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

Anonymous ACL submission

Abstract

Simplifying complex texts is essential for ensuring equitable access to information, especially for individuals with cognitive impairments. The Easy-to-Read (ETR) initiative offers a framework for making content accessible to the neurodivergent population, but the manual creation of 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 specificity of ETR constraints, we propose a multi-task learning (MTL) approach that trains models jointly on text summarization, text simplification, and ETR generation. We explore two different strategies: multi-task retrieval-augmented generation (RAG) for in-context learning, and MTL-LoRA for parameter-efficient fine-tuning. Our experiments with Mistral-7B and LLaMA-3-8B, based on ETR-fr, a new high-quality dataset, demonstrate the benefits of multi-task setups over single-task baselines across all configurations. Moreover, results show that the RAG-based strategy enables generalization in out-of-domain settings, while MTL-LoRA outperforms all learning strategies within in-domain configurations. Our code is publicly made available at https://anonymous.4open.science/r/ETR-MTL-C60E.

1 Introduction

Mental health and intellectual disabilities affect millions globally, posing serious challenges for equitable access to information (Maulik et al., 2011; Gustavsson et al., 2011). People with cognitive impairments often struggle with complex texts, limiting their participation in healthcare, education, and civic life. Despite international initiatives for inclusion,[1][2], accessible written content remains a major barrier for the neurodivergent population.

To address this issue, the Easy-to-Read (ETR) framework (Pathways, 2021) provides guidelines for producing cognitively accessible content. ETR prioritizes the use of clear and simple language, concise active sentences, consistent terminology, and supportive layout elements. It further necessitates collaboration between experts and individuals with cognitive impairments to validate accessibility, ensure adherence to guidelines, and meet the criteria for the 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 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 as illustrated in Figure 1.

In this paper, we address these gaps by introducing ETR-fr, the first dataset of 523 document-aligned text pairs fully compliant with the European ETR guidelines. We explore multi-task (MTL) learning to boost ETR generation by uniting 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

[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

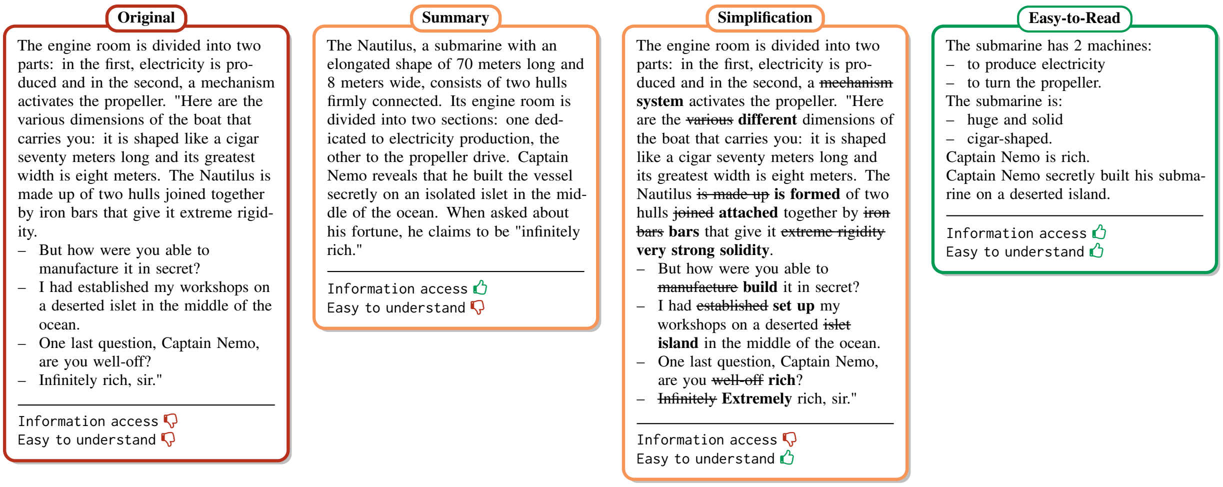


Figure 1: Different versions derived from a passage of *Twenty Thousand Leagues Under the Seas* by Jules Verne: from left to right, the original passage, an abstractive summary, a lexical simplification (crossed-out followed by words in bold indicate substitutions), and an Easy-to-Read generation targeting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

(PEFT) using MTL-LoRA (Yang et al., 2024). 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 28-point rubric from the European ETR guidelines, measuring clarity, coherence, and accessibility. Our experiments conducted on ETR-fr highlight the advantages of MTL setups over single-task baselines across all configurations. Furthermore, the results indicate that the 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, document-aligned dataset for ETR generation, fully compliant with European guidelines; (2) we benchmark multi-task ICL and PEFT approaches for ETR generation, introducing **MTL PEFT** to this task for the first time; (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 works support communication for users with cognitive impairments, often via dialogue agents (Martin and Nagalakshmi, 2024; Murillo-Morales et al., 2020; Huq et al., 2024; Wang et al., 2024). Much of the existing work has focused on dyslexia. For instance, Goodman et al. (2022) developed an email assistant based on LaMDA (Thoppilan et al., 2022), but found that the LLM's outputs lacked precision. In the French context, HECTOR (Todirascu et al., 2022) explored lexical and syntactic simplification, yielding mixed results. Efforts in other languages reveal similar challenges. In Finnish, Dmitrieva and Tiedemann (2024) aligned Easy-Finnish data with mBART (Liu et al., 2020) and FinGPT (Luukkonen et al., 2023), but reported poor alignment and partial compliance with ETR standards. For Spanish, ClearText (Espinosa-Zaragoza et al., 2023) uses ChatGPT to simplify administratives texts, however its corpus remains limited and prone to errors. Martínez et al. (2024) developed a sentence-level simplification dataset and fine-tuned LLaMA-2 (Touvron et al., 2023b), finding that translation-based methods suffer from semantic drift and domain mismatch.

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., 2023a; 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

	# 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	296402	34.88	29.28	1.68	1.56	27.53	23.74	65.38	71.35	31.97	12.79
OrangeSum	24401	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.

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., 2024) enable scalable adaptation of LLMs by modifying only a subset of parameters. LoRA leverage the intrinsic dimensionality of language models (Aghajanyan et al., 2021) to achieve strong performance with minimal computational overhead. However, LoRA-based strategies struggle in multi-task settings due to conflicting updates accross tasks (Wang et al., 2023). Alternatives like MultiLoRA (Wang et al., 2023) and MoELoRA (Liu et al., 2023b) seek to balance generalization with task specificity, but face challenges in task routing and mitigating interference. MTL-LoRA (Yang et al., 2024) addresses this by introducing both 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.

Description. ETR-fr consists of 523 paragraph-aligned text pairs. Table 1 outlines key dataset statistics, including KMRE readability score (Kandel and Moles, 1958), compression ratios, and lexical novelty. On average, the dataset yields a compression rate of 50.05%, with a reduction of 56.61 tokens and 2.17 sentences per pair. The average novelty rate, following Narayan et al. (2018), is 53.80%, reflecting the proportion of newly introduced unigrams in target texts. Readability improves by 7.51 KMRE points from source to target. The dataset is partitioned into fixed train, validation, and test subsets. The test set includes two books selected to maximize variation in linguistic attributes (e.g., sentence length, compression, novelty). The remaining nine books are divided into training and validation via stratified sampling.

ETR-fr-politic To assess generalization and robustness, we introduce ETR-fr-politic, an out-of-domain test set with 33 ETR-aligned paragraphs sampled from the 2022 French presidential election programs.[5] Compared to ETR-fr, the ETR-fr-politic dataset features shorter source texts (96.27 vs. 102.76 words) and fewer sentences (6.03 vs. 9.30), but yields longer rewritten outputs (62.85 vs. 46.15 words). Additionally, ETR-fr-politic exhibits higher novelty (63.78% vs. 53.80%) and significantly lower compression ratios (29.17% vs. 50.05%), indicating a greater degree of content expansion. While ETR-fr exhibits higher overall simplicity scores both before and after rewriting (91.43 and 98.94) compared to ETR-fr-politic (74.00 and

[4] http://www.yvelinedition.fr/Facile-a-lire
[5] https://www.cnccep.fr/candidats.html

87.74), the latter achieves a greater simplification gain, with a larger increase in KMRE (+13.75 vs. +7.51 points). Overall, ETR-fr-politic poses a more challenging and higher-novelty setting for evaluating ETR systems in politically sensitive, real-world rewriting contexts.[6]

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). By contrast, both ETR-fr and OrangeSum support document-level simplification, offering substantially longer inputs (102.76 and 375.98 words, respectively). 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, it 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 document-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

Our experiments leverage the ETR-fr dataset as the primary resource, supplemented by related rewriting tasks sourced from the OrangeSum summarization dataset and the lexical simplification dataset WikiLarge FR. 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). Note that foundation models are used for PEFT and their Instruct versions for ICL.

[6]Note that the documents on politics usually do not meet high-quality standards as evidenced by the François Baudez Publishing collection. Moreover, there are still difficult to gather as their repository is not centralized.

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 has been shown to be unsuitable for text simplification (Sulem et al., 2018; Xu et al., 2016; Tanprasert and Kauchak, 2021). To 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.

4.2 Multi-Task In-Context Learning

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 instruction, 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 B.

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 B (Figure 3 and Figure 4). 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. As baseline, we implement LoRA (Hu et al., 2022). LoRA approximates full fine-tuning by decomposing weight matrices into low-rank components. A weight matrix $\mathbf{W_0} \in \mathbb{R}^{d \times k}$ into two smaller 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. (2024) introduce MTL-LoRA. Given task input x_t, MTL-LoRA first applies a shared standard LoRA down-projection via matrix $\mathbf{A}$. 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, τ controls the softness of the weighting. Each Λ_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 from (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

Best models are selected based on the highest SRB score on the ETR-fr validation set, following a grid search hyperparameter tuning strategy.[7] To complement this analysis, all models are run five times with different seeds, and detailed average results can be found in Appendix C.

5.1 In-Domain Quantitative Results

ICL Performance. As shown in Table 2, 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.

[7] Hyperparameter tuning is detailed in Appendix A.

	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
Paramter-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

Table 2: **Performance comparison, on ETR-fr test set**, 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.

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 approach for high-quality in-domain ETR-fr.

5.2 Out-of-Domain Quantitative Results

ICL Performance. As shown in Table 3, 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,

Model	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
Paramter-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

Table 3: **Performance comparison, on ETR-fr-politic test set**, 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.

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 outputs are generated.

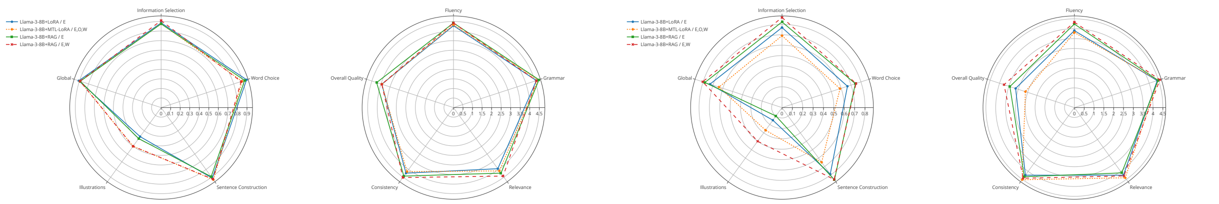
Impact of Fine-Tuning. While PEFT strategies generally lag behind RAG in terms of SRB and BERTScore, they offer stable and interpretable performance, with notably better compression ratios than zero-shot, CoT and most RAG-based strategies. The best PEFT model in terms of SRB, LLaMA-3-8B+LoRA trained 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 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,[8] to ensure content is understandable and appropriate for the target audience. We validated our approach through human evaluation with 10 native French speakers, seven NLP researchers and three linguists, who assessed outputs from the ETR-fr and ETR-politic

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



(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.

test sets.[9] 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 28 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 (1–5).

In-domain Results. Figures 2 present the human evaluation results.[10] 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, 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.

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.

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

[10] Overall scores are provided in a table in Appendix C.2.

7 Limitations

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

Misalignment with deployment contexts. 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 neurodivergent users remains untested.

Absence of direct end-user feedback. Human evaluation was conducted by proxy annotators, which limits insights into subjective usability, emotional response, and real-world accessibility, central concerns in ETR adoption.

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.

Simplification-centric task framing. Our formulation treats ETR as summarization and simplification. However, this may overlook strategies unique to ETR, such as intentional redundancy, explicit inference resolution, and narrative scaffolding, often crucial for accessibility.

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.

8 Impact and Ethical Considerations

Social and Ethical Challenge. Identifying limitations is essential for transparency and inclusive design. ETR generation impacts neurodivergent readers and intersects with accessibility, language rights, and communicative equity. As such, simplification systems must be evaluated not only on linguistic performance but on their potential to oversimplify or marginalize. By clarifying the limitations of our work, we aim to support responsible development and deployment. Acknowledging these boundaries also helps position ETR generation as a socio-technical task, one that demands sensitivity to both linguistic quality and lived experience.

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.

References

Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. 2021. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 7319–7328, Online. 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.

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.

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, and Alex Vaughan. 2024. The llama 3 herd of models. Preprint, arXiv:2407.21783.

Anders Gustavsson, Mikael Svensson, Frank Jacobi, Christer Allgulander, Jordi Alonso, Ettore Beghi, Richard Dodel, Mattias Ekman, Carlo Faravelli, Laura Fratiglioni, Brenda Gannon, David Hilton Jones, Poul Jennum, Albena Jordanova, Linus Jönsson, Korinna Karampampa, Martin Knapp, Gisela Kobelt, Tobias Kurth, Roselind Lieb, Mattias Linde, Christina Ljungcrantz, Andreas Maercker, Beatrice Melin, Massimo Moscarelli, Amir Musayev, Fiona Norwood, Martin Preisig, Maura Pugliatti, Juergen Rehm, Luis Salvador-Carulla, Brigitte Schlehofer, Roland Simon, Hans-Christoph Steinhausen, Lars Jacob Stovner, Jean-Michel Vallat, Peter Van den Bergh, Jim van Os, Pieter Vos, Weili Xu, Hans-Ulrich Wittchen, Bengt Jönsson, and Jes Olesen. 2011. Cost of disorders of the brain in europe 2010. European Neuropsychopharmacology, 21(10):718–779.

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. arXiv preprint. ArXiv:2310.06825 [cs].

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.

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. 2023a. 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 6th International Conference on Learning Representations (ICLR).

Qidong Liu, Xian Wu, Xiangyu Zhao, Yuanshao Zhu, Derong Xu, Feng Tian, and Yefeng Zheng. 2023b. When moe meets llms: Parameter efficient fine-tuning for multi-task medical applications. In Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. 2024. Dora: Weight-decomposed low-rank adaptation. In Proceedings of 41st International Conference on Machine Learning.

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 International Conference on Learning Representations.

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.

Lara J. Martin and Malathy Nagalakshmi. 2024. Bridging the Social & Technical Divide in Augmentative and Alternative Communication (AAC) Applications for Autistic Adults. Publisher: arXiv Version Number: 2.

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.

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.

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.

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, ICML'24. JMLR.org.

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.

R. Thoppilan, Daniel De Freitas, Jamie Hall, Noam M. Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, Yaguang Li, Hongrae Lee, H. Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, M. Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Yanqi Zhou, Chung-Ching Chang, I. Krivokon, W. Rusch, Marc Pickett, K. Meier-Hellstern, M. Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, J. Søraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Díaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, V. Kuzmina, Joseph Fenton, Aaron Cohen, R. Bernstein, R. Kurzweil, Blaise Aguera-Arcas, Claire Cui,

M. Croak, E. Chi, and Quoc Le. 2022. LaMDA: Language Models for Dialog Applications. ArXiv.

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, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 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.

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 H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

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.

Yaming Yang, Dilxat Muhtar, Yelong Shen, Yuefeng Zhan, Jianfeng Liu, Yujing Wang, Hao Sun, Denvy Deng, Feng Sun, Qi Zhang, Weizhu Chen, and Yunhai Tong. 2024. Mtl-lora: Low-rank adaptation for multi-task learning. CoRR, abs/2410.09437.

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 International Conference on Learning Representations.

A Implementation Details

A.1 MTL-LoRA

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. 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

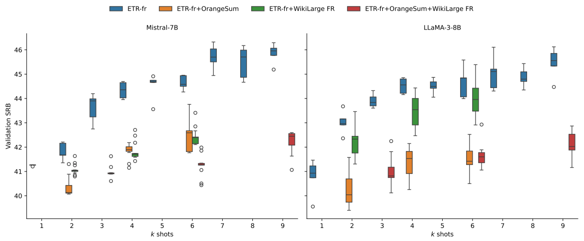


Figure 3: SRB performance score of Mistral-7B and LLaMA-3-8B on the ETR-fr validation set with varying number of in-context examples (k = 1–9) and task combinations.

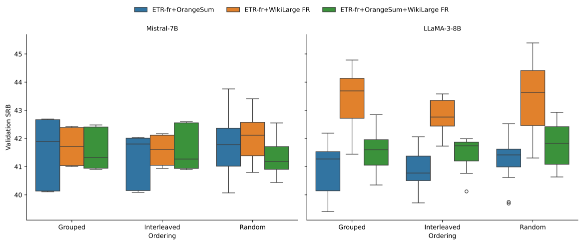


Figure 4: SRB performance of Mistral-7B and LLaMA-3-8B on the ETR-fr validation set under different example ordering strategies and task combination configurations.

MTL-LoRA configuration, sharpness of the weight distribution is fixed at 0.5 and the optimal n up-projections is selected among $\{1, 2, 3\}$. Best hyperparameters for PEFT methods are in Table 4

A.2 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 performance. We investigate three strategies for sequencing demonstrations in the prompt as mentioned in Section4.2: random, grouped and interleaved orderings.

The optimal hyperparameters for in-context learning are summarized in Table 5.

B In-Context Learning

Figure 5 illustrates examples of prompts used for zero-shot, chain-of-thought and RAG.

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

Figure 3 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 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 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 scores 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 scores 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 scores for LLaMA across configurations (e.g., up to 46.12) suggest it may have a higher performance ceiling, buy with more fluctuation.

B.2 Conclusion

In summary, increasing the in-context learning size (k) generally improves model performance. Task combination has mixed effects: beneficial in some configurations but detrimental in others, especially

			Batch size	lr	Acc. steps	Epochs	$\alpha = r$	Attn. matrices	n up proj.	τ
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 4: PEFT hyperparameter configurations selected based on SRB performance on the ETR-fr validation set.

			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 5: ICL hyperparameter configurations selected based on SRB performance on the ETR-fr validation set.

when too many tasks are combined. LLaMA-3-8B appears more sensitive to these changes than Mistral-7B, highlighting important considerations for prompt engineering.

B.2.1 Impact of the Tasks Ordering on ETR-fr Performance

Figure 4 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 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 scores 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 scores 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 scores, 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.

C Complementary Evaluation Results

C.1 Quantitative Results

The average performances of various methods on the ETR-fr and ETR-fr-politic test sets is presented in tables 6a and 6b, 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 instruction-tuned models (Mistral-7B, LlaMA-3-8B) 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 CoT-settings generally underperform across all metrics compared to RAG and PEFT. While CoTshows 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) PEFT models significantly outperform ICL approaches across the board. Both LoRA and MTL-LoRA configurations demonstrate strong improvements in fluency, simplicity, and informativeness. LlaMA-3-8B-MTL-LoRA shows the best overall performance, especially on metrics like SARI, BERT-F1, and Comp. ratio, reflecting its superior simplification quality and semantic fidelity. Multi-task LoRA (E+W) achieves the highest SARI (44.67), BERT-F1 (74.05), and compression ratio (56.11), indicating a well-balanced simplification that maintains semantic consistency while significantly reducing text length.

Out-of-Domain (ETR-fr-politic) Performance The performance gap between ICL and PEFT narrows slightly on the political subset, but PEFT models still maintain a strong advantage. RAG methods maintain their relative lead among ICL approaches, especially when enhanced with additional context (E+W and E+O+W), suggesting their better generalization ability. Interestingly, Zero-Shot LlaMA-3-8B achieves the highest novelty score (55.73), which may reflect increased variability but could also indicate decreased fidelity.

C.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 7 and 8.

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 8, RAG again performs competitively on both datasets. On ETR-fr, it achieves the highest ratings in grammar and coherence (both > 4.4), with 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 E and W. 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

Model	Method	Task	R-1	R-2	R-L	SARI	BERT-F1	SRB	Comp. ratio	Novelty
In Context Learning										
Mistral-7B	Zero-Shot	E	$23.96_{\pm 0.04}$	$7.08_{\pm 0.01}$	$16.25_{\pm 0.03}$	$37.07_{\pm 0.00}$	$69.75_{\pm 0.00}$	$29.17_{\pm 0.03}$	$-64.14_{\pm 0.00}$	$35.70_{\pm 0.00}$
	CoT	E	$23.53_{\pm 0.06}$	$7.23_{\pm 0.01}$	$16.20_{\pm 0.04}$	$37.39_{\pm 0.00}$	$68.80_{\pm 0.00}$	$29.12_{\pm 0.05}$	$-60.53_{\pm 0.00}$	$\underline{36.09}_{\pm 0.00}$
	RAG	E	$\underline{31.91}_{\pm 0.66}$	$\underline{10.77}_{\pm 0.65}$	$\underline{22.54}_{\pm 0.75}$	$\underline{40.14}_{\pm 0.57}$	$\underline{72.17}_{\pm 0.30}$	$\underline{36.08}_{\pm 0.80}$	$45.23_{\pm 1.17}$	$27.27_{\pm 0.58}$
		E,O	$30.36_{\pm 0.47}$	$9.61_{\pm 0.34}$	$21.80_{\pm 0.30}$	$39.49_{\pm 0.12}$	$71.07_{\pm 0.18}$	$35.19_{\pm 0.29}$	$\underline{47.99}_{\pm 1.91}$	$26.80_{\pm 0.84}$
		E,W	$30.46_{\pm 0.48}$	$9.93_{\pm 0.17}$	$21.72_{\pm 0.34}$	$38.76_{\pm 0.43}$	$71.57_{\pm 0.14}$	$34.96_{\pm 0.34}$	$35.08_{\pm 2.13}$	$23.32_{\pm 0.31}$
		E,O,W	$29.85_{\pm 0.04}$	$9.58_{\pm 0.03}$	$21.55_{\pm 0.05}$	$39.53_{\pm 0.00}$	$71.06_{\pm 0.00}$	$34.98_{\pm 0.05}$	$46.42_{\pm 0.00}$	$25.85_{\pm 0.00}$
LlaMA-3-8B	Zero-Shot	E	$24.90_{\pm 0.20}$	$8.16_{\pm 0.25}$	$17.10_{\pm 0.38}$	$38.48_{\pm 0.38}$	$70.15_{\pm 0.17}$	$30.38_{\pm 0.48}$	$-22.52_{\pm 2.47}$	$\mathbf{39.13}_{\pm 0.92}$
	CoT	E	$27.23_{\pm 0.91}$	$8.81_{\pm 0.21}$	$18.34_{\pm 0.57}$	$38.15_{\pm 0.23}$	$70.79_{\pm 0.52}$	$31.62_{\pm 0.65}$	$7.59_{\pm 4.82}$	$30.33_{\pm 1.75}$
	RAG	E	$\underline{33.05}_{\pm 0.72}$	$\underline{12.23}_{\pm 0.44}$	$\underline{23.77}_{\pm 0.68}$	$\underline{41.66}_{\pm 0.45}$	$\underline{72.59}_{\pm 0.38}$	$\underline{37.57}_{\pm 0.70}$	$\underline{43.36}_{\pm 2.62}$	$27.06_{\pm 0.29}$
		E,O	$30.77_{\pm 0.35}$	$10.85_{\pm 0.31}$	$22.10_{\pm 0.35}$	$39.84_{\pm 0.22}$	$71.13_{\pm 0.17}$	$35.54_{\pm 0.32}$	$24.36_{\pm 30.13}$	$25.02_{\pm 1.84}$
		E,W	$32.14_{\pm 0.56}$	$11.70_{\pm 0.34}$	$23.11_{\pm 0.19}$	$40.49_{\pm 0.32}$	$71.88_{\pm 0.18}$	$36.64_{\pm 0.24}$	$42.30_{\pm 1.59}$	$26.70_{\pm 0.92}$
		E,O,W	$30.53_{\pm 0.74}$	$10.67_{\pm 0.45}$	$21.65_{\pm 0.71}$	$39.24_{\pm 0.20}$	$71.21_{\pm 0.26}$	$35.00_{\pm 0.67}$	$31.18_{\pm 4.94}$	$24.08_{\pm 1.37}$
PEFT										
Mistral-7B	LoRA	E	$32.45_{\pm 0.03}$	$12.38_{\pm 0.02}$	$23.99_{\pm 0.05}$	$42.09_{\pm 0.00}$	$73.56_{\pm 0.00}$	$37.95_{\pm 0.04}$	$44.42_{\pm 0.00}$	$18.35_{\pm 0.00}$
	MTL-LoRA	E,O	$32.62_{\pm 0.04}$	$12.73_{\pm 0.01}$	$24.29_{\pm 0.04}$	$41.95_{\pm 0.00}$	$73.52_{\pm 0.00}$	$38.16_{\pm 0.03}$	$53.48_{\pm 0.00}$	$24.17_{\pm 0.00}$
		E,W	$32.68_{\pm 0.05}$	$\underline{12.91}_{\pm 0.01}$	$24.25_{\pm 0.03}$	$\underline{42.53}_{\pm 0.00}$	$\underline{73.90}_{\pm 0.00}$	$38.33_{\pm 0.03}$	$\underline{53.62}_{\pm 0.00}$	$\underline{24.99}_{\pm 0.00}$
		E,O,W	$\mathbf{33.60}_{\pm 0.05}$	$12.81_{\pm 0.05}$	$\underline{24.89}_{\pm 0.04}$	$42.25_{\pm 0.00}$	$73.62_{\pm 0.00}$	$\underline{38.74}_{\pm 0.03}$	$48.93_{\pm 0.00}$	$23.38_{\pm 0.00}$
LlaMA-3-8B	LoRA	E	$31.80_{\pm 0.03}$	$13.16_{\pm 0.09}$	$24.92_{\pm 0.18}$	$42.15_{\pm 0.01}$	$72.84_{\pm 0.17}$	$38.67_{\pm 0.17}$	$50.50_{\pm 0.28}$	$18.37_{\pm 0.88}$
	MTL-LoRA	E,O	$\underline{33.38}_{\pm 0.06}$	$13.16_{\pm 0.05}$	$24.20_{\pm 0.04}$	$43.06_{\pm 0.01}$	$73.88_{\pm 0.01}$	$38.42_{\pm 0.03}$	$50.90_{\pm 0.40}$	$23.25_{\pm 0.17}$
		E,W	$32.54_{\pm 0.05}$	$13.50_{\pm 0.06}$	$25.01_{\pm 0.06}$	$\mathbf{44.67}_{\pm 0.00}$	$\mathbf{74.05}_{\pm 0.00}$	$39.54_{\pm 0.05}$	$\mathbf{56.11}_{\pm 0.00}$	$\underline{33.05}_{\pm 0.00}$
		E,O,W	$32.78_{\pm 0.02}$	$\mathbf{13.67}_{\pm 0.03}$	$\mathbf{25.55}_{\pm 0.16}$	$43.58_{\pm 0.10}$	$73.33_{\pm 0.09}$	$\mathbf{39.62}_{\pm 0.09}$	$52.66_{\pm 1.00}$	$24.27_{\pm 0.21}$

(a) Performance on ETR-fr test set.

Model	Method	Task	R-1	R-2	R-L	SARI	BERT-F1	SRB	Comp. ratio	Novelty
In Context Learning										
Mistral-7B	Zero-Shot	E	$28.42_{\pm 0.12}$	$10.98_{\pm 0.07}$	$19.31_{\pm 0.03}$	$39.87_{\pm 0.00}$	$68.10_{\pm 0.00}$	$32.77_{\pm 0.03}$	$-309.24_{\pm 0.00}$	$48.37_{\pm 0.00}$
	CoT	E	$29.80_{\pm 0.03}$	$11.21_{\pm 0.05}$	$19.88_{\pm 0.08}$	$39.62_{\pm 0.00}$	$69.40_{\pm 0.00}$	$33.35_{\pm 0.07}$	$-261.30_{\pm 0.00}$	$\underline{50.85}_{\pm 0.00}$
	RAG	E	$\mathbf{40.19}_{\pm 0.63}$	$\underline{16.07}_{\pm 0.60}$	$28.25_{\pm 0.31}$	$\underline{41.40}_{\pm 0.46}$	$\underline{73.01}_{\pm 0.34}$	$\underline{40.96}_{\pm 0.35}$	$9.00_{\pm 3.96}$	$23.21_{\pm 2.39}$
		E,O	$37.49_{\pm 0.61}$	$14.50_{\pm 0.35}$	$26.38_{\pm 0.69}$	$39.46_{\pm 0.35}$	$72.27_{\pm 0.26}$	$38.92_{\pm 0.58}$	$14.26_{\pm 2.65}$	$17.57_{\pm 1.61}$
		E,W	$39.65_{\pm 0.19}$	$15.36_{\pm 0.35}$	$27.85_{\pm 0.38}$	$40.08_{\pm 0.36}$	$72.35_{\pm 0.29}$	$40.17_{\pm 0.23}$	$8.72_{\pm 1.73}$	$17.47_{\pm 1.68}$
		E,O,W	$39.14_{\pm 0.04}$	$15.96_{\pm 0.09}$	$\mathbf{28.40}_{\pm 0.11}$	$40.74_{\pm 0.00}$	$72.87_{\pm 0.00}$	$40.82_{\pm 0.07}$	$\underline{14.63}_{\pm 0.00}$	$18.33_{\pm 0.00}$
LlaMA-3-8B	Zero-Shot	E	$29.10_{\pm 0.40}$	$10.68_{\pm 0.35}$	$18.70_{\pm 0.41}$	$40.68_{\pm 0.48}$	$68.65_{\pm 0.11}$	$32.39_{\pm 0.51}$	$-178.23_{\pm 7.77}$	$\mathbf{55.73}_{\pm 1.07}$
	CoT	E	$31.15_{\pm 0.99}$	$10.47_{\pm 0.81}$	$19.54_{\pm 0.65}$	$39.80_{\pm 0.63}$	$69.66_{\pm 0.43}$	$33.09_{\pm 0.74}$	$-70.57_{\pm 8.09}$	$47.80_{\pm 1.71}$
	RAG	E	$37.68_{\pm 0.53}$	$14.46_{\pm 0.65}$	$26.09_{\pm 0.60}$	$42.05_{\pm 0.90}$	$73.01_{\pm 0.20}$	$39.57_{\pm 0.41}$	$1.47_{\pm 6.45}$	$41.78_{\pm 0.86}$
		E,O	$37.43_{\pm 2.11}$	$14.28_{\pm 0.89}$	$25.92_{\pm 1.42}$	$40.95_{\pm 0.90}$	$72.41_{\pm 0.61}$	$39.05_{\pm 1.37}$	$-7.72_{\pm 14.32}$	$31.85_{\pm 1.69}$
		E,W	$\underline{39.99}_{\pm 1.10}$	$\mathbf{16.27}_{\pm 0.61}$	$\underline{27.84}_{\pm 1.10}$	$\mathbf{42.41}_{\pm 0.43}$	$\mathbf{73.83}_{\pm 0.47}$	$\mathbf{41.06}_{\pm 0.96}$	$\underline{13.46}_{\pm 2.37}$	$36.72_{\pm 2.01}$
		E,O,W	$38.33_{\pm 1.46}$	$15.12_{\pm 1.08}$	$26.89_{\pm 1.10}$	$41.08_{\pm 0.94}$	$72.86_{\pm 0.51}$	$39.86_{\pm 1.13}$	$6.34_{\pm 7.54}$	$29.92_{\pm 0.48}$
PEFT										
Mistral-7B	LoRA	E	$35.10_{\pm 0.04}$	$12.28_{\pm 0.04}$	$25.97_{\pm 0.03}$	$38.04_{\pm 0.00}$	$70.28_{\pm 0.00}$	$37.96_{\pm 0.02}$	$21.55_{\pm 0.00}$	$11.79_{\pm 0.00}$
	MTL-LoRA	E,O	$29.29_{\pm 0.07}$	$11.02_{\pm 0.01}$	$21.90_{\pm 0.04}$	$38.68_{\pm 0.00}$	$69.22_{\pm 0.00}$	$34.90_{\pm 0.03}$	$\mathbf{36.68}_{\pm 0.00}$	$\underline{40.29}_{\pm 0.00}$
		E,W	$34.32_{\pm 0.06}$	$12.60_{\pm 0.07}$	$24.87_{\pm 0.11}$	$\underline{38.72}_{\pm 0.00}$	$70.54_{\pm 0.00}$	$37.40_{\pm 0.09}$	$22.51_{\pm 0.00}$	$19.10_{\pm 0.00}$
		E,O,W	$\underline{36.34}_{\pm 0.10}$	$\underline{13.24}_{\pm 0.02}$	$\underline{26.29}_{\pm 0.08}$	$38.39_{\pm 0.00}$	$\underline{70.97}_{\pm 0.00}$	$\underline{38.37}_{\pm 0.06}$	$18.33_{\pm 0.00}$	$10.55_{\pm 0.00}$
LlaMA-3-8B	LoRA	E	$34.65_{\pm 1.43}$	$13.34_{\pm 0.85}$	$26.40_{\pm 0.95}$	$\underline{39.70}_{\pm 0.35}$	$70.73_{\pm 0.99}$	$38.85_{\pm 0.90}$	$4.67_{\pm 2.97}$	$16.19_{\pm 0.11}$
	MTL-LoRA	E,O	$32.17_{\pm 0.52}$	$11.94_{\pm 0.23}$	$23.98_{\pm 0.22}$	$39.35_{\pm 0.44}$	$69.49_{\pm 0.21}$	$36.81_{\pm 0.06}$	$\underline{17.14}_{\pm 0.98}$	$\underline{20.01}_{\pm 0.62}$
		E,W	$\underline{37.58}_{\pm 0.12}$	$13.68_{\pm 0.05}$	$\underline{27.02}_{\pm 0.03}$	$38.26_{\pm 0.00}$	$\underline{71.30}_{\pm 0.00}$	$\underline{38.88}_{\pm 0.02}$	$8.45_{\pm 0.00}$	$6.44_{\pm 0.00}$
		E,O,W	$36.38_{\pm 0.22}$	$\underline{13.72}_{\pm 0.07}$	$25.75_{\pm 0.23}$	$36.19_{\pm 0.00}$	$70.94_{\pm 0.04}$	$37.24_{\pm 0.17}$	$8.76_{\pm 0.13}$	$2.04_{\pm 0.05}$

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

Table 6: **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 and LlaMA-3-8B 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.

like politics.

D Human Eval 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 es-

	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 7: **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 8: **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.

sential 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