# Facilitating Cognitive Accessibility with LLMs: A Multi-Task Approach to Easy-to-Read Text Generation

**François Ledoyen[1,2], Gaël Dias[1], Jérémie Pantin[1], Alexis Lechervy[1], Fabrice Maurel[1], Youssef Chahir[1]**
[1]Université Caen Normandie, ENSICAEN, CNRS, Normandie Univ, GREYC UMR 6072, F-14000 Caen, France
[2]Koena SAS, F-31450 Fourquevaux, France
**Correspondence:** ledoyenfrancois@gmail.com

## Abstract

Simplifying complex texts is essential to ensure equitable access to information, particularly for individuals with cognitive impairments. The Easy-to-Read (ETR) initiative provides a framework to make content more accessible for these individuals. However, manually creating such texts remains time-consuming and resource-intensive. In this work, we investigate the potential of large language models (LLMs) to automate the generation of ETR content. To address the scarcity of aligned corpora and the specific constraints of ETR, we propose a multi-task learning (MTL) approach that trains models jointly on text summarization, text simplification, and ETR generation. We explore two complementary strategies: multi-task retrieval-augmented generation (RAG) for in-context learning (ICL), and MTL-LoRA for parameter-efficient fine-tuning (PEFT). Our experiments with Mistral-7B and LLaMA-3-8B, conducted on ETR-fr, a new high-quality dataset, show that MTL-LoRA consistently outperforms all other strategies in in-domain settings, while the MTL-RAG-based approach achieves better generalization in out-of-domain scenarios. Our code is publicly available at `https://github.com/FrLdy/ETR-PEFT-Composition`.

## 1 Introduction

Mental health conditions and intellectual disabilities affect millions of people worldwide, creating significant challenges for equitable access to information (Maulik et al., 2011; Gustavsson et al., 2011). These individuals often struggle with complex texts, which limits their participation in healthcare, education, and civic life. Despite international initiatives for inclusion[1,2], accessible written content remains a major barrier to full participation.

While Easy-to-Read (ETR) (Pathways, 2021), text simplification (Paetzold and Specia, 2016), summarization (Rush et al., 2015), and plain language (Maaß, 2020) all aim to improve comprehension, they differ in purpose, audience, and methods. Text simplification rewrites content to enhance readability while preserving the original informational content (Gooding, 2022; Stajner, 2021). Summarization shortens the original text by extracting and presenting only the key points, often without rewording for greater clarity (Rush et al., 2015). Plain language addresses broad audiences, including people with limited literacy, by using clear structure and simple vocabulary, but it may still be too complex for individuals with cognitive impairments (Maaß, 2020). ETR, by contrast, is a rigorously standardized form of text adaptation developed specifically for individuals with intellectual disabilities. It requires strict adherence to Pathways (2021) guidelines, which mandate very short sentences, highly simplified vocabulary, visual supports, and obligatory end-user testing. The primary goal is to foster the autonomy of readers with cognitive impairments. Importantly, ETR materials must be co-created by subject-matter experts together with individuals with cognitive disabilities to ensure compliance with ETR standards and eligibility for European ETR certification[3].

However, ETR adoption remains limited due to the time-consuming and costly nature of manual adaptation, coupled with the lack of robust automated tools tailored to the linguistic and cognitive requirements of ETR content (Chehab et al., 2019). The potential of LLMs for improving accessibility (Freyer et al., 2024) is limited by the scarcity of high-quality, document-aligned ETR datasets. Existing resources, such as ClearSim (Espinosa-Zaragoza et al., 2023), are limited and only partially aligned, highlighting the broader challenge of constructing or recovering document-aligned corpora

[1]UN Sustainable Development Goals
[2]Leave No One Behind Principle

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

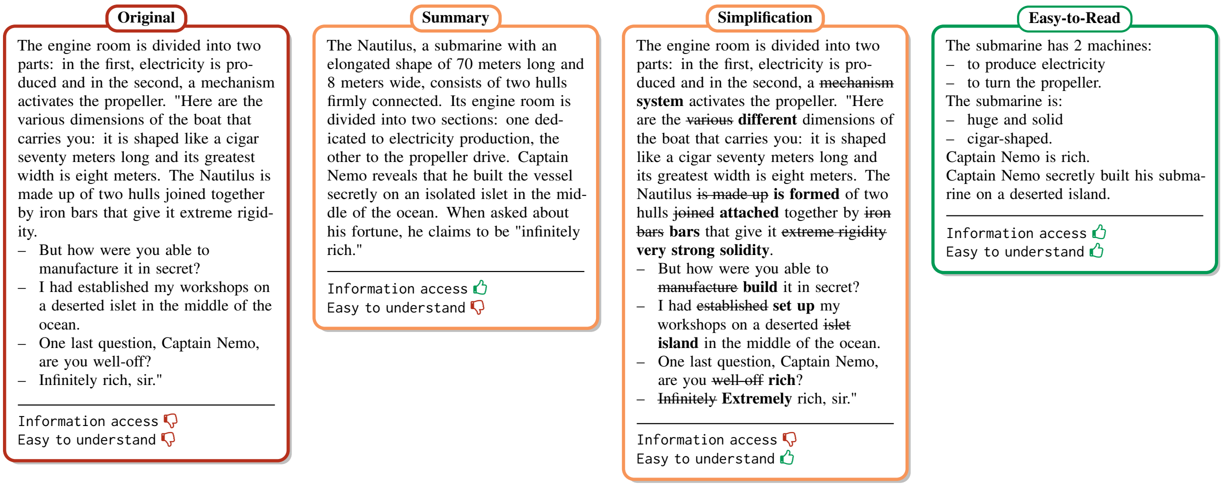



Figure 1: Different versions derived from a passage of *Twenty Thousand Leagues Under the Sea* by Jules Verne: from left to right, the original passage, an abstractive summary, a lexical simplification (crossed-out followed by words in bold indicate substitutions), and an Easy-to-Read transcription targeting readers with cognitive impairment.

suitable for model training. Consequently, prior studies (Martínez et al., 2024; Sun et al., 2023) have approached the ETR task by leveraging sentence simplification or summarization resources, which fall short of fully meeting ETR-specific requirements.

In this paper, we address these gaps by introducing ETR-fr, the first dataset of 523 paragraph-aligned text pairs fully compliant with the European ETR guidelines (Pathways, 2021). We explore multi-task learning (MTL) to boost ETR generation by combining summarization and simplification, traditionally applied in isolation. In particular, we evaluate two MTL strategies: in-context learning (ICL) via a multi-task variant of retrieval-augmented generation (RAG), and parameter-efficient fine-tuning (PEFT) using MTL-LoRA (Yang et al., 2025). Experiments are conducted on Mistral-7B (Jiang et al., 2023) and LLaMA-3-8B (Grattafiori et al., 2024), and compared against single-task baselines. The evaluation framework combines standard automatic metrics with detailed human assessment based on a 38-point rubric from the European ETR guidelines, measuring clarity, coherence, and accessibility. The experiments conducted on ETR-fr indicate that, in the majority of cases, MTL setups provide clear advantages over single-task baselines. Furthermore, the results indicate that the MTL-RAG-based strategy supports better generalization in out-of-domain scenarios, while MTL-LoRA consistently achieves superior performance in in-domain settings.

Our contributions are: (1) we release ETR-fr, the first high-quality, paragraph-aligned dataset for ETR generation, fully compliant with European guidelines and in the French language; (2) we benchmark MTL-RAG and MTL-LoRA approaches for ETR generation; (3) we propose a comprehensive evaluation combining automatic and human assessment based on official European ETR standards; (4) we evaluate model generalization to new domains, including political texts aimed at fostering civic engagement among individuals with cognitive disabilities.

## 2 Related Work

**Inclusive Text Generation.** Recent research has aimed to support communication for users with cognitive impairments, often through dialogue systems (Martin and Nagalakshmi, 2024; Murillo-Morales et al., 2020; Huq et al., 2024; Wang et al., 2024). Much of this work has focused on dyslexia. For example, Goodman et al. (2022) introduced an email writing assistant built on LaMDA (Thoppilan et al., 2022), but observed that its outputs often lacked precision. In the French context, HECTOR (Todirascu et al., 2022) investigated lexical and syntactic simplification, with mixed outcomes.

Similar challenges are observed across other languages. In German, several studies explore simplification for individuals with learning difficulties, though often without referencing the ETR framework (Hansen-Schirra et al., 2020; Anschütz et al., 2023; Deilen et al., 2023; Stodden et al., 2023). For English, relevant work includes Yaneva (2015). In

Finnish, Dmitrieva and Tiedemann (2024) trained mBART (Liu et al., 2020) and FinGPT (Luukkonen et al., 2023) on the automatically aligned Easy-Finnish dataset, though text pairings may be inaccurate and the data does not fully follow ETR guidelines. In Spanish, ClearText (Espinosa-Zaragoza et al., 2023) leverages ChatGPT to simplify administrative texts, though its corpus remains limited and prone to errors. Additionally, Martínez et al. (2024) constructed a sentence-level simplification dataset and fine-tuned LLaMA-2 (Touvron et al., 2023b), revealing that translation-based methods are vulnerable to semantic drift and domain mismatches.

**In-Context Learning (ICL).** ICL allows LLMs to learn tasks from examples without parameter updates (Brown et al., 2020; Chowdhery et al., 2023; OpenAI, 2023; Touvron et al., 2023a). Instruction tuning and Chain-of-Thought (CoT) prompting have been shown to improve task performance and reasoning (Liu et al., 2023; Wei et al., 2022; Yin et al., 2023). Tang et al. (2023) assessed ICL for controlled summarization, focusing on entity inclusion and length constraints. They observed that smaller models offered stronger controllability, while larger models achieved higher ROUGE scores. However, precise length control remained challenging. Prompt quality and exemplar selection critically affect ICL outcomes (Lu et al., 2022; Dong et al., 2024). Retrieval-augmented methods (Liu et al., 2022; Ram et al., 2023) have been proposed to improve exemplar selection. For simplification, Vadlamannati and Şahin (2023) have used metric-based selection (e.g., SARI, BERTScore) to improve output quality. Multi-task ICL and cross-task prompting (Bhasin et al., 2024; Shi et al., 2024; Chatterjee et al., 2024) further enhance generalization and stability, especially on unseen tasks, by leveraging format-aware prompts and semantically related exemplars.

**PEFT for Multi-Task Learning.** Parameter-efficient fine-tuning (PEFT) methods such as LoRA (Hu et al., 2022), QLoRA (Dettmers et al., 2023) and DoRA (Liu et al., 2024b) enable scalable adaptation of LLMs by modifying only a subset of parameters. LoRA leverages the intrinsic dimensionality of language models to achieve strong performance with minimal computational overhead. However, LoRA-based strategies struggle in multi-task settings due to conflicting updates across tasks (Wang et al., 2023). Alternatives such as MultiLoRA (Wang et al., 2023) and MoELoRA (Liu et al., 2024a) aim to balance generalization with task specificity, but still face challenges related to task routing and interference mitigation. To overcome these limitations, Yang et al. (2025) introduced MTL-LoRA, which combines shared and task-specific modules, achieving competitive results on GLUE (Wang et al., 2018) with fewer trainable parameters.

## 3 ETR-fr Dataset

While several datasets exist for text simplification and summarization (Gala et al., 2020; Hauser et al., 2022; Kamal Eddine et al., 2021; Liu et al., 2018), there remains a notable lack of high-quality, document-aligned corpora for ETR generation. To address this gap, we introduce the ETR-fr dataset, constructed from the François Baudez Publishing collection[4], which provides literature specifically designed for readers with cognitive impairments, following European ETR guidelines. A dataset sheet (Gebru et al., 2021), outlining the data collection methodology, preprocessing steps, and distribution details, is provided in Appendix D.

**Description.** ETR-fr consists of 523 paragraph-aligned text pairs in French. Table 1 outlines key dataset statistics, including KMRE readability score (Kandel and Moles, 1958), compression ratios (Kamal Eddine et al., 2021), and lexical novelty Narayan et al. (2018). On average, the dataset yields a compression rate of 50.05%, with a reduction of 56.61 words and 2.17 sentences per pair. The average novelty rate is 53.80%, reflecting the proportion of newly introduced unigrams in target texts. Readability improves by 7.51 KMRE points from source to target.

**Dataset Splits.** The dataset is partitioned into fixed train, validation, and test 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. This setup was used to test hard configurations for ETR generation and ensure non-thematic and lexical overlap.

**ETR-fr-politic.** To assess generalization and robustness, we introduce ETR-fr-politic, an out-of-domain test set with 33 ETR paragraphs sampled

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

| | # Examples | # Words | | # Sentences | | Sentence length | | KMRE ↑ | | Novelty (%) | Comp. ratio (%) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | source | target | source | target | source | target | source | target | | |
| **ETR-fr** | 523 | 102.76 | 46.15 | 9.30 | 7.13 | 12.57 | 7.89 | 91.43 | 98.94 | 53.80 | 50.05 |
| **Train** | 399 | 99.70 | 46.50 | 8.92 | 7.48 | 12.57 | 6.92 | 91.03 | 99.71 | 53.79 | 49.04 |
| **Dev** | 71 | 100.76 | 48.59 | 9.03 | 7.77 | 13.59 | 6.90 | 89.50 | 100.59 | 52.96 | 44.47 |
| **Test** | 53 | 128.47 | 40.26 | 12.51 | 10.34 | 11.16 | 3.97 | 97.02 | 103.67 | 55.01 | 65.19 |
| **ETR-fr-politic** | 33 | 96.27 | 62.85 | 6.03 | 6.42 | 16.69 | 11.84 | 74.00 | 87.74 | 63.78 | 29.17 |
| **WikiLarge FR** | 296,402 | 34.88 | 29.28 | 1.68 | 1.56 | 27.53 | 23.74 | 65.38 | 71.35 | 31.97 | 12.79 |
| **OrangeSum** | 24,401 | 375.98 | 34.00 | 17.15 | 1.86 | 22.77 | 21.68 | 69.80 | 68.32 | 38.24 | 89.16 |

Table 1: **Statistics across ETR-fr, ETR-fr-politic, and ETR-related tasks**, i.e. sentence simplification and text summarization with WikiLarge FR and OrangeSum. Results are reported on average per document.

from the 2022 French presidential election programs, which adhere to ETR guidelines[5] and manually aligned. Compared to ETR-fr test set, the ETR-fr-politic dataset features shorter source texts (96.27 vs. 128.47 words) and fewer sentences (6.03 vs. 12.51), but yields longer rewritten outputs (62.85 vs. 40.26 words). Additionally, ETR-fr-politic exhibits higher novelty (63.78% vs. 55.01%) and significantly lower compression ratios (29.17% vs. 65.19%), indicating a greater degree of content expansion. While ETR-fr test set exhibits higher overall simplicity scores both before and after rewriting (97.02 and 103.67) compared to ETR-fr-politic (74.00 and 87.74), the latter achieves a greater simplification gain, with a larger increase in KMRE (+13.74 vs. +6.65 points). Overall, ETR-fr-politic poses a more challenging and higher-novelty setting for evaluating ETR systems in politically sensitive, real-world rewriting contexts.

**ETR-fr vs. Related Tasks.** Table 1 compares ETR-fr with two gold-standard datasets on related tasks, respectively text simplification and summarization: WikiLarge FR (Cardon and Grabar, 2020) and OrangeSum (Kamal Eddine et al., 2021). While WikiLarge FR is larger (296K sentence pairs), it is limited to sentence-level simplification, with short inputs (34.88 words, 1.68 sentences on average). In contrast, ETR-fr and OrangeSum support transformations at the paragraph and document levels, respectively, providing significantly longer inputs of 102.76 and 375.98 words. ETR-fr demonstrates a balanced compression ratio (50.05%) higher than WikiLarge FR (12.79%) but lower than the extreme summarization found in OrangeSum (89.16%). Notably, ETR-fr offers the highest lexical richness and abstraction, evidenced by its top KMRE scores (91.43 source, 98.94 target) and novelty rate (53.80%). Simplified outputs also exhibit syntactic simplification, with shorter sentence lengths (7.89 words per sentence). In summary, while WikiLarge FR is suited for sentence-level simplification and OrangeSum for summarization, ETR-fr supports paragraph-level simplification, emphasizing lexical and structural transformation, making it well-suited for users with cognitive disabilities.

## 4 Multi-Task ETR Generation

### 4.1 Datasets, LLMs and Metrics

**Datasets.** Our experiments leverage the ETR-fr dataset as the primary resource, supplemented by related rewriting tasks sourced from the OrangeSum summarization dataset and the sentence simplification dataset WikiLarge FR.

**Models.** To evaluate the effectiveness of MTL for ETR transcription, we selected two recent LLMs that demonstrate strong generalization capabilities across a variety of NLP tasks : Llama3-8B (Grattafiori et al., 2024) and Mistral-7B (Jiang et al., 2023)[6]. Note that foundation models are used for PEFT and their Instruct versions for ICL.

**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 (Lin, 2004), along with BERTScore (Zhang et al., 2020). For simplification, we include SARI (Xu et al., 2016), the novelty ratio for new unigrams (Kamal Eddine et al., 2021). BLEU (Papineni et al., 2002) and KMRE, are excluded, as it is unsuitable for text simplification (Sulem et al., 2018; Xu et al., 2016; Tanprasert and Kauchak, 2021). To

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

[6] We evaluated the DeepSeek-R1-8B model. Its performance was notably lower than that of the other models. Results are reported in Table 5 from Appendix B.1

unify quality assessment of ETR texts, we propose SRB, a composite score combining SARI, ROUGE-L, and BERTScore-F1 via harmonic mean. This metric captures simplification, summarization, and meaning preservation for holistic ETR evaluation.

More details about metrics and models are available in Appendix A.

### 4.2 Multi-Task In-Context Learning

**Single-Task Baselines.** As baseline, we evaluate three single-task in-context learning strategies: zero-shot prompting (Kojima et al., 2022), chain-of-thought prompting (Wei et al., 2022), and retrieval-augmented generation (Lewis et al., 2020). In the zero-shot setting, the model is provided only with ETR task-specific instructions, without any examples, serving as a baseline to assess the model's ability to generalize purely from the prompt. To enhance reasoning in more complex tasks, we incorporate CoT prompting, which explicitly elicits intermediate reasoning steps in the prompt. For a fair and reproducible evaluation, we use consistent instruction-based prompt templates across all models, as detailed in Appendix C.

**Multi-Task RAG.** To enable few-shot multi-task ICL, we implement a multi-task RAG. Demonstrations from multiple tasks are retrieved and incorporated into the prompt. We explore three sequencing strategies for organizing demonstrations within the prompt context, which are listed as follows.

*Random Ordering:* Examples from all 3 tasks are interleaved in a fully randomized manner (e.g., $t_1, t_3, t_3, t_2, t_1, t_1, t_3, t_2, t_2$), serving as a baseline to assess robustness to prompt structure.

*Task-Grouped Ordering:* Examples are grouped by task, presenting all demonstrations from one task before moving to the next one (e.g., $t_1, t_1, t_1, t_2, t_2, t_2, t_3, t_3, t_3$). This structure emphasizes intra-task consistency.

*Task-Interleaved Ordering:* Examples alternate across tasks at each shot level, maintaining a round-robin pattern (e.g., $t_1, t_2, t_3, t_1, t_2, t_3, t_1, t_2, t_3$). This configuration aims to balance exposure across tasks within the prompt.

The impact of the number of shots per task and example orderings is shown in Appendix C (Figure 3a and Figure 3b). Note that to encode examples into dense vector representations, we use the `jina-embeddings-v3` (Sturua et al., 2024) model, and for distance computation, we employ the L2 distance metric.

### 4.3 Multi-Task PEFT

**LoRA Baseline.** As baseline, we implement LoRA (Hu et al., 2022). LoRA approximates full fine-tuning by decomposing weight matrices into low-rank components. To reduce dimensionality, the weight matrix $\mathbf{W}_0 \in \mathbb{R}^{d\times k}$ is approximated by the product of two lower-rank matrices: $\mathbf{B} \in \mathbb{R}^{d\times r}$ and $\mathbf{A} \in \mathbb{R}^{r\times k}$, with $r \ll \min(d, k)$. This low-rank update preserves the backbone while enabling efficient adaptation, such that $h = \mathbf{W_0}x + \frac{\alpha}{r}\mathbf{BA}x$. LoRA can be applied to each linear layer in the Transformer architecture, such as $\mathbf{W_Q}, \mathbf{W_K}, \mathbf{W_V}, \mathbf{W_O}$ matrices projections in the attention layers.

**MTL-LoRA.** Yang et al. (2025) introduce MTL-LoRA to face challenges related to task routing and interference mitigation. Given task input $x_t$, MTL-LoRA first applies a shared standard LoRA down-projection via matrix $\mathbf{A} \in \mathbb{R}^{r\times k}$. To retain task-specific information, it inserts a task-specific low-rank matrix $\Lambda_t \in \mathbb{R}^{r\times r}$ between the down- and up-projections, transforming $\mathbf{A}x_t$. Instead of a single shared up-projection, MTL-LoRA uses $n$ matrices $\mathbf{B}^i \in \mathbb{R}^{d\times r}$ to support diverse knowledge-sharing strategies. Outputs are combined via a weighted average, where weights $w_t \in \mathbb{R}^{n\times 1}$ are learned per task as in Equation 1.

$$h_t = \mathbf{W}x_t + \sum_{i=1}^{n} \frac{\exp(w_t^i/\tau)\mathbf{B}^i}{\sum_{j=1}^{n}\exp(w_t^j/\tau)}\Lambda_t\mathbf{A}x_t \quad (1)$$

Here, $\tau$ controls the softness of the weighting. Each $\Lambda_t$ is initialized as a diagonal identity matrix to ensure $\Delta\mathbf{W}_t = 0$ at start.

**MTL Loss for ETR Generation.** The model is trained to generate outputs conditioned on instructions. Given an instruction sequence $I = i_1, i_2, \ldots, i_m$ and a corresponding completion sequence $C = c_1, c_2, \ldots, c_n$, where $I$ may contain special prompt tokens (e.g., `<Input>` and `<Output>`), the full input is represented as $x = i_1, \ldots, i_m, c_1, \ldots, c_n$. The model is trained to autoregressively predict each token in $C$ conditioned on all preceding tokens in $I$ and $C$ as defined in Equation 2.

$$P(C|I) = \prod_{j=1}^{n} P(c_j \mid i_1, ..., i_m, c_1, ..., c_{j-1}) \quad (2)$$

Based on the findings of Huerta-Enochian and Ko (2024), the objective is to minimize the negative log-likelihood of the completion sequence given the instruction as defined in Equation 3.

$$\mathcal{L} = -\sum_{j=1}^{n} \log P(c_j \mid i_1, ..., i_m, c_1, ..., c_{j-1}) \quad (3)$$

To account for imbalance across different instruction-following tasks, we apply a task-specific weighting scheme during training. Let $N_t$ be the number of training examples for task $t$, and let $N = \sum_t N_t$ be the total number of training examples across all tasks. Each task's contribution to the overall loss is scaled by a factor $w_t = \frac{N_t}{N}$, such that the final loss is redefined in Equation 4.

$$\mathcal{L}_{MTL} = \sum_{t=1}^{T} w_t \times \mathcal{L}_t \quad (4)$$

## 5 Results

The top-performing models are chosen according to their highest SRB scores on the ETR-fr validation set, using a grid search strategy for hyperparameter tuning (see Appendix A for details). To complement this analysis, all models are run five times with different seeds, and detailed average results are in Appendix B.

### 5.1 In-Domain Quantitative Results

**ICL Performance.** As shown in Table 2a, ICL models evidence steady improvements when transitioning from zero-shot and CoT prompting to RAG-based prompting. For LlaMA-3-8B, RAG achieves the best results with ETR-fr only inputs (e.g., 33.43/12.99/24.38 ROUGE-1/2/L and 42.16 SARI), outperforming zero-shot by a large margin. Adding related tasks does not consistently improve performance under ICL, and in some cases, leads to reduced novelty and compression ratio.

**Impact of Fine-Tuning.** PEFT significantly outperforms ICL methods. The best overall performance is achieved by LlaMA-3-8B with MTL-LoRA fine-tuned on ETR-fr and WikiLarge FR, obtaining highest scores across SARI (44.67), BERTScore-F1 (74.05), SRB (39.60), and compression ratio (56.11), while maintaining strong novelty (33.05).

**LLM Comparison.** Across both prompting and fine-tuning paradigms, LlaMA-3-8B outperforms Mistral-7B in most metrics. For instance, with LoRA fine-tuning on ETR-fr, LlaMA-3-8B achieves higher ROUGE-L (25.04 vs. 24.02), SARI (42.15 vs. 42.09), and SRB (38.77 vs. 37.98). This suggests that the architectural or scale advantages of LlaMA-3-8B translate effectively into more efficient capabilities.

**Combination of Tasks.** Incorporating auxiliary tasks such as text summarization and simplification can provide complementary supervision, as seen in PEFT strategies. However, they do not yield performance gains in the ICL setting. Notably, MTL-LoRA with ETR-fr and WikiLarge FR for LlaMA-3-8B achieves the highest SARI and compression ratio, suggesting the relevance of sentence simplification data to the ETR generation task. However, inclusion of all three tasks does not universally yield the best results, and in some cases, introduces performance regressions in BERTScore and novelty. This implies that careful curation of task mixtures is essential to avoid dilution or conflict between training objectives. Overall, these results highlight that while RAG improves performance in ICL, parameter-efficient fine-tuning (particularly MTL-LoRA) remains the most effective method within the in-domain ETR-fr setting.

### 5.2 Out-of-Domain Quantitative Results

**ICL Performance.** As shown in Table 2b, among prompting strategies, RAG consistently outperforms zero-shot and CoT in all major content preservation metrics (ROUGE-1/2/L, BERTScore-F1) and the composite SRB score. On LlaMA-3-8B, using RAG with all three tasks (E,O,W) achieves the highest overall SRB score (41.52) and the best ROUGE-L (28.43), indicating its strong generalization and content fidelity. Moreover, it yields the highest SARI (42.63) and BERTScore-F1 (73.39), showcasing a balanced ability to simplify while preserving semantics. Interestingly, zero-shot exhibits extremely poor compression ratios, especially on Mistral-7B (-309.24), suggesting potential prompt misalignment or excessive hallucination. However, it achieves the highest novelty score (55.37) on LlaMA-3-8B, implying that despite poor content fidelity, more diverse lexical

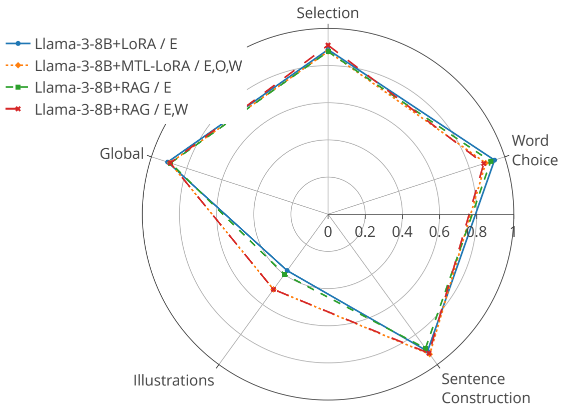
| | Method | Task | R-1 ↑ | R-2 ↑ | R-L ↑ | SARI ↑ | BERT-F1 ↑ | SRB ↑ | Comp. ratio | Novelty |
|---|---|---|---|---|---|---|---|---|---|---|
| **In-Context Learning** | | | | | | | | | | |
| **Mistral-7B** | **Zero-Shot** | **E** | 23.92 | 7.09 | 16.28 | 37.07 | 69.75 | 29.20 | −64.14 | 35.70 |
| | **CoT** | **E** | 23.58 | 7.22 | 16.17 | 37.39 | 68.80 | 29.10 | −60.53 | 36.09 |
| | **RAG** | **E** | 32.14 | 10.47 | 22.72 | 40.05 | 72.41 | 36.24 | 44.32 | 26.55 |
| | | **E,O** | 31.12 | 9.58 | 21.92 | 39.54 | 71.29 | 35.32 | 48.45 | 26.61 |
| | | **E,W** | 30.29 | 9.69 | 21.29 | 38.69 | 71.59 | 34.56 | 33.80 | 23.01 |
| | | **E,O,W** | 29.84 | 9.57 | 21.58 | 39.53 | 71.06 | 35.01 | 46.42 | 25.85 |
| **LlaMA-3-8B** | **Zero-Shot** | **E** | 24.94 | 8.23 | 17.37 | 38.59 | 70.29 | 30.70 | −21.56 | **38.73** |
| | **CoT** | **E** | 27.57 | 8.96 | 18.72 | 38.26 | 71.02 | 32.04 | 7.80 | 31.10 |
| | **RAG** | **E** | 33.43 | 12.99 | 24.38 | 42.16 | 72.58 | 38.21 | 46.18 | 27.14 |
| | | **E,O** | 31.10 | 10.87 | 22.37 | 39.94 | 71.27 | 35.81 | 39.22 | 24.29 |
| | | **E,W** | 33.03 | 11.62 | 23.28 | 40.59 | 72.14 | 36.83 | 41.89 | 25.26 |
| | | **E,O,W** | 29.35 | 9.97 | 20.54 | 39.03 | 70.84 | 33.93 | 25.94 | 23.69 |
| **Parameter-Efficient Fine-Tuning** | | | | | | | | | | |
| **Mistral-7B** | **LoRA** | **E** | 32.47 | 12.40 | 24.02 | 42.09 | 73.56 | 37.98 | 44.42 | 18.35 |
| | **MTL-LoRA** | **E,O** | 32.67 | 12.74 | 24.33 | 41.95 | 73.52 | 38.20 | 53.48 | 24.17 |
| | | **E,W** | 32.62 | 12.92 | 24.28 | 42.53 | 73.90 | 38.35 | 53.62 | 24.99 |
| | | **E,O,W** | **33.65** | 12.83 | 24.93 | 42.25 | 73.62 | 38.77 | 48.93 | 23.38 |
| **LlaMA-3-8B** | **LoRA** | **E** | 31.76 | 13.17 | 25.04 | 42.15 | 72.93 | 38.77 | 50.66 | 18.87 |
| | **MTL-LoRA** | **E,O** | 33.44 | 13.22 | 24.24 | 43.04 | 73.86 | 38.45 | 51.36 | 23.06 |
| | | **E,W** | 32.54 | 13.56 | 25.08 | **44.67** | **74.05** | 39.60 | **56.11** | 33.05 |
| | | **E,O,W** | 32.78 | **13.64** | **25.67** | 43.53 | 73.28 | **39.69** | 53.24 | 24.39 |

(a) **ETR-fr test set** (In-Domain).

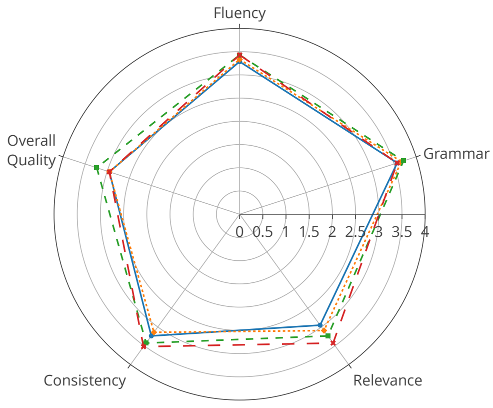
| | Method | Task | R-1 ↑ | R-2 ↑ | R-L ↑ | SARI ↑ | BERT-F1 ↑ | SRB ↑ | Comp. ratio | Novelty |
|---|---|---|---|---|---|---|---|---|---|---|
| **In-Context Learning** | | | | | | | | | | |
| **Mistral-7B** | **Zero-Shot** | **E** | 28.36 | 11.02 | 19.29 | 39.87 | 68.10 | 32.75 | −309.24 | 48.37 |
| | **CoT** | **E** | 29.78 | 11.22 | 19.90 | 39.62 | 69.40 | 33.37 | −261.30 | 50.85 |
| | **RAG** | **E** | 39.22 | 15.28 | 28.12 | 41.33 | 73.15 | 40.86 | 11.03 | 25.49 |
| | | **E,O** | 37.87 | 14.59 | 26.43 | 39.51 | 72.08 | 38.96 | 14.37 | 18.41 |
| | | **E,W** | 39.77 | 15.55 | 27.74 | 40.32 | 72.47 | 40.19 | 10.80 | 17.81 |
| | | **E,O,W** | 39.12 | 15.97 | 28.26 | 40.74 | 72.87 | 40.73 | 14.63 | 18.33 |
| **LlaMA-3-8B** | **Zero-Shot** | **E** | 29.60 | 10.84 | 18.83 | 40.55 | 68.68 | 32.50 | −180.74 | **55.37** |
| | **CoT** | **E** | 31.68 | 11.46 | 20.14 | 40.80 | 69.87 | 33.91 | −83.36 | 45.41 |
| | **RAG** | **E** | 37.48 | 13.98 | 26.94 | 41.05 | 73.18 | 39.92 | 11.37 | 41.63 |
| | | **E,O** | **40.53** | 15.15 | 27.47 | 41.14 | 72.75 | 40.29 | −12.56 | 31.01 |
| | | **E,W** | 39.72 | 16.02 | 26.83 | 41.99 | 73.32 | 40.15 | 13.75 | 35.70 |
| | | **E,O,W** | 40.12 | **16.55** | **28.43** | **42.63** | **73.39** | **41.52** | −4.79 | 30.08 |
| **Parameter-Efficient Fine-Tuning** | | | | | | | | | | |
| **Mistral-7B** | **LoRA** | **E** | 35.13 | 12.23 | 25.93 | 38.04 | 70.28 | 37.94 | 21.55 | 11.79 |
| | **MTL-LoRA** | **E,O** | 29.36 | 11.02 | 21.87 | 38.68 | 69.22 | 34.87 | **36.68** | 40.29 |
| | | **E,W** | 34.32 | 12.56 | 24.85 | 38.72 | 70.54 | 37.38 | 22.51 | 19.10 |
| | | **E,O,W** | 36.45 | 13.22 | 26.21 | 38.39 | 70.97 | 38.32 | 18.33 | 10.55 |
| **LlaMA-3-8B** | **LoRA** | **E** | 35.53 | 13.83 | 26.94 | 39.90 | 71.30 | 39.37 | 6.38 | 16.13 |
| | **MTL-LoRA** | **E,O** | 32.77 | 12.20 | 24.23 | 38.84 | 69.74 | 36.88 | 18.26 | 19.30 |
| | | **E,W** | 37.46 | 13.74 | 27.06 | 38.26 | 71.30 | 38.90 | 8.45 | 6.44 |
| | | **E,O,W** | 36.48 | 13.69 | 25.90 | 36.19 | 70.97 | 37.35 | 8.68 | 2.06 |

(b) **ETR-fr-politic test set** (Ouf-of-Domain).

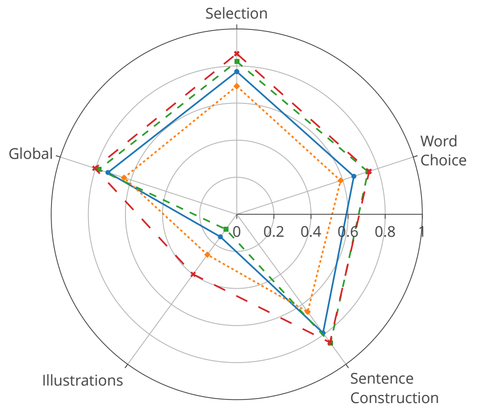
Table 2: **Performance comparison**, across ICL methods and PEFT strategies on three tasks: ETR-fr (E), OrangeSum (O) and WikiLarge FR (W). Best results are in **bold**, second-best are underlined.



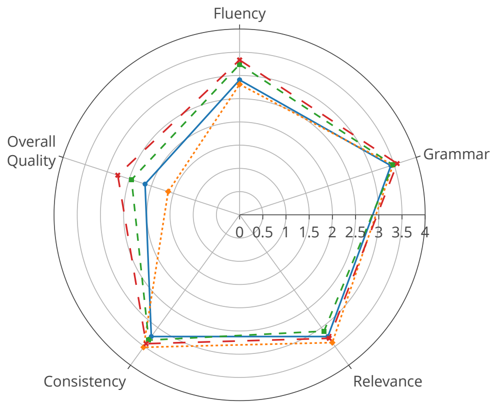

(a) Criteria on ETR-fr




(b) Quality on ETR-fr




(c) Criteria on ETR-fr-politic




(d) Quality on ETR-fr-politic

Figure 2: **Human evaluation of generation quality on ETR-fr and ETR-fr-politic** using their optimal ICL and MTL configurations. Subfigures (a) and (c) show average scores based on the ETR guideline criteria. Subfigures (b) and (d) present average human ratings for text generation quality.

outputs are generated.

**Impact of Fine-Tuning.** While PEFT strategies generally lag behind RAG in terms of SRB and BERTScore, they offer notably better compression ratios than zero-shot, CoT and most RAG-based strategies. The best PEFT model in terms of SRB, LLaMA-3-8B+LoRA trained solely on ETR-fr, achieves a relatively low compression ratio (6.38), indicating only moderate summarization. However, this comes at the expense of lower ROUGE, SARI, and BERTScore metrics compared to RAG-based approaches. Additionally, MTL-LoRA configurations do not demonstrate performance improvements over single-task LoRA in out-of-domain (OOD) settings, particularly on LlaMA-3-8B, suggesting a tendency toward overspecialization on the target task of ETR derived from children's books.

**Combination of Tasks.** Prompting or training with multiple datasets (E,O,W) can improve OOD generalization. LLaMA-3-8B+RAG and Mistral-7B+RAG show substantial gains across all metrics compared to single-task prompting, confirming the benefits of multi-domain exposure in OOD settings. This situation is mitigated for the PEFT strategy, where performance improvement is backbone-dependent. While Mistral-7B+MTL-LoRA steadily benefits from concurrent learning achieving best results in terms of SRB with its (E,O,W) configuration, overall best results with LLaMA-3-8B are obtained with a single task setting.

### 5.3 Human Evaluation

Manual evaluation is essential for assessing ETR text quality and compliance with European guidelines, which include 57 weighted questions covering clarity, simplicity, and accessibility,[7] to ensure content is understandable and appropriate for

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

the target audience. We validated our approach through human evaluation with 10 native French speakers, 7 NLP researchers and 3 linguists, all volunteers, who assessed outputs from the ETR-fr and ETR-politic test sets[8]. We evaluated outputs generated by two model configurations: (1) Llama-3-8B+RAG augmented with ETR-fr (E) and WikiLarge FR (W), and (2) Llama-3-8B+MTL-LoRA trained on ETR-fr, OrangeSum (O), and WikiLarge FR, alongside their respective single-task variants. These models were chosen as the best performing ones, respectively for ICL and PEFT, for in-domain settings. The evaluation was performed on 6 source documents (3 from ETR-fr and 3 from ETR-fr-politic test sets). Each annotator reviewed 24 outputs, resulting in 60 samples per model and a total of 240 different samples evaluated. The assessment prioritized the most critical ETR guideline criteria, including information selection, sentence construction, word choice, and illustrations, covering 38 detailed questions (see Table 9 in Appendix). Additionally, we assessed general text generation quality metrics such as Fluency, Grammar/Spelling, Relevance, Textual Coherence, and Overall Perceived Quality, through additional five questions. ETR criteria were rated on a binary scale (respected, not respected, not applicable), whereas human judgments used a 5-point Likert scale (0–4).

**In-domain Results.** Figures 2 presents the human evaluation results and overall scores are provided in a table in Appendix B.2. On ETR-fr, all methods perform well with respect to the European ETR guidelines. LoRA achieves the highest overall validation rate of 0.91, particularly excelling in word choice and sentence construction. MTL-LoRA+(E,O,W) shows the best results for sentence construction, while RAG+(E,W) outperforms other models in information selection. In terms of text generation quality, single-task RAG leads with an overall score of 4.24, driven by strong performance in fluency, grammar, and coherence. While MTL-LoRA+(E,O,W) and LoRA are competitive across individual criteria, with MTL-LoRA+(E,O,W) scoring best on 3 out of 4 dimensions, their overall quality scores are comparable (3.95). Although automatic metrics indicate improved performance in multi-task settings, human evaluation results are more mixed, revealing no clear advantage for single- versus multi-task strategies, except in the Illustrations dimension.

[8]All evaluators received training and were blind to model development to prevent bias.

**Out-of-domain Results** Overall performance declines on the more challenging ETR-fr-politic, yet RAG+(E,W) remains the most robust across both ETR criteria and text quality evaluations, underscoring the value of the multi-task setting. Specifically, RAG+(E,W), trained on a broader mix of tasks combining ETR and sentence simplification, achieves a total validation rate of 0.80 for ETR guidelines and an overall quality score of 3.76. In contrast, MTL-LoRA+(E,O,W) exhibits the sharpest drop in quality (2.62), indicating difficulties in managing politically nuanced content, although it still outperforms the single-task configuration in 3 out of 5 evaluation dimensions. Furthermore, in terms of European ETR compliance, MTL-LoRA+(E,O,W) struggles to generalize in out-of-domain settings, showing improvement only in the Illustrations criterion.

## 6 Conclusion

In this paper, we introduced ETR-fr, the first dataset fully compliant with the European ETR guidelines targeting neurodivergent populations, and explored multi-task learning to improve ETR generation with LLMs. Our experiments show that multi-task setups, particularly RAG for ICL and MTL-LoRA for PEFT, consistently improve performance in both in-domain and OOD settings according to automatic metrics. While human evaluation reveals more nuanced outcomes, it nonetheless confirms the benefits of multi-task learning across a broad range of ETR criteria and text quality dimensions.

## 7 Limitations

The development of ETR generation models introduces important constraints and considerations that reflect the complexity of cognitive accessibility and language model behavior.

**Untested Practical Utility for Users with Disabilities** While our evaluation combines automatic and human assessments, it does not simulate usage in real-world settings such as assistive reading tools or educational platforms. Thus, the practical utility of outputs for users with intellectual disabilities remains untested. We aim to support the responsible co-construction of experiments accordingly with ETR inclusion requirements. Acknowledging these boundaries also helps position ETR generation as a sociotechnical task, one that demands sensitivity to both linguistic quality and lived experience.

**No explicit modeling of cognitive load.** Though our models optimize for readability and fluency, they do not account for cognitive effort. Even simplified outputs may challenge users when processing abstract or ambiguous content.

**ETR guidelines as a fixed supervision target.** We use European ETR guidelines as a normative framework. While they offer structure, rigid adherence may exclude culturally specific or individualized accessibility strategies, limiting generalization.

**Susceptibility to hallucinations.** As with most generative models, hallucinations and factual drift remain concerns, especially with RAG-based systems. This is particularly risky for audiences who may interpret outputs literally or depend on high textual reliability.

**Underexplored ethical considerations.** The automation of content adaptation for cognitive impaired users raises ethical questions around oversimplification, loss of nuance, and possible reinforcement of stereotypes. These dimensions are not addressed in the current evaluation, though they are central to responsible deployment.

## 8 Impact and Ethical Considerations

**Risks of Oversimplification.** Simplified language is not neutral, it involves choices about what meaning is retained or lost. In some cases, simplification may erase nuance, flatten perspective, or reinforce harmful stereotypes. This tension is particularly acute for readers who engage with language differently.

**Toward Responsible Design.** Mitigating risks requires human-in-the-loop systems, participatory evaluation involving end users, and adaptation strategies that go beyond surface-level clarity. ETR guidelines should be viewed as a starting point, not a universal solution.

**Positioning ETR as a Research Problem.** ETR remains underexplored in NLP. By introducing aligned data, task-specific metrics, and a critical lens on modeling assumptions, we aim to establish it as a standalone task, one that demands linguistic sensitivity, practical design, and participatory validation.

## Acknowledgments

We express our gratitude to François Baudez Publishing for generously granting free access to the ETR book collection, which enabled the construction of our dataset.

We also acknowledge the annotators for their thorough and precise contributions to the manual evaluation, which were instrumental in ensuring the validity and robustness of our results.

Finally, this work was carried out using the computing resources provided by CRIANN (Normandy, France).

## References

Miriam Anschütz, Joshua Oehms, Thomas Wimmer, Bartłomiej Jezierski, and Georg Groh. 2023. Language models for German text simplification: Overcoming parallel data scarcity through style-specific pre-training. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1147–1158, Toronto, Canada. Association for Computational Linguistics.

Harmon Bhasin, Timothy Ossowski, Yiqiao Zhong, and Junjie Hu. 2024. How does multi-task training affect transformer in-context capabilities? investigations with function classes. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 169–187, Mexico City, Mexico. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, and Arvind Neelakantan. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Rémi Cardon and Natalia Grabar. 2020. French Biomedical Text Simplification: When Small and Precise Helps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 710–716, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Anwoy Chatterjee, Eshaan Tanwar, Subhabrata Dutta, and Tanmoy Chakraborty. 2024. Language models can exploit cross-task in-context learning for data-scarce novel tasks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11568–11587, Bangkok, Thailand. Association for Computational Linguistics.

Nael Chehab, Hadmut Holken, and Mathilde Malgrange. 2019. Simples - etude recueil des besoins falc. Technical report, SYSTRAN and EPNAK and EPHE and CHArt-LUTIN.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, and Paul Barham. 2023. Palm: Scaling language modeling with pathways. *J. Mach. Learn. Res.*, 24:240:1–240:113.

Silvana Deilen, Sergio Hernández Garrido, Ekaterina Lapshinova-Koltunski, and Christiane Maaß. 2023. Using ChatGPT as a CAT tool in easy language translation. In *Proceedings of the Second Workshop on Text Simplification, Accessibility and Readability*, pages 1–10, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115.

Anna Dmitrieva and Jörg Tiedemann. 2024. Towards Automatic Finnish Text Simplification. In *Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context @ LREC-COLING 2024*, pages 39–50, Torino, Italia. ELRA and ICCL.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128, Miami, Florida, USA. Association for Computational Linguistics.

Isabel Espinosa-Zaragoza, José Abreu-Salas, Paloma Moreda, and Manuel Palomar. 2023. Automatic Text Simplification for People with Cognitive Disabilities: Resource Creation within the ClearText Project. In *Proceedings of the Second Workshop on Text Simplification, Accessibility and Readability*, pages 68–77, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Nils Freyer, Hendrik Kempt, and Lars Klöser. 2024. Easy-read and large language models: on the ethical dimensions of llm-based text simplification. *Ethics and Information Technology*, 26(3).

Núria Gala, Anaïs Tack, Ludivine Javourey-Drevet, Thomas François, and Johannes C. Ziegler. 2020. Alector: A Parallel Corpus of Simplified French Texts with Alignments of Misreadings by Poor and Dyslexic Readers. In *12th Language Resources and Evaluation Conference*, pages 1353–1361, Marseille, France. European Language Resources Association.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM*, 64(12):86–92.

Sian Gooding. 2022. On the ethical considerations of text simplification. In *Ninth Workshop on Speech and Language Processing for Assistive Technologies (SLPAT-2022)*, pages 50–57, Dublin, Ireland. Association for Computational Linguistics.

Steven M. Goodman, Erin Buehler, Patrick Clary, Andy Coenen, Aaron Donsbach, Tiffanie N. Horne, Michal Lahav, Robert MacDonald, Rain Breaw Michaels, Ajit Narayanan, Mahima Pushkarna, Joel Riley, Alex Santana, Lei Shi, Rachel Sweeney, Phil Weaver, Ann Yuan, and Meredith Ringel Morris. 2022. LaMPost: Design and Evaluation of an AI-assisted Email Writing Prototype for Adults with Dyslexia. In *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '22, pages 1–18, New York, NY, USA. Association for Computing Machinery.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models.

Anders Gustavsson, Mikael Svensson, Frank Jacobi, Christer Allgulander, Jordi Alonso, Ettore Beghi, Richard Dodel, Mattias Ekman, Carlo Faravelli, Laura Fratiglioni, Brenda Gannon, et al. 2011. Cost of disorders of the brain in europe 2010. *European Neuropsychopharmacology*, 21(10):718–779.

Lasse Hansen, Ludvig Renbo Olsen, and Kenneth Enevoldsen. 2023. Textdescriptives: A python package for calculating a large variety of metrics from text. *Journal of Open Source Software*, 8(84):5153.

Silvia Hansen-Schirra, Walter Bisang, Arne Nagels, Silke Gutermuth, Julia Fuchs, Liv Borghardt, Silvana Deilen, Anne-Kathrin Gros, Laura Schiffl, and Johanna Sommer. 2020. *Intralingual Translation into Easy Language - or how to reduce cognitive processing costs*, page 197–225. Easy - Plain - Accessible. Frank & Timme.

Renate Hauser, Jannis Vamvas, Sarah Ebling, and Martin Volk. 2022. 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. European Language Resources Association.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International conference on learning representations*.

Mathew Huerta-Enochian and Seung Yong Ko. 2024. Instruction fine-tuning: Does prompt loss matter? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22771–22795, Miami, Florida, USA. Association for Computational Linguistics.

Syed Mahmudul Huq, Rytis Maskeliūnas, and Robertas Damaševičius. 2024. Dialogue agents for artificial intelligence-based conversational systems for cognitively disabled: a systematic review. *Disability and Rehabilitation: Assistive Technology*, 19(3):1059–1078.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Moussa Kamal Eddine, Antoine Tixier, and Michalis Vazirgiannis. 2021. BARThez: a Skilled Pretrained French Sequence-to-Sequence Model. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9369–9390, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Liliane Kandel and Abraham Moles. 1958. Application de l'indice de Flesch à la langue francaise. *Cahiers Etudes de Radio-Télévision*, 19.

J. P. Kincaid and And Others. 1975. 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. ERIC Number: ED108134.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.

Ariel N. Lee, Cole J. Hunter, and Nataniel Ruiz. 2023. Platypus: Quick, Cheap, and Powerful Refinement of LLMs.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for gpt-3? In *Proceedings of Deep Learning Inside Out: The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures, DeeLIO@ACL 2022, Dublin, Ireland and Online, May 27, 2022*, pages 100–114. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9).

Peter J. Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating wikipedia by summarizing long sequences. In *International Conference on Learning Representations*.

Qidong Liu, Xian Wu, Xiangyu Zhao, Yuanshao Zhu, Derong Xu, Feng Tian, and Yefeng Zheng. 2024a. When MOE Meets LLMs: Parameter Efficient Fine-tuning for Multi-task Medical Applications. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 1104–1114, New York, NY, USA. Association for Computing Machinery.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. 2024b. DoRA: Weight-decomposed low-rank adaptation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 32100–32121. PMLR.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual Denoising Pre-training for Neural Machine Translation. *Transactions of the Association for Computational Linguistics*, 8:726–742. Place: Cambridge, MA Publisher: MIT Press.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

Risto Luukkonen, Ville Komulainen, Jouni Luoma, Anni Eskelinen, Jenna Kanerva, Hanna-Mari Kupari, Filip Ginter, Veronika Laippala, Niklas Muennighoff, Aleksandra Piktus, Thomas Wang, Nouamane Tazi, Teven Scao, Thomas Wolf, Osma Suominen, Samuli Sairanen, Mikko Merioksa, Jyrki Heinonen, Aija Vahtola, Samuel Antao, and Sampo Pyysalo. 2023. FinGPT: Large Generative Models for a Small Language. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2710–2726, Singapore. Association for Computational Linguistics.

Christiane Maaß. 2020. *Easy language–plain language–easy language plus: Balancing comprehensibility and acceptability*. Frank & Timme.

Lara J Martin and Malathy Nagalakshmi. 2024. Bridging the social & technical divide in augmentative and alternative communication (aac) applications for autistic adults. *arXiv preprint arXiv:2404.17730*.

Paloma Martínez, Alberto Ramos, and Lourdes Moreno. 2024. Exploring large language models to generate Easy to Read content. *Frontiers in Computer Science*, 6. Publisher: Frontiers.

Pallab K. Maulik, Maya N. Mascarenhas, Colin D. Mathers, Tarun Dua, and Shekhar Saxena. 2011. Prevalence of intellectual disability: A meta-analysis of population-based studies. *Research in Developmental Disabilities*, 32(2):419–436.

Tomas Murillo-Morales, Peter Heumader, and Klaus Miesenberger. 2020. Automatic Assistance to Cognitive Disabled Web Users via Reinforcement Learning on the Browser. In *Computers Helping People with Special Needs*, pages 61–72, Cham. Springer International Publishing.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

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

Gustavo Paetzold and Lucia Specia. 2016. Unsupervised lexical simplification for non-native speakers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Pathways. 2021. Information for all: European standards for making information easy to read and understand.

Clifton Poth, Hannah Sterz, Indraneil Paul, Sukannya Purkayastha, Leon Engländer, Timo Imhof, Ivan Vulić, Sebastian Ruder, Iryna Gurevych, and Jonas Pfeiffer. 2023. Adapters: A Unified Library for Parameter-Efficient and Modular Transfer Learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 149–160, Singapore. Association for Computational Linguistics.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A Neural Attention Model for Abstractive Sentence Summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, Lisbon, Portugal. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Zhenmei Shi, Junyi Wei, Zhuoyan Xu, and Yingyu Liang. 2024. Why larger language models do in-context learning differently? In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 44991–45013. PMLR.

Sanja Stajner. 2021. Automatic text simplification for social good: Progress and challenges. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2637–2652, Online. Association for Computational Linguistics.

Regina Stodden, Omar Momen, and Laura Kallmeyer. 2023. DEplain: A German parallel corpus with intralingual translations into plain language for sentence and document simplification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16441–16463, Toronto, Canada. Association for Computational Linguistics.

Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Nan Wang, and Han Xiao. 2024. jina-embeddings-v3: Multilingual embeddings with task lora. *Preprint*, arXiv:2409.10173.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018. Simple and Effective Text Simplification Using Semantic and Neural Methods. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 162–173, Melbourne, Australia. Association for Computational Linguistics.

Renliang Sun, Zhixian Yang, and Xiaojun Wan. 2023. Exploiting summarization data to help text simplification. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 39–51, Dubrovnik, Croatia. Association for Computational Linguistics.

Yuting Tang, Ratish Puduppully, Zhengyuan Liu, and Nancy Chen. 2023. In-context learning of large language models for controlled dialogue summarization: A holistic benchmark and empirical analysis. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 56–67, Singapore. Association for Computational Linguistics.

Teerapaun Tanprasert and David Kauchak. 2021. Flesch-kincaid is not a text simplification evaluation metric. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 1–14, Online. Association for Computational Linguistics.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Amalia Todirascu, Rodrigo Wilkens, Eva Rolin, Thomas François, Delphine Bernhard, and Núria Gala. 2022. HECTOR: A Hybrid TExt SimplifiCation TOol for Raw Texts in French. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4620–4630, Marseille, France. European Language Resources Association.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, et al. 2023b. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint*. ArXiv:2307.09288 [cs].

Subhadra Vadlamannati and Gözde Şahin. 2023. Metric-based in-context learning: A case study in text simplification. In *Proceedings of the 16th International Natural Language Generation Conference*, pages 253–268, Prague, Czechia. Association for Computational Linguistics.

Ashish Vaswani, Samy Bengio, Eugene Brevdo, Francois Chollet, Aidan N. Gomez, Stephan Gouws, Llion Jones, Łukasz Kaiser, Nal Kalchbrenner, Niki Parmar, Ryan Sepassi, Noam Shazeer, and Jakob Uszkoreit. 2018. Tensor2tensor for neural machine translation. *CoRR*, abs/1803.07416.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Xi Wang, Procheta Sen, Ruizhe Li, and Emine Yilmaz. 2024. Simulated Task Oriented Dialogues for Developing Versatile Conversational Agents. In *Advances in Information Retrieval*, pages 157–172, Cham. Springer Nature Switzerland.

Yiming Wang, Yu Lin, Xiaodong Zeng, and Guannan Zhang. 2023. Multilora: Democratizing lora for better multi-task learning. *arXiv preprint arXiv:2311.11501*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Sander Wubben, Antal van den Bosch, and Emiel Krahmer. 2012. Sentence Simplification by Monolingual Machine Translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1015–1024, Jeju Island, Korea. Association for Computational Linguistics.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing Statistical Machine Translation for Text Simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415. Place: Cambridge, MA Publisher: MIT Press.

Victoria Yaneva. 2015. Easy-read documents as a gold standard for evaluation of text simplification output. In *Proceedings of the Student Research Workshop*, pages 30–36, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.

Yaming Yang, Dilxat Muhtar, Yelong Shen, Yuefeng Zhan, Jianfeng Liu, Yujing Wang, Hao Sun, Weiwei Deng, Feng Sun, Qi Zhang, Weizhu Chen, and Yunhai Tong. 2025. Mtl-lora: Low-rank adaptation for multi-task learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(20):22010–22018.

Fan Yin, Jesse Vig, Philippe Laban, Shafiq Joty, Caiming Xiong, and Chien-Sheng Wu. 2023. Did you read the instructions? rethinking the effectiveness of task definitions in instruction learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3063–3079, Toronto, Canada. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

# A Implementation Details

## A.1 Multi-Task Methods

**Finetuning.** LLMs are trained for 6 epochs maximum, using the AdamW optimizer (Loshchilov and Hutter, 2019) 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 4, with 4 steps gradient accumulation and training tasks are randomly sampled. 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 selection is performed to maximize SRB (see §A.3). According to experimental findings, LoRA and MTL-LoRA hyperparameters are set to $r = 128$ and $attn\_matrices = W_{QKVO}$. Moreover, we chose $\alpha = r$ to keep a 1:1 ratio so as not to overpower the backbone (Lee et al., 2023). For MTL-LoRA configuration, sharpness of the weight distribution is fixed at 0.5 and the optimal $n$ up-projections is selected among $\{1, 2, 3\}$. We rely on the implementation provided by `Adapters` library (Poth et al., 2023) for all PEFT methods. Best hyperparameters for PEFT methods are in Table 3.

**MTL-RAG.** To facilitate few-shot multi-task learning within the in-context learning framework, we develop a multi-task extension of Retrieval-Augmented Generation (RAG). Our approach retrieves demonstrations from various tasks and integrates them into the prompt. We conduct experiments using 1, 2, and 3 examples per task, analyzing how the ordering of tasks and examples within the prompt influences the performance. We investigate three strategies for sequencing demonstrations in the prompt as mentioned in Section 4.2: random, grouped and interleaved orderings. Detailled results are in Appendix C.

The optimal hyperparameters for in-context learning are summarized in Table 4.

## A.2 Models

We utilize the following instruct models for In-Context Learning (ICL):

- `Llama-3.1-8B-Instruct`
- `Mistral-7B-Instruct-v0.3`

For experiments involving Parameter-Efficient Fine-Tuning (PEFT), we employ the following base models:

- `Llama-3.1-8B`
- `Mistral-7B-v0.3`
- `DeepSeek-R1-Distill-Llama-8B`

## A.3 Metrics

**Text Descriptive Statistics.** To compute the descriptive statistics presented in Table 1, such as word count, sentence length, compression ratio, KMRE, and others, we employ the `TextDescriptives` (Hansen et al., 2023) and `textacy` Python libraries, both of which use the `fr_core_news_md-3.8.0` model from `SpaCy`.

**ROUGE** (Recall-Oriented Understudy for Gisting Evaluation) (Lin, 2004) is a widely used metric for assessing the quality of automatically generated summaries by measuring n-gram and sequence overlap with reference texts. Specifically, we report the F1-scores for ROUGE-1 (ROUGE-1), ROUGE-2 (ROUGE-2), and ROUGE-L (ROUGE-L), which capture overlap of unigrams, bigrams, and longest common subsequences, respectively. The F1-score represents the harmonic mean of precision and recall. For evaluation, we use Hugging Face's interface to Google's official implementation.

**BERTScore** (Zhang et al., 2020) is based on the contextual word representations generated by BERT-like encoders. Unlike traditional metrics like BLEU or ROUGE, which rely on exact lexical matches, BERTScore uses embeddings to capture finer semantic similarities, offering more flexibility with respect to context and greater robustness to word reordering and synonyms. For each word in the generated text, BERTScore finds the most similar word in the reference text using cosine similarities of their representations. The goal of this step is to align the words in the generated text with those in the reference text. These similarity scores for the aligned word pairs are then aggregated to obtain recall, precision, and F1-score. For reproducibility, we use the Hugging Face's wrapper coupled with `bert-base-multilingual-cased` model.

**SARI** (Sentence-level Accuracy Rating for Text Simplification) (Xu et al., 2016) is commonly used to evaluate sentence and text simplification. Unlike other metrics like BLEU or ROUGE, which focus primarily on lexical similarity to reference texts, SARI takes into account three key aspects of simplification: content preservation (`keep`), information addition (`add`), and information deletion (`del`). For

| | | | Batch size | lr | Acc. steps | Epochs | $\alpha = r$ | Attn. matrices | n up proj. | $\tau$ |
|---|---|---|---|---|---|---|---|---|---|---|
| LlaMA-3-8B | LoRA | E | 4 | $1 \cdot 10^{-4}$ | 4 | 6 | 128 | $W_{QKVO}$ | - | - |
| | MTL-LoRA | E,O,W | 4 | $1 \cdot 10^{-4}$ | 4 | 6 | 128 | $W_{QKVO}$ | 3 | 0.5 |
| | | E,O | 4 | $1 \cdot 10^{-4}$ | 4 | 6 | 128 | $W_{QKVO}$ | 3 | 0.5 |
| | | E,W | 4 | $1 \cdot 10^{-4}$ | 4 | 6 | 128 | $W_{QKVO}$ | 3 | 0.5 |
| Mistral-7B | LoRA | E | 4 | $1 \cdot 10^{-4}$ | 4 | 6 | 128 | $W_{QKVO}$ | - | - |
| | MTL-LoRA | E,O,W | 4 | $1 \cdot 10^{-4}$ | 4 | 6 | 128 | $W_{QKVO}$ | 3 | 0.5 |
| | | E,O | 4 | $5 \cdot 10^{-5}$ | 4 | 6 | 128 | $W_{QKVO}$ | 3 | 0.5 |
| | | E,W | 4 | $1 \cdot 10^{-4}$ | 4 | 6 | 128 | $W_{QKVO}$ | 3 | 0.5 |

Table 3: PEFT hyperparameter configurations chosen based on SRB performance on the ETR-fr validation set. E, O, and W refer to ETR-fr, OrangeSum, and WikiLarge FR, respectively.

| | | | k | Ordering |
|---|---|---|---|---|
| Mistral-7B | Zero-Shot | E | - | - |
| | CoT | E | - | - |
| | RAG | E | 7 | Random |
| | | E,O | 3 | Random |
| | | E,W | 3 | Random |
| | | E,O,W | 3 | Interleaved |
| LlaMA-3-8B | Zero-Shot | E | - | - |
| | CoT | E | - | - |
| | RAG | E | 9 | Random |
| | | E,O | 3 | Random |
| | | E,W | 3 | Random |
| | | E,O,W | 2 | Random |

Table 4: ICL hyperparameter configurations selected based on SRB performance on the ETR-fr validation set. Here, E denotes ETR-fr, O denotes OrangeSum, and W denotes WikiLarge FR.

each word or n-gram generated, SARI evaluates whether the word should be kept, added, or deleted by comparing it with its source and the ground truth. The mathematical expression of SARI is the average of the F1-score of these three measures.

$$\text{SARI} = \frac{F1_{\text{keep}} + F1_{\text{add}} + F1_{\text{del}}}{3}$$

For evaluation, we use Hugging Face's interface, which is adapted from TensorFlow's `tensor2tensor` implementation (Vaswani et al., 2018).

**KMRE** (Kandel-Moles Reading Ease) (Kandel and Moles, 1958) is the French adaptation of the Flesch-Kincaid Reading Ease (FKRE) (Kincaid and Others, 1975), originally designed for English. It measures the complexity of French texts based on sentence length and word length without the need for comparison with a reference text:

$$\text{KMRE} = 207 - 1.015 \left(\frac{\#\text{words}}{\#\text{sentences}}\right) - 73.6 \left(\frac{\#\text{syllables}}{\#\text{words}}\right)$$

KMRE, like the FKRE, is theoretically bounded between 0 and 100. However, it can exceed 100 in rare cases, particularly when the text contains very short sentences and simple, monosyllabic words. This is often the case in ETR documents, which are specifically designed for ease of reading.

Moreover, Wubben et al. (2012) advises not to use this metric alone, as it does not account for grammar quality or meaning preservation. This is why we pair it with BERTScore, ROUGE, and SARI, and we do not monitor it for hyperparameter tuning.

**SRB** is proposed to measure the quality of a ETR text by aggregating metrics related to ETR transcription characteristics, *i.e.* simplification, summarization, and meaning preservation. To do this, we compute the harmonic mean of SARI, ROUGE-L, and BERTScore-F1:

$$\text{SRB} = \frac{3}{\frac{1}{\text{SARI}} + \frac{1}{\text{R-L}} + \frac{1}{\text{BERTScore-F1}}}$$

**Novelty** is used to evaluate abstractiveness, measured by the percentage of n-grams in the generated text that do not appear in the source document (See et al., 2017; Kamal Eddine et al., 2021). We report only novel 1-grams, excluding stop words (commonly used words in a language).

**Compression ratio** is the proportion of the document that has been removed. A higher compression ratio indicates more reduction, meaning the summary is more condensed compared to the original document.

$$\text{Comp. Ratio} = 1 - \frac{\#\text{words in ETR}}{\#\text{words in source}}$$

## B Complementary Evaluation Results

### B.1 Quantitative Results

Average performances of various methods on ETR-fr and ETR-fr-politic test sets are presented in tables 5a and 5b, respectively. These results compare In-Context Learning (ICL) techniques, such as Zero-shot, Chain-of-Thought (CoT), and Retrieval-Augmented Generation (RAG), against Parameter-Efficient Fine-Tuning (PEFT) methods including LoRA and MTL-LoRA. Evaluations are conducted across different LLM models (Mistral-7B, LlaMA-3-8B and DeepSeek-R1-8B) and task combinations (E: ETR-fr, O: OrangeSum, W: WikiLarge FR). Metrics such as ROUGE (R-1, R-2, R-L), SARI, BERTScore-F1, SRB, Compression Ratio, and Novelty are used to provide a comprehensive performance overview.

The experimental results clearly highlight the performance benefits of both retrieval augmentation and fine-tuning approaches, particularly under multi task settings.

**In-Context Learning (ICL).** Zero-Shot and CoTsettings generally underperform across all metrics compared to RAG and PEFT. While CoT shows a slight improvement in novelty and informativeness over Zero-Shot, gains are marginal. RAG consistently improves performance over basic prompting, especially on the main ETR-fr test set. For both Mistral-7B and LlaMA-3-8B, RAG with task combinations (E, E+O, E+W, E+O+W) achieves substantial boosts in ROUGE and SARI scores. Notably, RAG yields the highest performance in most individual metrics under the ICL category.

**Parameter-Efficient Fine-Tuning (PEFT)** models consistently outperform in-context learning (ICL) methods across all evaluation metrics. Both the LoRA and MTL-LoRA setups yield notable gains in fluency, simplicity, and informativeness. Among them, LlaMA-3-8B-MTL-LoRA achieves the best overall performance, excelling in metrics such as SARI, BERTScore-F1, and compression ratio, highlighting its effectiveness in producing simplified text that remains semantically faithful. The MTL-LoRA+(E+W) variant records the highest scores for SARI (44.67), BERTScore (74.05), and compression ratio (56.11), suggesting a well-balanced approach that preserves meaning while substantially reducing text length. Additionally, we report results for the DeepSeek-R1-8B model; however, its performance is consistently lower than other LLM configurations, regardless of the fine-tuning strategy applied.

**Out-of-Domain (ETR-fr-politic) Performance.** On the political subset, the performance gap between ICL and PEFT narrows slightly; however, PEFT models continue to demonstrate a clear advantage. Among the ICL methods, RAG-based approaches retain their relative lead, particularly when augmented with additional context (E+W and E+O+W), indicating stronger generalization capabilities. Notably, the zero-shot LlaMA-3-8B model achieves the highest novelty score (55.73), which could signal greater output diversity, though it might also suggest reduced fidelity. Similar to previous findings, DeepSeek-R1-8B consistently underperforms compared to other LLM configurations, regardless of the fine-tuning method used.

### B.2 Human Evaluation

We conduct a comprehensive human evaluation on two datasets, ETR-fr and ETR-fr-politic, assessing the generated explanations along dimensions guided by the ETR framework and general language quality metrics. Results are reported in Tables 6 and 7.

**Explanation Criteria (ETR dimensions).** On ETR-fr, all methods exhibit strong performance across information selection, word selection, and sentece construction construction (scores $> 0.88$), with the LoRA method slightly outperforming others in word selection ($0.94$) and overall global quality ($0.91$). Illustration quality, however, remains a consistent weakness across methods, with high variance indicating instability or inconsistent strategy for visual grounding.

For the more challenging ETR-fr-politic, overall scores decrease across all explanation criteria. Notably, RAG with joint training on E and W achieves the best global score ($0.80$), outperforming LoRA and MTL-LoRA. While RAG maintains high scores in information selection and sentece construction illustration scores remain low across the board, underscoring the difficulty of generating coherent examples or analogies in politically sensitive domains.

**General Language Quality.** As shown in Table 7, RAG again performs competitively on both datasets. On ETR-fr, it achieves the highest ratings in grammar and coherence (both $> 4.4$), with

| | Method | Task | R-1 ↑ | R-2 ↑ | R-L ↑ | SARI ↑ | BERT-F1 ↑ | SRB ↑ | Comp. ratio | Novelty |
|---|---|---|---|---|---|---|---|---|---|---|
| | **In Context Learning** | | | | | | | | | |
| **Mistral-7B** | **Zero-Shot** | **E** | $23.96_{\pm0.04}$ | $7.08_{\pm0.01}$ | $16.25_{\pm0.03}$ | $37.07_{\pm0.00}$ | $69.75_{\pm0.00}$ | $29.17_{\pm0.03}$ | $-64.14_{\pm0.00}$ | $35.70_{\pm0.00}$ |
| | **CoT** | **E** | $23.53_{\pm0.06}$ | $7.23_{\pm0.01}$ | $16.20_{\pm0.04}$ | $37.39_{\pm0.00}$ | $68.80_{\pm0.00}$ | $29.12_{\pm0.05}$ | $-60.53_{\pm0.00}$ | $\underline{36.09}_{\pm0.00}$ |
| | **RAG** | **E** | $\underline{31.91}_{\pm0.66}$ | $\underline{10.77}_{\pm0.65}$ | $\underline{22.54}_{\pm0.75}$ | $\underline{40.14}_{\pm0.57}$ | $\underline{72.17}_{\pm0.30}$ | $\underline{36.08}_{\pm0.80}$ | $45.23_{\pm1.17}$ | $27.27_{\pm0.58}$ |
| | | **E,O** | $30.36_{\pm0.47}$ | $9.61_{\pm0.34}$ | $21.80_{\pm0.30}$ | $39.49_{\pm0.12}$ | $71.07_{\pm0.18}$ | $35.19_{\pm0.29}$ | $\underline{47.99}_{\pm1.91}$ | $26.80_{\pm0.84}$ |
| | | **E,W** | $30.46_{\pm0.48}$ | $9.93_{\pm0.17}$ | $21.72_{\pm0.34}$ | $38.76_{\pm0.43}$ | $71.57_{\pm0.14}$ | $34.96_{\pm0.34}$ | $35.08_{\pm2.13}$ | $23.32_{\pm0.31}$ |
| | | **E,O,W** | $29.85_{\pm0.04}$ | $9.58_{\pm0.03}$ | $21.55_{\pm0.05}$ | $39.53_{\pm0.00}$ | $71.06_{\pm0.00}$ | $34.98_{\pm0.05}$ | $46.42_{\pm0.00}$ | $25.85_{\pm0.00}$ |
| **LlaMA-3-8B** | **Zero-Shot** | **E** | $24.90_{\pm0.20}$ | $8.16_{\pm0.25}$ | $17.10_{\pm0.38}$ | $38.48_{\pm0.38}$ | $70.15_{\pm0.17}$ | $30.38_{\pm0.48}$ | $-22.52_{\pm2.47}$ | $\mathbf{39.13}_{\pm0.92}$ |
| | **CoT** | **E** | $27.23_{\pm0.91}$ | $8.81_{\pm0.21}$ | $18.34_{\pm0.57}$ | $38.15_{\pm0.23}$ | $70.79_{\pm0.52}$ | $31.62_{\pm0.65}$ | $7.59_{\pm4.82}$ | $30.33_{\pm1.75}$ |
| | **RAG** | **E** | $\underline{33.05}_{\pm0.72}$ | $\underline{12.23}_{\pm0.44}$ | $\underline{23.77}_{\pm0.68}$ | $\underline{41.66}_{\pm0.45}$ | $\underline{72.59}_{\pm0.38}$ | $\underline{37.57}_{\pm0.70}$ | $\underline{43.36}_{\pm2.62}$ | $27.06_{\pm0.29}$ |
| | | **E,O** | $30.77_{\pm0.35}$ | $10.85_{\pm0.31}$ | $22.10_{\pm0.35}$ | $39.84_{\pm0.22}$ | $71.13_{\pm0.17}$ | $35.54_{\pm0.32}$ | $24.36_{\pm30.13}$ | $25.02_{\pm1.84}$ |
| | | **E,W** | $32.14_{\pm0.56}$ | $11.70_{\pm0.34}$ | $23.11_{\pm0.19}$ | $40.49_{\pm0.32}$ | $71.88_{\pm0.18}$ | $36.64_{\pm0.24}$ | $42.30_{\pm1.59}$ | $26.70_{\pm0.92}$ |
| | | **E,O,W** | $30.53_{\pm0.74}$ | $10.67_{\pm0.45}$ | $21.65_{\pm0.71}$ | $39.24_{\pm0.20}$ | $71.21_{\pm0.26}$ | $35.00_{\pm0.67}$ | $31.18_{\pm4.94}$ | $24.08_{\pm1.37}$ |
| | **PEFT** | | | | | | | | | |
| **Mistral-7B** | **LoRA** | **E** | $32.45_{\pm0.03}$ | $12.38_{\pm0.02}$ | $23.99_{\pm0.05}$ | $42.09_{\pm0.00}$ | $73.56_{\pm0.00}$ | $37.95_{\pm0.04}$ | $44.42_{\pm0.00}$ | $18.35_{\pm0.00}$ |
| | **MTL-LoRA** | **E,O** | $32.62_{\pm0.04}$ | $12.73_{\pm0.01}$ | $24.29_{\pm0.04}$ | $41.95_{\pm0.00}$ | $73.52_{\pm0.00}$ | $38.16_{\pm0.03}$ | $53.48_{\pm0.00}$ | $24.17_{\pm0.00}$ |
| | | **E,W** | $32.68_{\pm0.05}$ | $\underline{12.91}_{\pm0.01}$ | $24.25_{\pm0.03}$ | $\underline{42.53}_{\pm0.00}$ | $\underline{73.90}_{\pm0.00}$ | $38.33_{\pm0.03}$ | $\underline{53.62}_{\pm0.00}$ | $\underline{24.99}_{\pm0.00}$ |
| | | **E,O,W** | $\mathbf{33.60}_{\pm0.05}$ | $12.81_{\pm0.05}$ | $\underline{24.89}_{\pm0.04}$ | $42.25_{\pm0.00}$ | $73.62_{\pm0.00}$ | $\underline{38.74}_{\pm0.03}$ | $48.93_{\pm0.00}$ | $23.38_{\pm0.00}$ |
| **LlaMA-3-8B** | **LoRA** | **E** | $31.80_{\pm0.03}$ | $13.16_{\pm0.09}$ | $24.92_{\pm0.18}$ | $42.15_{\pm0.01}$ | $72.84_{\pm0.17}$ | $38.67_{\pm0.17}$ | $50.50_{\pm0.28}$ | $18.37_{\pm0.88}$ |
| | **MTL-LoRA** | **E,O** | $\underline{33.38}_{\pm0.06}$ | $13.16_{\pm0.05}$ | $24.20_{\pm0.04}$ | $43.06_{\pm0.01}$ | $73.88_{\pm0.01}$ | $38.42_{\pm0.03}$ | $50.90_{\pm0.40}$ | $23.25_{\pm0.17}$ |
| | | **E,W** | $32.54_{\pm0.05}$ | $13.50_{\pm0.06}$ | $25.01_{\pm0.06}$ | $\mathbf{44.67}_{\pm0.00}$ | $\mathbf{74.05}_{\pm0.00}$ | $39.54_{\pm0.05}$ | $\underline{56.11}_{\pm0.00}$ | $\underline{33.05}_{\pm0.00}$ |
| | | **E,O,W** | $32.78_{\pm0.02}$ | $\mathbf{13.67}_{\pm0.03}$ | $\mathbf{25.55}_{\pm0.16}$ | $43.58_{\pm0.10}$ | $73.33_{\pm0.09}$ | $\mathbf{39.62}_{\pm0.09}$ | $52.66_{\pm1.00}$ | $24.27_{\pm0.21}$ |
| **DeepSeek-R1-8B** | **LoRA** | **E** | $20.45_{\pm0.65}$ | $7.72_{\pm0.29}$ | $15.40_{\pm0.13}$ | $41.29_{\pm0.04}$ | $66.02_{\pm0.26}$ | $28.76_{\pm0.16}$ | $-4.61_{\pm3.83}$ | $21.86_{\pm0.29}$ |
| | **MTL-LoRA** | **E,O** | $23.70_{\pm0.32}$ | $8.86_{\pm0.04}$ | $18.18_{\pm0.33}$ | $42.91_{\pm0.06}$ | $66.72_{\pm0.24}$ | $32.15_{\pm0.37}$ | $\underline{8.57}_{\pm1.08}$ | $27.92_{\pm0.86}$ |
| | | **E,W** | $\underline{25.38}_{\pm0.11}$ | $\underline{9.35}_{\pm0.05}$ | $\underline{18.52}_{\pm0.07}$ | $\underline{43.06}_{\pm0.03}$ | $\underline{68.08}_{\pm0.14}$ | $\underline{32.64}_{\pm0.08}$ | $-0.52_{\pm2.52}$ | $\underline{36.16}_{\pm0.30}$ |
| | | **E,O,W** | $22.70_{\pm0.10}$ | $7.93_{\pm0.01}$ | $16.59_{\pm0.02}$ | $42.94_{\pm0.00}$ | $67.18_{\pm0.00}$ | $30.47_{\pm0.02}$ | $-9.35_{\pm0.00}$ | $29.50_{\pm0.00}$ |

(a) Performance on ETR-fr test set.

| | Method | Task | R-1 ↑ | R-2 ↑ | R-L ↑ | SARI ↑ | BERT-F1 ↑ | SRB ↑ | Comp. ratio | Novelty |
|---|---|---|---|---|---|---|---|---|---|---|
| | **In Context Learning** | | | | | | | | | |
| **Mistral-7B** | **Zero-Shot** | **E** | $28.42_{\pm0.12}$ | $10.98_{\pm0.07}$ | $19.31_{\pm0.03}$ | $39.87_{\pm0.00}$ | $68.10_{\pm0.00}$ | $32.77_{\pm0.03}$ | $-309.24_{\pm0.00}$ | $48.37_{\pm0.00}$ |
| | **CoT** | **E** | $29.80_{\pm0.03}$ | $11.21_{\pm0.05}$ | $19.88_{\pm0.08}$ | $39.62_{\pm0.00}$ | $69.40_{\pm0.00}$ | $33.35_{\pm0.07}$ | $-261.30_{\pm0.00}$ | $\underline{50.85}_{\pm0.00}$ |
| | **RAG** | **E** | $\mathbf{40.19}_{\pm0.63}$ | $\underline{16.07}_{\pm0.60}$ | $28.25_{\pm0.31}$ | $\underline{41.40}_{\pm0.46}$ | $\underline{73.01}_{\pm0.34}$ | $\underline{40.96}_{\pm0.35}$ | $9.00_{\pm3.96}$ | $23.21_{\pm2.39}$ |
| | | **E,O** | $37.49_{\pm0.61}$ | $14.50_{\pm0.35}$ | $26.38_{\pm0.69}$ | $39.46_{\pm0.35}$ | $72.27_{\pm0.26}$ | $38.92_{\pm0.58}$ | $14.26_{\pm2.65}$ | $17.57_{\pm1.61}$ |
| | | **E,W** | $39.65_{\pm0.19}$ | $15.36_{\pm0.35}$ | $27.85_{\pm0.38}$ | $40.08_{\pm0.36}$ | $72.35_{\pm0.29}$ | $40.17_{\pm0.23}$ | $8.72_{\pm1.73}$ | $17.47_{\pm1.68}$ |
| | | **E,O,W** | $39.14_{\pm0.04}$ | $15.96_{\pm0.09}$ | $\underline{28.40}_{\pm0.11}$ | $40.74_{\pm0.00}$ | $72.87_{\pm0.00}$ | $40.82_{\pm0.07}$ | $\underline{14.63}_{\pm0.00}$ | $18.33_{\pm0.00}$ |
| **LlaMA-3-8B** | **Zero-Shot** | **E** | $29.10_{\pm0.40}$ | $10.68_{\pm0.35}$ | $18.70_{\pm0.41}$ | $40.68_{\pm0.48}$ | $68.65_{\pm0.11}$ | $32.39_{\pm0.51}$ | $-178.23_{\pm7.77}$ | $\mathbf{55.73}_{\pm1.07}$ |
| | **CoT** | **E** | $31.15_{\pm0.99}$ | $10.47_{\pm0.81}$ | $19.54_{\pm0.65}$ | $39.80_{\pm0.63}$ | $69.66_{\pm0.43}$ | $33.09_{\pm0.74}$ | $-70.57_{\pm8.09}$ | $47.80_{\pm1.71}$ |
| | **RAG** | **E** | $37.68_{\pm0.53}$ | $14.46_{\pm0.65}$ | $26.09_{\pm0.60}$ | $42.05_{\pm0.90}$ | $73.01_{\pm0.20}$ | $39.57_{\pm0.41}$ | $1.47_{\pm6.45}$ | $41.78_{\pm0.86}$ |
| | | **E,O** | $37.43_{\pm2.11}$ | $14.28_{\pm0.89}$ | $25.92_{\pm1.42}$ | $40.95_{\pm0.90}$ | $72.41_{\pm0.61}$ | $39.05_{\pm1.37}$ | $-7.72_{\pm14.32}$ | $31.85_{\pm1.69}$ |
| | | **E,W** | $\underline{39.99}_{\pm1.10}$ | $\mathbf{16.27}_{\pm0.61}$ | $\underline{27.84}_{\pm1.10}$ | $\mathbf{42.41}_{\pm0.43}$ | $\mathbf{73.83}_{\pm0.47}$ | $\mathbf{41.06}_{\pm0.96}$ | $\underline{13.46}_{\pm2.37}$ | $36.72_{\pm2.01}$ |
| | | **E,O,W** | $38.33_{\pm1.46}$ | $15.12_{\pm1.08}$ | $26.89_{\pm1.10}$ | $41.08_{\pm0.94}$ | $72.86_{\pm0.51}$ | $39.86_{\pm1.13}$ | $6.34_{\pm7.54}$ | $29.92_{\pm0.48}$ |
| | **PEFT** | | | | | | | | | |
| **Mistral-7B** | **LoRA** | **E** | $35.10_{\pm0.04}$ | $12.28_{\pm0.04}$ | $25.97_{\pm0.03}$ | $38.04_{\pm0.00}$ | $70.28_{\pm0.00}$ | $37.96_{\pm0.02}$ | $21.55_{\pm0.00}$ | $11.79_{\pm0.00}$ |
| | **MTL-LoRA** | **E,O** | $29.29_{\pm0.07}$ | $11.02_{\pm0.01}$ | $21.90_{\pm0.04}$ | $38.68_{\pm0.00}$ | $69.22_{\pm0.00}$ | $34.90_{\pm0.03}$ | $\mathbf{36.68}_{\pm0.00}$ | $\underline{40.29}_{\pm0.00}$ |
| | | **E,W** | $34.32_{\pm0.06}$ | $12.60_{\pm0.07}$ | $24.87_{\pm0.11}$ | $\underline{38.72}_{\pm0.00}$ | $70.54_{\pm0.00}$ | $37.40_{\pm0.09}$ | $22.51_{\pm0.00}$ | $19.10_{\pm0.00}$ |
| | | **E,O,W** | $\underline{36.34}_{\pm0.10}$ | $\underline{13.24}_{\pm0.02}$ | $\underline{26.29}_{\pm0.08}$ | $38.39_{\pm0.00}$ | $\underline{70.97}_{\pm0.00}$ | $\underline{38.37}_{\pm0.06}$ | $18.33_{\pm0.00}$ | $10.55_{\pm0.00}$ |
| **LlaMA-3-8B** | **LoRA** | **E** | $34.65_{\pm1.43}$ | $13.34_{\pm0.85}$ | $26.40_{\pm0.95}$ | $\underline{39.70}_{\pm0.35}$ | $70.73_{\pm0.99}$ | $38.85_{\pm0.90}$ | $4.67_{\pm2.97}$ | $16.19_{\pm0.11}$ |
| | **MTL-LoRA** | **E,O** | $32.17_{\pm0.52}$ | $11.94_{\pm0.23}$ | $23.98_{\pm0.22}$ | $39.35_{\pm0.44}$ | $69.49_{\pm0.21}$ | $36.81_{\pm0.06}$ | $\underline{17.14}_{\pm0.98}$ | $\underline{20.01}_{\pm0.62}$ |
| | | **E,W** | $\underline{37.58}_{\pm0.12}$ | $13.68_{\pm0.05}$ | $\underline{27.02}_{\pm0.03}$ | $38.26_{\pm0.00}$ | $\underline{71.30}_{\pm0.00}$ | $\underline{38.88}_{\pm0.02}$ | $8.45_{\pm0.00}$ | $6.44_{\pm0.00}$ |
| | | **E,O,W** | $36.38_{\pm0.22}$ | $\underline{13.72}_{\pm0.07}$ | $25.75_{\pm0.23}$ | $36.19_{\pm0.00}$ | $70.94_{\pm0.04}$ | $37.24_{\pm0.17}$ | $8.76_{\pm0.13}$ | $2.04_{\pm0.05}$ |
| **DeepSeek-R1-8B** | **LoRA** | **E** | $23.89_{\pm0.27}$ | $7.57_{\pm0.30}$ | $18.48_{\pm0.30}$ | $39.34_{\pm0.32}$ | $63.60_{\pm0.24}$ | $31.49_{\pm0.34}$ | $-50.45_{\pm2.83}$ | $24.56_{\pm1.15}$ |
| | **MTL-LoRA** | **E,O** | $26.81_{\pm1.84}$ | $8.41_{\pm0.40}$ | $19.60_{\pm1.15}$ | $39.03_{\pm0.16}$ | $65.02_{\pm0.42}$ | $32.58_{\pm1.08}$ | $\underline{-38.60}_{\pm0.76}$ | $\underline{25.44}_{\pm0.10}$ |
| | | **E,W** | $26.53_{\pm0.79}$ | $9.77_{\pm0.86}$ | $18.97_{\pm0.73}$ | $\underline{39.47}_{\pm0.49}$ | $65.37_{\pm0.23}$ | $32.14_{\pm0.83}$ | $-49.42_{\pm0.94}$ | $21.95_{\pm0.85}$ |
| | | **E,O,W** | $\underline{29.83}_{\pm0.04}$ | $\underline{11.18}_{\pm0.04}$ | $\underline{21.13}_{\pm0.07}$ | $36.58_{\pm0.00}$ | $\underline{67.35}_{\pm0.00}$ | $\underline{33.51}_{\pm0.06}$ | $-46.02_{\pm0.00}$ | $4.32_{\pm0.00}$ |

(b) Performance on ETR-fr-politic test set.

Table 5: **Performance comparison across prompting methods (zero-shot, Chain-of-Thought, RAG) and fine-tuning strategies (LoRA, Multi-task LoRA)** on three tasks: ETR-fr (E), OrangeSum (O) and WikiLarge FR (W), using Mistral-7B, LlaMA-3-8B and DeepSeek-R1-8B models. Metrics: ROUGE-1/2/L, SARI, BERTScore-F1, composite SRB score, compression ratio, and lexical novelty. Results are presented as mean ± standard deviation. Best overall results are shown in **bold**, and best results for each model are underlined.

strong fluency and relevance. MTL-LoRA slightly improves grammaticality, but this does not translate to gains in perceived overall quality.

In the political domain, quality metrics decline, consistent with the ETR scores. RAG trained on E and W maintains robust fluency and coherence, achieving the best overall quality score (3.76). In contrast, MTL-LoRA's performance degrades notably in global quality (2.62), despite competitive scores in coherence and relevance, suggesting potential trade-offs introduced by multitask learning in more nuanced domains.

**Summary.** These results highlight RAG's robustness across both explanation and linguistic quality metrics, particularly when trained jointly on ETR and sentence simplification tasks. The consistent underperformance in illustration generation across all models indicates a need for future work on grounded or multimodal explanation strategies, especially in high-stakes domains like politics.

### B.3 Comparison of Ground Truth and Generated ETR Outputs

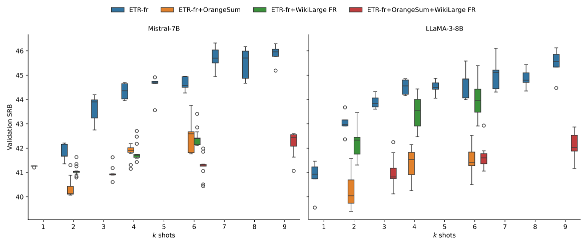
The table 8 presents a detailed comparison of different model configurations (Mistral-7B and LlaMA-8B), training methods (RAG, LoRA, MTL-LoRA), and task combinations (ETR, summarization and simplification). Metrics include the average number of words and sentences, sentence length, KMRE (higher is better), novelty, and compression ratio.

Overall, models trained with MTL-LoRA tend to generate more concise outputs while maintaining strong performance in terms of KMRE. For instance, LlaMA-8B + MTL-LoRA (E,W) achieves the highest KMRE score (102.98) and the highest novelty (33.05), indicating its ability to produce informative and diverse content.

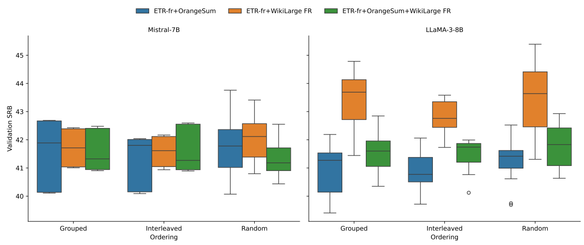
RAG-based methods generally generate longer texts, with higher sentence lengths (up to 11.07 words on average for LlaMA-8B + RAG (E,O,W)), but often at the expense of novelty. This suggests that RAG relies more heavily on retrieved content, which may reduce the originality of generated text.

Compared to the ground truth, the generated texts generally contain more words and exhibit equal or greater sentence lengths. Notably, the MTL-LoRA configurations achieve higher compression ratios, highlighting their ability to effectively condense information. While no method fully replicates the characteristics of the test set, defined by its notably short sentences and high compression. LlaMA-8B MTL-LoRA trained on Wikilarge (W) and ETR-fr (E) yields outputs that most closely resemble the test set in terms of both compression and sentence structure.

## C In-Context Learning Hyperparameters Effects




(a) SRB performance under varying number of in-context examples ($k \in [1; 9]$) and task combinations.




(b) SRB performance under different example ordering strategies and task combination configurations.

Figure 3: Comparison of SRB performance on the ETR-fr validation set across different in-context settings and ordering strategies.

Figure 5 illustrates examples of prompts used for zero-shot (Fig. 5a), chain-of-thought (Fig. 5c) and few-shot (Fig. 5b).

### C.1 Impact of the Number of Shots on ETR-fr Performance

Figure 3a presents the performance of LLaMA-3-8B and Mistral-7B on the French text simplification benchmark (ETR-fr) across varying numbers of in-context learning (ICL) examples ($k = 1$ to 9) and under different training configurations.

**LLaMA-3-8B Performance.** For the LLaMA-3-8B model, performance generally increases with larger $k$ values. The basic task ETR-fr alone yields steadily rising median SRB scores, from 40.93 at $k = 1$ to 45.96 at $k = 9$. The incorporation of auxiliary datasets (OrangeSum and WikiLarge FR) leads to varied results. For instance, combining ETR-fr with WikiLarge FR at $k = 2$ raises the median from 42.96 to 42.33, while the three-dataset

| | Method | Task | Informations | Words | Sentences | Illustrations | Global |
|---|---|---|---|---|---|---|---|
| | **ETR-fr** | | | | | | |
| **LlaMA-3-8B** | **LoRA** | **E** | $0.89_{\pm 0.08}$ | $0.94_{\pm 0.04}$ | $0.91_{\pm 0.05}$ | $0.38_{\pm 0.40}$ | $0.91_{\pm 0.04}$ |
| | **MTL-LoRA** | **E,O,W** | $0.88_{\pm 0.06}$ | $0.89_{\pm 0.07}$ | $0.93_{\pm 0.04}$ | $0.50_{\pm 0.65}$ | $0.89_{\pm 0.04}$ |
| | **RAG** | **E** | $0.88_{\pm 0.07}$ | $0.92_{\pm 0.05}$ | $0.89_{\pm 0.04}$ | $0.40_{\pm 0.52}$ | $0.89_{\pm 0.04}$ |
| | | **E,W** | $0.91_{\pm 0.05}$ | $0.88_{\pm 0.07}$ | $0.92_{\pm 0.04}$ | $0.50_{\pm 0.44}$ | $0.89_{\pm 0.04}$ |
| | **ETR-fr-politic** | | | | | | |
| **LlaMA-3-8B** | **LoRA** | **E** | $0.77_{\pm 0.14}$ | $0.66_{\pm 0.11}$ | $0.79_{\pm 0.11}$ | $0.15_{\pm 0.24}$ | $0.73_{\pm 0.08}$ |
| | **MTL-LoRA** | **E,O,W** | $0.69_{\pm 0.13}$ | $0.59_{\pm 0.11}$ | $0.65_{\pm 0.12}$ | $0.27_{\pm 0.27}$ | $0.64_{\pm 0.08}$ |
| | **RAG** | **E** | $0.82_{\pm 0.09}$ | $0.74_{\pm 0.10}$ | $0.86_{\pm 0.07}$ | $0.10_{\pm 0.23}$ | $0.78_{\pm 0.05}$ |
| | | **E,W** | $0.87_{\pm 0.06}$ | $0.75_{\pm 0.09}$ | $0.85_{\pm 0.08}$ | $0.40_{\pm 0.37}$ | $0.80_{\pm 0.06}$ |

Table 6: **Human evaluation of generations based on ETR guideline criteria**, comparing various methods on the ETR-fr and ETR-fr-politic test sets using their optimal ICL and MTL configurations. Each method is evaluated along four explanation dimensions: Informations (information selection), Words (lexical choice), Sentences (sentence construction), Illustrations, and Global representing the overall quality score. Training tasks are abbreviated as E (ETR-fr), O (OrangeSum), and W (WikiLarge FR). Reported scores are means with 95% confidence intervals.

| | Method | Task | Fluency | Grammar | Relevance | Coherence | Overall Quality |
|---|---|---|---|---|---|---|---|
| | **ETR-fr** | | | | | | |
| **LlaMA-3-8B** | **LoRA** | **E** | $4.29_{\pm 0.26}$ | $4.57_{\pm 0.23}$ | $3.95_{\pm 0.39}$ | $4.24_{\pm 0.32}$ | $3.95_{\pm 0.37}$ |
| | **MTL-LoRA** | **E,O,W** | $4.33_{\pm 0.33}$ | $4.67_{\pm 0.22}$ | $4.10_{\pm 0.38}$ | $4.14_{\pm 0.39}$ | $3.95_{\pm 0.44}$ |
| | **RAG** | **E** | $4.43_{\pm 0.27}$ | $4.71_{\pm 0.21}$ | $4.24_{\pm 0.38}$ | $4.43_{\pm 0.34}$ | $4.24_{\pm 0.35}$ |
| | | **E,W** | $4.43_{\pm 0.23}$ | $4.57_{\pm 0.23}$ | $4.43_{\pm 0.34}$ | $4.52_{\pm 0.27}$ | $3.95_{\pm 0.34}$ |
| | **ETR-fr-politic** | | | | | | |
| **LlaMA-3-8B** | **LoRA** | **E** | $3.90_{\pm 0.52}$ | $4.43_{\pm 0.42}$ | $4.24_{\pm 0.43}$ | $4.24_{\pm 0.45}$ | $3.14_{\pm 0.62}$ |
| | **MTL-LoRA** | **E,O,W** | $3.81_{\pm 0.45}$ | $4.48_{\pm 0.34}$ | $4.40_{\pm 0.38}$ | $4.52_{\pm 0.23}$ | $2.62_{\pm 0.55}$ |
| | **RAG** | **E** | $4.24_{\pm 0.38}$ | $4.48_{\pm 0.34}$ | $4.10_{\pm 0.35}$ | $4.33_{\pm 0.30}$ | $3.45_{\pm 0.44}$ |
| | | **E,W** | $4.33_{\pm 0.33}$ | $4.57_{\pm 0.23}$ | $4.29_{\pm 0.29}$ | $4.43_{\pm 0.27}$ | $3.76_{\pm 0.40}$ |

Table 7: **Human ratings of fluency, grammar, relevance, coherence, and overall quality** for different methods evaluated on the ETR-fr and ETR-fr-politic test sets, using their optimal ICL and MTL configurations. Training tasks are abbreviated as E (ETR-fr), O (OrangeSum), and W (WikiLarge FR). Scores are reported as means with 95% confidence intervals.

combination at $k = 6$ has a lower median of 41.60 compared to 44.84 for ETR-fr alone. This suggests diminishing returns or even negative interference when too many tasks are combined.

**Mistral-7B Performance.** The Mistral-7B model demonstrates a similar trend of improved performance with increasing $k$ values for the ETR-fr task. Median SRB rise from 41.26 at $k = 1$ to 45.96 at $k = 9$. However, Mistral exhibits less variation across configurations. The inclusion of OrangeSum and WikiLarge FR improves SRB modestly, and the three-dataset combination remains slightly below the single-task performance. For example, at $k = 6$, ETR-fr alone achieves a median of 44.58, whereas the triple combination achieves only 41.28.

**Comparative Insights.** When comparing both models, LLaMA-3-8B tends to show greater gains from dataset combinations than Mistral-7B, although it also experiences more variance. For both models, the highest performances are obtained when using ETR-fr alone at higher $k$ values, indicating that overloading the prompt context with multiple tasks may dilute performance. Moreover, the higher maximum SRB for LLaMA across configurations (e.g., up to 46.12) suggest it may have

| | Method | Tasks | # Words | # Sentences | Sentence length | KMRE ↑ | Novelty | Comp. ratio |
|---|---|---|---|---|---|---|---|---|
| | *Ground Truth* | *Test Set* | 40.26 | 8.91 | 4.64 | 102.99 | 55.01 | 65.19 |
| **Mistral-7B** | **RAG** | **E** | 66.38 | 7.70 | 8.76 | 99.77 | 26.55 | 44.32 |
| | | **E,O** | 60.91 | 6.13 | 10.05 | 97.21 | 26.61 | 48.45 |
| | | **E,W** | 80.74 | 7.83 | 10.67 | 97.37 | 23.01 | 33.80 |
| | | **E,O,W** | 62.45 | 6.15 | 10.25 | 97.62 | 25.85 | 46.42 |
| **LlaMA-8B** | **RAG** | **E** | 63.72 | 7.87 | 8.38 | 101.70 | 27.14 | 46.18 |
| | | **E,O** | 74.19 | 7.57 | 9.92 | 97.45 | 24.29 | 39.22 |
| | | **E,W** | 69.72 | 7.64 | 9.49 | 100.34 | 25.26 | 41.89 |
| | | **E,O,W** | 87.17 | 8.40 | 11.07 | 97.48 | 23.69 | 25.94 |
| **Mistral-7B** | **LoRA** | **E** | 65.55 | 9.26 | 7.73 | 101.20 | 18.35 | 44.42 |
| | **MTL-LoRA** | **E,O** | 56.75 | 8.25 | 7.38 | 102.61 | 24.17 | 53.48 |
| | | **E,W** | 54.08 | 9.28 | 6.46 | 104.23 | 24.99 | 53.62 |
| | | **E,O,W** | 60.08 | 8.81 | 7.23 | 101.80 | 23.38 | 48.93 |
| **LlaMA-8B** | **LoRA** | **E** | 56.96 | 8.64 | 7.62 | 100.93 | 18.87 | 50.66 |
| | **MTL-LoRA** | **E,O** | 60.08 | 9.87 | 7.00 | 100.84 | 23.06 | 51.36 |
| | | **E,W** | 50.09 | 9.19 | 6.50 | 102.98 | 33.05 | 56.11 |
| | | **E,O,W** | 54.06 | 8.77 | 7.42 | 101.35 | 24.39 | 53.24 |

Table 8: Comparison of different model configurations (Mistral-7B and LlaMA-8B) and training methods (RAG, LoRA, MTL-LoRA) across various task combinations (E: ETR-fr, O: OrangeSum, W: WikiLarge FR). The metrics include word count, sentence count, average sentence length, KMRE (higher is better), novelty, and compression ratio. Ground truth statistics from the test set are also provided for reference.

a higher performance ceiling, but with more fluctuation.

### C.2 Impact of the Tasks Ordering on ETR-fr Performance

Figure 3b presents the impact of task ordering on model performance under different multi-task training configurations. For both models, three types of example ordering are compared: *grouped*, *interleaved*, and *random*. Each ordering is evaluated with different training task combinations, such as ETR-fr+OrangeSum, ETR-fr+WikiLarge FR, and ETR-fr+OrangeSum+WikiLarge FR.

**LLaMA-3-8B Performance.** For LLaMA-3-8B, performance consistently improves when WikiLarge FR data is added to the training set. The configuration using only ETR-fr+WikiLarge FR yields the highest SRB scores across all ordering methods, particularly under the random strategy, which achieves the highest maximum score (45.39). Overall, grouped and random orderings tend to result in higher median and upper-quartile SRB compared to interleaved ordering, indicating that the sequential arrangement of examples plays a role in performance.

**Mistral-7B Performance.** For Mistral-7B, the impact of training set composition is similarly positive, with improvements observed upon including WikiLarge FR. However, the differences among the three ordering strategies are more subtle. grouped and interleaved yield very similar statistics, with slight advantages in median SRB depending on the training data. The highest maximum score for Mistral-7B (43.76) occurs under the random strategy with the ETR-fr+OrangeSum dataset, although this configuration does not have the most consistent results across runs.

**Comparative Insights.** Comparing the two models, LLaMA-3-8B generally outperforms Mistral-7B in terms of median and maximum SRB, particularly when trained with ETR-fr and WikiLarge FR. Mistral-7B demonstrates more stable performance with narrower score ranges but slightly lower central tendency metrics. These results suggest that while both models benefit from enriched prompts, LLaMA-3-8B exhibits greater potential for high-end performance when paired with appropriate example ordering and task combinations.

## D ETR-fr Dataset Sheet

The dataset description follows the recommendations and template proposed by Gebru et al. (2021).

**Motivation**

### For what purpose was the dataset created?

The ETR-fr dataset was created to address the lack of high-quality, document-aligned corpora suitable for generating Easy-to-Read (ETR) text. It supports the task of generating cognitively accessible texts for individuals with cognitive impairments by providing paragraph-aligned text pairs that follow the European ETR guidelines. This dataset enables the training and evaluation of automatic systems for ETR generation in French, targeting the linguistic and cognitive accessibility requirements typically overlooked by existing simplification or summarization.

### Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

The dataset was constructed by the authors of the this paper on ETR-fr.

## Composition

### What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?

Each instance in the ETR-fr dataset consists of a pair of paragraph-aligned French texts: a source text and its corresponding Easy-to-Read (ETR) version. These are designed to support document-level simplification, emphasizing both lexical and structural transformation.

### How many instances are there in total (of each type, if appropriate)?

The dataset contains 523 paragraph-aligned text pairs. Additionally, an out-of-domain subset, ETR-fr-politic, includes 33 paragraph pairs from 2022 French presidential election programs.

### What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features?

Each instance consists of "raw" French text paragraphs: a complex source text and its corresponding simplified (ETR) version. These are aligned at the paragraph level and include natural language text only.

### Is there a label or target associated with each instance?

Yes. The target is the simplified (ETR-compliant) version of the source paragraph, forming a supervised text-to-text pair for generation tasks.

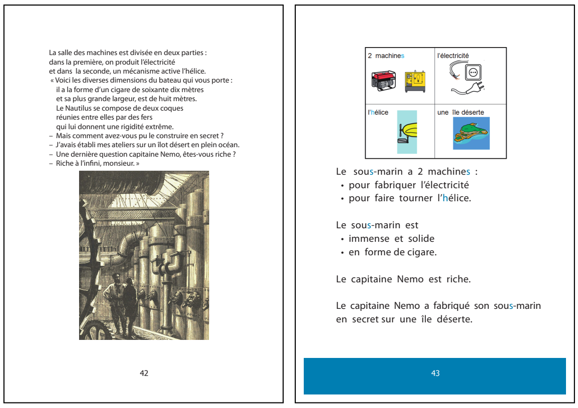



Figure 4: Extract of the ETR book *Twenty Thousand Leagues Under the Seas* by Jules Verne from François Baudez Publishing. **Left page** is the original text with an illustration. **Right page** is the ETR transcription with the main information plus its captioned *vignettes*.

### Is any information missing from individual instances?

The pictograms present with the original texts have not been extracted.

### Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?

No such relationships exist or are made explicit in this dataset.

### Are there recommended data splits (e.g., training, development/validation, testing)?

Yes. The dataset is divided into training (399 pairs), validation (71 pairs), and test (53 pairs) subsets. The test set comprises two distinct books chosen to ensure diversity in linguistic features such as text length, structure, and readability. The remaining books were split into training and validation sets using a stratified approach to minimize thematic and lexical overlap. Additionally, the ETR-fr-politic test set (33 pairs) was introduced to assess model generalization on out-of-domain content not seen during training.

### Are there any errors, sources of noise, or redundancies in the dataset?

No specific mention of noise or redundancy issues is made in the source document.

### Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?

The dataset is self-contained, it does not rely on external resources.

**Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals non-public communications)?**
No. All texts are from published sources and are intended for public consumption.

**Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?**
No such content is reported or expected in the dataset.

**Does the dataset relate to people?**
No. The dataset is composed of literary and political texts and does not contain personal information.

## Collection Process

**How was the data associated with each instance acquired?**
The data are directly observable from published ETR books. Each ETR version is produced by a pair of trained transcribers working collaboratively, in accordance with the European Easy-to-Read guidelines (Pathways, 2021), to obtain official ETR certification.

**What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?**
To collect the data from ETR books, we first obtained the PDF versions and manually curated them to identify pairs of pages containing the original text and its corresponding ETR version. The textual content was then extracted using the Python library `pypdfium2`[9].

**If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?**
The dataset is not sampled from a larger set; it includes the complete collection of available aligned texts selected for the study.

**Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?**
Unknown for the mannual book transcrptions. The data collection was carried out by the main author of this paper as part of their research work.

**Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)?**
The exact creation dates of the original books are unknown. However, the dataset itself was constructed between May 2023 and June 2023.

**Were any ethical review processes conducted (e.g., by an institutional review board)?**
No ethical review.

**Does the dataset relate to people?**
No.

## Preprocessing/cleaning/labeling

**Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?**

Manual cleaning was performed to remove chapter titles from the original texts, as these were not present in the corresponding ETR versions.

**Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?**
Yes. The raw data is provided alongside the cleaned version.

**Is the software used to preprocess/clean/label the instances available?**

- `pypdfium2`: `https://github.com/pypdfium2-team/pypdfium2`
- `cleantext`: `https://pypi.org/project/cleantext/`

## Uses

**Has the dataset been used for any tasks already?**
No.

**What (other) tasks could the dataset be used for?**
This dataset could also be used for text classification and style transfer.

[9] `https://github.com/pypdfium2-team/pypdfium2`

**Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?**
No.

**Are there tasks for which the dataset should not be used?**
No.

## Distribution

**How will the dataset will be distributed (e.g., tarball on website, API, GitHub)**
The dataset will be available on GitHub repository.

**When will the dataset be distributed?**
The dataset will be released pending agreement from the ETR books publisher.

**Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?**
The dataset will be released under a custom license, subject to approval from the ETR books publisher. Redistribution and use will be permitted for research purposes only, with appropriate citation. No commercial use will be allowed without explicit permission.

**Have any third parties imposed IP-based or other restrictions on the data associated with the instances?**

No.

**Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?**

No restrictions.

## Maintenance

**Who will be supporting/hosting/maintaining the dataset?**
The dataset will be maintained by the primary author of the paper.

**How can the owner/curator/manager of the dataset be contacted (e.g., email address)?**
By submitting an issue on the dataset's GitHub repository.

**Is there an erratum?**
Yes, errata can be reported and tracked via GitHub issues.

**Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?**
Yes, updates will be handled by the repository maintainer on GitHub. Users can receive update notifications by subscribing to the repository.

**If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)?**
This dataset does not contain or pertain to any personal data.

**Will older versions of the dataset continue to be supported/hosted/maintained?**
Yes, previous versions will remain available in the "Releases" section of the GitHub repository.

**If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?**
Yes, contributors may open a GitHub issue and submit a pull request. They should mention the maintainer and clearly describe their proposed changes, which will then be reviewed and validated before being merged.

# E Human Evaluation Questions

Table 9 presents a comprehensive set of human evaluation questions based on the ETR European guidelines, organized into four key categories: Information Choice, Sentence Construction and Word Choice, Illustrations, and Overall Quality. Each category includes multiple criteria designed to assess the clarity, structure, and accessibility of information provided in a text. For example, the Information Choice section evaluates whether essential information is prioritized, logically ordered, and clearly grouped. Sentence Construction and Word Choice emphasizes linguistic simplicity, clarity, and consistency, discouraging complex vocabulary, metaphors, or abbreviations unless adequately explained. The Illustrations section assesses the use of relatable examples to clarify abstract ideas, while the Quality section covers fluency, grammar, factual correctness, coherence, and other aspects of textual integrity. These criteria serve as a structured framework to ensure texts are understandable, reader-friendly, and fit for purpose.

| Information Choice | Code | Description |
|---|---|---|
| Information Choice | CI3 | Providing too much information can create confusion. Only important information should be given. Is this criterion met? |
| | CI4 | Are the pieces of information placed in an order that is easy to follow and understand? |
| | CI5 | Is the main information easy to find? |
| | CI6 | Are pieces of information about the same topic grouped together? |
| | CI8 | Are important pieces of information repeated? |
| Sentence construction and word choice | CPM1 | Are the sentences short? |
| | CPM2 | Are the words easy to understand? |
| | CPM3 | Are difficult words clearly explained when you use them? |
| | CPM4 | Are difficult words explained more than once? |
| | CPM5 | Is the language used the most suitable for the people who will use the information? |
| | CPM6 | Is the same word used throughout the document to describe the same thing? |
| | CPM7 | Difficult and abstract ideas like metaphors should not be used. Is this criterion met? |
| | CPM8 | Uncommon words in a foreign language should not be used. Is this criterion met? |
| | CPM9 | Contracted words, like text messaging slang, should not be used. Is this criterion met? |
| | CPM10 | Does the author address directly the people for whom the information is intended? |
| | CPM11 | Can you easily identify to whom or what the pronouns correspond? |
| | CPM12 | Are positive sentences rather than negative ones used whenever possible? |
| | CPM13 | Is the active voice used instead of the passive voice whenever possible? |
| | CPM14 | Is the punctuation simple? |
| | CPM15 | Are bullets or numbers used instead of lists of words separated by commas? |
| | CPM16 | Are numbers written in digits (1, 2, 3) rather than words? |
| | CPM17 | Acronyms should be avoided or explained when used. Is this criterion met? |
| | CPM18 | Abbreviations should not be used. Is this criterion met? |
| | CPM19 | Are dates written out in full? |
| | CPM20 | The use of percentages or large numbers should be limited and always explained. Is this criterion met? |
| | CPM21 | Special characters should not be used. Is this criterion met? |
| Illustrations | I1 | Are there examples to illustrate complex ideas? |
| | I2 | Are examples, as much as possible, drawn from everyday life? |
| Quality | CA1 | Language fluency |
| | CA2 | Grammar / Spelling |
| | CA3 | Factual accuracy |
| | CA4 | Textual coherence |
| | CA5 | Presence of copies from the original text? |
| | CA6 | Presence of chaotic repetitions? |
| | CA7 | Presence of hallucinations? |
| | CA8 | Overall perceived quality |

Table 9: Evaluation criteria, extracted from ETR European guidelines, for information clarity, sentence construction, illustrations, and quality.

```
Rewrite this text by following the principles of clarity and accessibility below:
- Provide only essential information. Avoid information overload.
- Present the information in a logical and easy-to-follow order.
- Highlight the main message right from the start.
- Group related information together.
- Repeat important information if it helps understanding.
- Use short and simple sentences.
- Choose easy-to-understand words.
- Clearly explain difficult words, and repeat the explanation if needed.
- Use language appropriate for the intended audience.
- Use the same word to refer to the same thing throughout the text.
- Avoid abstract ideas, metaphors, and complex comparisons.
- Don’t use foreign or obscure words without explanation.
- Avoid contractions and texting-style language.
- Speak directly to the reader in a clear and accessible way.
- Ensure that pronouns are always clear and unambiguous.
- Prefer positive phrasing over negative.
- Use the active voice as much as possible.
- Choose simple punctuation.
- Use bullet points or numbers for lists, not commas.
- Write numbers as digits (e.g., 1, 2, 3), not in words.
- Explain acronyms the first time they appear.
- Don’t use unexplained abbreviations.
- Write dates out in full for better clarity.
- Limit use of percentages or large numbers, and explain them simply.
- Don’t use unnecessary special characters.
- Use concrete examples to explain complex ideas.
- Prefer examples from everyday life.
###Input: <input_text>
###Output:
```

(a) Zero Shot Prompt

```
Rewrite this text by following the principles of clarity and accessibility below:
- Provide only essential information. Avoid information overload.
- Present the information in a logical and easy-to-follow order.
- Highlight the main message right from the start.
- Group related information together.
- Repeat important information if it helps understanding.
- Use short and simple sentences.
- Choose easy-to-understand words.
- Clearly explain difficult words, and repeat the explanation if needed.
- Use language appropriate for the intended audience.
- Use the same word to refer to the same thing throughout the text.
- Avoid abstract ideas, metaphors, and complex comparisons.
- Don’t use foreign or obscure words without explanation.
- Avoid contractions and texting-style language.
- Speak directly to the reader in a clear and accessible way.
- Ensure that pronouns are always clear and unambiguous.
- Prefer positive phrasing over negative.
- Use the active voice as much as possible.
- Choose simple punctuation.
- Use bullet points or numbers for lists, not commas.
- Write numbers as digits (e.g., 1, 2, 3), not in words.
- Explain acronyms the first time they appear.
- Don’t use unexplained abbreviations.
- Write dates out in full for better clarity.
- Limit use of percentages or large numbers, and explain them simply.
- Don’t use unnecessary special characters.
- Use concrete examples to explain complex ideas.
- Prefer examples from everyday life.
###Exemple 1
Task: <task_name>
Input: <example_input>
Output: <example_output>
...
Complete the following example:
Task: ETR
Input: <input_text>
Output:
```

(b) Few Shot Prompt

```
1. Analyze the text to identify what can be simplified or clarified.
2. Briefly note the points that need improvement (syntax, vocabulary, structure...).
3. Rewrite the text by applying the following guidelines:
- Provide only essential information. Avoid information overload.
- Present the information in a logical and easy-to-follow order.
- Highlight the main message right from the start.
- Group related information together.
- Repeat important information if it helps understanding.
- Use short and simple sentences.
- Choose easy-to-understand words.
- Clearly explain difficult words, and repeat the explanation if needed.
- Use language appropriate for the intended audience.
- Use the same word to refer to the same thing throughout the text.
- Avoid abstract ideas, metaphors, and complex comparisons.
- Don't use foreign or obscure words without explanation.
- Avoid contractions and texting-style language.
- Speak directly to the reader in a clear and accessible way.
- Ensure that pronouns are always clear and unambiguous.
- Prefer positive phrasing over negative.
- Use the active voice as much as possible.
- Choose simple punctuation.
- Use bullet points or numbers for lists, not commas.
- Write numbers as digits (e.g., 1, 2, 3), not in words.
- Explain acronyms the first time they appear.
- Don't use unexplained abbreviations.
- Write dates out in full for better clarity.
- Limit use of percentages or large numbers, and explain them simply.
- Don't use unnecessary special characters.
- Use concrete examples to explain complex ideas.
- Prefer examples from everyday life.
Start by reasoning step by step, then finish by providing the final version.
###Input: <input_text>
###Output:
```

(c) Chain of Thought Prompt

Figure 5: Zero Shot, Chain of Thought and Few Shot Prompts