Translate or Simplify First: An Analysis of Cross-lingual Text Simplification in English and French

Anonymous ACL submission

Abstract

Cross-Lingual Text Simplification (CLTS) aims to make content more accessible across languages by simultaneously addressing both linguistic complexity and translation. This study investigates the effectiveness of different prompting strategies for CLTS between English and French using large language models (LLMs). We examine five distinct prompting systems: a direct prompt instructing the LLM to perform both translation and simplification simultaneously, two chain-of-thought (CoT) approaches that either translate-then-simplify or simplify-then-translate within a single prompt, and two pipeline approaches that perform the same operations in separate, consecutive prompts. These systems are evaluated across a diverse set of five corpora of different genres (Wikipedia and medical texts) using seven state-of-the-art LLMs. Output quality is assessed through a multi-faceted evaluation framework comprising automatic metrics, comprehensive linguistic feature analysis, and human evaluation of simplicity and meaning preservation. Our findings reveal that while direct prompting consistently achieves the highest BLEU scores, indicating meaning fidelity, Translate-then-Simplify approaches demonstrate the highest simplicity, as measured by the linguistic features.[1]

1 Introduction

Text simplification (TS) reduces linguistic complexity while preserving meaning to improve accessibility for diverse audiences, such as language learners and individuals with reading difficulties. Traditionally, TS focuses on monolingual transformations like lexical substitution and syntactic restructuring (Siddharthan, 2014). However, the global demand for accessible content has created a growing need for Cross-Lingual Text Simplification (CLTS), which simultaneously translates and simplifies text across languages.

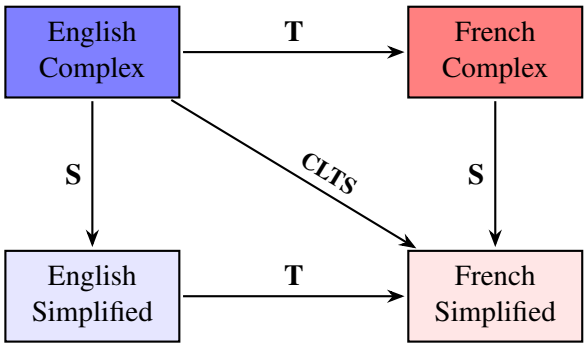

Figure 1: Conceptual diagram of text transformation tasks. Arrows represent the core operations: Translation (**T**), Simplification (**S**), and their combined operation, Cross-lingual text Simplification (**CLTS**).

CLTS presents unique challenges that go beyond the individual tasks of translation and simplification. While machine translation systems have achieved remarkable performance in recent years, they typically prioritize accuracy and fluency over accessibility (Leiter et al., 2024). Similarly, monolingual text simplification systems excel at reducing complexity within a single language but do not address the needs of multilingual audiences. The intersection of these two tasks requires careful consideration of how simplification strategies interact with translation processes, as linguistic complexity manifests differently across languages due to varying grammatical structures, vocabulary distributions, and cultural contexts (Miestamo, 2008).

CLTS can be conceptualized as a direct transformation from complex text in one language to simplified text in another. However, this single-step operation can alternatively be decomposed into sequential operations: either translating first then simplifying (T→S), or simplifying first then translating (S→T). Crucially, while Figure 1 depicts these three paths as distinct routes to the

[1] All the data and code will be made available upon publication.

same destination, they do not necessarily produce equivalent outputs. Each sequence of operations introduces different transformations, potential error propagation patterns, and linguistic trade-offs that may significantly impact the quality and characteristics of the final simplified text. This complexity is further compounded by the phenomenon of *translationese* - the distinct linguistic patterns that emerge in translated texts, which often include implicit simplification effects such as reduced lexical variety (Volansky et al., 2014).

Traditional automatic TS approaches rely on explicit training data and struggle with cross-lingual adaptation (Alva-Manchego et al., 2021). The advent of large language models offers a powerful alternative, demonstrating zero-shot capabilities in both translation and simplification. However, the optimal way to leverage LLMs for cross-lingual simplification remains an open question. Different prompting strategies may lead to varying trade-offs between translation quality and simplification effectiveness, and the order of operations (translate-then-simplify versus simplify-then-translate) may significantly impact the final output quality.

Recent advances in prompt engineering have shown that the way tasks are presented to LLMs can dramatically influence their performance (Wei et al., 2023). Chain-of-thought prompting, which encourages models to break down complex tasks into intermediate steps, has proven effective across various natural language processing tasks. Similarly, pipeline approaches that explicitly separate different sub-tasks may offer advantages over single-step approaches. However, the effectiveness of these different prompting strategies for CLTS has not been systematically investigated.

Furthermore, while automatic evaluation metrics provide valuable insights into system performance, they may not capture all aspects of simplification quality. Different metrics may favor different types of simplification strategies, and their effectiveness may vary depending on the target language. The linguistic features that characterize effective simplification may also differ across languages, highlighting the need for comprehensive analysis beyond traditional evaluation. To address these limitations, we incorporate human annotation to provide a more nuanced assessment of meaning preservation, and the actual degree of simplification achieved.

This study presents the first comprehensive evaluation of CLTS between English and French that systematically compares five different prompting strategies across state-of-the-art language models. Specifically, we focus on bidirectional cross-lingual simplification: from English to French and from French to English, allowing us to investigate directional asymmetries in simplification strategies and their effectiveness. Our evaluation encompasses seven different language models, applied to five corpora, using a robust evaluation framework that combines automatic metrics, linguistic feature analysis, and human judgment. Table 1 illustrates example outputs from three of these approaches applied to an English source text from the ASSET corpus (Alva-Manchego et al., 2020).

This study advances the field through two pivotal contributions: First, we provide the first systematic comparison of various prompting strategies for CLTS between English and French, revealing which approaches are most effective under different conditions with multiple automatic evaluation metrics, extensive linguistic analysis, and human evaluation. Second, we provide a comprehensive evaluation protocol for CLTS in English and French, including multiple datasets and multi-faceted evaluations adapted to the task, which can be extended to additional language pairs.

2 Related Work

2.1 Cross-Lingual Simplification

While monolingual text simplification has seen significant advancements, cross-lingual simplification remains an understudied area. Early work by Mallinson et al. (2020) demonstrated the feasibility of zero-shot knowledge transfer from English to German. More recently, the MultiCochrane dataset (Joseph et al., 2023) provided parallel sentence-aligned data for medical simplification across English, French, Spanish, and Farsi, facilitating research on how simplification strategies transfer across specific language pairs. These works have established that cross-lingual simplification benefits from understanding language-specific transfer patterns and operation ordering, though the optimal strategies for combining translation and simplification with current LLMs remain underexplored.

2.2 LLMs for (Cross-lingual) Text Generation

Recent advances in LLMs have demonstrated remarkable capabilities in zero-shot text generation tasks. Brown et al. (2020) established the foundational principle with GPT-3, showing that large-scale pre-trained models can perform various NLP

Original Text (English):

A few animals have chromatic response, changing color in changing environments, either seasonally (ermine, snowshoe hare) or far more rapidly with chromatophores in their integument (the cephalopod family).

French Simplification	English Translation
Direct Simplification: Quelques animaux changent de couleur pour s'adapter à leur environnement, soit au fil des saisons (hermine, lièvre d'Amérique), soit plus rapidement grâce à des cellules spéciales dans leur peau (les céphalopodes).	**Translation:** Some animals change color to adapt to their environment, either with the seasons (ermine, snowshoe hare), or more quickly thanks to special cells in their skin (cephalopods).
CoT T>S Simplification: Certains animaux changent de couleur pour s'adapter à leur environnement. Ce changement peut être saisonnier (comme chez l'hermine ou le lièvre d'Amérique) ou très rapide grâce à des cellules spéciales dans leur peau (comme chez les calmars ou les poulpes).	**Translation:** Some animals change color to adapt to their environment. This change can be seasonal (as in ermine or snowshoe hare) or very rapid thanks to special cells in their skin (as in squid or octopus).
Pipe S>T Simplification: Certains animaux changent de couleur pour se fondre dans leur environnement, soit saisonnièrement (comme l'hermine), soit rapidement grâce à des cellules cutanées spéciales (comme les pieuvres).	**Translation:** Some animals change color to blend into their surroundings, either seasonally (like stoats) or rapidly using special skin cells (like octopuses).

Table 1: Original English text from the ASSET corpus with multiple French simplifications generated by Gemini 2.5 Flash-Lite using 3 different prompting techniques. Abbreviations: CoT = Chain-of-Thought, T>S = Translate then Simplify, S>T = Simplify then Translate, Pipe = Pipeline.

tasks without task-specific fine-tuning through appropriate prompting strategies. Building on this foundation, Wei et al. (2023) introduced chain-of-thought prompting, demonstrating that encouraging models to generate intermediate reasoning steps dramatically improves performance on complex reasoning tasks, and establishing the critical importance of prompt engineering strategies for LLM performance. In the context of cross-lingual tasks, Wang et al. (2023) explored zero-shot cross-lingual summarization, proposing effective prompting strategies that we adapt in our work by replacing summarization instructions with simplification objectives. Our work builds on these insights by systematically investigating how different prompting approaches affect CLTS performance across multiple models and domains.

2.3 Evaluation of Text Simplification

Evaluating text simplification quality presents unique challenges that differ from traditional NLP tasks, as simplification involves complex trade-offs between meaning preservation, readability improvement, and linguistic adequacy (Alva-Manchego et al., 2021). This has led to the development of specialized automatic metrics and feature-based approaches tailored specifically for assessing simplification effectiveness.

2.3.1 Automatic Evaluation of Simplification

The evaluation of text simplification has traditionally relied on machine translation metrics such as BLEU (Papineni et al., 2002). SARI (Xu et al., 2016) was introduced to explicitly reward additions, deletions, and retentions, and is now widely used for simplification evaluation. Later studies by Sulem et al. (2018) confirmed that BLEU penalizes legitimate simplification operations and may negatively correlate with simplicity, validating the need for specialized metrics like SARI, though BLEU remains valuable for assessing fluency. More recently, semantic similarity metrics such as BERTScore (Zhang et al., 2020) have been adopted to better capture meaning preservation in text generation tasks. Surveys highlight that no single metric fully captures simplification quality and recommend combining multiple measures (Grabar and Saggion, 2022; Alva-Manchego et al., 2021).

2.3.2 Feature-Based Evaluation of Simplicity

Several studies have emphasized the role of linguistic features in assessing text simplicity. Vajjala and Meurers (2016) showed that lexical and syntactic features are effective predictors of sentence readability and can distinguish simplified from complex sentences. Recent work has expanded this direction by combining transformer-based models with diverse linguistic features, to improve the explainability of simplification systems (Qiao et al., 2022). Kreutz et al. (2024) introduced BATS (BenchmArking Text Simplicity), a comprehensive framework that implements 37 literature-derived features across multiple dimensions. We adopt the majority of BATS features in our evaluation framework, applying them to assess cross-lingual simplification, a novel extension beyond the original monolingual focus of BATS. These findings support our approach of integrating linguistic features to complement automatic metrics in evaluating CLTS, providing a strong foundation for understanding how different prompting strategies affect specific textual properties.

2.3.3 Human Evaluation of Simplification

Despite the efficiency of linguistic features, human evaluation remains the gold standard for assessing text simplification, as it directly measures the actual impact on accessibility and communicative success (Alva-Manchego et al., 2021). Traditional human evaluation frameworks focus on three primary dimensions: *meaning preservation* (adequacy), *grammaticality* (fluency), and *simplicity* (Štajner et al., 2014). However, obtaining reliable human judgments in simplification is challenging due to the inherent subjectivity of "simplicity" which can vary significantly depending on the target audience's profile (Yaneva et al., 2016). In the cross-lingual context, human judgment is particularly vital to account for the interplay between translation accuracy and readability, which automatic metrics may overlook. We therefore incorporate human evaluation to validate whether improvements in automatic scores and linguistic features translate to genuine gains in accessibility and fidelity for bilingual audiences.

3 Prompting Strategies in CLTS

CLTS can be approached through various operational strategies that differ in three fundamental dimensions: (1) whether to perform CLTS directly or sequentially, (2) if sequential, which operation to perform first, and (3) whether to execute operations within a single prompt or across separate prompts. This section examines these strategic choices and their implications for CLTS system design.

CLTS involves two distinct transformations: converting text from one language to another (translation) and reducing linguistic complexity while preserving meaning (simplification). These operations can be combined in different ways:

Direct Approach: Performing both translation and simplification together in a single prompt, instructing the model to produce simplified text directly in the target language without explicit intermediate steps.

Sequential Approaches: Decomposing the task into two explicit operations performed in sequence, which can follow two different orderings:

Translate-then-Simplify (T→S): First converting text to the target language, then simplifying it.

Simplify-then-Translate (S→T): First simplifying in the source language, then translating the simplified version.

Implementation Methods: Sequential approaches can be implemented in two distinct ways:

Chain-of-Thought (CoT): Both operations specified within a single prompt, with explicit instructions to perform them sequentially.

Pipeline: Operations performed through separate consecutive prompts, with the output of the first serving as input to the second.

This yields five distinct prompting strategies, all explored in this work:

- Direct prompting
- CoT T→S (sequential, single prompt)
- CoT S→T (sequential, single prompt)
- Pipeline T→S (sequential, separate prompts)
- Pipeline S→T (sequential, separate prompts)

Each strategic choice involves distinct **trade-offs**: (i) Parallel Processing, where the model, without guidance, can optimize both objectives simultaneously, vs. Sequential Processing, where the task decomposition is clear but can propagate errors; (ii) Simplifying First, which can facilitate translation but removes information early in the process, vs. Translating First, which preserves information but can make simplification harder in the target language; (iii) Single Prompt, which allows better integration but could increase cognitive load vs. Multiple Prompts, which can address each task separately but do not optimize coordination.

Feature Category	Features
Lexical Features	lexical richness; infrequent words ratio; long words ratio; content words ratio; average word length
Syntactic and Structural Features	words before main verb; noun phrases ratio; relative clauses ratio; appositions ratio; conditional clauses ratio; conjunctions ratio; passive voice ratio; syntactic tree depth; sentences number; words per sentence; short sentences ratio
Readability Features	Flesch Reading Ease; Flesch Kincaid Grade; syllables ratio
Named Entity Features	max same entity distances; unique entities; unique entities average; avg same entity distance; entity to token ratio; unique entities to total num of entities; consecutive entity distance
Grammatical Features	modifiers ratio; negations ratio; past perfect verbs; past tense verbs; punctuation ratio; third person pronouns ratio

Table 2: Comprehensive linguistic features used for CLTS analysis.

Our systematic evaluation addresses key questions about these strategic choices: (1) Which prompting strategy yields the best performance across different evaluation metrics (BLEU, SARI, semantic similarity)? (2) How do different strategies differ in the specific linguistic transformations they perform? (3) Does human evaluation confirm the trade-off between meaning preservation and simplicity across different prompting strategies?

4 Experimental Setup

This section describes the datasets, models, and preprocessing procedures in our CLTS experiments.

4.1 Datasets

We evaluate our approaches on five corpora that represent different domains and language pairs (Details about corpus sizes are in Appendix A):

ASSET: An English Wikipedia dataset with multiple human simplifications crowdsourced for each sentence (Alva-Manchego et al., 2020).

WIKI-AUTO: A large-scale English dataset automatically aligned from English Wikipedia and Simple English Wikipedia (Jiang et al., 2020).

MultiCochrane: The first multilingual, sentence-aligned medical text simplification dataset, which includes complex and simplified texts in English, Spanish, French, and Farsi (Joseph et al., 2023).

CLEAR: A French medical corpus of comparable texts from sources like encyclopedias and drug leaflets, with a subset of manually aligned parallel sentences (Grabar and Cardon, 2018).

WikiLarge-Fr: A French version of WikiLarge (Zhang and Lapata, 2017), translated into French using Google Translate (Ormaechea and Tsourakis, 2024).

4.2 Data Preprocessing

Current text simplification datasets are primarily designed for monolingual settings and lack the parallel structures necessary for evaluating cross-lingual simplification across both source and target languages. To address this limitation, we used machine translation to create the required cross-lingual evaluation data for monolingual corpora. Specifically, for each original-simplified text pair in a source language, we translated both texts to the target language using machine translation, creating the original text and reference simplified texts in the target language. The `MultiCochrane` dataset, which already provides cross-lingual parallel data, did not require this translation step. To ensure data quality and semantic consistency across all corpora, we applied a filtering process. We used Sentence-BERT (Reimers and Gurevych, 2019) to compute semantic similarity scores between original texts and their corresponding simplified versions. We filtered out all sentence pairs with semantic similarity scores below 0.6, ensuring that simplified versions maintain sufficient semantic overlap with their original counterparts while allowing for meaningful simplification transformations.

4.3 Models

We evaluate our strategies using seven state-of-the-art LLMs, encompassing both proprietary and open-source architectures. Proprietary models include **GPT-3.5-Turbo**, **GPT-4o-Mini**, and **Gemini 2.5 Flash-Lite**. We also use open-source models including **DeepSeek-Chat**, **Aya101** (Üstün et al., 2024), **Mistral-NeMo**, and a fine-tuned **mT5** baseline (Xue et al., 2021) to ensure the reproducibility of our findings and promote transparent benchmarking in the research community. Detailed documentation and resource links for each model are provided in Appendix B.

Corpus	Method	GPT-3.5-turbo			GPT-4o-mini			Gemini 2.5 Flash-Lite			DeepSeek-Chat		
		BLEU	SARI	Cam.	BLEU	SARI	Cam.	BLEU	SARI	Cam.	BLEU	SARI	Cam.
ASSET	Direct	**66.091**	39.184	**0.839**	**70.140**	39.173	**0.837**	**68.127**	39.712	**0.826**	65.181	40.491	0.825
	CoT T>S	57.105	42.558	0.805	58.775	43.911	0.799	47.134	41.445	0.763	63.789	41.531	0.827
	CoT S>T	59.474	41.748	0.795	60.496	44.282	0.805	48.923	40.505	0.744	67.152	40.417	0.824
	Pipe T>S	51.888	42.855	0.815	60.017	43.464	0.814	52.500	42.021	0.767	55.914	43.183	0.801
	Pipe S>T	48.027	43.675	0.807	60.384	44.191	0.814	54.242	42.097	0.765	52.280	42.423	0.775
WikiAuto	Direct	**30.966**	32.237	**0.793**	**32.406**	33.281	**0.795**	**29.950**	33.240	**0.785**	29.356	34.152	0.788
	CoT T>S	21.887	33.361	0.756	21.366	34.439	0.756	16.635	32.866	0.728	28.952	34.768	0.786
	CoT S>T	23.786	33.477	0.760	24.317	35.000	0.768	17.257	32.740	0.712	29.691	33.991	0.787
	Pipe T>S	22.024	33.703	0.768	25.634	35.264	0.777	18.879	33.654	0.735	22.399	34.641	0.765
	Pipe S>T	20.271	**34.252**	0.761	25.465	35.412	0.774	19.751	33.835	0.735	20.012	34.074	0.747
MultiCoch.	Direct	10.778	32.259	0.590	**13.462**	33.166	0.664	**12.931**	34.780	**0.663**	11.958	36.774	0.663
	CoT T>S	10.047	36.425	0.637	9.544	38.680	0.635	7.191	38.866	0.608	11.288	36.602	0.660
	CoT S>T	12.732	34.558	0.652	10.096	38.985	0.648	8.262	39.624	0.614	12.136	36.507	0.663
	Pipe T>S	9.883	37.929	0.650	10.451	37.794	0.653	7.918	39.049	0.623	9.430	39.766	0.644
	Pipe S>T	9.564	38.928	0.652	10.627	39.205	0.657	8.891	40.276	0.632	8.680	40.270	0.639

Table 3: Automatic Evaluation Metrics for Text Simplification Systems (English Corpora). **Bold** indicates that the best score is significantly higher than all others for the same corpus, model, and metric. **Abbreviations:** CoT = Chain-of-Thought, T>S = Translate then Simplify, S>T = Simplify then Translate, Pipe = Pipeline, Cam. = CamembertScore, MultiCoch. = MultiCochrane.

5 Methodology

This section details our specific approaches for CLTS, including prompting strategies, evaluation metrics, and analysis methods.

5.1 Prompting Strategies

We investigate five distinct prompting strategies (prompts shown for French as target language; the same structure applies when English is the target):

1. Direct Prompting: A single-step approach where the model is instructed to perform both translation and simplification simultaneously: *"Please simplify the following text in French: "*.

2. Chain-of-Thought Translate-then-Simplify: A single prompt that explicitly guides the model through a two-step reasoning process, first translating and then simplifying: *"Please first translate the following text to French and then simplify the translated text in French: "*.

3. Chain-of-Thought Simplify-then-Translate: A single prompt that reverses the order, first simplifying in the source language and then translating: *"Please first simplify the following text and then translate the simplification to French: "*.

4. Pipeline Translate-then-Simplify: A two-step pipeline where translation and simplification are performed in separate consecutive prompts:
Step 1: *"Please translate the following text to French: "*.
Step 2: *"Please simplify the following text in French: "*.

5. Pipeline Simplify-then-Translate: A two-step pipeline with reversed order:
Step 1: *"Please simplify the following text: "*.
Step 2: *"Please translate the following text to French: "*.

5.2 Evaluation Metrics

We employ a comprehensive evaluation framework using multiple automatic metrics to assess different aspects of cross-lingual simplification quality:

BLEU (Papineni et al., 2002): Measures n-gram overlap between generated and reference texts, providing an indication of translation quality.

SARI (Xu et al., 2016): Evaluates simplification quality by comparing n-gram additions, deletions, and retentions against references and the source

Semantic Similarity Metrics: To capture meaning preservation across languages, we use language-specific semantic similarity scores: **BERTScore** (Zhang et al., 2020), for texts simplified from French to English, and **CamemBERTScore** for texts simplified from English to French, using CamemBERT (Martin et al., 2020) instead of BERT (Devlin et al., 2019) embeddings.

We conduct separate statistical analyses (t-test, $p<0.05$) for each combination of corpus, model,

		GPT-3.5-turbo			GPT-4o-mini			Gemini 2.5 Flash-Lite			DeepSeek-Chat		
Corpus	**Method**	**BLEU**	**SARI**	**BERT**	**BLEU**	**SARI**	**BERT**	**BLEU**	**SARI**	**BERT**	**BLEU**	**SARI**	**BERT**
WikiLargeFR	Direct	**34.516**	36.250	**0.648**	33.856	36.479	0.612	**33.486**	34.658	**0.642**	**34.655**	36.156	0.651
	CoT T>S	23.030	36.074	0.584	19.801	35.511	0.534	19.157	35.087	0.556	31.415	36.804	0.643
	CoT S>T	31.978	35.459	0.624	32.749	36.854	0.605	25.722	35.032	0.589	32.494	36.032	0.651
	Pipe T>S	19.771	35.676	0.566	23.832	37.248	0.569	17.943	34.381	0.538	19.146	35.159	0.559
	Pipe S>T	24.280	36.295	0.590	26.784	37.777	0.578	29.635	34.725	0.612	29.167	35.955	0.617
Clear	Direct	26.776	35.989	**0.634**	27.993	35.737	0.651	**27.250**	34.363	**0.639**	27.932	35.237	0.641
	CoT T>S	13.873	33.002	0.536	12.580	32.527	0.536	11.364	31.179	0.509	24.826	35.428	0.630
	CoT S>T	25.688	34.374	0.609	26.481	35.799	0.632	19.969	33.738	0.571	26.889	35.623	0.634
	Pipe T>S	11.338	31.829	0.506	15.541	34.330	0.569	10.507	30.402	0.479	11.392	31.585	0.502
	Pipe S>T	19.110	35.395	0.574	21.845	36.632	0.612	23.257	34.119	0.608	23.808	35.623	0.612

Table 4: Automatic Evaluation Metrics for Text Simplification Systems (English to French). **Bold** indicates scores significantly higher than other configurations. **Abbreviations:** CoT = Chain-of-Thought, T>S = Translate then Simplify, S>T = Simplify then Translate, Pipe = Pipeline, BERT = BERTScore.

and evaluation metric to identify statistically significant differences between prompting strategies.

5.3 Feature-based Analysis

To gain deeper insights into the characteristics of successful CLTS, we extract and analyze a comprehensive set of linguistic features from the outputs of each prompting strategy. These features were designed to capture various linguistic and structural properties, which are detailed in Table 2. We apply t-tests to compare these linguistic features between the outputs of different prompting strategies for each model and corpus combination. This analysis helps identify which linguistic transformations are most characteristic of effective cross-lingual simplification and how these patterns vary across different experimental conditions.

5.4 Human Evaluation

To validate our automatic evaluation, we conducted a human annotation study on the first 70 texts from each corpus using GPT-3.5-turbo outputs. For all corpora, in-house human annotators, highly proficient in both English and French, compared these outputs against the original source texts. Additionally, for the WikiAuto and CLEAR corpora, we extended the study to include a comparison against the translated versions of the texts. Most comparisons were reviewed by two annotators, with several instances receiving three. Participants assessed meaning preservation using 0 to 5 intensity scales to measure both the scale of information addition and the scale of information removal. Simplicity was rated on a -2 to 2 scale for overall simplicity. The full guidelines appear in Appendix C. This framework allowed us to quantify the specific trade-offs between text simplification and semantic fidelity inherent in each cross-lingual strategy.

6 Results

6.1 Automatic Evaluation Metrics Results

Performance of the different prompting approaches across datasets is provided in Table 3 (English-to-French) and Table 4 (French-to-English). As `Mistral-NeMo` and `Aya101` models generally yielded low performance and are therefore less relevant to the strategy investigation, we present their results in Appendix D together with the **mT5** results (Appendix E).

BLEU Score Performance: Direct prompting consistently achieves the highest BLEU scores across both translation directions and all closed-source models, significantly outperforming other approaches on general domain corpora. Notably, translate-first approaches consistently show the poorest BLEU performance in the French-to-English direction, suggesting significant operational ordering challenges for this language pair.

SARI Score Performance: Multi-step approaches dominate SARI scores across both directions and most models, with Pipeline Simplify-then-Translate emerging as the consistent leader. However, in almost all cases, the differences are not statistically significant.

Semantic Similarity Performance: Analyses across both translation directions reveal distinct performance patterns and a notable directional asymmetry. For English-to-French tasks, direct prompting achieves significantly superior CamemBERT scores on general domain datasets, suggesting strong semantic preservation for encyclopedic

Corpus	Strategy	Simplicity	Add.	Rem.
Asset	Direct	-0.029	0.557	0.579
	CoT $T \rightarrow S$	0.036	0.586	1.029
	CoT $S \rightarrow T$	0.086	0.579	1.050
	Pipe $T \rightarrow S$	0.071	0.736	0.771
	Pipe $S \rightarrow T$	0.143	0.757	0.871
MultiCochrane	Direct	0.064	0.514	0.707
	CoT $T \rightarrow S$	0.164	0.493	0.636
	CoT $S \rightarrow T$	0.186	0.507	0.650
	Pipe $T \rightarrow S$	0.507	0.521	1.250
	Pipe $S \rightarrow T$	0.643	0.514	1.064
WikiLarge FR	Direct	0.150	0.521	0.529
	CoT $T \rightarrow S$	0.364	0.536	0.750
	CoT $S \rightarrow T$	0.350	0.521	0.693
	Pipe $T \rightarrow S$	0.464	0.521	0.786
	Pipe $S \rightarrow T$	0.436	0.536	0.786
WikiAuto (Src)	Direct	-0.057	0.521	0.521
	CoT $T \rightarrow S$	0.257	0.550	0.721
	CoT $S \rightarrow T$	0.336	0.521	0.807
	Pipe $T \rightarrow S$	**0.814**	**1.000**	**1.300**
	Pipe $S \rightarrow T$	0.507	0.600	0.907
WikiAuto (Trn)	Direct	0.550	0.457	0.433
	CoT $T \rightarrow S$	0.657	0.438	0.767
	CoT $S \rightarrow T$	0.850	0.419	0.867
	Pipe $T \rightarrow S$	0.800	**0.700**	0.621
	Pipe $S \rightarrow T$	0.752	0.519	0.505
CLEAR (Src)	Direct	0.314	0.536	0.586
	CoT $T \rightarrow S$	0.429	0.650	0.986
	CoT $S \rightarrow T$	0.343	0.529	0.664
	Pipe $T \rightarrow S$	0.471	**0.843**	**1.324**
	Pipe $S \rightarrow T$	0.436	0.586	0.907
CLEAR (Trn)	Direct	0.457	0.064	0.121
	CoT $T \rightarrow S$	0.671	0.136	0.371
	CoT $S \rightarrow T$	0.450	0.086	0.164
	Pipe $T \rightarrow S$	0.829	0.279	0.436
	Pipe $S \rightarrow T$	0.621	0.143	0.236

Table 5: Human Evaluation Averages. Simp.: Simplicity; Add.: Info. Addition Scale; Rem.: Info. Removal Scale; Src: Source; Trn: Translation. **Bold:** significant.

content. Conversely, in the French-to-English direction, while direct prompting remains superior, translate-first approaches yield the lowest BERT scores, indicating substantial semantic drift when French text is translated prior to simplification.

The **mT5** baseline model achieved substantially lower BLEU and semantic scores across all datasets compared to the prompted LLMs, though it showed competitive SARI performance on some corpora.

6.2 Features-Based Analysis Results

Our linguistic analysis, grounded in the feature set detailed in Table 2, revealed distinct patterns in how each prompting strategy achieves simplification, differentiating their impact across lexical, syntactic, and structural dimensions. Results of selected features appear in Appendix F.

The **Direct Prompting** approach, consistently demonstrated the least significant changes in Syntactic and Structural Features (e.g., minimal reduction in syntactic tree depth and words per sentence). This approach produced outputs with higher complexity metrics, including elevated Flesch-Kincaid grade (Kincaid et al., 1975), and lower Flesch Reading Ease (Flesch, 1948), indicating higher complexity, and increased syntactic tree depth. These results indicate that direct prompting tends to generate more complex and formal language structures, implying a prioritization of fluency and translation fidelity over deep structural transformations.

In contrast, the **CoT** methods excelled at targeted complexity reduction by leveraging explicit intermediate steps. Specifically, **CoT translate→simplify** generally resulted in simplified outputs, with higher Flesch Reading Ease scores (indicating better readability), reduced syntactic complexity, more sentences, higher short sentence ratios, and shorter sentence lengths. The **pipeline** approaches showed intermediate characteristics, with **pipeline simplify→translate** producing more complex outputs than **pipeline translate→simplify** (higher syntactic tree depth, words per sentence and Flesch-Kincaid grade). This indicates that the order of operations in pipeline methods significantly influences the final output characteristics.

6.3 Human Evaluation Results

As shown in Table 5, Human annotations reveal a clear trade-off: Pipeline approaches received the highest simplicity ratings (0.071 to 0.829), while Direct prompting was rated lowest (-0.057 to 0.550). This increased simplicity in pipeline outputs is accompanied by high information addition and removal rates. In contrast, Direct prompting showed the lowest scales of information modification, indicating that it remains the most faithful to the original source text.

7 Conclusion

Our comprehensive evaluation of CLTS in English and French reveals that no single prompting strategy universally outperforms others across all evaluation dimensions, rather the optimal approach depends critically on the specific objectives and constraints of the simplification task. For applications prioritizing translation fidelity and lexical accuracy, Direct Prompting emerges as the most effective strategy, consistently achieving the highest BLEU scores across both language directions and multiple corpora. However, when accessibility and readability are paramount, translate-then-simplify approaches demonstrate clear advantages, achieving the most effective syntactic simplification while maintaining moderate lexical richness.

Limitations

While this study provides comprehensive insights into cross-lingual text simplification strategies, several limitations highlight directions for future work. First, our evaluation is restricted to the English-French language pair. Future work should explore the generalizability of our findings to other language combinations, particularly those with greater typological differences or lower-resource contexts. Second, some nuanced aspects of simplification quality such as cultural appropriateness, or actual reader comprehension are not captured by the automatic evaluation metrics (BLEU, SARI, BERTScore, and CamemBERTScore) and feature-based analysis used in the paper. Finally, while we evaluate multiple state-of-the-art LLMs, the rapid evolution of these models means our findings represent a snapshot of current capabilities. Emerging models or fine-tuning approaches may yield different patterns and will thus require novel evaluation and analysis, for which we provide tools in this paper.

References

Fernando Alva-Manchego, Louis Martin, Antoine Bordes, Carolina Scarton, Benoît Sagot, and Lucia Specia. 2020. ASSET: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4668–4679, Online. Association for Computational Linguistics.

Fernando Alva-Manchego, Louis Martin, Carolina Scarton, and Lucia Specia. 2019. EASSE: Easier automatic sentence simplification evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 49–54, Hong Kong, China. Association for Computational Linguistics.

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2021. The (un)suitability of automatic evaluation metrics for text simplification. *Computational Linguistics*, 47(4):861–889.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Natalia Grabar and Rémi Cardon. 2018. CLEAR – simple corpus for medical French. In *Proceedings of the 1st Workshop on Automatic Text Adaptation (ATA)*, pages 3–9, Tilburg, the Netherlands. Association for Computational Linguistics.

Natalia Grabar and Horacio Saggion. 2022. Evaluation of automatic text simplification: Where are we now, where should we go from here. In *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, pages 453–463, Avignon, France. ATALA.

Matthew Honnibal and Ines Montani. 2020. spacy: Industrial-strength natural language processing in python. *To appear*.

Chao Jiang, Mounica Maddela, Wuwei Lan, Yang Zhong, and Wei Xu. 2020. Neural CRF model for sentence alignment in text simplification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7943–7960, Online. Association for Computational Linguistics.

Sebastian Joseph, Kathryn Kazanas, Keziah Reina, Vishnesh Ramanathan, Wei Xu, Byron Wallace, and Junyi Jessy Li. 2023. Multilingual simplification of medical texts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16662–16692, Singapore. Association for Computational Linguistics.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Defense Technical Information Center (DTIC) Document.

Christin Kreutz, Fabian Haak, Björn Engelmann, and Philipp Schaer. 2024. BATS: BenchmArking text simplicity. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11968–11989, Bangkok, Thailand. Association for Computational Linguistics.

Christoph Leiter, Piyawat Lertvittayakumjorn, Marina Fomicheva, Wei Zhao, Yang Gao, and Steffen Eger. 2024. Towards explainable evaluation metrics for machine translation. *J. Mach. Learn. Res.*, 25:75:1–75:49.

Edward Loper and Steven Bird. 2002. Nltk: the natural language toolkit. *CoRR*, cs.CL/0205028.

Jonathan Mallinson, Rico Sennrich, and Mirella Lapata. 2020. Zero-shot crosslingual sentence simplification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5109–5126, Online. Association for Computational Linguistics.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. Camembert: a tasty french language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Matti Miestamo. 2008. Grammatical complexity in cross-linguistic perspective. *Language Complexity: Typology, Contact, Change*, pages 23–42.

Lucía Ormaechea and Nikos Tsourakis. 2024. Automatic text simplification for french: model fine-tuning for simplicity assessment and simpler text generation. *International Journal of Speech Technology*, 27:957–976.

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.

Yu Qiao, Xiaofei Li, Daniel Wiechmann, and Elma Kerz. 2022. (Psycho-)linguistic features meet transformer models for improved explainable and controllable text simplification. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 125–146, Abu Dhabi, United Arab Emirates (Virtual). Association for Computational Linguistics.

Thilina C. Rajapakse, Andrew Yates, and Maarten de Rijke. 2024. Simple transformers: Open-source for all. In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, SIGIR-AP 2024, page 209–215, New York, NY, USA. Association for Computing Machinery.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Advaith Siddharthan. 2014. A survey of research on text simplification. *ITL - International Journal of Applied Linguistics*, 165:259–298.

Sanja Štajner, Ruslan Mitkov, and Horacio Saggion. 2014. One step closer to automatic evaluation of text simplification systems. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, pages 1–10, Gothenburg, Sweden. Association for Computational Linguistics.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018. BLEU is not suitable for the evaluation of text simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 738–744, Brussels, Belgium. Association for Computational Linguistics.

Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15894–15939, Bangkok, Thailand. Association for Computational Linguistics.

Sowmya Vajjala and Detmar Meurers. 2016. Readability-based sentence ranking for evaluating text simplification. *Preprint*, arXiv:1603.06009.

Pauli Virtanen, Ralf Gommers, Travis Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan Walt, Matthew Brett, Joshua Wilson, K. Millman, Nikolay Mayorov, Andrew Nelson, Eric Jones, Robert Kern, Eric Larson, and Yoshiki Vázquez-Baeza. 2020. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature Methods*, 17:1–12.

Vered Volansky, Noam Ordan, and Shuly Wintner. 2014. On the features of translationese. *Digital Scholarship in the Humanities*, 30.

Jiaan Wang, Yunlong Liang, Fandong Meng, Beiqi Zou, Zhixu Li, Jianfeng Qu, and Jie Zhou. 2023. Zero-shot cross-lingual summarization via large language models. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 12–23, Singapore. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Wes McKinney. 2010. Data Structures for Statistical Computing in Python. In *Proceedings of the 9th Python in Science Conference*, pages 56 – 61.

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.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. *Preprint*, arXiv:2010.11934.

Victoria Yaneva, Irina Temnikova, and Ruslan Mitkov. 2016. Evaluating the readability of text simplification output for readers with cognitive disabilities. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 293–299, Portorož, Slovenia. European Language Resources Association (ELRA).

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *Proceedings of ICLR*.

Xingxing Zhang and Mirella Lapata. 2017. Sentence simplification with deep reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 584–594, Copenhagen, Denmark. Association for Computational Linguistics.

A Corpus Statistics

	Asset	WikiAuto	MultiCochrane	Clear	WikiLarge FR
Train set size	20,000	576,126	28,998	4,110	264,258
Test set size	359	4,690	264	389	1,067

Table 6: Corpus statistics showing dataset sizes for training and evaluation

B Model Documentation and Accessibility

To ensure reproducibility and provide full transparency regarding the specific versions and technical documentation of the models used, we provide the official resource links for each LLM evaluated:

- GPT-3.5-turbo: https://platform.openai.com/docs/models/gpt-3.5-turbo
- GPT-4o-mini: https://platform.openai.com/docs/models/gpt-4o-mini
- Gemini 2.5 Flash-lite: https://deepmind.google/models/gemini/flash-lite/
- Deepseek-chat: https://api-docs.deepseek.com/news/news251201
- Mistral-NeMo: https://mistral.ai/news/mistral-nemo
- Aya101: https://cohere.com/research/aya

C Human Annotation Guidelines

Annotators were presented with a side-by-side view of the original text or its translated version (Version A) and the simplified LLM response (Version B). They were provided with the following prompt: *"Given Versions A and B, please answer the following questions:"*

- **Simplicity:** *"Is B simpler than A?"*
  (Scale: -2 to 2, where -2 is much more complex and 2 is much simpler).
- **Information Addition:** *"Does Version B add information compared to Version A?"*
  (Scale: 0 to 5, where 0 is no addition and 5 is significant addition).
- **Information Removal:** *"Does Version B remove information compared to Version A?"*
  (Scale: 0 to 5, where 0 is no removal and 5 is total loss of original meaning).

D Open Source Models Results

Corpus	Method	Aya101			Mistral-NeMo		
		BLEU	**SARI**	**Sem.**	**BLEU**	**SARI**	**Sem.**
ASSET	Direct	54.122	39.664	0.790	56.403	38.583	0.779
	CoT T>S	56.220	38.466	0.796	65.880	38.943	0.824
	CoT S>T	56.365	38.829	0.800	64.821	37.911	0.828
	Pipe T>S	55.914	39.069	0.787	54.213	40.693	0.768
	Pipe S>T	54.200	38.456	0.761	46.919	39.556	0.734
WikiAuto	Direct	25.099	32.884	0.755	23.626	31.502	0.740
	CoT T>S	26.910	32.419	0.763	28.735	32.083	0.776
	CoT S>T	26.611	32.555	0.763	30.577	31.926	0.781
	Pipe T>S	25.651	32.502	0.755	21.068	32.814	0.733
	Pipe S>T	24.285	33.290	0.732	16.224	31.928	0.697
MultiCochrane	Direct	9.575	34.560	0.631	10.444	34.202	0.618
	CoT T>S	10.132	33.023	0.625	12.864	32.815	0.659
	CoT S>T	9.880	33.241	0.629	12.933	31.318	0.654
	Pipe T>S	9.662	33.306	0.629	9.663	37.238	0.620
	Pipe S>T	9.475	33.836	0.620	6.112	37.128	0.559
WikiLarge FR	Direct	31.978	34.326	0.626	28.092	34.691	0.614
	CoT T>S	33.916	33.736	0.639	31.621	35.616	0.634
	CoT S>T	33.203	33.859	0.639	35.374	34.641	0.652
	Pipe T>S	32.701	33.841	0.626	16.907	32.979	0.522
	Pipe S>T	29.479	34.727	0.601	26.923	34.794	0.598
Clear	Direct	25.570	35.159	0.593	20.941	34.389	0.596
	CoT T>S	26.272	33.815	0.600	23.338	35.003	0.609
	CoT S>T	26.439	34.838	0.607	28.221	34.149	0.638
	Pipe T>S	26.476	34.797	0.600	10.176	30.247	0.461
	Pipe S>T	23.152	33.915	0.549	20.295	34.284	0.578

Table 7: Automatic Evaluation Metrics for additional models Aya101 and Mistral-NeMo across all corpora.
Abbreviations: Sem.: semantic similarity metrics(BERTScore or CamemBertScore)

E mT5 Results

Metric	Asset	WikiAuto	MultiCochrane	Clear	WikiLarge FR
Bleu	9.678	11.213	2.529	3.214	8.700
Sari	28.817	37.525	35.436	24.708	29.658
Semantic Score	0.331	0.496	0.269	-0.004	0.132

Table 8: mT5 model evaluation results across different corpora and metrics. Semantic Score uses CamembertScore for English→French simplification and BertScore for French→English simplification.

F Linguistic Features Results

Model	Prompt	Lexical Richness	Tree Depth	Sent. #	Words/Sent.	Short S. Ratio	FK Grade	FR Ease
GPT-3.5	direct	0.903	4.621	1.036	24.549	0.046	10.13	60.60
	CoT t→s	0.924	4.217	1.047	20.337	0.121	8.420	66.75
	CoT s→t	0.923	4.309	1.022	20.382	0.097	8.572	65.86
	pipe t→s	0.910	4.507	1.142	21.943	0.078	8.830	65.99
	pipe s→t	0.905	4.663	1.237	21.320	0.063	8.659	66.81
GPT-4o mini	direct	0.908	4.643	1.019	23.500	0.054	9.508	63.08
	CoT t→s	0.924	4.014	1.031	18.595	0.117	7.272	71.92
	CoT s→t	0.918	4.284	1.019	20.766	0.102	8.403	67.45
	pipe t→s	0.917	4.348	1.014	20.692	0.082	8.233	68.14
	pipe s→t	0.913	4.454	1.006	21.682	0.070	8.813	65.87
Gemini Lite	direct	0.917	4.368	1.028	22.017	0.057	9.014	64.26
	CoT t→s	0.930	3.816	1.178	16.794	0.156	6.585	73.74
	CoT s→t	0.937	3.813	1.064	16.450	0.205	6.919	71.11
	pipe t→s	0.935	3.730	1.047	16.967	0.200	6.769	72.69
	pipe s→t	0.931	3.897	1.011	17.521	0.201	7.667	67.24
DeepSeek	direct	0.924	4.329	1.008	21.831	0.065	9.192	62.96
	CoT t→s	0.893	4.284	1.237	19.958	0.092	8.329	66.34
	CoT s→t	0.922	4.312	1.011	21.706	0.056	9.272	62.30
	pipe t→s	0.932	4.100	1.014	19.458	0.102	8.015	67.81
	pipe s→t	0.927	4.067	1.003	18.760	0.155	8.165	66.21
Aya101	direct	0.903	4.591	1.014	23.944	0.067	9.642	63.03
	CoT t→s	0.906	4.599	1.003	24.260	0.064	9.836	62.19
	CoT s→t	0.901	4.635	1.014	24.050	0.058	9.732	62.48
	pipe t→s	0.903	4.588	1.019	23.675	0.070	9.647	62.73
	pipe s→t	0.913	4.331	1.047	21.088	0.066	8.695	65.47
Mistral-Nemo	direct	0.918	4.148	1.008	21.231	0.104	9.040	63.24
	CoT t→s	0.905	4.563	1.003	23.386	0.067	9.714	62.02
	CoT s→t	0.903	4.646	1.003	24.230	0.060	9.992	61.05
	pipe t→s	0.925	4.100	1.008	19.529	0.167	8.332	65.97
	pipe s→t	0.930	3.852	1.014	16.972	0.216	7.367	68.76

Table 9: Results of selected features for the **ASSET** Corpus

Model	Prompt	Lexical Richness	Tree Depth	Sent. #	Words/Sent.	Short S. Ratio	FK Grade	FR Ease
GPT-3.5	direct	0.896	4.680	1.103	25.155	0.061	9.872	63.26
	CoT t→s	0.918	4.245	1.093	20.963	0.120	8.391	67.61
	CoT s→t	0.914	4.258	1.072	20.816	0.107	8.267	68.71
	pipe t→s	0.899	4.450	1.270	21.642	0.088	8.449	68.21
	pipe s→t	0.892	4.565	1.391	20.638	0.093	8.234	68.43
GPT-4o mini	direct	0.904	4.477	1.038	25.030	0.059	9.506	64.80
	CoT t→s	0.928	3.886	1.084	18.895	0.130	6.981	73.90
	CoT s→t	0.912	4.251	1.071	21.744	0.088	8.244	69.47
	pipe t→s	0.920	4.163	1.047	21.739	0.084	8.109	69.87
	pipe s→t	0.912	4.358	1.041	22.664	0.081	8.737	67.27
Gemini Lite	direct	0.914	4.279	1.061	22.739	0.086	8.855	66.04
	CoT t→s	0.928	3.697	1.294	16.572	0.213	6.102	76.39
	CoT s→t	0.933	3.687	1.107	16.851	0.218	6.768	72.56
	pipe t→s	0.937	3.656	1.082	17.169	0.204	6.571	73.87
	pipe s→t	0.930	3.846	1.044	18.262	0.178	7.366	70.20
DeepSeek	direct	0.917	4.325	1.058	22.760	0.070	8.988	65.29
	CoT t→s	0.899	4.230	1.307	20.255	0.104	7.924	69.05
	CoT s→t	0.915	4.348	1.049	22.893	0.072	9.095	64.84
	pipe t→s	0.930	4.028	1.063	19.915	0.109	7.764	69.78
	pipe s→t	0.923	4.091	1.017	19.943	0.127	8.100	68.16
Aya101	direct	0.894	4.535	1.035	25.067	0.059	9.499	64.98
	CoT t→s	0.892	4.585	1.014	26.364	0.055	9.886	63.92
	CoT s→t	0.891	4.575	1.021	25.939	0.058	9.808	64.10
	pipe t→s	0.894	4.550	1.035	25.405	0.067	9.625	64.55
	pipe s→t	0.911	4.226	1.052	21.452	0.089	8.446	67.20
Mistral-Nemo	direct	0.913	4.231	1.024	22.310	0.107	8.936	65.16
	CoT t→s	0.899	4.540	1.018	25.031	0.074	9.720	63.86
	CoT s→t	0.893	4.638	1.016	26.272	0.060	10.06	63.04
	pipe t→s	0.920	4.030	1.040	20.086	0.167	7.922	69.35
	pipe s→t	0.928	3.726	1.029	17.539	0.229	7.156	70.67

Table 10: Results of selected features for the **WikiAuto** Corpus.
Abbreviations: Tree Depth: Syntactic Tree Depth; Sent. #: Number of Sentences; Words/Sent.: Average Words per Sentence; Short S. Ratio: Ratio of Short Sentences; FK Grade: Flesch Kincaid Grade Level; FR Ease: Flesch Reading Ease.

Model	Prompt	Lexical Richness	Tree Depth	Sent. #	Words/Sent.	Short S. Ratio	FK Grade	FR Ease
GPT-3.5	direct	0.887	5.239	1.019	27.964	0.045	13.00	45.19
	CoT t→s	0.886	5.110	1.011	25.250	0.078	11.47	53.39
	CoT s→t	0.881	5.473	1.038	27.400	0.055	12.45	49.54
	pipe t→s	0.892	5.189	1.034	24.388	0.078	11.17	54.44
	pipe s→t	0.884	5.367	1.045	25.040	0.057	11.42	53.50
GPT-4o mini	direct	0.880	5.561	1.027	29.733	0.044	12.94	48.67
	CoT t→s	0.905	4.731	1.030	22.093	0.083	9.474	62.62
	CoT s→t	0.893	5.208	1.057	25.237	0.045	10.91	57.16
	pipe t→s	0.895	5.080	1.038	25.197	0.068	10.99	56.42
	pipe s→t	0.893	5.303	1.049	26.479	0.051	11.85	52.28
Gemini Lite	direct	0.884	5.394	1.053	28.131	0.064	12.67	48.31
	CoT t→s	0.910	4.735	1.182	21.048	0.119	8.742	65.71
	CoT s→t	0.906	4.655	1.091	20.994	0.124	9.639	59.64
	pipe t→s	0.918	4.515	1.102	20.350	0.128	9.089	61.82
	pipe s→t	0.911	4.723	1.061	21.063	0.117	10.06	56.59
DeepSeek	direct	0.898	5.390	1.027	26.896	0.055	12.46	48.25
	CoT t→s	0.883	5.375	1.159	25.347	0.068	11.77	50.87
	CoT s→t	0.895	5.318	1.015	26.890	0.053	12.56	47.75
	pipe t→s	0.913	4.894	1.038	22.053	0.081	10.27	56.70
	pipe s→t	0.912	4.833	1.011	21.763	0.089	10.58	54.82
Aya101	direct	0.855	5.674	1.015	30.741	0.042	11.94	52.44
	CoT t→s	0.858	5.705	1.008	31.131	0.047	12.15	51.52
	CoT s→t	0.859	5.534	1.011	30.748	0.042	12.06	51.87
	pipe t→s	0.858	5.557	1.008	30.852	0.038	12.15	51.32
	pipe s→t	0.863	5.379	1.064	27.949	0.057	11.35	52.97
Mistral-NeMo	direct	0.890	5.258	1.027	25.972	0.062	11.58	51.71
	CoT t→s	0.872	5.697	1.015	29.877	0.036	12.94	48.36
	CoT s→t	0.864	5.875	1.008	31.786	0.040	13.42	47.61
	pipe t→s	0.895	4.962	1.030	23.629	0.110	10.62	56.41
	pipe s→t	0.913	4.470	1.034	18.607	0.217	8.927	60.20

Table 11: Results of selected features for the **MultiCochrane** Corpus

Model	Prompt	Lexical Richness	Tree Depth	Sent. #	Words/Sent.	Short S. Ratio	FK Grade	FR Ease
GPT-3.5	direct	0.931	6.298	1.051	20.658	0.102	11.77	42.83
	CoT t→s	0.948	5.820	1.036	18.769	0.096	10.06	53.12
	CoT s→t	0.934	6.226	1.021	20.299	0.138	11.95	41.00
	pipe t→s	0.939	5.987	1.103	18.866	0.087	9.211	58.87
	pipe s→t	0.932	6.344	1.059	20.144	0.103	11.09	47.23
GPT-4o mini	direct	0.940	6.149	1.005	20.425	0.105	11.92	41.21
	CoT t→s	0.959	5.463	1.013	17.311	0.150	9.435	54.53
	CoT s→t	0.939	5.954	1.010	20.042	0.104	11.27	45.25
	pipe t→s	0.953	5.668	1.013	18.578	0.126	10.19	50.95
	pipe s→t	0.939	5.817	1.005	18.692	0.123	11.09	44.93
Gemini Lite	direct	0.945	5.763	1.010	19.347	0.134	11.68	40.79
	CoT t→s	0.959	5.195	1.080	16.524	0.216	8.796	57.53
	CoT s→t	0.957	5.134	1.033	16.350	0.215	10.68	43.54
	pipe t→s	0.969	4.730	1.018	15.013	0.294	8.640	56.08
	pipe s→t	0.955	5.483	1.013	17.675	0.188	11.46	39.76
DeepSeek	direct	0.946	5.910	1.003	19.721	0.121	12.09	38.81
	CoT t→s	0.946	5.794	1.026	19.144	0.116	11.44	42.72
	CoT s→t	0.949	5.805	1.000	19.350	0.131	12.00	38.85
	pipe t→s	0.971	4.823	1.000	15.095	0.272	9.530	50.32
	pipe s→t	0.955	5.625	1.000	18.177	0.159	11.73	38.99
Aya101	direct	0.914	6.033	1.031	20.144	0.128	11.60	42.92
	CoT t→s	0.916	6.290	1.021	21.205	0.099	12.12	41.16
	CoT s→t	0.920	6.211	1.036	20.641	0.115	12.04	40.75
	pipe t→s	0.915	6.180	1.013	20.707	0.109	11.93	41.51
	pipe s→t	0.896	5.882	1.185	18.240	0.157	10.60	47.54
Mistral-NeMo	direct	0.956	5.501	1.010	17.415	0.185	11.06	42.68
	CoT t→s	0.948	5.733	1.005	18.937	0.170	11.55	41.73
	CoT s→t	0.938	6.013	1.005	20.120	0.148	12.36	37.78
	pipe t→s	0.978	4.337	1.008	12.578	0.407	8.878	50.70
	pipe s→t	0.959	5.470	1.005	17.284	0.224	11.45	39.79

Table 12: Results of selected features for the **CLEAR** Corpus

Abbreviations: Tree Depth: Syntactic Tree Depth; Sent. #: Number of Sentences; Words/Sent.: Average Words per Sentence; Short S. Ratio: Ratio of Short Sentences; FK Grade: Flesch Kincaid Grade Level; FR Ease: Flesch Reading Ease.

Model	Prompt	Lexical Richness	Tree Depth	Sent. #	Words/Sent.	Short S. Ratio	FK Grade	FR Ease
GPT-3.5	direct	0.889	6.001	1.084	22.467	0.063	9.743	59.26
	CoT t→s	0.911	5.531	1.080	19.455	0.107	8.233	66.11
	CoT s→t	0.900	5.739	1.056	21.303	0.101	9.517	59.25
	pipe t→s	0.906	5.536	1.211	18.536	0.107	7.531	69.65
	pipe s→t	0.890	5.772	1.134	20.547	0.084	8.875	62.87
GPT-4o mini	direct	0.899	5.858	1.039	22.434	0.075	9.712	59.01
	CoT t→s	0.929	5.202	1.063	18.278	0.148	7.642	68.06
	CoT s→t	0.900	5.670	1.045	21.714	0.088	9.274	61.04
	pipe t→s	0.918	5.604	1.053	20.180	0.099	8.437	64.94
	pipe s→t	0.905	5.456	1.040	20.034	0.103	8.850	61.78
Gemini Lite	direct	0.913	5.405	1.067	20.482	0.107	9.254	58.87
	CoT t→s	0.922	4.899	1.225	16.634	0.191	7.097	69.10
	CoT s→t	0.927	4.916	1.100	17.408	0.191	8.013	63.53
	pipe t→s	0.940	4.482	1.069	15.886	0.250	7.039	68.13
	pipe s→t	0.923	5.155	1.054	18.791	0.138	8.710	60.24
DeepSeek	direct	0.911	5.704	1.036	21.523	0.082	9.653	57.86
	CoT t→s	0.910	5.591	1.103	20.468	0.090	9.034	60.90
	CoT s→t	0.912	5.643	1.029	21.418	0.091	9.619	57.93
	pipe t→s	0.938	4.915	1.031	17.223	0.177	7.976	63.88
	pipe s→t	0.921	5.421	1.019	20.150	0.115	9.286	58.57
Aya101	direct	0.880	5.961	1.026	23.169	0.069	9.960	58.67
	CoT t→s	0.878	6.016	1.020	23.821	0.067	10.19	57.88
	CoT s→t	0.880	5.918	1.022	23.506	0.069	10.07	58.23
	pipe t→s	0.883	5.943	1.036	22.828	0.078	9.862	58.75
	pipe s→t	0.886	5.843	1.078	21.178	0.095	9.349	60.24
Mistral-NeMo	direct	0.925	5.316	1.040	19.404	0.128	8.971	59.55
	CoT t→s	0.911	5.624	1.037	20.787	0.099	9.404	58.67
	CoT s→t	0.893	5.853	1.030	22.761	0.080	10.03	57.17
	pipe t→s	0.947	4.460	1.043	15.199	0.292	7.194	65.72
	pipe s→t	0.922	5.325	1.036	19.397	0.142	9.171	58.10

Table 13: Results of selected features for the **WikiLarge-FR** Corpus
Abbreviations: Tree Depth: Syntactic Tree Depth; Sent. #: Number of Sentences; Words/Sent.: Average Words per Sentence; Short S. Ratio: Ratio of Short Sentences; FK Grade: Flesch Kincaid Grade Level; FR Ease: Flesch Reading Ease.

G Used Python Packages

In our implementation, we utilize the following packages:

- deep-translator(https://github.com/nidhaloff/deep-translator)
- easse (Alva-Manchego et al., 2019)
- NLTK (Loper and Bird, 2002)
- Pandas (Wes McKinney, 2010)
- Pyphen (https://github.com/Kozea/Pyphen)
- SciPy (Virtanen et al., 2020)
- sentence-transformers (Reimers and Gurevych, 2020)
- simpletransformers (Rajapakse et al., 2024)
- SpaCy (Honnibal and Montani, 2020)
- textstat (https://github.com/textstat)