScore-F1|SARI |Compression-Ratio|Novelty| |———————|———-|———|———|———-|———|———————|———-| |mBARThez+LoRA |32.88 |11.81 |23.10 |73.73 |41.48|56.52 |16.89 | |mBARThez+MTL-LoRA|33.01 |11.37 |22.97 |73.68 |39.94|50.10 |8.69 |

R3/Dataset Construction and Certification: Our role was limited to extracting the book content in accordance with the publisher's agreement. Each ETR book from François Baudez Publishing presents original text and its ETR version on facing pages, ensuring proper alignment. After automatic extraction, we manually verified all pairs for accuracy. The books follow strict ETR transcription protocols and involve individuals with intellectual disabilities in proofreading, enabling them to carry the ETR label[37]. Since our dataset is directly derived from these labeled, unmodified texts, we consider it certified.

R3/Generalization from a Single Out-of-Domain Dataset: We recognize that the term "generalization" may overstate findings based on one out-of-domain dataset. The current evaluation is limited by the lack of a broader ETR benchmark. Even if such a benchmark existed in another language, automatic translation would not suffice, as ETR requires specialized transcription sessions involving people with intellectual disabilities. Nonetheless, our results are promising. The mBARThez+LoRA model, although trained only on children's literature (ETR-fr), performs well on the ETR-politic test set. It outperforms larger LLMs in both automatic (Table 4) and manual evaluations (Figure 2). Notably, its scores are comparable across domains: ROUGE-L of 23.10 and BERTScore of 73.73 on ETR-fr, versus 28.11 and 71.31 on ETR-politic. As mentioned earlier, our ongoing platform development and planned manual evaluations with cognitively impaired users will help build a broader evaluation dataset to better assess generalization capabilities.

R3/Manual Evaluation Reliability: As shown by Bayerl and Paul (https://aclanthology.org/J11-4004/), the inter-annotator agreement (IAA) tends to decrease as the number of categories increases by construction. So, the low agreement is attributed to the very large number of criteria (>30) that had to be evaluated by the coders. Note that we are the first work to include IAA for ETR generation. Although the evaluators are not ETR specialists, they were trained in ETR and in applying the evaluation criteria by a subject-matter expert. Nevertheless, we recognize that this process can be improved, as discussed in §1,2 of the "Limitations" section.

R3/Model Parameter Count and Training Cost: We did not include detailed analysis on the impact of parameter count on performance due to space constraints, though such content could be added as an appendix. Our results indicate that in parameter-efficient fine-tuning (PEFT), performance improves with larger LoRA ranks and full adaptation of attention matrices (WQ,WK,WV,WO). In prefix tuning, short prefixes and large bottlenecks yield better results. Overall, performance tends to increase with the number of trained parameters.