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

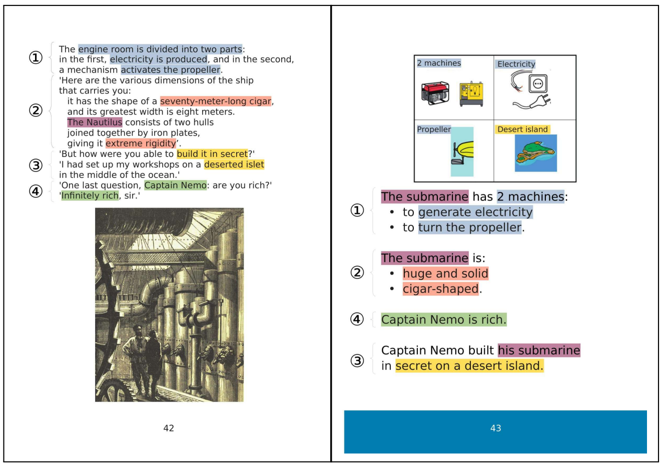
François Ledoyen[a,c,*], **Gaël Dias**[a], **Alexis Lechervy**[a], **Jérémie Pantin**[a], **Fabrice Maurel**[a], **Youssef Chahir**[a], **Elisa Gouzonnat**[b], **Mélanie Berthelot**[b], **Stanislas Moravac**[b], **Armony Altinier**[c] and **Amy Khairalla**[c]

[a]Université Caen Normandie, ENSICAEN, CNRS, Normandie Univ, GREYC UMR 6072, F-14000 Caen, France
[b]Université Caen Normandie, CRISCO UR 4255, F-14000 Caen, France
[c]Koena SAS, F-31450 Fourquevaux, France

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 lightweight learning of 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 generative baselines. To ensure high-quality and accessible outputs, we introduce an evaluation framework based on automatic metrics supplemented by human assessments. The latter is conducted using a 36-question evaluation form that is aligned with the guidelines. Overall results show that PLMs perform comparably to LLMs and adapt effectively to out-of-domain texts. Code and datasets are available at https://github.com/FrLdy/ETR-fr.

1 Introduction

Reflecting the priorities of global initiatives such as the United Nations Sustainable Development Goals[1] and the Leave No One Behind Principle[2], ensuring accessibility for individuals with cognitive impairments is crucial to fostering autonomy, self-determination, and full citizenship. Individuals with intellectual disabilities deserve equal rights to participate in society, to make informed choices, and to fully engage 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 [32]. In Europe alone, 4.2 million individuals are affected, while in France, between 650,000 and 700,000 people live with intellectual disabilities that limit their ability to comprehend written materials [12].



① The engine room is divided into two parts:
in the first, electricity is produced, and in the second,
a mechanism activates the propeller.
② 'Here are the various dimensions of the ship
that carries you:
it has the shape of a seventy-meter-long cigar,
and its greatest width is eight meters.
The Nautilus consists of two hulls
joined together by iron plates,
giving it extreme rigidity'.
③ 'But how were you able to build it in secret?'
'I had set up my workshops on a deserted islet
in the middle of the ocean.'
④ 'One last question, Captain Nemo: are you rich?'
'Infinitely rich, sir.'

42

① The submarine has 2 machines:
- to generate electricity
- to turn the propeller.

② The submarine is:
- huge and solid
- cigar-shaped.

④ Captain Nemo is rich.

③ Captain Nemo built his submarine
in secret on a desert island.

43

Figure 1. Extract of the Easy-to-Read book *Twenty Thousand Leagues Under the Sea* 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.

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 [36]. 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 [7].

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 parameter-efficient fine-tuning (PEFT) 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

[*] Corresponding Author. Email: ledoyenfrancois@gmail.com
[1] https://sdgs.un.org/goals
[2] https://unsdg.un.org/2030-agenda/universal-values/leave-no-one-behind
[3] https://www.cnsa.fr/

guidelines. To address these challenges, we introduce ETR-fr, the first dataset specifically designed for ETR text generation, tailored to users with cognitive disabilities. This dataset comprises 523 aligned text pairs and fully complies with European ETR guidelines. We develop generative models using PEFT strategies, such as prefix-tuning [23] and Low-Rank Adaptation (LoRA) [14] applied to pretrained language models like mBART [26] and mBARThez [18], as well as large language models like Mistral-7B [17] and Llama-2-7B [40]. 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 automatic 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 parallel dataset fully compliant with European ETR guidelines.
- Implementation of baselines for ETR generation based on PEFT strategies, such as prefix-tuning and LoRA, applied to PLMs and LLMs backbones.
- Comprehensive evaluation framework using intrinsic metrics from text simplification and summarization, reinforced by a 36-question manual assessment based on European ETR guidelines.
- Investigation of the model's ability to generalize ETR generation from our ETR-fr to 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 [36] (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 the 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 and using bullet lists while avoiding lengthy paragraphs

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 [36].

3 Related Work

Automating ETR generation could significantly streamline document creation and bridge the digital divide. However, research in this area remains scarce, except for very few studies mainly conducted in Europe [5, 31]. In contrast, related fields such as text simplification [1, 24] and text summarization [44] 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 [30, 33], with dialogue agents being a widely explored solution [15].

Within the context of inclusive text generation, Goodman et al. [11] introduced an email-writing interface based on LaMDA LLM [38], offering features such as summarization, subject line generation, and text revision. However, human evaluations show that current LLMs still lack accuracy and quality for dyslexic users, highlighting the need for further research. In French, the Hector system [39] integrates word embeddings with rule-based methods for adapting text to be dyslexia-friendly. While syntactic transformations improve readability, results show a decline in performance at the discourse and lexical levels.

Within the specific domain of ETR generation, Dmitrieva and Tiedemann [8] have created the Finnish-Easy dataset, which aligns news articles with their Easy Finnish[4] (*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 [28]. Similarly, the ClearText project [9] 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. However, the project plans to expand the corpora to 18,000 texts, 15,000 generated by ChatGPT and 3,000 transcribed by experts. More recently, Martínez et al. [31] 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 the 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

Although several datasets exist for French text simplification and summarization, such as Alector [10], OrangeSum [18], and multilingual corpora [13], there remains a lack of high-quality, document-aligned corpora specifically designed for ETR text generation. This gap is particularly noticeable for the French language.

To address this, we introduce the ETR-fr dataset, constructed from the *Facile à Lire et à Comprendre* (Easy-to-Read-and-Understand)[6]

[4] Easy Finnish is a form of Finnish where the language has been adapted so that it is easier to read and understand in terms of content, vocabulary and structure.

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

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

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 ↑	**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 ↓	**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 (%)		$53.80_{\pm 16.14}$	$17.84_{\pm 8.72}$	$38.24_{\pm 19.71}$	$54.74_{\pm 16.55}$	

collection published by François Baudez Publishing[7]. This collection consists of eleven children's books adapted according to European guidelines for cognitive accessibility. Each book presents the original version on the left page and its ETR transcription on the right, as illustrated in Figure 1.

From these books, we extracted 523 aligned page pairs (source, target), where the source corresponds to the original text and the target to its ETR version. These alignments form the core of the ETR-fr dataset.

ETR-fr Characteristics Table 1 summarizes the characteristics of the dataset, including readability metrics, compression ratios, and novelty rates. We use two readability indicators: KMRE [19], a French adaptation of the Flesch-Kincaid Reading Ease formula [20], and LIX [3]. KMRE produces a score from 0 (very difficult) to 100 or more (very easy), based on sentence and word lengths. LIX measures difficulty based on average sentence length and the proportion of long words (more than six letters), with typical values ranging from 20 (easy) to 60 (difficult). Novelty [34] indicates the proportion of new unigrams introduced in the target text.

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 is 53.80%. The KMRE score improves by 7.51 points on average, showing a measurable gain in readability from source to ETR output.

Comparison with Related Datasets To better contextualize ETR-fr, we compare it with other French-language datasets: Alector [10], designed for text simplification, and OrangeSum [18], built for summarization. As shown in Table 1, OrangeSum features a high compression rate (89.16%) but reduced readability in its target texts. Alector presents minimal compression (6.84%) but improves readability by 6.69 KMRE points. ETR-fr offers a more balanced profile, combining moderate compression (50.05%), a readability improvement of 7.51 KMRE points, and higher novelty (53.80%) than both OrangeSum (38.24%) and Alector (17.84%).

We also compare ETR-fr to foreign-language ETR-style datasets: Easy-Finnish [8] and ClearSim [9]. These corpora are designed for broader audiences and focus on various text types. Easy-Finnish covers news articles, and ClearSim includes administrative texts. Easy-Finnish demonstrates a high compression rate (75.40%) and a novelty score similar to ETR-fr (≃ 54%). However, both datasets exhibit lower accessibility, with LIX readability scores significantly higher than those of ETR-fr: +33.85 and +25.53 for source texts, and +31.23 and +18.41 for target texts. Additionally, ClearSim does not include reliable compression and novelty statistics due to misalignment in its text pairs.

ETR-fr Splits The ETR-fr dataset is divided into training, validation, and test sets, as described in Table 2. Two books are selected for the test set to maximize diversity in sentence structure, length, compression, novelty, and readability. The remaining nine books are split into training and validation subsets using a stratified approach.

ETR-fr-politic Test Set The participation of persons with disabilities in political and public life is enshrined in the United Nations Convention on the Rights of Persons with Disabilities, which France has ratified. Since 2021, candidates for the French presidential election have been required to submit an ETR version of their electoral programs.

To assess the robustness of ETR models and their ability to generalize across diverse and previously unseen domains, we evaluate them on a test set specifically focused on political election texts. It is important to note that political texts were not part of the training data, making this evaluation a critical measure of model generalization.

To this end, we introduce an out-of-domain test set, ETR-fr-politic, comprising 33 paragraph pairs manually extracted from the ETR-

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

Table 2. Statistics for the ETR-fr dataset (Train/Validation/Test) and the ETR-fr-politic test set. Results are given on average with corresponding standard deviation over documents.

	ETR-fr						**ETR-fr-politic**	
	Train		**Validation**		**Test**		**Test**	
	source	**target**	**source**	**target**	**source**	**target**	**source**	**target**
Num. of texts	399		71		53		33	
Num. of words	$99.70_{\pm 39.25}$	$46.50_{\pm 16.80}$	$100.76_{\pm 48.12}$	$48.59_{\pm 17.20}$	$128.47_{\pm 52.54}$	$40.26_{\pm 14.38}$	$96.27_{\pm 56.34}$	$62.85_{\pm 30.04}$
Num. of sentences	$8.92_{\pm 4.73}$	$7.48_{\pm 3.42}$	$9.03_{\pm 5.21}$	$7.77_{\pm 3.91}$	$12.51_{\pm 6.60}$	$10.34_{\pm 3.81}$	$6.42_{\pm 3.17}$	$6.09_{\pm 2.87}$
Sentence length	$12.57_{\pm 4.53}$	$6.92_{\pm 2.91}$	$13.59_{\pm 10.53}$	$6.90_{\pm 2.30}$	$11.16_{\pm 2.86}$	$3.97_{\pm 0.88}$	$15.68_{\pm 6.32}$	$11.47_{\pm 7.21}$
KMRE ↑	$91.03_{\pm 8.67}$	$99.71_{\pm 9.43}$	$89.50_{\pm 13.49}$	$100.59_{\pm 10.30}$	$97.02_{\pm 5.48}$	$103.67_{\pm 10.71}$	$75.03_{\pm 11.15}$	$88.12_{\pm 11.34}$
Compression (%)	$49.04_{\pm 20.12}$		$44.47_{\pm 22.10}$		$65.19_{\pm 14.18}$		$29.17_{\pm 22.48}$	
Novelty (%)	$53.79_{\pm 16.32}$		$52.96_{\pm 16.24}$		$55.01_{\pm 14.80}$		$63.78_{\pm 13.85}$	

labeled versions of the 2022 French presidential election programs[8]. These paragraphs have been carefully aligned with their original versions, allowing for precise quantitative evaluation of generated texts. A detailed overview of the dataset is provided in Table 2.

Compared to the ETR-fr test set, ETR-fr-politic contains fewer texts (33 vs. 53), and its source texts are shorter in both word count (96.27 vs. 128.47) and sentence count (6.42 vs. 12.51). However, its target texts are longer, averaging 62.85 words compared to 40.26 in ETR-fr. The ETR-fr test set exhibits higher readability, with KMRE scores of 97.02 (source) and 103.67 (target), versus 75.03 and 88.12 in ETR-fr-politic. Furthermore, ETR-fr has a higher compression ratio (65.19%) and a slightly lower novelty rate (55.01%) compared to ETR-fr-politic (29.17% and 63.78%, respectively). These differences highlight the more complex and varied nature of the political texts.

Summary In summary, the ETR-fr dataset fills an important gap in French-language NLP resources by providing a high-quality, document-aligned corpus tailored for readers with cognitive impairments. It effectively bridges simplification and summarization, improving readability while maintaining moderate compression and incorporating a high level of novel content. Its structure and evaluation design make it well-suited as a benchmark for training and evaluating ETR generation systems.

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 pre-trained language models (PLMs) and LLMs. Our approach also incorporates a two-step pipeline combining text simplification and summarization, mimicking a human-expert strategy.

5.1 *Expert-Centric Configuration*

We introduce an expert-centric pipeline motivated by the lack of established ETR benchmarks and inspired by manual transcription practices. This approach replicates the traditional two-step process used by experts, where summarization precedes simplification. Following the methodology proposed by Blinova et al. [4], our pipeline first applies a document-level summarization model, using BARThez trained on OrangeSum [18], and then simplifies the output with the MUSS model [29], which performs sentence-level simplification using default control tokens. Since neither model is fine-tuned on ETR-fr, this setup allows us to evaluate the zero-shot performance of task-specific models on ETR-fr.

5.2 *Parameter-Efficient Fine-Tuning*

To conduct ETR generation, we also investigate 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 [26] and mBARThez [18]. Additionally, we explore the performance of LLMs, namely Mistral-7B [17] and Llama-2-7B [40] 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 (LoRA) [14]. 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 [42].

Prefix-tuning introduces a lightweight set of trainable vectors that are prepended to the key and value inputs of the Transformer multi-head attention mechanism [41]. Formally, for each attention head i, prefix-tuning prepends learned vectors $P_K^i \in \mathbb{R}^{\rho \times d_{\text{head}}}$ and $P_V^i \in \mathbb{R}^{\rho \times d_{\text{head}}}$, each of length ρ, to the projected keys and values, respectively. The resulting attention computation for the i-th head is expressed as:

$$\text{head}_i = \text{Attention}\big(QW_Q^i, [P_K^i; KW_K^i], [P_V^i; VW_V^i]\big) \qquad (1)$$

where $Q, K, V \in \mathbb{R}^{L \times d_{\text{model}}}$ denote the query, key, and value matrices derived from an input sequence of length L, and $W_Q^i, W_K^i, W_V^i \in \mathbb{R}^{d_{\text{model}} \times d_{\text{head}}}$ are the frozen projection matrices associated with the i-th attention head. The notation $[;]$ denotes the concatenation function. To enhance the stability of prefix optimization, the number of trainable parameters is increased by employing a dedicated two-layer feed-forward network for re-parameterizing the prefix associated with each attention type. This network features an intermediate hidden dimension h_{MLP} and enables a richer parameterization of the prefix vectors.

LoRA offers an efficient alternative to full fine-tuning by introducing a low-rank decomposition of the linear model's weight matrices. Instead of updating the full weight matrix $W_0 \in \mathbb{R}^{d \times k}$, LoRA expresses it as a sum of the original weights and a trainable low-rank perturbation. Specifically, the update is represented by two smaller matrices: $B \in \mathbb{R}^{d \times r}$ and $A \in \mathbb{R}^{r \times k}$, with $r \ll \min(d, k)$. Here,

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

Table 3. Performance of expert-centric and fine-tuned models on the ETR-fr test set (FT: full fine-tuning, PT: prefix-tuning). BARThez* denotes BARThez fine-tuned on OrangeSum [18]. Scores are averaged over 5 runs (except pipelines), with standard deviation. Best results are in bold, except for novelty and compression, where values closest to the ETR-fr test set (Table 2) are highlighted.

		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$_{\pm 0.58}$	**23.10**$_{\pm 0.54}$	72.51$_{\pm 0.23}$	**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$_{\pm 4.08}$	12.72$_{\pm 1.28}$
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$_{\pm 1.35}$	**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$_{\pm 0.00}$	34.70$_{\pm 0.00}$	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	**32.88**$_{\pm 0.29}$	**11.81**$_{\pm 0.31}$	**23.10**$_{\pm 0.29}$	**73.73**$_{\pm 0.14}$	41.48$_{\pm 0.34}$	**104.21**$_{\pm 0.20}$	56.52$_{\pm 0.80}$	16.89$_{\pm 1.40}$

d and k denote the input and output dimensions of the layer, respectively. The low-rank component is scaled by a factor α to control the magnitude of the update, ensuring minimal interference with the pre-trained backbone:

$$h = W_0x + \frac{\alpha}{r}BAx \tag{2}$$

LoRA can be seamlessly integrated into Transformer architectures by applying it to each linear transformation, including the attention projection matrices W_Q, W_K, W_V, and W_O.

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, ROUGE-L [25], and BERTScore [45]. For simplification, we include SARI [43], Kandel-Moles Readability Estimate (KMRE) [19], and novelty ratio for unigrams [18]. BLEU is excluded, as it is unsuitable for text simplification [43, 37].

5.4 *Experimental Setup*

All PLMs are trained for 30 epochs, while LLMs are trained for 5 epochs, using the AdamW optimizer [27] 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 a hyperparameter search policy using grid search.

In particular for prefix-tuning, we explore prefix length $\rho \in \{10, 50, 150, 250, 500\}$ and re-parametrization MLP hidden size $h_{\text{MLP}} \in \{256, 512, 1024, 2048\}$.

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 $attn\_matrices \in \{W_Q, W_K, W_V, W_O, W_{QK}, W_{QV}, W_{KV}, W_{QKVO}\}$. To keep a 1:1 ratio so as not to overpower the backbone, we choose $\alpha = r$ [22].

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, 31]. 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. In the expert-centric pipelines, MUSS achieves the best ROUGE-1 (28.11) and ROUGE-2 (8.87) but shows low compression (6.62) and novelty (15.00), indicating a conservative style. BARThez performs moderately (ROUGE-1: 22.85). The combined pipelines trade fidelity for abstraction: MUSS+BARThez yields the highest compression (75.62), best SARI (37.56), and greatest novelty (37.48), though with weaker ROUGE (20.15/5.36/13.58) and BERTScore (66.85).

For fine-tuned models, PEFT methods outperform full fine-tuning, aligning with the findings of [42]. 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,

Table 4. Performance metrics for fine-tuned models on ETR-fr, tested on the ETR-fr-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	$33.16_{\pm 1.34}$	$12.04_{\pm 0.84}$	$25.00_{\pm 0.92}$	$69.45_{\pm 0.53}$	$39.39_{\pm 0.40}$	$79.66_{\pm 0.39}$	$7.90_{\pm 4.60}$	$15.33_{\pm 1.98}$
Llama-2-7B	PT	$24.64_{\pm 3.04}$	$8.90_{\pm 1.42}$	$19.44_{\pm 2.03}$	$65.35_{\pm 1.46}$	$37.74_{\pm 2.17}$	$81.89_{\pm 1.01}$	$-20.17_{\pm 19.57}$	$22.54_{\pm 3.44}$
	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	$28.58_{\pm 0.79}$	$9.72_{\pm 1.42}$	$21.20_{\pm 1.60}$	$67.94_{\pm 0.49}$	$\mathbf{40.42}_{\pm 0.77}$	$\mathbf{86.98}_{\pm 1.73}$	$46.24_{\pm 3.13}$	$39.03_{\pm 6.68}$
	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	$36.79_{\pm 0.68}$	$14.43_{\pm 0.72}$	$26.95_{\pm 0.65}$	$71.11_{\pm 0.35}$	$39.23_{\pm 0.60}$	$81.92_{\pm 0.80}$	$37.86_{\pm 2.43}$	$12.58_{\pm 3.57}$
	LoRA	$\mathbf{38.12}_{\pm 0.32}$	$\mathbf{14.73}_{\pm 0.67}$	$\mathbf{28.11}_{\pm 0.40}$	$\mathbf{71.31}_{\pm 0.32}$	$40.35_{\pm 0.37}$	$81.58_{\pm 0.50}$	$\mathbf{35.37}_{\pm 1.30}$	$16.74_{\pm 2.20}$

delivers competitive performance, with ROUGE-L scores of 19.97 and 20.53, respectively.

Among the fine-tuned models, mBART with LoRA exhibits the best compression ratio (61.34) (closest to the test split reference), while maintaining strong ROUGE-1 (29.60) and ROUGE-2 (10.22) scores. The PLM 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.

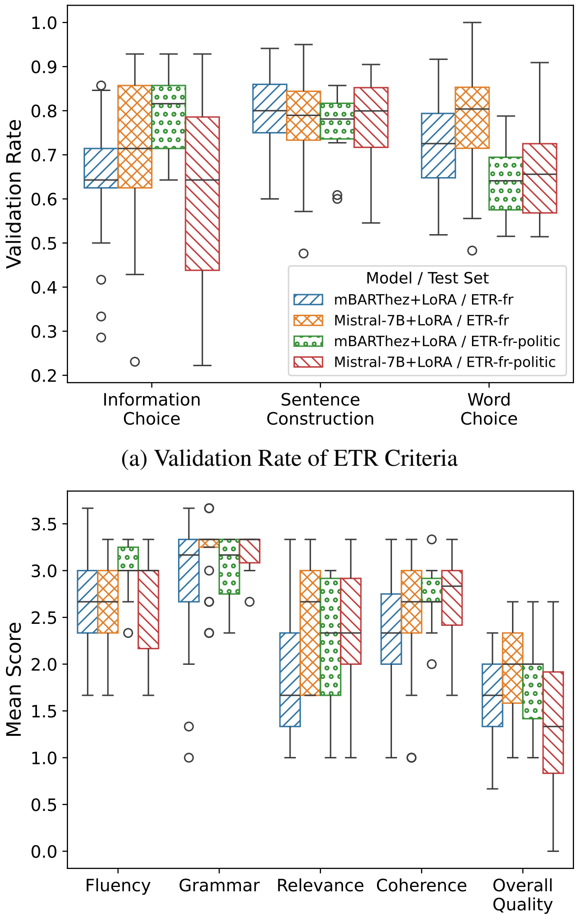
6.2 *Out-of-Domain Quantitative Results*

Table 4 illustrates the performance of fine-tuned models on ETR-fr when evaluated on ETR-fr-politic test set. 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-6B 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. By following these standards, 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[9] across the ETR-fr and ETR-fr-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[10], and consisting of 29 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 an additional 8 individual questions. ETR criteria are assessed using a binary scale (respected,



(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.

[9] The linguist-experts, all second-year Masters students in Language Studies, received dedicated training sessions to prepare for the evaluation task. Their participation was voluntary and uncompensated, and they were kept unaware of the model development to ensure unbiased assessments.

[10] Results for Illustrations are not presented, as this criterion did not apply to most of the evaluated texts.

not respected), while human judgments are rated on a 5-point Likert scale (04).

For each model, annotators were assigned to evaluate 20 texts from ETR-fr and 10 from ETR-fr-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 $\alpha = 0.07$[11] [21].

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-fr-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 mBARThez 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 the highest scores for most criteria on the ETR-fr test set, though mBARThez+LoRA performs equally well in 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 the highest results on ETR-fr-politic while maintaining strong performance on ETR-fr.

7 Limitations and Perspectives

The automatic evaluation of text generation models remains an open issue [16]. 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.

The low inter-annotator agreement observed in §6.3 may be explained by the high number of ETR criteria (>30), which is known to reduce agreement levels [2], as well as the abstract nature of these criteria [6], which introduces subjectivity. Improved formalization or targeted annotator training, especially with disabled users, could help mitigate this variability.

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, Martínez et al. [31] demonstrate 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) [35] 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 optimization of language models. This process would involve annotation tasks where users rank generated texts by 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 individuals with cognitive impairments, 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 BERTScore, 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 lightweight approach generalizes better, achieving the highest results on the political test set while maintaining strong performance on the original task.

Acknowledgements

We are grateful to François Baudez Publishing for generously providing free access to the ETR book collection to build our dataset.

Finally, the present work was performed using computing resources of CRIANN (Normandy, France).

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] P. S. Bayerl and K. I. Paul. What determines inter-coder agreement in manual annotations? a meta-analytic investigation. *Computational Linguistics*, 37(4):699–725, Dec. 2011. doi: 10.1162/COLI_a_00074. URL https://aclanthology.org/J11-4004/.

[3] 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.

[4] 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.

[5] 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.

[6] 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*, Nř 171(3):171–201, 2020.

[7] 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.

[8] 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.

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

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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/.

[14] 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.

[15] S. M. Huq, R. Maskelinas, and R. Damaeviius. 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.

[16] H. Jamet, Y. R. Shrestha, and M. Vlachos. Difficulty Estimation andăSimplification ofăFrench 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.

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

[18] 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.

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

[20] J. P. Kincaid et al. Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. Technical report, National Technical Information Service, Springfield, Virginia 22151 (AD-A006 655/5GA, MF $2, Feb. 1975. ERIC Number: ED108134.

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

[22] 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.

[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] 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.

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

[28] R. Luukkonen, V. Komulainen, J. Luoma, A. Eskelinen, J. Kanerva, et 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.

[29] 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.

[30] L. J. Martin and M. Nagalakshmi. Aging up aac: An introspection on augmentative and alternative communication applications for autistic adults. *arXiv preprint arXiv:2404.17730*, 2025. doi: 10.48550/arXiv.2404.17730. URL https://arxiv.org/abs/2404.17730. Version 3, last revised 4 Aug 2025.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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].

[36] 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/.

[37] E. Sulem, O. Abend, and A. Rappoport. Simple and Effective Text Simplification Using Semantic and Neural Methods. In *56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 162–173, 2018. doi: 10.18653/v1/P18-1016.

[38] R. Thoppilan, D. D. Freitas, Hall, et al. LaMDA: Language Models for Dialog Applications. *ArXiv*, 2022. URL https://www.semanticscholar.org/paper/LaMDA%3A-Language-Models-for-Dialog-Applications-Thoppilan-Freitas/b3848d32f7294ec708627897833c4097eb4d8778.

[39] 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.

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

[41] 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.

[42] 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.

[43] 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.

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

[45] 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.