

TSAR 2025

**The Fourth Workshop on Text Simplification, Accessibility
and Readability**

Proceedings of the Workshop

November 4-9, 2025

©2025 Association for Computational Linguistics

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
317 Sidney Baker St. S
Suite 400 - 134
Kerrville, TX 78028
USA
Tel: +1-855-225-1962
acl@aclweb.org

ISBN 979-8-89176-176-6

Introduction

The organisers are pleased to present the proceedings of the 4th edition of the Workshop on Text Simplification, Accessibility and Readability (TSAR), hosted at The 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP 2025), in Suzhou, China.

The Text Simplification, Accessibility, and Readability (TSAR) workshop aims at bringing together researchers, developers and industries of assistive technologies, public organizations representatives, and other parties interested in the problem of making information more accessible to all citizens. We will discuss recent trends and developments in the area of automatic text simplification, automatic readability assessment, language resources and evaluation for text simplification, etc. This year the workshop was organised around two key tracks. The main track was of general interest to the audience and covered topics surrounding empirical research on text simplification, accessibility and readability. The second track was in relation to a shared-task. Participants were tasked with generating simplifications of texts that conform to a specified target readability level, balancing reduced linguistic complexity with meaning preservation and fluency. Participants were then encouraged to submit system description papers. All papers belonging to both tracks are listed below.

Main Track

- Template-Based Text-to-Image Alignment for Language Accessibility: A Study on Visualizing Text Simplifications
- Document-level Simplification and Illustration Generation: Multimodal Coherence
- Medical Text Simplification: From Jargon Detection to Jargon-Aware Prompting
- Readability Reconsidered: A Cross-Dataset Analysis of Reference-Free Metrics
- Evaluating Health Question Answering Under Readability-Controlled Style Perturbations
- A Multi-Agent Framework with Diagnostic Feedback for Iterative Plain Language Summary Generation from Cochrane Medical Abstracts
- Efficient On-Device Text Simplification for Firefox with Synthetic Data Fine-Tuning

TSAR-2025 Shared-Task(*)

- Findings of the TSAR 2025 Shared Task on Readability-Controlled Text Simplification
- OneNRC@TSAR2025 Shared Task: Small Models for Readability Controlled Text Simplification
- GRIPF at TSAR 2025 Shared Task: Towards controlled CEFR level simplification with the help of inter-model interactions
- ITU NLP at TSAR 2025 Shared Task: A Three-Stage Prompting Approach for CEFR-Oriented Text Simplification
- STARLING at TSAR 2025 Shared Task: Leveraging Alternative Generations for Readability Level Adjustment in Text Simplification
- taskGen at TSAR 2025 Shared Task: Exploring prompt strategies with linguistic knowledge
- EasyJon at TSAR 2025 Shared Task: Evaluation of Automated Text Simplification with LLM-as-a-Judge

- HULAT-UC3M at TSAR 2025 Shared Task: A Prompt-Based Approach using Lightweight Language Models for Readability-Controlled Text Simplification
- UoL-UPF at TSAR 2025 Shared Task: A Generate-and-Select Approach for Readability-Controlled Text Simplification
- Uniandes at TSAR 2025 Shared Task: Multi-Agent CEFR Text Simplification with Automated Quality Assessment and Iterative Refinement
- EhiMeNLP at TSAR 2025 Shared Task: Candidate Generation via Iterative Simplification and Reranking by Readability and Semantic Similarity
- OUNLP at TSAR 2025 Shared Task: Multi-Round Text Simplifier via Code Generation
- HIT-YOU at TSAR 2025 Shared Task: Leveraging Similarity-Based Few-Shot Prompting, Round-Trip Translation, and Self-Refinement for Readability-Controlled Text Simplification
- SQUIREL at TSAR 2025 Shared Task: CEFR-Controlled Text Simplification with Prompting and Reinforcement Fine-Tuning
- Archaeology at TSAR 2025 Shared Task: Teaching Small Models to do CEFR Simplifications
- HOPE at TSAR 2025 Shared Task: Balancing Control and Complexity in Readability-Controlled Text Simplification
- Know-AI at TSAR 2025 Shared Task: Difficulty-aware Text Simplification System

All submissions were peer-reviewed by the members of the program committee which includes distinguished specialists in text simplification, accessibility, and readability.

The workshop is held in-person, with online attendance for authors who were unable to attend due to constraints beyond the organisers control.

We would like to thank the members of the program committee for their timely help in reviewing the submissions and all the authors for submitting their papers to the workshop. We also thank the EMNLP 2025 workshop chairs for their kind support in delivering the workshop and producing these proceedings.

TSAR Organizing Committee

Matthew Shardlow,
 Fernando Alva-Manchego,
 Kai North,
 Regina Stodden,
 Horacio Saggion,
 Nouran Khallaf,
 Akio Hayakawa

Organizing Committee

Organizing Committee

Matthew Shardlow, Manchester Metropolitan University, UK

Fernando Alva-Manchego, Cardiff University, UK

Kai North, Cambium Assessment, USA

Regina Stodden, University of Bielefeld, Germany

Horacio Saggion, Universitat Pompeu Fabra, Spain

Nouran Khallaf, University of Leeds, UK

Akio Hayakawa, Universitat Pompeu Fabra, Spain

Program Committee

Program Committee

Kai North, Cambium Assessment, USA
Matthew Shardlow, Manchester Metropolitan University, UK
Regina Stodden, Heinrich Heine University, Germany
Jaap Kamps, University of Amsterdam, Netherlands
Fernando Alva-Manchego, Cardiff University, UK
Dave Kauchak, Pomona College, USA
Akio Hayakawa, Universitat Pompeu Fabra, Spain
Yingqiang Gao, University of Zurich, Switzerland
Lourdes Moreno, Universidad Carlos III de Madrid, Spain
Joseph Marvin Imperial, University of Bath, UK
Nouran Khallaf, University of Leeds, UK
Horacio Saggion, Universitat Pompeu Fabra, Spain
Anna Dmitrieva, University of Helsinki, Finland
Rémi Cardon, HULAT - UC3M, Spain
Michael Gille, Hochschule für angewandte Wissenschaften Hamburg, Germany
Martina Miliari, Università di Pisa, Italy
Thomas Francois, Université catholique de Louvain, Belgium
Natalia Grabar, STL CNRS Université Lille 3, France
Reno Kriz, Johns Hopkins University, USA
Liana Ermakova, HCTI, Université de Bretagne Occidentale, France
Tomoyuki Kajiwara, Ehime University, Japan
Marcos Zampieri, George Mason University, USA
Ekaterina Kochmar, MBZUAI, UAE
Dennis Aumiller, Heidelberg University, Germany
Zihao Li, University of Manchester, UK
Daniele Schicchi, CNR ITD, Italy
Mounica Maddela, Bloomberg, USA
Sarah Ebling, University of Zurich, Switzerland
Sandaru Seneviratne, The Australian National University, Australia
Jasper Degraeuwe, Ghent University, Belgium
Yannick Parmentier, LORIA - Université de Lorraine, France
Tadashi Nomoto, National Institute of Japanese Literature, Japan
Raquel Hervás, Universidad Complutense de Madrid, Spain
Tannon Kew, University of Zurich, Switzerland
Giulia Venturi, ILC-CNR, Italy
Sowmya Vajjala, National Research Council, Canada
Freya Hewett, Humboldt Institut für Internet & Gesellschaft, Germany
Piotr Przybyła, Universitat Pompeu Fabra / Polish Academy of Sciences, Spain & Poland
Maja Popovic, ADAPT Centre @ DCU, Ireland
Arne Jönsson, Linköping University, Sweden
Christina Niklaus, University of St. Gallen, Switzerland
Silvana Deilen, University of Hildesheim, Germany

Keynote Talk

Controllable Text Simplification

Tomoyuki Kajiwara
Ehime University
November 09, 2025 –

Abstract: Text simplification is a task of paraphrasing a given text in an easy-to-understand manner to assist a variety of people in language comprehension. To fully benefit from text simplification, consideration of individual differences in language ability is crucial.

Over the past decade, much effort has been devoted to studying methods for text simplification that adapt expressions according to the target audience. This presentation focuses on approaches to personalizing text simplification, specifically controlling readability and editing operations.

Bio: Tomoyuki Kajiwara received the B.S. and M.S. degrees in engineering from the Nagaoka University of Technology, Japan, in 2013 and 2015, respectively, and the Ph.D. degree in engineering from the Tokyo Metropolitan University, Japan, in 2018. From 2018 to 2020, he was a Specially-Appointed Assistant Professor with the Osaka University. He is currently an Assistant Professor with the Ehime University. His research interests include natural language processing, paraphrasing, and quality estimation.

Table of Contents

<i>Template-Based Text-to-Image Alignment for Language Accessibility A Study on Visualizing Text Simplifications</i>	
Belkiss Souayed, Sarah Ebling and Yingqiang Gao	1
<i>Document-level Simplification and Illustration Generation Multimodal Coherence</i>	
Yuhang Liu, Mo Zhang, Zhaoyi Cheng and Sarah Ebling	19
<i>Medical Text Simplification From Jargon Detection to Jargon-Aware Prompting</i>	
Taiki Papandreou, Jan Bakker and Jaap Kamps	36
<i>Readability Reconsidered A Cross-Dataset Analysis of Reference-Free Metrics</i>	
Catarina Belem, Parker Glenn, Alf Samuel, Anoop Kumar and Daben Liu	47
<i>Evaluating Health Question Answering Under Readability-Controlled Style Perturbations</i>	
Md Mushfiqur Rahman and Kevin Lybarger	70
<i>A Multi-Agent Framework with Diagnostic Feedback for Iterative Plain Language Summary Generation from Cochrane Medical Abstracts</i>	
Felipe Arias Russi, Carolina Salazar Lara and Ruben Manrique	87
<i>Efficient On-Device Text Simplification for Firefox with Synthetic Data Fine-Tuning</i>	
Pablo Romero, Zihao Li and Matthew Shardlow	105
<i>Findings of the TSAR 2025 Shared Task on Readability-Controlled Text Simplification</i>	
Fernando Alva-Manchego, Regina Stodden, Joseph Marvin Imperial, Abdullah Barayan, Kai North and Harish Tayyar Madabushi	116
<i>OneNRC@TSAR2025 Shared Task Small Models for Readability Controlled Text Simplification</i>	
Sowmya Vajjala	131
<i>GRIPF at TSAR 2025 Shared Task Towards controlled CEFR level simplification with the help of inter-model interactions</i>	
David Alfter and Sebastian Gombert	137
<i>ITU NLP at TSAR 2025 Shared Task A Three-Stage Prompting Approach for CEFR-Oriented Text Simplification</i>	
Kutay Arda Dinç, Fatih Bektaş and Gülşen Eryiğit	149
<i>STARLING at TSAR 2025 Shared Task Leveraging Alternative Generations for Readability Level Adjustment in Text Simplification</i>	
Piotr Przybyła	155
<i>taskGen at TSAR 2025 Shared Task Exploring prompt strategies with linguistic knowledge</i>	
Juan Cruz Oviedo, Elisabet Comelles Pujadas, Laura Alonso Alemany and Jordi Atserias Batalla	160
<i>EasyJon at TSAR 2025 Shared Task Evaluation of Automated Text Simplification with LLM-as-a-Judge</i>	
Paul-Gerhard Barbu, Adrianna Lipska-Dieck and Lena Lindner	173
<i>HULAT-UC3M at TSAR 2025 Shared Task A Prompt-Based Approach using Lightweight Language Models for Readability-Controlled Text Simplification</i>	
Jesus M. Sanchez-Gomez, Lourdes Moreno, Paloma Martínez and Marco Antonio Sanchez-Escudero	183

<i>UoL-UPF at TSAR 2025 Shared Task A Generate-and-Select Approach for Readability-Controlled Text Simplification</i>	
Akio Hayakawa, Nouran Khallaf, Horacio Saggion and Serge Sharoff	193
<i>Uniandes at TSAR 2025 Shared Task Multi-Agent CEFR Text Simplification with Automated Quality Assessment and Iterative Refinement</i>	
Felipe Arias Russi, Kevin Cohen Solano and Ruben Manrique	211
<i>EhiMeNLP at TSAR 2025 Shared Task Candidate Generation via Iterative Simplification and Reranking by Readability and Semantic Similarity</i>	
Rina Miyata, Koki Horiguchi, Risa Kondo, Yuki Fujiwara and Tomoyuki Kajiwara	217
<i>OUNLP at TSAR 2025 Shared Task Multi-Round Text Simplifier via Code Generation</i>	
Cuong Huynh and Jie Cao	223
<i>HIT-YOU at TSAR 2025 Shared Task Leveraging Similarity-Based Few-Shot Prompting, Round-Trip Translation, and Self-Refinement for Readability-Controlled Text Simplification</i>	
Mao Shimada, Kexin Bian, Zhidong Ling and Mamoru Komachi	231
<i>SQUREL at TSAR 2025 Shared Task CEFR-Controlled Text Simplification with Prompting and Reinforcement Fine-Tuning</i>	
Daria Sokova, Anastasiia Bezobrazova and Constantin Orasan	242
<i>Archaeology at TSAR 2025 Shared Task Teaching Small Models to do CEFR Simplifications</i>	
Rares-Alexandru Roscan and Sergiu Nisioi	251
<i>HOPE at TSAR 2025 Shared Task Balancing Control and Complexity in Readability-Controlled Text Simplification</i>	
Sujal Maharjan and Astha Shrestha	261
<i>Know-AI at TSAR 2025 Shared Task Difficulty-aware Text Simplification System</i>	
Yiheng Wu, Anisia Katinskaia, Jue Hou and Roman Yangarber	266

Template-Based Text-to-Image Alignment for Language Accessibility: A Study on Visualizing Text Simplifications

Belkiss Souayed^{UZH} Sarah Ebling^{UZH} Yingqiang Gao^{†UZH}

^{UZH}Department of Computational Linguistics, University of Zurich, Switzerland

belkiss.souayed@uzh.ch

{ebling, yingqiang.gao}@cl.uzh.ch

Warning: This paper includes AI-generated images that may cause visual discomfort.

Abstract

Individuals with intellectual disabilities often have difficulties in comprehending complex texts. While many text-to-image models prioritize aesthetics over accessibility, it is not clear how visual illustrations relate to text simplifications (TS) generated from them. This paper presents a structured vision-language model (VLM) prompting framework for generating accessible images from simplified texts. We designed five prompt templates, i.e., *Basic Object Focus*, *Contextual Scene*, *Educational Layout*, *Multi-Level Detail*, and *Grid Layout*, each following distinct spatial arrangements while adhering to accessibility constraints such as object count limits, spatial separation, and content restrictions. Using 400 sentence-level simplifications from four established TS datasets (OneStopEnglish, SimPA, Wikipedia, and ASSET), we conducted a two-phase evaluation: Phase 1 assessed prompt template effectiveness with CLIPScores, and Phase 2 involved human annotation of generated images across ten visual styles by four accessibility experts. Results show that the *Basic Object Focus* prompt template achieved the highest semantic alignment, indicating that visual minimalism enhances language accessibility. Expert evaluation further identified *Retro* style as the most accessible and Wikipedia as the most effective data source. Inter-annotator agreement varied across dimensions, with *Text Simplicity* showing strong reliability and *Image Quality* proving more subjective. Overall, our framework offers practical guidelines for accessible content generation and underscores the importance of structured prompting in AI-generated visual accessibility tools.

 Dataset  Code

[†]Corresponding author.



Figure 1: Example image generated based on the simplified text “I will never forget the wonderful memories he has given us, like that magical night in Moscow.” (Style: Artistic, Dataset: OneStopEnglish).

1 Introduction

Individuals with intellectual disabilities often have difficulties understanding complex texts (Alva-Manchego et al., 2020; Yawiloeng, 2022). While text simplification (TS) improves readability, it is frequently insufficient on its own. Research shows that visual support, recommended by Easy-to-Read guidelines, can significantly enhance comprehension (Madina et al., 2023). However, most text-to-image models prioritize aesthetics over accessibility, which can cause cluttered, abstract, or semantically misaligned visuals generated from text inputs.

Recent studies have explored image retrieval for simplified text as a potential solution (Geislinger et al., 2023; Singh et al., 2023). However, none have systematically investigated structured prompting for accessible image generation in the TS context. In this work, we address this gap by introducing a template-based framework that explicitly enforces accessibility constraints, such as

controlled object count, spatial separation, and the absence of embedded text, directly within prompts for querying vision-language models (VLMs).

Using 400 text simplification pairs from four datasets (ASSET, OneStopEnglish, SimPA, and Wikipedia), we generated and evaluated 4,000 images across five prompt templates and ten visual styles. Our study aims to answer the following research questions (RQs):

- **RQ1:** How can template-based prompting improve the accessibility of images generated from simplified texts?
- **RQ2:** Which visual styles and data sources are most effective for accessible image generation?
- **RQ3:** How do expert annotations compare to automatic evaluation metrics in assessing accessibility?

Our results show that the *Basic Object Focus* template performs best, supporting visual minimalism. Experts rate *Retro* style and *Wikipedia*-sourced simplifications as most accessible. We also find weak correlation between CLIPScores and human judgments, underscoring the need for human-centered evaluation in accessible and inclusive AI.

2 Related Work

2.1 Visual-Aided Text Simplification

As a natural language processing (NLP) task to improve language accessibility, TS actively modifies complex texts to improve readability for diverse target group persons, including persons with cognitive disabilities (Espinosa-Zaragoza et al., 2023; Gao et al., 2025), foreign language learners (Degraeuwe and Saggion, 2022; Li et al., 2025) and others. TS can operate at the lexical, syntactic, and discourse levels (Alva-Manchego et al., 2020; Zhong et al., 2020). Beyond language, accessible communication increasingly benefits from multimodal aids, since simplified text alone may not suffice for users who have reading difficulties. Studies show that pairing text with visual components improves overall comprehension and engagement (Yawiloeng, 2022) and mirrors those principles in Easy-to-Read guidelines (Madina et al., 2023), which recommend simple text with visual illustrations for better language accessibility. However, producing accessible visuals remains a technical challenge, motivating research interest in automatic image retrieval and text-to-image generation in a TS context (Marturi and Elwazzan, 2025). Our study builds on this line

of work by investigating how images generated through VLMs from text simplifications can support language accessibility through both semantic simplicity and visual clarity.

2.2 Retrieval-based Language Accessibility

Early multimodal approaches retrieve images from databases to support the language accessibility. Geislinger et al. (2023) combined NLP approaches with eye-tracking to provide real-time visual support for difficult terms, while Singh et al. (2023) optimized textbook enrichment by retrieving web images using CLIP-based similarity (Radford et al., 2021). Such systems improved perceived educational outcomes, but they remain limited by the availability of accessible visuals in existing databases, especially for abstract or personalized concepts. These challenges further motivate research in text-to-image alignment for language accessibility.

2.3 Datasets and Text-to-Image Benchmarks

Recent advances in VLMs such as CLIP (Radford et al., 2021) and DALL·E (Betker et al., 2023) have driven progress in text-to-image alignment studies. Several datasets have been proposed to either train image captioning models or evaluate the alignment quality.

Benchmark datasets like MS COCO (Lin et al., 2014) and Flickr30K (Plummer et al., 2015) focus on short image captions but are not specific to accessibility for simplified text. MOTIF (Wang et al., 2022) pairs simplified sentences with illustrative images to support second-language learners, though it remains limited to multimodal retrieval rather than generation. Our work addresses this gap by generating new images directly from text simplifications, enabling scalable visual support.

Anschütz et al. (2024) present the first systematic study of text-to-image for Easy-to-Read German content. Using 80 structured prompts translated from German, authors generated 2,240 images across seven models and evaluated them with automated metrics FID (Heusel et al., 2017), CLIP-Score (Radford et al., 2021), and TIFA (Hu et al., 2023).

While previous works have investigated text-to-image alignment in the context of image captioning and alignment evaluation, to the best of our knowledge, no prior work has proposed a structured prompting strategy that explicitly enforces accessibility constraints such as controlled object

count, spatial separation, and the exclusion of abstract or textual elements. Likewise, no previous study has compared the accessibility effects of different visual styles in a systematic way. In this work, we address these research gaps by introducing a template-based prompting framework aligned with accessibility principles and conducting a large-scale experiment: 4,000 images generated from 400 simplified sentences, spanning five prompt templates and ten visual styles. In contrast to fixed-format prompts, our pipeline transforms naturally simplified sentences into structured prompts while preserving semantic content. We combine automatic metrics with expert annotations across multiple dimensions, enabling the first human evaluation of how prompt design and visual style influence cognitively accessible image generation.

3 Methodology

Our study follows a two-phase pipeline. In **Phase 1**, we developed and evaluated five prompt templates used by VLMs for generating visuals, each implementing unique accessibility constraints while sharing core principles such as:

- **Controlled object counts:** Defining the quantity of visual components present in the generated images;
- **Spatial separation:** Checking if visual components are physically well-arranged;
- **Exclusion of text or abstract elements:** Examining whether VLMs render texts in generated images, which can cause additional confusion.

In **Phase 2**, we take the best-performing template and apply it to produce a large-scale multimodal dataset. We generated 4,000 images from 400 simplified sentences, spanning ten visual styles, with style-specific prompt adjustments (see Appendix C.2 for an illustration of how a single sentence generates multiple style-specific prompts). Four expert annotators then assessed the resulting text–image pairs across six dimensions using pre-defined criteria designed to capture both quantitative and qualitative accessibility aspects.

While we acknowledge that images are linked to textual units of varying granularity, operating on the sentence level throughout was a means of isolating the effect of text unit size.

3.1 Prompt Templates

Our initial trials with direct prompts (e.g., “*Generate an accessible image for this simplified sentence: [sent]*”) produced generic, inconsistent re-

sults. Without structural guidance, the model often failed to enforce key accessibility needs such as clear object focus, reduced background clutter, or stylistic appropriateness. Although the templates were structurally pre-defined, GPT-4 (Achiam et al., 2023) was necessary to map sentence semantics into visual prompts. It interpreted each simplified sentence to decide which objects and relations to depict, how to represent abstract ideas concretely, and which visual style to apply. This semantic-to-visual prompt generation goes beyond simple template filling or rule-based rewriting. We concluded that such direct prompts lacked the precision needed to control the layout structure and enforce necessary visual requirements such as selecting the most suitable stylistic representation, maintaining a clear depiction of primary objects without clutter, and minimizing background noise to reduce cognitive load.

To address this, we adopted structured templates that embed accessibility constraints directly into the prompt. Template design was grounded in W3C Web Accessibility Initiative (WAI) image tutorials¹, which classify accessible images as *informative* (conveying concepts visually), *complex* (layered diagrams), or *grouped* (collections representing unified information). Each template isolates a specific layout logic, e.g., alignment, sequencing, or hierarchy, to assess its impact on visual comprehension, reflecting formats common in educational and assistive contexts:

- **Basic Object Focus:** Removes all spatial context to measure the effect of object isolation. Isolates between one and three objects on a plain background to maximize clarity and minimize cognitive load;
- **Contextual Scene:** Presents simple real-world layouts (e.g., items on a shelf) to test how minimal grounding supports understanding. Situates objects in simple real-world settings (e.g., items in a room) to provide minimal contextual grounding;
- **Educational Layout:** Introduces sequencing and flow, mimicking instructional visuals like timelines. Uses simple relations such as arrows to support instructional use;
- **Multi-Level Detail:** Inspired by textbook diagrams, adds foreground–background layering to

¹W3C Web Accessibility Initiative (WAI) image tutorials: guidance on *informative images*, *complex images*, and *groups of images*. Available at: <https://www.w3.org/WAI/tutorials/images/>

explore hierarchy and layered perception. Layers a central object with two to three related subobjects to explore hierarchical perception;

- **Grid Layout:** Simulates classification interfaces (e.g., icon grids) to examine whether symmetry aids clarity. Organizes items in a 2×2 or 3×3 structure, simulating classification interfaces and testing whether symmetry aids clarity.

All five templates share a unified set of accessibility constraints (see Appendix B for detailed specifications). To ensure fair comparison and accessibility, the generation process enforced baseline constraints: (1) between three to five distinct objects per image; (2) sufficient spatial separation between objects; (3) avoidance of text, numbers, or motion effects; (4) exclusion of abstract, metaphorical, or culturally biased elements; and (5) preference for plain or neutral backgrounds. These requirements operationalize accessibility principles and reduce potential bias.

To identify which prompt template best supports cognitively accessible image generation, we generated 100 images per template (500 total) from simplified sentences. Prompts were constructed with GPT-4 (Achiam et al., 2023) and images generated using DALL·E 3 (Betker et al., 2023). Outputs were assessed using CLIPScore (ViT-L/14@336px; Hessel et al. (2021)) to measure text-to-image alignment. We chose these models for their high prompt fidelity, built-in safety filters, and consistency across large batches. As the focus was on evaluating prompt design rather than comparing VLMs or model-specific optimization, alternative architectures were not considered.

To guide template selection, we developed a composite scoring system to balance accuracy and robustness rather than optimize a single metric. Weights were empirically chosen to prevent unstable templates from dominating, ensuring selection of templates that perform reliably across diverse simplification contexts rather than excelling on a narrow subset of inputs. Given a template t , the composite score is computed as:

$$\text{Composite}(t) = 0.4\mu_t + 0.2C_t + 0.2S_t + 0.1B_t + 0.1(1 - W_t),$$

where μ_t is the mean CLIPScore (40%), C_t is the consistency (20%), S_t is the success rate (20%), B_t is the fraction of best-performing cases (10%), and W_t is the fraction of worst-performing cases (10%). All components were normalized to [0, 1] before aggregation.

Template	CLIP	Composite
Basic Object Focus	0.211	5.31
Contextual Scene	0.210	4.95
Educational Layout	0.202	4.63
Multi-Level Detail	0.201	4.50
Grid Layout	0.199	4.39

Table 1: Phase 1 results (100 prompts × 5 templates). Basic Object Focus ranked highest and was selected for Phase 2.

As shown in Table 1, the *Basic Object Focus* template achieved the highest text-to-image alignment, consistency, and overall composite score. This suggests that reducing background clutter and emphasizing a small number of core objects is most effective for supporting accessibility. Based on these findings, *Basic Object Focus* was selected as the foundation for Phase 2, where we scaled generation to 4,000 images across ten visual styles and conduct expert evaluations to address RQ2 and RQ3.

3.2 Dataset Compilation

We compiled a text-to-image TS corpus by sampling from four established text simplification datasets: OneStopEnglish (Vajjala and Lučić, 2018), SimPA (Scarton et al., 2018), Wikipedia (Sun et al. (2020), without context), and ASSET (Alva-Manchego et al., 2020). These corpora were selected for their complementary domains and simplification strategies, covering news, public administration, encyclopedic text, and web content, each provides parallel complex-simplified sentence pairs. Table A1 in Appendix A summarizes their main features.

The four source datasets are all sentence-aligned, offering multiple simplifications per complex text or varying types of simplification methods. This led to the necessity of adopting a consistent sampling strategy to support uniformity and methodological coherence across all datasets. Specifically, among the multiple simplifications available for each complex sentence, we retained only one per original to ensure a one-to-one mapping between the simplification and its visually generated counterpart. We also ensured to randomly draw the exact number of samples from each corpus in order to create a balanced dataset which reflects the distinct simplification approaches and domains covered within each source text.

We sampled a balanced subset of 400 pairs (100

per dataset). The sampling procedure used random selection to avoid bias and ensure that sampled instances reflected the overall linguistic variability and quality of the source datasets. We deliberately avoided cherry-picking instances we believed to be ideal for generation or applying quality filtering. By “ideal” or “quality”, we refer to sentences with simple syntax and clear meaning as opposed to abstract or ambiguous ones that are harder to visualize. This approach ensured our evaluation methodology reflects those real-world difficulties of accessible image generation across diverse text features.

The 400 sentence pairs were stored in a JSON format, where each line represents one structured data entry. In addition to the complex and its simplified counterpart, we annotated each pair with relevant metadata fields aimed to support traceability and data hierarchy (see Appendix C.1 for a complete example). This dataset served as the basis for prompt formulation, image generation, and expert annotation.

As all TS datasets were human-produced and sentence-aligned, we conducted minimal data preprocessing including removal of formatting artifacts (particularly in Wikipedia) and standardizing all samples into JSON format with metadata for dataset source, domain, and token counts. We focused on filtering overly short (<10 tokens) or long (>55 tokens) sentence-level simplifications to ensure balanced text complexity. After preprocessing, the final dataset exhibited an average complex sentence length of 26.2 tokens and simplified sentence length of 23.9, with an overall reduction of 2.3 tokens (8.8%).

Following Phase 1, we refined *Basic Object Focus* to preserve minimalism while improving reliability: exactly four objects, at least 30% spacing between objects, and a 10% cap on size variation to maintain equal prominence. A pilot trial on 20 samples yielded a CLIPScore of 0.3465, a 64% increase over the Phase 1 score (0.2108), confirming the benefit of explicit spacing and uniformity constraints.

We then generated 4,000 prompts and produced high-resolution images with DALL·E 3 using a robust, checkpointed pipeline (asynchronous batching, retries, and traceable file naming). The intrinsic safety moderation of DALL·E 3 blocked a small subset of requests, primarily those containing dense named entities or historical and military references, which the system may have misinter-

preted as potentially promoting violence, political propaganda, or disinformation.

3.3 Evaluation Methods

We conducted automatic and human evaluation using both automatically and expert-annotated accessibility, clarity, and style-related effects. Specifically, we used:

- **Inter-annotator agreement (IAA):** Measured using Krippendorff’s α (Krippendorff, 1970) for human annotations;
- **Style identification:** Evaluated via Recall@3;
- **Human-computer correlation:** Measured using Pearson correlation coefficients;
- **Composite accessibility scores:** Combining expert ratings into weighted indices of effectiveness.

Statistical analysis included descriptive and comparative assessments across datasets, styles, and evaluation dimensions. Four expert annotators participated in this study, all with academic and professional expertise in accessibility and inclusive communication. Their backgrounds involve specialization in barrier-free communication, Easy Language, audio description, subtitling, and text simplification. Several have contributed to national research projects on accessible technologies and have practical experience designing and evaluating content for people with cognitive disabilities. All annotators signed informed consent forms and received detailed annotation guidelines. We compensated the expert annotators with a fair hour rate of 120 Swiss Francs.

The annotation study was originally planned with 4,000 generated images but reduced to 2,000 due to cost and workload constraints. The final set ensured each simplified sentence had images in all 10 styles, maintaining equal style representation. All images were renamed adopting a standardized numerical convention to ensure unbiased annotation, as original filenames contained style information that could influence decisions. A shared IAA set of 200 images (20 texts \times 10 styles) was annotated by all four experts to assess agreement. The remaining 1,800 images were split evenly so each annotator received 200 IAA images plus 450 unique ones (650 images per annotator, 2,600 total annotations planned). In the end 976 annotations were completed by the four experts.

The annotation interface was implemented

through a customized Label Studio² configuration that presents the evaluation framework in a clear format optimized for expert assessment. The interface displays the simplified text at the top, followed by the image presented with zoom controls for detailed examination. The evaluation questions are organized into collapsible sections corresponding to distinct domains. The configuration also ensured that annotation data was stored in structured formats suitable for subsequent statistical analysis. The interface included progress tracking which allowed the experts to track their progress and resume annotation across multiple sessions as needed. The complete set of evaluation questions is provided in Appendix D for reference.

4 Results

A total of 976 annotations were collected from four experts (A, K, L, M), corresponding to 37.5% of the initially planned 2,600 (Table 2). Despite lower coverage, the annotated sample provided sufficient data for meaningful analysis. Distribution across datasets was balanced (Table 3), ensuring findings are not skewed toward one text source.

Expert	# Assigned	Completion Rate
Expert A	650	38.5%
Expert K	650	38.5%
Expert L	650	30.8%
Expert M	650	42.5%
Total	2,600	37.5%

Table 2: Expert assignment and completion rates.

Dataset	# Annotation	Percentage
Wikipedia	280	28.7%
SimPA	270	27.7%
ASSET	214	21.9%
OneStop	212	21.7%
Total	976	100.0%

Table 3: Source distribution in completed dataset.

4.1 Scoring Overview

Experts rated images across six dimensions with a maximum total of 100 points. Table 4 summarizes mean scores and contributions. *Ethics* dominated in expert evaluations (31.4%), followed by

²Apache-2.0 license, available at <https://github.com/HumanSignal/label-studio/>

Text Quality (21.6%). Visual dimensions such as *Image Simplicity* and *Text-Image Alignment* contributed less, reflecting both interpretive challenges and model limitations.

Dimension	Scale	Mean	Contribution
Image Simplicity	0–15	4.56	9.7%
Image Quality	0–15	6.34	13.5%
Text Simplicity	0–15	5.65	12.1%
Text Quality	0–15	10.13	21.6%
Ethics	0–20	14.74	31.4%
Text-Image Alignment	0–20	5.49	11.7%

Table 4: Six evaluation dimensions used during expert annotation.

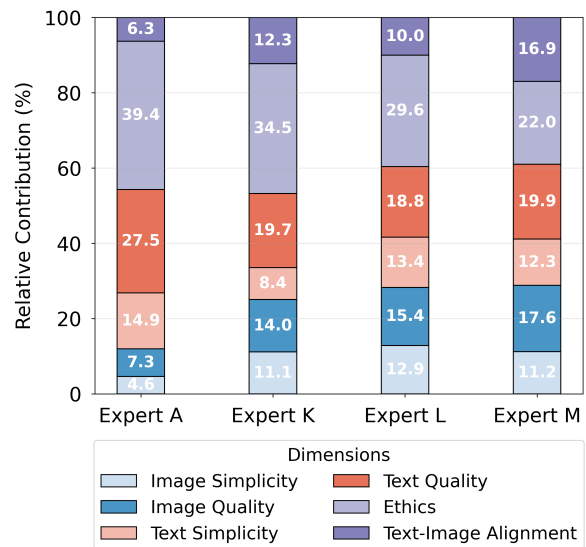


Figure 2: Relative contribution of each evaluation dimension per expert.

We observed that experts differed in evaluation style: Expert A was highly critical of visual dimensions but more lenient on *Ethics*; Expert K was generous overall, with the highest mean scores; Expert L was most conservative, assigning the lowest averages; Expert M showed consistency, prioritizing *Image Quality* and *Text-Image Alignment*. Figure 2 illustrates the composition of expert ratings.

4.2 Inter-Annotator Agreement

IAA was measured with Krippendorff’s Alpha across dimensions (Table 5). *Text Simplicity* achieved the highest agreement (up to $\alpha = 0.599$), suggesting a shared understanding of linguistic complexity. Agreement was weaker for *Image Simplicity* and *Text-Image Alignment*, and nearly absent for *Image Quality* and *Ethics*, reflecting subjective ratings.

Dimension	4 Experts	3+ Experts	2+ Experts
Text Simplicity	0.599	0.458	0.570
Image Simplicity	0.486	0.374	0.282
Text-Image Alignment	0.379	0.267	0.115
Text Quality	0.086	0.228	0.223
Ethics	0.018	0.053	-0.214
Image Quality	0.013	-0.007	0.006

Table 5: Krippendorff’s Alpha across dimensions, reported for all and subgroup of experts.

4.3 Style Recognition Performance

To study the influence of image styles for accessibility, we asked the experts to identify three image styles from the ten pre-defined style categories. Recall@3 averaged 47.3%, with significant variance across experts (Table 6). Styles such as *3D Rendered* and *Retro* were easily recognized, while *Artistic* and *Technical* proved to be the most difficult ones (Table 7).

Expert	# Correct	# Total	Recall@3
Expert A	126	250	0.504
Expert M	139	276	0.504
Expert K	120	250	0.480
Expert L	77	200	0.385
Average	462	976	0.473

Table 6: Recall@3 for style recognition.

Style	# Images	Recall@3	Tier
3D Rendered	103	81.6%	Easy
Retro	100	76.0%	Easy
Cartoon	109	69.7%	Medium
Geometric	94	55.3%	Medium
Realistic	86	44.2%	Hard
Storybook	96	43.8%	Hard
Digital Art	97	37.1%	Hard
Minimalistic	88	34.1%	Hard
Artistic	100	18.0%	Very Hard
Technical	103	9.7%	Very Hard

Table 7: Style recognition difficulty comparison.

4.4 Human-Computer Correlation

We compared CLIPScores with expert *Text-Image Alignment* ratings. We observed weak but statistically significant Pearson correlation ($r = 0.17$, $p < 0.001$), improving after per-expert standardization (Table 8). We observed that Expert A aligned most closely with CLIPScore ($r = 0.251$). In addition, we found that *Text-Image Alignment* varied by

style, with strongest results for *3D Rendered* and *Artistic*, and weakest for *Technical*.

Analysis Level	r	p -value	Sig.
Overall (Raw)	0.133	<0.001	***
Per-Expert Standard.	0.173	<0.001	***
Expert A	0.251	<0.001	***
Expert M	0.152	0.011	*
Expert L	0.146	0.040	*
Expert K	0.133	0.036	*

Table 8: Human-Computer correlation (Pearson r).

4.5 Expert-Specific Scoring Patterns



(a) Expert A (Top)

“Kazuma Watanabe (born 10 August 1986) is a Japanese football player. He plays for Yokohama F. Marinos and Japan national team.”



(b) Expert K (Top)

“Bangui makes textiles, food products, beer, shoes, and soap.”



(c) Expert L (Top)

“Shunsuke Iwanuma (born 2 June 1988) is a Japanese football player. He plays for Consadole Sapporo.”



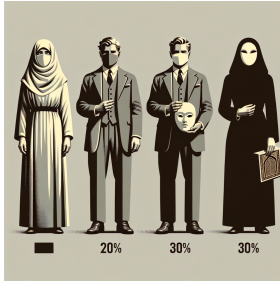
(d) Expert M (Top)

“Franklin Delano Roosevelt or FDR, was the 32nd President of the United States.”

Figure 3: Highest-rated images by each expert, with corresponding simplified sentences.

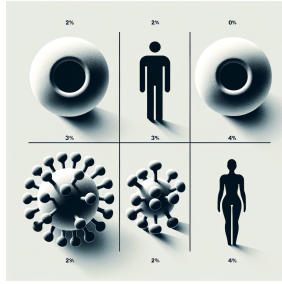
To further understand individual expert ratings, we conducted per-expert analysis on how they rate different evaluation dimensions, with a focus on how individual experts subjectively rate different evaluation dimensions.

We found that *Image Quality* and *Ethics* were the two dimensions with the lowest agreement. The main reason is that each expert used the scale differently. For example, Expert A was very strict



(a) Expert A (Bottom)

“In Mask fetishism is persons wants to see another person wearing mask or taking off a mask.”



(b) Expert K (Bottom)

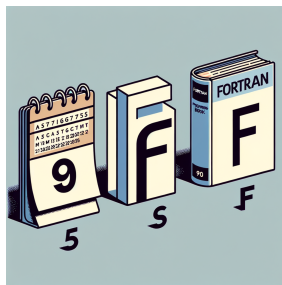
“Cytomegalovirus (from the Greek cyto-, cell, and megalo-, large) is a viral genus of the Herpesviruses group in humans...”

Figure 4: Lowest-rated images by Expert A and K.



(c) Expert L (Bottom)

“One of the boyfriends wrote to me and said, Listen, she’s not mad but Cynthia found out.”



(d) Expert M (Bottom)

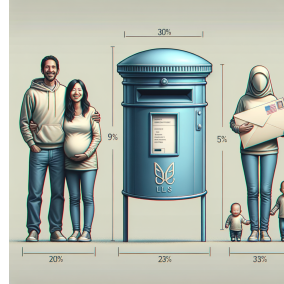
“Since Fortran 90, the capitalization has been abandoned. The published formal standards use Fortran.”

Figure 5: Lowest-rated images by Expert L and M.

with visuals but very generous in Ethics, while Expert M was the opposite. This systematic rating behavior explains why the same images received very different scores.

Figure 3 to Figure 6 present concrete examples of high and low scores given by experts. While some of these images might be potentially biased, these cases show that even with advanced VLMs, generating accessible visual interpretation for text simplification remains a challenging task, as it might further introduce confusion and misinterpretation of the text.

Building on these expert-annotated examples, we emphasize the importance of systematically examining the cultural and ethical biases present in VLMs when generating visuals for text simplification. Our findings suggest that even when accessibility constraints are technically enforced, subtle cultural cues, stereotypes, or contextual omissions can still emerge, potentially undermining the inclusiveness of the generated content. Future research should therefore not only focus on improving vi-



Case 1 (0125)

“Letterbox Service ...”
Percentages not linked to text, cluttered layout, weak link between people and the concept. Some experts gave very low Alignment/Quality, others still rewarded Ethics.



Case 2 (0944)

“He traveled over 200,000 miles ...”
Vintage style with many props, missing legs, low contrast, and the number not shown. Experts disagreed on Simplicity and Quality.

Figure 6: Examples where experts strongly disagreed.

sual alignment and linguistic adequacy, but also on auditing and mitigating such biases through participatory evaluation and the inclusion of diverse user perspectives. Addressing these challenges is essential for developing truly inclusive, transparent, and socially responsible AI systems that can serve users with a broad range of linguistic and cognitive needs.

5 Discussion

5.1 Style and Dataset Impact for Accessibility

To determine which styles and datasets best support accessibility, we combined expert ratings into composite accessibility scores. For visual styles, the score weighted *Text–Image Alignment* (60%), *Image Simplicity* (25%), and *Image Quality* (15%). For datasets, we focused on *Text Quality* (50%) and *Text Simplicity* (50%). Scores were scaled to a 0–100 range and averaged across experts.

Visual styles. *Retro* achieved the highest overall accessibility, performing strongly in both alignment and simplicity. *Realistic* style also scored well, driven by high image quality and familiarity, while *3D Rendered* provided the most consistent results. By contrast, abstract or heavily stylized categories such as *Geometric*, *Minimalistic*, and *Artistic* ranked lowest, often due to clutter or ambiguous representation. These results suggest that concrete and familiar visuals are more effective for accessibility than abstract or decorative designs.

Data sources. Wikipedia and ASSET ranked highest, with Wikipedia benefiting from factual clarity and wide coverage, and ASSET offering

more consistency due to its accessibility-focused design. OneStopEnglish and SimPA performed less well, with lower simplicity scores and more complex phrasing. This indicates that general-purpose and collaboratively simplified corpora provide stronger material for accessible image generation.

The findings highlight that both style and dataset choice substantially affect accessibility outcomes. *Retro* and *Realistic* visuals, paired with simplified Wikipedia content, emerged as the most effective combination. Conversely, highly abstract styles or bureaucratic text sources may hinder accessibility, underscoring the need to carefully select both visual style and textual input when designing accessible multimodal content.

5.2 Evaluation Agreement

The evaluation highlights both strengths and weaknesses of template-based image generation for accessibility. The 100-point scoring framework showed that *Ethics* and *Text Quality* consistently received the highest ratings, together accounting for more than half of the total, likely because these dimensions had clearer definitions. By contrast, *Image Simplicity* and *Text-Image Alignment* contributed less and showed greater variability, pointing to interpretive difficulties and model limitations. Expert behavior differed noticeably: some applied stricter standards while others were more generous, underlining the need for normalization when aggregating scores.

Inter-annotator agreement further confirmed these differences. *Text Simplicity* achieved the highest reliability, indicating a shared understanding of linguistic complexity. However, dimensions such as *Image Quality* and *Ethics* had very low or even negative agreement values, reflecting systematic disagreement rooted in subjective interpretations of what counts as “quality” or “ethical”. These results suggest that clearer task guidelines, calibration, or refined criteria will be necessary to improve reproducibility in future studies.

Style recognition performance also highlighted challenges. Experts correctly identified the intended style in less than half of cases, which is reasonable given overlapping visual features across categories. Distinctive styles such as *3D Rendered* or *Retro* were easier to recognize, whereas *Artistic* and *Technical* were consistently difficult, revealing both model inconsistencies and unclear category boundaries. Style recognition therefore appears

cognitively demanding and not fully reliable for accessibility-focused evaluations.

6 Conclusion

This work investigated template-based prompting for generating cognitively accessible images from simplified text. The *Basic Object Focus* template proved most effective, showing that minimalism and object isolation enhance clarity. Wikipedia and ASSET emerged as the most suitable datasets, while *Retro* and *Realistic* visuals supported accessibility better than abstract styles. Expert evaluations highlighted systematic disagreement on subjective dimensions, and CLIPScores showed only weak alignment with human judgments. Overall, our findings demonstrate that structured prompting improves accessibility but current VLMs remain limited, underscoring the need for better models, clearer annotation protocols, and continued integration of human expertise.

Limitations

Our study faced several constraints. Only 976 of the planned 2,000 annotations were collected, reducing statistical power and inter-annotator coverage. The expert panel was small (four annotators), making results sensitive to individual biases, and subjective dimensions such as *Image Quality* and *Ethics* showed strong disagreement. Despite prompt constraints, the model often produced cluttered layouts or text artifacts, and style recognition was hindered by ambiguous category boundaries. Finally, CLIPScore correlated only weakly with human judgments, raising concerns about its reliability for accessibility evaluation. These limitations highlight the need for clearer guidelines, improved models, and more diverse expert input in future work.

Ethical Statement

All expert annotators involved in the human evaluation provided informed consent and received fair compensation for their participation. The evaluation framework included ethical safeguards designed to detect and flag potentially discriminatory or culturally insensitive content. Additionally, filtering mechanisms were applied during image generation to mitigate harmful or inappropriate outputs, and selected filtering results were documented to assess potential model biases.

The authors acknowledge that some example images presented in this work may still reflect unintended biases or misleading representations toward certain ethnic or cultural groups. These instances are discussed transparently to emphasize the importance of continuous bias assessment in developing inclusive AI systems.

Lay Summary

Individuals with cognitive disabilities or reading difficulties often face challenges when processing complex visual and textual information. Current text-to-image models tend to prioritize photorealism or artistic creativity rather than accessibility, leaving a gap for users who require simplified and structured visuals.

Our study addresses this issue by integrating text simplification with structured prompting to generate cognitively accessible images. We designed five visual templates that control object number, spatial arrangement, and content complexity, while deliberately avoiding textual or abstract elements that may cause confusion.

This work contributes practical guidelines for producing accessible visual content. The proposed image styles and prompt templates can support policymakers, AI researchers, and assistive technology designers in creating visual materials that are both inclusive and cognitively accessible. By aligning text-to-image generation with accessibility principles, our approach ensures more equitable support for individuals with cognitive disabilities.

Acknowledgments

This work was supported by the Swiss Innovation Agency Innosuisse, Flagship Inclusive Information and Communication Technology (IICT), funding no. PFFS-21-47. We also thank our experts Luisa Carrer, Martin Kapuss, Alexa Lintner at Zurich University of Applied Sciences (ZHAW), and Katrin Andermatt at traduko, for their valuable contributions to the study. Finally, we thank the anonymous reviewers for their constructive feedback on our work.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.

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.

Miriam Anschutz, Tringa Sylaj, and Georg Groh. 2024. Images Speak Volumes: User-Centric Assessment of Image Generation for Accessible Communication. In *Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024)*, pages 27–40.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, and 1 others. 2023. Improving Image Generation with Better Captions. *Computer Science*. <https://cdn.openai.com/papers/dalle-3.pdf>, 2(3):8.

Jasper Degraeuwe and Horacio Saggion. 2022. Lexical Simplification in Foreign Language Learning: Creating Pedagogically Suitable Simplified Example Sentences. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 98–110.

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.

Yingqiang Gao, Kaede Johnson, David Froehlich, Luisa Carrer, and Sarah Ebling. 2025. Evaluating the Effectiveness of Direct Preference Optimization for Personalizing German Automatic Text Simplifications for Persons with Intellectual Disabilities. *arXiv preprint arXiv:2507.01479*.

Robert Geislinger, Ali Ebrahimi Pourasad, Deniz Gül, Daniel Djahangir, Seid Muhie Yimam, Steffen Remus, and Chris Biemann. 2023. Multi-Modal Learning Application - Support Language Learners with NLP Techniques and Eye-Tracking. In *Proceedings of the 1st Workshop on Linguistic Insights from and for Multimodal Language Processing*, pages 6–11.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Roman Le Bras, and Yejin Choi. 2021. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. *arXiv preprint arXiv:2104.08718*.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. *Advances in neural information processing systems*, 30.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith.

2023. TIFA: Accurate and Interpretable Text-to-Image Faithfulness Evaluation with Question Answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20406–20417.
- Klaus Krippendorff. 1970. Estimating the Reliability, Systematic Error and Random Error of Interval Data. *Educational and Psychological Measurement*, 30(1):61–70.
- Guanlin Li, Yuki Arase, and Noel Crespi. 2025. Aligning Sentence Simplification with ESL Learner’s Proficiency for Language Acquisition. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 492–507.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Computer vision—ECCV 2014: 13th European conference, zurich, Switzerland, September 6–12, 2014, proceedings, part v 13*, pages 740–755. Springer.
- Margot Madina, Itziar Gonzalez-Dios, and Melanie Siegel. 2023. Easy-to-Read in Germany: A Survey on its Current State and Available Resources. *arXiv preprint arXiv:2306.03189*.
- Krishna Chaitanya Marturi and Heba Elwazzan. 2025. LLM-Guided Planning and Summary-Based Scientific Text Simplification: DS@GT at CLEF 2025 SimpleText. *arXiv preprint arXiv:2508.11816*.
- Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning Transferable Visual Models from Natural Language Supervision. In *International conference on machine learning*, pages 8748–8763. PmlR.
- Carolina Scarton, Gustavo Paetzold, and Lucia Specia. 2018. Simpa: A Sentence-Level Simplification Corpus for the Public Administration Domain. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.
- Janvijay Singh, Vilém Zouhar, and Mrinmaya Sachan. 2023. Enhancing Textbooks with Visuals from the Web for Improved Learning. *arXiv preprint arXiv:2304.08931*.
- Renliang Sun, Zhe Lin, and Xiaojun Wan. 2020. On The Helpfulness of Document Context To Sentence Simplification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1411–1423.
- Sowmya Vajjala and Ivana Lučić. 2018. On-eStopEnglish Corpus: A New Corpus for Automatic Readability Assessment and Text Simplification. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications*, pages 297–304.
- Xintong Wang, Florian Schneider, Özge Alacam, Praatek Chaudhury, and Chris Biemann. 2022. MOTIF: Contextualized Images for Complex Words to Improve Human Reading. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2468–2477.
- Rattana Yawiloeng. 2022. Using Instructional Scaffolding and Multimodal Texts to Enhance Reading Comprehension: Perceptions and Attitudes of EFL Students. *Journal of Language and Linguistic Studies*, 18(2):877–894.
- Yang Zhong, Chao Jiang, Wei Xu, and Junyi Jessy Li. 2020. Discourse Level Factors for Sentence Deletion in Text Simplification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 9709–9716.

A Overview of Dataset Sources

Dataset	Domain	Size	Year	Level
ASSET	Wikipedia	23,590 simplifications for 2,359 original sentences	2020	Sentence
OneStopEnglish	News	189 articles (567 texts) at three reading levels	2018	Document & Sentence
SimPA	Web	1,100 sentences with 3 lexical and 1 syntactic simplification each	2018	Sentence
Wikipedia (w/o context)	Wikipedia	110K (with context) and 41K (without context)	2020	Sentence

Table A1: Overview of Text Simplification Datasets.

B Prompt Templates used for Image Generation

This appendix provides the complete set of prompt templates developed for structured text-to-image generation. Each template implements specific accessibility constraints while maintaining semantic alignment with simplified text inputs.

B.1 Basic Object Focus Template

Basic Object Focus - Specific Prompt Instructions

- Do not align or group objects (arrange them with neutral positioning).
- Avoid any suggestion of scene, narrative, or sequence.
- Ensure all objects are visually equal.
- No object should stand out more than the others.
- Background must be uniform and simple (e.g., white or gray).
- Emphasize maximum spacing between all objects.

B.2 Contextual Scene Template

Contextual Scene - Specific Prompt Instructions

- Arrange all objects in a straight, horizontal line.
- Use a single perspective, no variation in object size or depth.
- Maintain equal size across all objects to avoid depth illusion.
- Include one minimal environmental element (e.g., surface, wall) when needed.
- Keep at least 20% spacing between each object to preserve separation.

B.3 Educational Layout Template

Educational Layout - Specific Prompt Instructions

- Arrange objects in a strict left-to-right horizontal sequence.
- Visually connect each object to the next with a line or arrow.
- Gradually reduce object size from left to right by 10–15%.
- Include a visible numeric marker (1, 2, 3...) near each object.

- Limit the maximum object count to 4 to maintain consistency.
- Narrow spacing slightly with each subsequent object to guide visual flow.

B.4 Multi-Level Detail Template

Multi-Level Detail - Specific Prompt Instructions

- Place objects across exactly three spatial layers: foreground, midground, and background.
- Foreground objects must be 2× larger than midground objects.
- Midground objects must be 2× larger than background objects.
- Each layer must use a unique lightness or brightness level.
- Position layers vertically: foreground at the bottom, background at the top.
- Avoid horizontal alignment across layers to emphasize separation.

B.5 Grid Layout Template

Grid Layout - Specific Prompt Instructions

- Choose a 2×2 or 3×3 grid structure, depending on object count.
- Place one object per cell, centered precisely.
- Use equal-sized cells with clearly defined, thick borders.
- Ensure all objects are the same size and prominence.
- Maintain at least 25% margin around each object within its cell.
- Do not allow diagonal, overlapping, or asymmetrical arrangements.

B.6 General Accessibility Constraints

All templates shared the following baseline requirements to ensure cognitive accessibility:

- **Object count control:** 3–5 distinct objects per image to avoid cognitive overload
- **Spatial separation:** Minimum spacing requirements to enhance visual clarity
- **Content restrictions:** Exclusion of text, numbers, abstract elements, or cultural bias
- **Background simplicity:** Plain or neutral backgrounds to minimize distraction
- **Visual equality:** Balanced prominence across objects unless explicitly specified

C Example JSON Entry

C.1 Example from the Dataset

This appendix shows the structure of our compiled text-to-image TS corpus. Each entry contains the original and simplified sentence pairs along with metadata for traceability and analysis. This is an example from the final dataset of the 96th entry sampled from the ASSET corpus:

Dataset Structure Example

```
{
  "id": "asset_069",
  "dataset": "ASSET",
  "domain": "Wikipedia",
  "original": "The Odyssey is an ancient Greek epic poem attributed to Homer.",
  "simplified": "The Odyssey is an old Greek poem about Homer.",
  "length_original": 12,
  "length_simplified": 10
}
```

C.2 Multiple Style Prompts for Single Sentence

This appendix illustrates how our template-based framework generates style-specific prompts from a single simplified sentence. The example demonstrates the systematic application of accessibility constraints across ten distinct visual styles while maintaining semantic consistency.

Multiple Style Prompts for Single Sentence

```
{
  "index": 71,
  "id": "wikipedia_387",
  "simplified_text": "Originally, a pie made with any kind of meat and mashed potato was called a cottage pie.",
  "dataset_source": "Wikipedia",
  "template_prompts": [
    {
      "style": "Cartoon",
      "prompt": "Generate a cartoon-style image with a light gray background. Include four distinct objects: A whole, uncooked piece of meat (such as a steak or a chicken drumstick). A knife and a fork, indicating the meat is ready to be cut and cooked. A bowl of raw, unpeeled potatoes. Each object should be scaled to similar sizes with no more than 10% variation..."
    },
    {
      "style": "Realistic",
      "prompt": "Create a realistic image with a light gray background, showcasing four distinct objects: A piece of raw meat (like a steak or chicken breast) symbolizing 'any kind of meat'. A fresh, whole potato. A bowl of mashed potatoes. A traditional cottage or small house..."
    },
    {
      "style": "Artistic",
      "prompt": "Produce an image in an Artistic style featuring the following elements in a clear and simple layout against a light-gray background: A classic pie dish, A piece of meat, A potato, A small cottage. Arrange with 30% spacing between objects..."
    },
    {

```

```

"style": "Minimalistic",
"prompt": "Generate a minimalistic image with four objects: an empty pie
dish, uncooked minced meat, ready-to-eat mashed potato, and a simple
cottage representation. The pie dish should be empty, the minced meat
should be uncooked and the mashed potato should look ready to eat.
Each object should be scaled to similar sizes with no more than 10%
variation..."
},
{
"style": "Digital Art",
"prompt": "Create a digital art image consisting of four distinct
objects: a pie, a piece of meat, a potato, and a simple cottage.
Arrange these objects on a light gray background with a minimum of
30% spacing between them. Maintain consistent sizing within 10%
variation..."
},
{
"style": "3D Rendered",
"prompt": "Generate an image in a 3D rendered style on a neutral
light-grey background. The image should contain four distinct
objects: a pie dish, raw meat, a potato, and a cottage. Apply 30%
minimum spacing and maintain size consistency within 10% variation..."
},
{
"style": "Geometric",
"prompt": "Generate a geometric style image composed of four distinct
objects: a circular pie, a slice of pie, a piece of meat, and a
mashed potato. Use simple geometric shapes and maintain 30% spacing
between elements..."
},
{
"style": "Retro",
"prompt": "Generate an image in a Retro style that depicts the following
scene: A whole pie with a distinguishable crust on a light gray
background. The pie should be depicted in a simple, stylized way with
clear boundaries and vintage aesthetic..."
},
{
"style": "Storybook",
"prompt": "Generate an image in a Storybook style on a solid light grey
background. The picture should include four distinct objects: a whole
uncooked pie, a separate piece of uncooked meat, a pile of mashed
potatoes, and a small cottage. Each object should be scaled to a
similar size with a 10% variation allowance..."
},
{
"style": "Technical",
"prompt": "Create an image on a light gray background. The image should
contain four distinct objects: a pie dish, a piece of meat, a potato,
and a mashed potato. Use technical illustration style with clear,
precise lines and minimal shading..."
}
]
}

```

This example demonstrates how the refined template specifications are consistently applied while allowing for style-specific adaptations. The same simplified text generates distinct visual approaches while maintaining consistent accessibility principles.

D Evaluation Questions

This appendix reproduces the full set of evaluation questions as presented to experts in our customized Label Studio interface.

Image Accessibility Evaluation Questions

This evaluation form is designed to assess the accessibility of images for people with cognitive disabilities. The questions focus on image clarity, simplicity, quality, and alignment with simplified text. Please provide thorough and thoughtful responses to help improve accessibility standards.

Image Characteristics

Very Low (0-3)	Low (4-6)	Medium (7-9)	High (10-12)	Very High (13-15)
-------------------	--------------	-----------------	-----------------	----------------------

Image Simplicity: (0-15)

Evaluate how easily viewers can understand what the image represents. Consider whether the main elements are clear, distinguishable, and convey their meaning without requiring extensive interpretation.

Image Quality: (0-15)

Assess the visual clarity, resolution, and professional appearance of the image. Consider whether the image looks polished, well-composed, and appropriate for educational materials.

Text Characteristics

Text Simplicity: (0-15)

Rate how easy the text is to read for people with cognitive disabilities. Focus on basic readability: short sentences, simple words, clear structure, and no complex language. A high score means the text uses plain language that most people can understand without help.

Text Quality: (0-15)

Rate how well the text delivers its message. Focus on content: accuracy, organization, completeness, and purpose. A high score means the text effectively teaches or explains its topic, regardless of the language level used.

Alignment and Ethics

Very Low (0-4)	Low (5-8)	Medium (9-12)	High (13-16)	Very High (17-20)
-------------------	--------------	------------------	-----------------	----------------------

Ethics: (0-20)

Evaluate whether the image is free from bias, stereotypes, or harmful content. Consider if the image represents people and concepts in a fair, respectful, and inclusive way that avoids reinforcing negative stereotypes.

Text-Image Alignment: (0-20)

Assess how well the image represents and supports the simplified text. Consider whether the image accurately illustrates the key concepts in the text and provides visual support that enhances understanding.

Additional Checks

Contains no text: (Yes/No)

Indicate whether the image is free of embedded text or captions. Select 'Yes' if the image contains no text elements, or 'No' if it includes text within the image itself.

Well-separated objects: (Yes/No)

Evaluate whether elements in the image are clearly distinguishable from each other. Select 'Yes' if objects are well-defined with clear boundaries, or 'No' if elements blend together or are difficult to distinguish.

Image Style (Select up to 3)

- 3D Rendered
- Artistic
- Cartoon
- Digital Art
- Geometric
- Minimalistic
- Realistic
- Retro
- Storybook
- Technical

Style Descriptions:

- **3D Rendered:** Computer-generated images with three-dimensional depth, shading, and perspective.
- **Artistic:** Stylized images with expressive, creative, or painterly qualities.
- **Cartoon:** Simplified, often outlined drawings with bright colors and exaggerated features.
- **Digital Art:** Images created using digital tools with a modern, clean appearance.
- **Geometric:** Composed primarily of basic shapes like circles, squares, and triangles.
- **Minimalistic:** Simple designs with limited elements, colors, and details.
- **Realistic:** Images that closely resemble photographs or real-life objects and scenes.
- **Retro:** Designs that mimic or reference visual styles from past decades.
- **Storybook:** Whimsical, illustrative style similar to children's book illustrations.
- **Technical:** Diagram-like images with precise lines and instructional qualities.

Concerns (Select all that apply)

- Discriminatory content
- Culturally insensitive
- Potentially triggering
- Too complex for target audience
- Misleading representation
- Poor accessibility

Concerns Descriptions:

- **Discriminatory content:** The image contains stereotypes, biases, or content that discriminates against groups based on race, gender, ability, etc.
- **Culturally insensitive:** The image inappropriately represents cultural elements or shows disrespect toward specific cultural practices or symbols.
- **Potentially triggering:** The image contains elements that might cause distress, anxiety, or traumatic responses in vulnerable viewers.
- **Too complex for target audience:** The image contains too many elements or complicated visual information that may confuse people with cognitive disabilities.
- **Misleading representation:** The image does not accurately represent the concept described in the text or could lead to misunderstanding.

• **Poor accessibility:** The image has characteristics that make it difficult to perceive or understand, such as low contrast, cluttered layout, etc.

Additional Notes

Please provide any additional feedback about this image's accessibility, clarity, or appropriateness for people with cognitive disabilities.

Document-level Simplification and Illustration Generation: Multimodal Coherence

Yuhang Liu^{1*} Mo Zhang^{1*} Zhaoyi Cheng¹ Sarah Ebling^{1†}

¹University of Zurich

{yuhang.liu3, mo.zhang, zhaoyi.cheng}@uzh.ch, ebling@cl.uzh.ch

Abstract

We present a novel multimodal system for document-level text simplification and automatic illustration generation, aimed at enhancing information accessibility for individuals with cognitive impairments. While prior research has primarily focused on sentence- or paragraph-level simplification, and text-to-image generation for narrative contexts, this work addresses the unique challenges of simplifying long-form documents and generating semantically aligned visuals. The pipeline consists of four stages: (1) Document-Level Text Simplification, (2) discourse-aware segmentation using large language models, (3) visually grounded description generation via abstraction, and (4) controlled image synthesis using state-of-the-art diffusion models, including DALL-E 3 and FLUX.1-dev. We further incorporate stylistic constraints to ensure visual coherence, and we conduct a human evaluation measuring comprehension, semantic alignment, and visual clarity. Experimental results demonstrate that our system effectively combines simplified text and visual content.

1 Introduction

Simplified language aims to enhance information accessibility for individuals with cognitive impairments, learning disabilities, and others who may have difficulty comprehending standard-language written texts (Bredel and Maaß, 2016). Existing research primarily focuses on transforming complex sentences or paragraphs into more comprehensible variants, particularly for readers with low literacy skills and non-native speakers (Al-Thanyyan and Azmi, 2021). While automatic text simplification technologies have made significant progress in recent years, the role of visual aids in supporting textual comprehension is comparatively underresearched. Studies have shown that incorporating

illustrations into simplified texts can further enhance understanding among individuals with cognitive disabilities (Lin et al., 2009; Winberg and Saletta, 2018; Sutherland and Isherwood, 2016). Most existing work has focused on sentence- or paragraph-level simplification and accompanying image generation (Zhang et al., 2024; Shou et al., 2023; Anschütz et al., 2024). With the growing contextual understanding capabilities of large language models (LLMs), document-level simplification and illustration generation for longer texts become possible. Our goal is to leverage these capabilities to build a pipeline for document-level text simplification and corresponding illustration generation, aiming to improve information comprehension for persons with cognitive impairments.

2 Related Work

In recent years, advances in natural language processing have significantly accelerated the development of automatic text simplification technologies. Concurrently, researchers have begun to explore multimodal approaches to further enhance the comprehensibility of simplified texts, especially through image generation. Illustrating text with images is an effective strategy to support comprehension. Visual elements not only help capture the reader’s attention but also concretize abstract concepts, thereby reducing cognitive load (Glenberg and Langston, 1992; Guo et al., 2020; Wang and Zewen, 2023). The recent progress in text-to-image generation (T2I) has made the automated realization of this idea increasingly feasible. Diffusion-based models (Ho et al., 2020) have emerged as the dominant paradigm in T2I and have achieved remarkable breakthroughs. State-of-the-art models such as OpenAI’s DALL-E 3 (Betker et al., 2023), Google’s Imagen 3 (Baldridge et al., 2024), and Stability AI’s Stable Diffusion 3 (Esser et al., 2024) can generate high-resolution, semantically relevant,

*Equal contribution.

†Corresponding author. **Email:** ebling@cl.uzh.ch

and visually creative images from complex textual descriptions. Existing research has demonstrated the use of DALL·E 3 and Stable Diffusion 3, to generate high-quality visual content that is semantically well-aligned with Easy-to-Read (E2R) textual materials (Anschütz et al., 2024). These advances are largely attributed to the models’ pretraining on massive text-image paired datasets, which enable them to learn nuanced mappings between textual semantics and visual representations.

Our work builds upon these powerful diffusion models as the foundation for visual generation. While most existing systems are still confined to sentence- or paragraph-level T2I generation, generating contextually appropriate illustrations from long-form documents requires document-level understanding and scene planning capabilities. With the rapid evolution of LLMs, new frameworks are emerging that allow for these models to manage long-document processing and orchestrate image generation. In such settings, LLMs serve as directors or scriptwriters that structure the narrative and guide visual synthesis (Gado et al., 2025; Leandro et al., 2024). However, these systems are primarily tailored for narrative storytelling. In contrast, generating illustrations for informational documents demands greater factual accuracy and lower semantic ambiguity. Our research aims to bridge this gap by developing a multimodal system that integrates an LLM with a diffusion-based image generation model, enabling more accurate document-level text simplification and illustration. The proposed system features a structure-aware text simplification module and a semantically aligned image generator. Through semantic optimization and cross-modal feedback mechanisms, our method enhances the coherence between text and images and improves cognitive accessibility.

3 Method

To generate visual content for complex documents, we propose a multi-stage generative pipeline. This pipeline first decomposes a document into semantically coherent units, then translates these units into visually grounded descriptions, and finally renders them into images. Our approach integrates the advanced capabilities of LLMs for complex text processing with state-of-the-art T2I models for visual synthesis. The entire framework consists of four key stages: (1) Document-Level Text Simplification, (2) semantic document segmentation, (3)

visually grounded description generation, and (4) controlled image generation.

3.1 Document-Level Text Simplification

For each discourse segment, GPT-4o rewrites the passage into Easy-to-Read German under fidelity constraints: preserve named entities, numbers, and domain terms; avoid deletions that remove obligations or eligibility; keep sentences short and syntax simple; and prohibit invented facts. Outputs are returned in JSON. We enforce tokenizer-aware limits (less than 20k characters per call) and run sanity checks for numeric consistency and entity preservation. Evaluation uses expert ratings on Simplicity (Q4), Semantic Adequacy (Q5), and Fluency (Q6) described in Section 5.

3.2 Semantic Document Segmentation

Real-world documents are rarely monolithic in topic; rather, they typically exhibit inherent discourse structures involving shifts in themes, scenes, or arguments (Grosz and Sidner, 1986). Long documents, often exceeding several thousand words, pose challenges for direct image generation, resulting in overgeneralization or omission of critical details in the images. To address this, our first step involves re-segmenting the source document into shorter sub-paragraphs, each expressing a single idea or thematic unit. Each segment is then paired with a corresponding illustration. Prior work in document-level text simplification has similarly emphasized the importance of managing information hierarchy and discourse structure, often through explicit structural analysis or summarization (Crippwell et al., 2023; Fang et al., 2025; Blinova et al., 2023). Our method automates this decomposition process, forming the foundation for downstream visual generation.

We employ GPT-4o as a zero-shot, discourse-aware segmenter. This decision is motivated by the emergent capabilities of LLMs to perform complex structural tasks without task specific fine-tuning. Unlike traditional unsupervised approaches that rely on shallow lexical cohesion signals, LLMs can exploit deep semantic and world knowledge to detect more nuanced topic boundaries (Mu et al., 2024). Using carefully designed prompts, we guide the model to function as an advanced textual analyzer. Because GPT-4o exhibits degraded performance when processing input text with long contexts, its effectiveness decreases as the length of the input increases (Karpinska et al., 2024;

Ma et al., 2024). To accommodate this context-length constraint, we implement a tokenizer-aware, sentence-preserving segmentation procedure that limits each model input to fewer than 20,000 characters. Specifically, we first compute the input length using a byte-pair-encoding (BPE) (Sennrich et al., 2016) tokenizer consistent with the model’s tokenization scheme; if the character count exceeds 20,000, we partition the document into sub-documents. For sentence boundary detection, we adopt “Segment any Text” (Frohmann et al., 2024). To ensure machine-readability and robust integration with downstream components, we enforce output in JSON format. Under our tokenizer-aware JSON prompting, the model produced consistent discourse segments with low formatting error rates, which was sufficient for downstream components.

3.3 Visually Grounded Description Generation

Narrative language in documents often differs significantly from the concrete, descriptive phrasing required by T2I models to generate high-quality images (Saharia et al., 2022). Using raw text snippets as prompts frequently results in vague or abstract outputs. To bridge this semantic-to-visual gap, we introduce an intermediate transformation step, which we conceptualize as cross-modal abstraction. The goal is to distill the essential, visually representable elements from each text segment. This aligns with the broader goals of multimodal learning, where shared representation spaces enable meaningful alignment between textual and visual modalities.

We again leverage GPT-4o, configuring it as a text-to-text transformation agent for this task. Using few-shot prompting, we embed examples that guide the model to learn the desired input-output mapping without parameter updates, a practical benefit in our setting (few-shot prompting without task-specific finetuning) (Zhang and Xu, 2024). The prompt explicitly decomposes the task into two steps: (1) internal summarization to extract key entities and actions, and (2) translation of this summary into a visual scene description. We adopt a two-step prompt (key-entity summary → scene description) to make the transformation explicit; this yielded clearer, more actionable descriptions for image generation in our pilot settings (Wei et al., 2022).

A crucial component of our prompting strategy is the imposition of faithfulness constraints. We

explicitly instruct the model to avoid hallucinations—i.e., adding objects, attributes, or details not present in the source text. By including directives such as “Do not invent or alter details not mentioned in the original text,” we aim to minimize semantic drift and ensure that the final image is a faithful visual rendering of the textual document content.

3.4 Controlled Image Generation

Once visually grounded descriptions are prepared, the final step is image synthesis. The choice of T2I model architecture and the stylization of prompts play pivotal roles in determining the visual clarity, aesthetic quality, and suitability of the generated content for the target audience. For instance, illustrations for children’s storybooks require an entirely different visual style than those in technical manuals. Recent advances in T2I, particularly diffusion models and their Transformer-based successors, have enabled unprecedented levels of photorealism and fine-grained style control (Betker et al., 2023).

In this study, we experiment with two state-of-the-art T2I models to explore architectural diversity: OpenAI’s DALL·E 3 and Black Forest Labs’s FLUX.1-dev. DALL·E 3 is renowned for its high fidelity to complex prompts, attributed largely to its use of a powerful language model to preprocess and enrich textual input prior to image generation (Betker et al., 2023). In contrast, FLUX.1-dev represents a new generation of diffusion/Transformer hybrids, making it a valuable comparative baseline.

To ensure stylistic consistency and interpretability across illustrations generated for the same document, we apply a stylistic modifier to each prompt. As an initial case study, we target the generation of “clear and concise cartoon-style illustrations.” This stylistic choice is widely recognized for its ability to convey information clearly by abstracting over details, making it particularly suitable for educational and simplified communication contexts. The final prompt structure for image generation is a simple concatenation: “A clear and concise cartoon-style illustration depicting: <VISUAL DESCRIPTION>”.

4 Experiments

4.1 Materials

We utilized the WebCorpus dataset (Battisti and Ebling, 2020), which is specifically designed for

automatic readability assessment and text simplification tasks in the German language. The corpus comprises approximately 6,200 documents and nearly 211,000 sentences, collected from the websites of governmental bodies, professional institutions, and non-profit organizations across Germany, Austria, and Switzerland, covering a total of 92 distinct domains. The data includes both HTML webpages and PDF files, with content dated between late 2018 and early 2019. In addition to providing parallel corpora and monolingual simplified German texts, the dataset is characterized by the preservation of text structure, typographic information, and embedded image content, which were structurally extracted using an HTML parser and PDFlib tools.

For our experiments, we constructed a specialized evaluation subset from the WebCorpus to test and assess our proposed method. We selected twelve parallel documents available in PDF format based on the following criteria:

Human-Generated Simplifications with Illustrations These documents not only contain simplified versions created by human experts but also include illustrations accompanying specific paragraphs in the simplified texts. As illustrated in Figure 1, these images are designed to visually explain or complement the core ideas of the corresponding textual segments.

Topical and Stylistic Diversity To ensure the generalizability of our evaluation, we deliberately selected documents covering a range of topics such as legal aid, public health guidelines, and social welfare application procedures. These documents also exhibit considerable variation in both the complexity of the source texts and the visual styles of the illustrations.

Although modest in scale, this twelve-document subset offers high-quality human annotations and rich internal diversity. It provides a rigorous and controlled experimental setting for end-to-end evaluation, allowing us to verify the full pipeline, from text segmentation and visual description generation to final image synthesis, and to conduct direct comparisons with the original simplified documents.

4.2 Experimental Procedure

The experimental procedure of this study adheres to the three-stage generation framework defined in Section 3, with the aim of validating the effectiveness of our proposed method in an end-to-

end manner. The implementation consists of the following steps: We first extracted the plain text (.txt) versions of twelve selected PDF documents from the WebCorpus dataset as the original input texts. Following the method outlined in Section 3.1, we employed GPT-4o as a document segmentation module to automatically decompose each document into a series of semantically coherent sub-paragraphs. For each sub-paragraph, we invoked the GPT-4o model again based on the visual anchoring description strategy described in Section 3.2. Through chain-of-thought reasoning, the model distilled the abstract narrative of each sub-paragraph into concrete, renderable visual scene descriptions. These descriptions were generated under faithfulness constraints to prevent hallucinations or factual distortions. The generated visual descriptions were fed into both the DALL-E 3 and FLUX.1-dev image generation models. To ensure stylistic consistency and interpretability of illustrations throughout the document, we prefixed each description with a prompt specifying “a clear and concise cartoon-style illustration,” as specified in Section 3.3. This approach yielded stylistically aligned images corresponding to each textual segment.

4.3 Evaluation

To conduct a comprehensive and reliable evaluation of the generated outputs, we employed expert human assessment. Four domain experts with backgrounds in simplified language were recruited to participate in the study. We developed a structured online questionnaire in which experts rated and qualitatively assessed the generated outputs. Original human-designed illustrations from the source documents were shown only as qualitative context when licensing permitted and were not included in the quantitative analysis. The evaluation focused on four key dimensions, including support for comprehension, semantic alignment with the input text, visual coherence, and style match. The primary goal of this assessment was to quantify the effectiveness of our method in terms of faithfulness, clarity, and aesthetic quality. The evaluation involved four text passages, each paired with two images generated by DALL-E 3 and FLUX.1-dev, resulting in eight image-text combinations. Three experimental conditions were considered: text only, text + image from FLUX.1-dev, and text + image from DALL-E 3. To ensure balanced exposure across conditions, we adopted a Latin square de-

sign.

Each evaluator completed a three-part evaluation. First, a brief pre-questionnaire collected background information such as years of professional experience and domain expertise. In the main evaluation, participants reviewed all conditions and rated four dimensions for each sample: (1) support for comprehension, (2) semantic alignment, (3) visual coherence, and (4) style match. The detailed definitions of the four dimensions in the Appendix A. Comprehension accuracy was additionally measured via multiple-choice questions as an objective check separate from the four subjective ratings.

Finally, a short post-questionnaire confirmed evaluators’ understanding of the task and the rating criteria. Responses were collected using 5-point Likert scales for subjective measures (e.g., alignment and simplicity) and accuracy scores for comprehension questions. Each evaluator was required to complete a total of 99 questions. These included 80 questions pertaining to the evaluation of eight images, 12 questions assessing simplified texts, three questions concerning evaluators’ background information, and four open-ended questions eliciting overall evaluations and feedback. The full survey design is provided in the appendix.

5 Results

5.1 Text Simplification Evaluation

We collected Likert ratings on a five-point scale from four expert evaluators along three criteria: Simplicity (Q4), Semantic Adequacy (Q5), and Fluency (Q6). Non-numeric “Other” entries were treated as missing for averaging and are reported separately in the distribution table. Table 1 summarizes, for each evaluator, the overall distribution of assigned scores across all texts and criteria, including a separate count for “Other” and the total numeric score. Table 2 reports per-criterion means with the effective sample size for each evaluator, together with the total numeric score and the overall mean across all available numeric ratings.

Fluency shows consistently high evaluations. Evaluator 4 attains the highest fluency mean and also the highest total score and overall mean. Simplicity varies more strongly by evaluator. Evaluator 3 tends to assign higher simplicity with a mean of 4.00, while Evaluator 2 assigns lower values with a mean of 2.25, which suggests different expectations for ease of reading. Semantic adequacy concentrates around mid to high values, with Eval-

uator 4 again providing the most favorable adequacy judgments. “Other” responses appear only for Evaluator 3 on Q6 and are excluded from all mean calculations by design. In aggregate, the results indicate robust perceived fluency, moderate to high semantic adequacy, and evaluator-dependent variation in perceived simplicity.

ID	1	2	3	4	5	Other	Total score
1	0	3	3	5	1	0	40
2	3	1	2	6	0	0	35
3	0	2	3	4	0	3	29
4	0	3	0	4	5	0	47

Table 1: Per-rater distribution of ratings across all texts and all three criteria. Columns “1–5” give counts of numeric ratings. “Other” counts non-numeric entries. “Total score” sums all numeric ratings for the rater.

ID	Q4 mean	Q5 mean	Q6 mean	Total score	Overall mean
1	2.75	3.00	4.25	40	3.33
2	2.25	2.50	4.00	35	2.92
3	4.00	2.50	3.00	29	3.22
4	3.00	4.00	4.75	47	3.92

Table 2: Per-rater means by criterion (Q4–Q6), along with each rater’s total score (sum of all numeric ratings over Q4–Q6 and all texts) and overall mean (average of all available numeric ratings).

5.2 Illustration Generation Evaluation

We collected 64 image-related Likert responses per rater. Covering semantic alignment, support for comprehension, visual coherence, and stylistic appropriateness, under two models (FLUX.1-dev and DALL•E 3). Table 3 summarizes the per-rater tallies by score category, including non-numeric Others. Ratings for both systems concentrated in the 3–4 range, indicating that most illustrations were perceived as broadly supportive yet rarely exceptional; 5s were occasional, while 1s were rare. Raters 1–2 slightly preferred FLUX.1-dev (total differences within six points), whereas Raters 3–4 favored DALL•E 3, with Rater 4 assigning thirteen maximum scores to DALL•E 3.

Across all evaluators, ratings for both image generation systems were concentrated in the 3–4 range, with score 4 being the most frequently assigned. This indicates that the majority of generated images were perceived as broadly supportive of text comprehension and semantically adequate, but rarely outstanding. Scores of 5 were assigned only occasionally, reflecting the fact that few images were

ID	Model	Total	1	2	3	4	5	Others	Mean
1	flux.1-dev	92	6	6	7	12	1	0	3.29
	DALL-E 3	90	4	9	9	9	1	0	3.26
2	flux.1-dev	74	6	7	6	9	0	4	2.96
	DALL-E 3	69	8	9	5	7	0	3	2.84
3	flux.1-dev	58	11	5	7	4	0	5	2.39
	DALL-E 3	69	7	9	8	5	0	3	2.91
4	flux.1-dev	105	0	7	7	10	6	2	3.75
	DALL-E 3	116	2	9	1	7	13	0	3.97

Table 3: Aggregated Likert ratings (1–5) and non-numeric responses (*Others*) for image evaluation tasks across four evaluators, including the mean numeric score per model.

judged as fully satisfactory across all evaluative dimensions. Conversely, scores of 1 were rare, suggesting that completely inadequate outputs occurred only sporadically. Differences across evaluators were evident. Evaluators 1 and 2 awarded higher totals to flux.1-dev, the score difference between the two models was relatively small, with the total discrepancy remaining within six points. In contrast, Evaluators 3 and 4 assigned higher totals to DALL-E 3, with Evaluator 4 giving thirteen maximum ratings (5), far more than for flux.1-dev.

The questionnaire design allowed ratings to capture multiple facets of image quality. We conducted a closer comparison of evaluators’ ratings across these different dimensions and found consistent patterns. On average, both systems received the highest scores on supporting text comprehension and semantic alignment with the text, where the majority of judgments fell between 3 and 4. By contrast, lower ratings (1–2) were more frequently observed in dimensions such as visual coherence and stylistic appropriateness, reflecting instances where images were perceived as misaligned in style or insufficiently coherent, even if they captured the general semantics of the text. This distribution indicates that the models were more successful in generating images that conveyed the intended meaning than in ensuring stylistic naturalness and visual consistency. In analyzing the responses to the question “Do you think this image was created by AI or manually?”, we found that most evaluators misclassified the images generated by the FLUX.1-dev model as manually created rather than

AI-generated. By contrast, the majority of evaluators correctly identified the images produced by DALL-E 3 as AI-generated. We also found a consistent issue emerged across specific combinations of the images with many texts, as shown in Figure 1. In these cases, all four experts independently highlighted the same concern: When the generated images contained too much text or overly intricate visual patterns, they became cognitively overwhelming.

6 Discussion and Conclusion

Our study indicates that automatic simplification produced texts that readers judged as fluent and largely faithful to source meaning, while perceived ease-of-reading is more sensitive to individual evaluator standards. On the visual side, Both image generation models were generally effective in providing semantically supportive illustrations, though their perceived utility varied across evaluators and dimensions. Ratings tended to cluster around the mid-scale (3–4), suggesting that while AI-generated visuals achieved a baseline adequacy, they rarely reached the level of high-quality human-created illustrations. This implies that the models captured textual meaning with reasonable reliability but seldom produced outputs regarded as exemplary in terms of clarity, stylistic appropriateness, or overall communicative effectiveness. Divergent preferences among evaluators, some favoring FLUX.1-dev and others DALL-E 3 highlight that judgments of quality are not solely determined by semantic accuracy, but are also shaped by individual aesthetic expectations and tolerance for stylistic variation. Particularly noteworthy was the tendency of participants to misclassify outputs from FLUX.1-dev as manually created, suggesting that its visual naturalness may enhance perceived authenticity. While such naturalness is promising for accessibility and engagement, it also raises potential concerns about transparency and user trust in contexts where it is important to distinguish human- from machine-generated content.

Qualitative comments about images containing too much text point to avoidable extraneous cognitive load. Images that incorporated excessive textual content or visually dense layouts were often judged as distracting, thereby diminishing rather than enhancing comprehension. This observation underscores the importance of maintaining visual simplicity, especially when designing for audi-

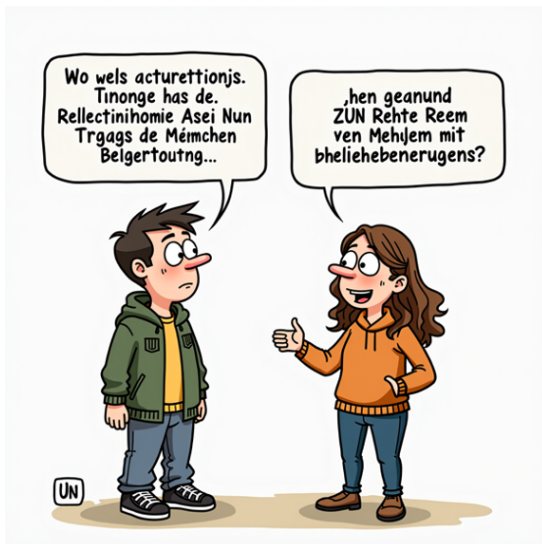


Figure 1: Examples of undesirable illustrations, where images contain excessive textual elements that make them confusing for the viewer.

ences with cognitive impairments. For such users, the trade-off between semantic fidelity and visual clarity becomes particularly critical; cluttered or overly detailed images may undermine the intended benefits of simplified language materials. These findings point to the necessity of introducing explicit design constraints in text-to-image workflows. Mechanisms such as filtering strategies, prompt engineering techniques that enforce minimalism, or post-processing methods to eliminate superfluous elements could help align outputs more effectively with accessibility goals.

The pattern of high fluency and adequate semantics in text, combined with mid-scale visual ratings, points to a practical synthesis: when textual simplification reliably preserves meaning and flow, illustrations function best as lightweight scaffolds rather than dense carriers of information. Simplified text can shoulder the primary communicative load, while images should reinforce key entities, relations, or processes without introducing visual clutter. This aligns with our finding that evaluators penalize visually dense layouts: if text is already fluent and semantically adequate, adding heavy captioning or intricate scene details may yield diminishing returns or even harm comprehension. Therefore, downstream design should prioritize (i) simplicity-first visual layouts, (ii) restrained use of textual overlays inside images, and (iii) explicit alignment between each image and a small set of core propositions in the simplified text.

Future research should extend beyond expert-based evaluations to incorporate direct feedback from end users, particularly individuals with cognitive impairments, in order to ensure that the generated visuals truly enhance comprehension and accessibility.

7 Limitations

While this study provides valuable evidence for the role of AI-generated images in supporting simplified text comprehension, several limitations remain. (1) The language scope of this study was limited to German, and all texts were drawn from a single corpus of simplified expository materials. This may constrain the generalizability of our findings to other languages, genres, or cultural contexts. Expanding to multilingual or narrative datasets could uncover additional design considerations. (2) Visual complexity was identified as a recurring issue, but our analysis relied on qualitative judgments rather than formal cognitive load metrics. The absence of behavioral or physiological measures (e.g., comprehension scores, reading time, or gaze data) limits our ability to precisely quantify the cognitive effects of visual detail. (3) The image generation models used (DALL-E 3 and FLUX.1-dev) were not fine-tuned for the simplified language setting or for accessibility-related constraints. As a result, the outputs may occasionally include dense textual overlays or unnecessary visual embellishments. Future research should explore prompt engineering

techniques and post-processing methods to explicitly control for simplicity and semantic salience.

8 Lay Summary

This paper presents a practical method to make long and complex documents easier to understand, especially for readers with cognitive impairments. Unlike most tools that simplify single sentences, our approach operates at the document level and adds supportive illustrations. First, a large language model (GPT-4o) divides a document into short, coherent segments and rewrites them in simpler language while preserving meaning. Next, the model drafts faithful visual descriptions for each segment, and state-of-the-art image generators (DALL-E 3 and FLUX.1-dev) produce clear, consistent cartoon-style illustrations that align with the simplified text.

We evaluated the pipeline on a curated subset of real German public-information documents and asked four experts in simplified language to review the outputs. Their ratings clustered around the middle of the 5-point scale, indicating that the images generally helped comprehension but were not uniformly excellent. Reviewers also noted that illustrations overloaded with on-image text can increase cognitive load and reduce clarity, underscoring the value of minimal, consistent visuals.

Pairing document-level simplification with faithful, stylistically coherent illustrations appears promising for making public-facing materials, such as health guidance or social-service instructions more accessible. Future work will expand user studies with target populations and further constrain visual design to keep images simple, readable, and trustworthy.

References

- Suha S Al-Thanyyan and Aqil M Azmi. 2021. [Automated text simplification: A survey](#). *ACM Computing Surveys (CSUR)*, 54(2):1–36.
- Miriam Anschütz, Tringa Sylaj, and Georg Groh. 2024. Images speak volumes: User-centric assessment of image generation for accessible communication. *arXiv preprint arXiv:2410.03430*.
- Jason Baldridge, Jakob Bauer, Mukul Bhutani, Nicole Brichtova, Andrew Bunner, Lluís Castrejon, Kelvin Chan, Yichang Chen, Sander Dieleman, Yuqing Du, and 1 others. 2024. Imagen 3. *arXiv preprint arXiv:2408.07009*.
- Alessia Battisti and Sarah Ebling. 2020. [A corpus for automatic readability assessment and text simplification of German](#). In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*, pages 3306–3314, Marseille, France. European Language Resources Association.
- James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, and 1 others. 2023. Improving image generation with better captions. *Computer Science*. <https://cdn.openai.com/papers/dall-e-3.pdf>, 2(3):8.
- Sofia Blinova, Xinyu Zhou, Martin Jaggi, Carsten Eickhoff, and Seyed Ali Bahrainian. 2023. Simsum: Document-level text simplification via simultaneous summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9927–9944.
- Ursula Bredel and Christiane Maaß. 2016. *Leichte Sprache: Theoretische Grundlagen? Orientierung für die Praxis*. Duden.
- Liam Cripwell, Joël Legrand, and Claire Gardent. 2023. Document-level planning for text simplification. In *17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 993–1006. Association for Computational Linguistics.
- Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, and 1 others. 2024. [Scaling rectified flow transformers for high-resolution image synthesis](#). In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*. PMLR.
- Dengzhao Fang, Jipeng Qiang, Yi Zhu, Yunhao Yuan, Wei Li, and Yan Liu. 2025. Progressive document-level text simplification via large language models. *arXiv preprint arXiv:2501.03857*.
- Markus Frohmann, Igor Sterner, Ivan Vulić, Benjamin Minixhofer, and Markus Schedl. 2024. Segment any text: A universal approach for robust, efficient and adaptable sentence segmentation. *arXiv preprint arXiv:2406.16678*.
- Mohamed Gado, Towhid Taliee, Muhammad Memon, Dmitry Ignatov, and Radu Timofte. 2025. Vist-gpt: Ushering in the era of visual storytelling with llms? *arXiv preprint arXiv:2504.19267*.
- A. Glenberg and William E. Langston. 1992. [Comprehension of illustrated text: Pictures help to build mental models](#). *Journal of Memory and Language*, 31:129–151.
- Barbara J Grosz and Candace L Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational linguistics*, 12(3):175–204.

- Daibao Guo, Shuai Zhang, Katherine Landau Wright, and E. McTigue. 2020. [Do you get the picture? a meta-analysis of the effect of graphics on reading comprehension](#). *AERA Open*, 6.
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851.
- Marzena Karpinska, Katherine Thai, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. [One thousand and one pairs: A "novel" challenge for long-context language models](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- Jorge Leandro, Sudha Rao, Michael Xu, Weijia Xu, Nebojsa Jojic, Chris Brockett, and Bill Dolan. 2024. Geneva: Generating and visualizing branching narratives using llms. In *2024 IEEE Conference on Games (CoG)*, pages 1–5. IEEE.
- Y. Lin, Ting-Fang Wu, Ya-Hui Tasi, Hui-Ching Chen, and Ming-Chung Chen. 2009. [The effect of different representations on reading digital text for students with cognitive disabilities](#). *Br. J. Educ. Technol.*, 40:764–770.
- Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiao wen Dong, Pan Zhang, Liangming Pan, Yu-Gang Jiang, Jiaqi Wang, Yixin Cao, and Aixin Sun. 2024. [Mmlongbench-doc: Benchmarking long-context document understanding with visualizations](#). *ArXiv*, abs/2407.01523.
- Yida Mu, Chun Dong, Kalina Bontcheva, and Xingyi Song. 2024. [Large language models offer an alternative to the traditional approach of topic modelling](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*, pages 10160–10171.
- Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, and 1 others. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. [Neural machine translation of rare words with subword units](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.
- Mike Zheng Shou, Di Zhang, Chunhua Shen, Zhongyuan Wang, Lele Cheng, Yan Li, Weijia Wu, Yefei He, Tingting Gao, and Zhuang Li. 2023. [Paragraph-to-image generation with information-enriched diffusion model](#). *ArXiv*, abs/2311.14284.
- R. Sutherland and T. Isherwood. 2016. [The evidence for easy-read for people with intellectual disabilities: A systematic literature review](#). *Journal of Policy and Practice in Intellectual Disabilities*, 13:297–310.
- Yuanzhe Wang and Zheng Zewen. 2023. [An eye-tracking based study: The role of images and explanatory texts in reading comprehension](#). *International Journal of Frontiers in Sociology*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.
- Jennifer Winberg and Meredith Saletta. 2018. [Leveled texts for adults with intellectual or developmental disabilities: A pilot study](#). *Focus on Autism and Other Developmental Disabilities*, 34:118 – 127.
- Chao Zhang and Shangqing Xu. 2024. [Misconfidence-based demonstration selection for llm in-context learning](#). *ArXiv*, abs/2401.06301.
- Yangqianhui Zhang, Qinghao Hu, Liang Zhao, Chunjiang Fu, Tengtu Chen, and Pingda Huang. 2024. [Multi-sentence complementarily generation for text-to-image synthesis](#). *IEEE Transactions on Multimedia*, 26:8323–8332.

A Operational Definitions.

We use the following operational definitions for rated dimensions: *support for comprehension* = perceived helpfulness of the image for understanding the passage’s main message; *semantic alignment* = fidelity of depicted entities, attributes, and relations to the source passage; *visual coherence* = absence of distracting artifacts or clutter and a legible, uncluttered composition within the image; *style match* = appropriateness and consistency of the visual style with the document’s genre and Easy-to-Read conventions.

B Prompt Templates

We include here the prompts used for text simplification, text segmentation and visual description generation. All prompts are designed to enforce output constraints (JSON formatting) and minimize semantic drift in multimodal generation.

B.1 Prompt for Text Simplification

You are a professional editor specializing in document-level text simplification for broader accessibility.
Your goals are: (1) preserve meaning and factual correctness, (2) increase readability and accessibility, and (3) maintain discourse-level coherence across sections.

Follow the steps and constraints below exactly. Do not hallucinate or omit essential information.

0) Controls (adjust these before running)
Target audience: general adult readers without domain expertise.
Readability target: approximately B1–B2 (plain language); avoid jargon unless defined.
Lexical simplicity: prefer high-frequency, concrete words; define any necessary technical terms briefly.
Syntactic simplicity: prefer simple main clauses; split long/complex sentences (>25–30 words).
Style: neutral, clear, consistent; no rhetorical questions; active voice where appropriate.

1) Plan at the Document Level (no output yet)
Produce a hidden plan to guide rewriting (do not include the plan in the final answer):
Section map: list sections/paragraphs and their main points.
Entity & timeline register: people, organizations, quantities, dates, and events; ensure consistency of names/abbreviations across the document.
Discourse links: for each section, note how it connects to the previous one (cause → effect, problem → solution, comparison, sequence, contrast).
Risk items: legal/medical/financial claims; numbers, percentages, dates, and units that must remain exact.

2) Rewrite Rules (apply throughout)
Meaning preservation: keep all factual statements, numbers, dates, and units; do not invent content; do not change scope or evidential hedges.
Sentence-level operations: (a) split long sentences; (b) delete redundancy and filler; (c) paraphrase rare idioms and nominalizations; (d) reorder for subject–verb proximity.
Lexical operations: replace rare words with common alternatives; define unavoidable terms in-line the first time they appear.
Coreference & cohesion: resolve ambiguous pronouns; repeat a short, clear noun phrase when needed; add explicit connectives (e.g., “However,” “As a result,” “In addition”) to preserve coherence across sentences and sections.
Structure: keep informative headings; convert dense lists into bullets or tables where it improves clarity; keep citations/references but simplify their surrounding prose.
Safety & integrity: never remove warnings, limitations, or risk qualifiers; never alter quoted material; keep figure/table references consistent.

3) Self-Review Checklist (enforce before finalizing)
Confirm all items; if any fail, revise and re-check:
Meaning preservation: Each paragraph answers the same questions as the source; all numbers/dates/units/entities match the original.
Readability & simplicity: Average sentence length reduced; complex clauses minimized; jargon defined or replaced.
Document coherence: Section openings include bridging phrases; topic flow is consistent; pronouns are unambiguous.
Style consistency: Tone and tense are consistent; active voice is used where natural; no rhetorical filler.

Table 4: Prompt used for text simplification

B.2 Prompt for Text Segmentation

This prompt guides the model to identify thematic and semantic boundaries in expository texts and return machine-readable subparagraphs. The output format is constrained to valid JSON to ensure compatibility with downstream modules in our pipeline.

```
Please analyze the following text and split it into coherent subparagraphs based on thematic and semantic boundaries. Follow these rules strictly:  
1. Output MUST be valid JSON format only  
2. Use numbered keys starting from "1"  
3. Ensure all strings are properly quoted  
4. Escape any internal double quotes  
5. Do NOT include any additional text or explanations  
6. Maintain original content integrity  
  
Text to process: {text}  
  
Output format example:  
{ "1": "first subparagraph text", "2": "second subparagraph text" }
```

Table 5: Prompt used for text segmentation

B.3 Prompt for Visual Description Generation

This prompt constrains the model to produce faithful visual descriptions of the segmented subparagraphs. A key element is the explicit instruction to avoid hallucination, ensuring that no visual elements are introduced beyond the source text.

```
Please generate visual descriptions for each subparagraph following these steps:  
1. Create a brief summary highlighting main content  
2. Convert summaries into visual descriptions suitable for image generation. Ensure that the visual description faithfully represents the original text without adding or altering objects, attributes, or details not present in the source. Maintain semantic accuracy while simplifying the expression for better clarity.  
3. Return ONLY JSON with subparagraph numbers and visual description. Do not include any additional text or explanations.  
  
Output format example:  
{ "1": "Peaceful countryside with green fields and cottages", "2": "Busy city street with neon lights" }
```

Table 6: Prompt used for visual description generation

C Example of WebCorpus

As part of the WebCorpus dataset, we include authentic examples of German documents written in *Leichte Sprache* (easy-to-read German). One representative source is the newsletter series *Bericht aus Genf* ¹.

¹<https://www.bodys-wissen.de/bericht-aus-genf.html>

Example of simplified German text from WebCorpus

Bericht aus Genf Nr. 8 / 2014 Newsletter von Theresia Degener Mitglied im Ausschuss für den UN-Vertrag über die Rechte von Menschen mit Behinderungen Begrüßung Dieser Info-Brief ist über die 12. Sitzung von unserer Arbeits-Gruppe in Genf. Vor dem Treffen habe ich gedacht: Das ist das letzte Mal für mich. Die Mitglieder in der Arbeits-Gruppe arbeiten immer 4 Jahre mit. Und ich bin schon 4 Jahre dabei. Aber im Juni ist etwas Schönes passiert: Es waren Wahlen für die Arbeits-Gruppe. Und ich wurde wieder-gewählt. Das bedeutet: Ich darf noch einmal 4 Jahre in der Arbeits-Gruppe mitmachen. Darüber freue ich mich sehr. Seit November gibt es eine Sonder-Bericht-Erstatteerin für die Rechte von Menschen mit Behinderungen. Sie arbeitet für den Menschen-Rechts-Rat bei den Vereinten Nationen. Das ist die Aufgabe von der Sonder-Bericht-Erstatteerin: Sie schreibt Berichte für den Menschenrechts-Rat: Wie geht es Menschen mit Behinderungen auf der ganzen Welt. Die Sonder-Bericht-Erstatteerin heißt: Catalina Devantas. Sie kennt sich sehr gut aus mit Menschen-Rechten. Sie ist Rollstuhl-Fahrerin und Mutter von drei Kindern. Auf dem Bild bin ich mit Catalina Devantas und ihren Kindern und mit Silvia Judith Quan-Chang. Sie ist auch in der Arbeits-Gruppe. Bei der 12. Sitzung hatten wir einen Gast aus Mexiko. Er heißt Ricardo Adair Rubles. Er ist ein Mann mit Lern-Schwierigkeiten. Ricardo Adair Rubles ist sehr mutig. Das hat er der Arbeits-Gruppe erzählt: Er kämpft gegen ein Gesetz in Mexiko. In dem Gesetz steht: Menschen mit Lern-Schwierigkeiten dürfen nicht selbst entscheiden. Das ist schlecht. Jeder Mensch darf selbst entscheiden. Manchmal braucht ein Mensch Unterstützung. Dann muss der Mensch die Unterstützung bekommen. Ich wünsche Ihnen eine schöne Weihnachts-Zeit und ein gutes Jahr 2015! Theresia Degener. Ein Tipp zum Lesen: Der Bericht ist in Leichter Sprache. Aber manchmal stehen auch schwere Wörter in dem Bericht. Diese Wörter sind dann dick und blau. Die Wörter werden im Text erklärt.

Table 7: An example of simplified German text, extracted from the first two pages of Bericht aus Genf 8 Leichte Sprache.pdf

D Segmentation and Visual Descriptions for the WebCorpus Example

This appendix reports the output of our processing pipeline applied to the simplified German document in Appendix B.

ID	Segment (Simplified German)	Visual Description
1	Bericht aus Genf Nr. 8 / 2014 Newsletter von Theresia Degener Mitglied im Ausschuss für den UN-Vertrag über die Rechte von Menschen mit Behinderungen Begrüßung Dieser Info-Brief ist über die 12. Sitzung von unserer Arbeits-Gruppe in Genf. Vor dem Treffen habe ich gedacht: Das ist das letzte Mal für mich. Die Mitglieder in der Arbeits-Gruppe arbeiten immer 4 Jahre mit. Und ich bin schon 4 Jahre dabei. Aber im Juni ist etwas Schönes passiert: Es waren Wahlen für die Arbeits-Gruppe. Und ich wurde wieder-gewählt. Das bedeutet: Ich darf noch einmal 4 Jahre in der Arbeits-Gruppe mitmachen. Darüber freue ich mich sehr.	Theresia Degener freut sich über ihre Wiederwahl in die Arbeits-Gruppe.
2	Seit November gibt es eine Sonder-Bericht-Erstatteerin für die Rechte von Menschen mit Behinderungen. Sie arbeitet für den Menschen-Rechts-Rat bei den Vereinten Nationen. Das ist die Aufgabe von der Sonder-Bericht-Erstatteerin: Sie schreibt Berichte für den Menschenrechts-Rat: Wie geht es Menschen mit Behinderungen auf der ganzen Welt. Die Sonder-Bericht-Erstatteerin heißt: Catalina Devantas. Sie kennt sich sehr gut aus mit Menschen-Rechten. Sie ist Rollstuhl-Fahrerin und Mutter von drei Kindern. Auf dem Bild bin ich mit Catalina Devantas und ihren Kindern und mit Silvia Judith Quan-Chang. Sie ist auch in der Arbeits-Gruppe.	Catalina Devantas, eine Rollstuhlfahrerin und Mutter, arbeitet als Sonder-Bericht-Erstatteerin.
3	Bei der 12. Sitzung hatten wir einen Gast aus Mexiko. Er heißt Ricardo Adair Rubles. Er ist ein Mann mit Lern-Schwierigkeiten. Ricardo Adair Rubles ist sehr mutig. Das hat er der Arbeits-Gruppe erzählt: Er kämpft gegen ein Gesetz in Mexiko. In dem Gesetz steht: Menschen mit Lern-Schwierigkeiten dürfen nicht selbst entscheiden. Das ist schlecht. Jeder Mensch darf selbst entscheiden. Manchmal braucht ein Mensch Unterstützung. Dann muss der Mensch die Unterstützung bekommen.	Ricardo Adair Rubles kämpft mutig gegen ein Gesetz in Mexiko.
4	Ich wünsche Ihnen eine schöne Weihnachts-Zeit und ein gutes Jahr 2015! Theresia Degener Ein Tipp zum Lesen: Der Bericht ist in Leichter Sprache. Aber manchmal stehen auch schwere Wörter in dem Bericht. Diese Wörter sind dann dick und blau. Die Wörter werden im Text erklärt.	Theresia Degener wünscht frohe Weihnachten und gibt einen Lesetipp.

Table 8: Segmentation (IDs 1–4) and automatically generated visual descriptions for the simplified German text from appendix B

E Generated Images from Visual Descriptions

This appendix presents generated images corresponding to the visual descriptions in Appendix C. We compare two text-to-image models, DALL·E 3 and FLUX.1-dev, across four descriptions (IDs 1–4).

DALL·E 3



FLUX.1-dev

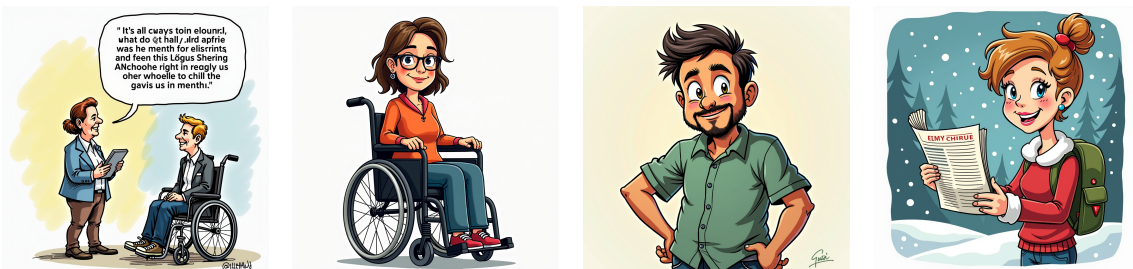


Figure 2: Comparison grid of generated images using DALL E 3 and FLUX.1-dev for four visual descriptions from Appendix C (IDs 1–4). Columns map left-to-right to description IDs 1, 2, 3, and 4.

F Questionnaire for Image Evaluation

This appendix presents the full bilingual questionnaire used to evaluate AI-generated and manually created images accompanying simplified texts. It consists of a brief pre-questionnaire about participants' background, a main questionnaire covering text comprehension, text-image alignment, and image quality, followed by a short post-study comparison and an open feedback item. Items explicitly marked for conditions (b) and (c) apply only when a text is shown with an image. Items without such a marker apply to all presentation conditions, including text-only. For check items, respondents select exactly one option unless otherwise stated; free-text fields are provided for comments.

Legend of Conditions / Legende der Bedingungen. (a) Text only / Nur Text (b) Text + AI-generated image / Text + KI-generiertes Bild (c) Text + manually created image / Text + manuell erstelltes Bild

Pre-Questionnaire / Vorbefragung

Question 1 / Frage 1. How many years of experience do you have working as a simplified language expert? / Wie viele Jahre Erfahrung haben Sie als Expert:in für vereinfachte Sprache? (select one / eine Option wählen)

- 0–1 years / 0–1 Jahr
- 1–2 years / 1–2 Jahre
- 3–5 years / 3–5 Jahre
- More than 5 years / Mehr als 5 Jahre
- Other (please specify) / Andere (bitte angeben): _____

Question 2 / Frage 2. How many years of experience do you have evaluating images as part of simplified language? / Wie viele Jahre Erfahrung haben Sie in der Bewertung von Bildern im Kontext vereinfachter Sprache? (select one / eine Option wählen)

- 0–1 years / 0–1 Jahr
- 1–2 years / 1–2 Jahre
- 3–5 years / 3–5 Jahre
- More than 5 years / Mehr als 5 Jahre
- Other (please specify) / Andere (bitte angeben): _____

Question 3 / Frage 3. What is your work setting in this field? / In welchem Arbeitsverhältnis sind Sie in diesem Bereich tätig? (select one / eine Option wählen)

- Freelancer / Freiberuflich
- Employee at a company / Angestellt in einem Unternehmen
- Employee at a research institute / Angestellt in einem Forschungsinstitut
- Employee at an association / public sector organization / Angestellt in einem Verband / einer öffentlichen Einrichtung
- Other (please specify) / Andere (bitte angeben): _____

**Main Questionnaire: Evaluation of AI-Generated Images and Text-Image Combination /
Hauptfragebogen: Bewertung von KI-generierten Bildern und Text-Bild-Kombinationen**

Section 1 / Abschnitt 1: Overall Text Comprehension / Textverständnis (all conditions a–c / alle Bedingungen a–c) Question 4 / Frage 4. How simple is this text? / Wie einfach ist dieser Text?

- 1 = Very difficult / Sehr schwierig
- 2 = Somewhat difficult / Eher schwierig
- 3 = Neutral / Neutral
- 4 = Somewhat easy / Eher einfach
- 5 = Very easy / Sehr einfach
- Other (please specify) / Andere (bitte angeben): _____

Question 5 / Frage 5. Is the text semantically adequate? / Ist der Text semantisch angemessen?

- 1 = Not at all / Überhaupt nicht
- 2 = Mostly not / Meistens nicht
- 3 = Partially / Teilweise
- 4 = Mostly / Größtenteils
- 5 = Completely / Vollständig
- Other (please specify) / Andere (bitte angeben): _____

Question 6 / Frage 6. Is the text fluent / grammatical? / Ist der Text flüssig / grammatikalisch korrekt?

- 1 = Not at all / Überhaupt nicht
- 2 = Mostly not / Meistens nicht
- 3 = Partially / Teilweise
- 4 = Mostly / Größtenteils
- 5 = Completely / Vollständig
- Other (please specify) / Andere (bitte angeben): _____

Section 2 / Abschnitt 2: Text-Image Alignment / Text-Bild-Übereinstimmung (conditions b–c / Bedingungen b–c) Question 7 / Frage 7. Does the image enhance the understanding of the text? / Unterstützt das Bild das Verständnis des Textes?

- 1 = Not at all / Überhaupt nicht
- 2 = Mostly not / Meistens nicht
- 3 = Partially / Teilweise
- 4 = Mostly / Größtenteils
- 5 = Completely / Vollständig
- Other (please specify) / Andere (bitte angeben): _____

Question 8 / Frage 8. How well does the text align with the image (meaning, message)? / Wie gut stimmt der Text in Bedeutung und Botschaft mit dem Bild überein?

- 1 = Not aligned at all / Überhaupt nicht übereinstimmend
- 2 = Mostly not aligned / Meistens nicht übereinstimmend
- 3 = Partially aligned / Teilweise übereinstimmend
- 4 = Mostly aligned / Größtenteils übereinstimmend
- 5 = Completely aligned / Vollständig übereinstimmend
- Other (please specify) / Andere (bitte angeben): _____

Question 9 / Frage 9. Which type of image do you think was used? / Was glauben Sie, welche Art von Bild verwendet wurde?

- AI-generated / KI-generiert
- Manually created / Manuell erstellt
- Unsure / Unsicher
- Other (please specify) / Andere (bitte angeben): _____

Section 3 / Abschnitt 3: Image Quality Evaluation / Bildqualitätsbewertung (conditions b–c / Bedingungen b–c) Question 10 / Frage 10. How visually coherent are the images? / Wie visuell kohärent ist das Bild?

- 1 = Not coherent at all / Überhaupt nicht kohärent
- 2 = Mostly not coherent / Meistens nicht kohärent
- 3 = Partially coherent / Teilweise kohärent
- 4 = Mostly coherent / Größtenteils kohärent
- 5 = Completely coherent / Vollständig kohärent
- Other (please specify) / Andere (bitte angeben): _____

Question 11 / Frage 11. What is the function of the image relative to the text? / Welche Funktion hat das Bild in Bezug auf den Text?

- “Expansion” / Expansion
- “Exemplification” / Exemplifikation
- “Explication” / Explikation
- “Condensation” / Kondensation
- Other (please specify) / Andere (bitte angeben): _____

Reference within item / Referenz im Item: see (??).

Question 12 / Frage 12. How well does the image fulfill this function relative to the text? / Wie gut erfüllt das Bild diese Funktion in Bezug auf den Text?

- 1 = Not at all / Überhaupt nicht
- 2 = Mostly not / Meistens nicht
- 3 = Partially / Teilweise
- 4 = Mostly / Größtenteils
- 5 = Completely / Vollständig
- Other (please specify) / Andere (bitte angeben): _____

Question 13 / Frage 13. How well does the image style match the text? / Wie gut passt der Bildstil zum Text?

- 1 = Not natural or pleasing at all / Überhaupt nicht natürlich oder ansprechend
- 2 = Mostly not natural or pleasing / Größtenteils nicht natürlich oder ansprechend
- 3 = Partially natural and pleasing / Teilweise natürlich und ansprechend
- 4 = Mostly natural and pleasing / Größtenteils natürlich und ansprechend
- 5 = Completely natural and pleasing / Vollständig natürlich und ansprechend
- Other (please specify) / Andere (bitte angeben): _____

Post-study Questionnaire / Nachbefragung

Section 4 / Abschnitt 4: Comparison of Image Conditions / Vergleich der Bildbedingungen (for b & c / für b & c) Question 14 / Frage 14. Which type of image do you find more useful? / Welche Art von Bild empfinden Sie als nützlicher?

- AI-generated images / KI-generierte Bilder
- Manually created images / Manuell erstellte Bilder
- No significant difference / Kein signifikanter Unterschied

Question 15 / Frage 15. Which type of image do you find more visually appealing? / Welche Art von Bild empfinden Sie als visuell ansprechender?

- AI-generated images / KI-generierte Bilder
- Manually created images / Manuell erstellte Bilder
- No significant difference / Kein signifikanter Unterschied

Question 16 / Frage 16. Which type of image best supports comprehension of the text? / Welche Art von Bild unterstützt das Textverständnis am besten?

- AI-generated images / KI-generierte Bilder
- Manually created images / Manuell erstellte Bilder
- No significant difference / Kein signifikanter Unterschied

Section 5 / Abschnitt 5: Open Feedback / Offenes Feedback

Question 17 / Frage 17. Do you have any comments or suggestions on the text, images, or their combination? / Haben Sie Kommentare oder Anregungen zum Text, zu den Bildern oder zu deren Kombination?

- Response / Antwort: _____

Medical Text Simplification: From Jargon Detection to Jargon-Aware Prompting

Taiki Papandreou and Jan Bakker and Jaap Kamps

Institute for Logic, Language and Computation (ILLC)

University of Amsterdam

Amsterdam, The Netherlands

taiki.papandreou-lazos@student.uva.nl, j.bakker@uva.nl, kamps@uva.nl

Abstract

Jargon identification is critical for improving the accessibility of biomedical texts, yet models are often evaluated on isolated datasets, leaving open questions about generalization. After reproducing MedReadMe’s jargon detection results and extending evaluation to the PLABA dataset, we find that transfer learning across datasets yields only modest gains, largely due to divergent annotation objectives. Through manual re-annotation we show that aligning labeling schemes improves cross-dataset performance. Building on these findings, we evaluate several jargon-aware prompting strategies for LLM-based medical text simplification. Explicitly highlighting jargon in prompts does not consistently improve simplification quality. When gains occur, they often trade off against readability and are model-dependent. Human evaluation indicates that simple prompting can be as effective as more complex, jargon-aware instructions. We release code to facilitate further research: https://github.com/taikilazos/thesis_codebase.

1 Introduction

Medical text simplification is crucial for improving health literacy by making technical content accessible to lay readers, with jargon handling being a central challenge. In this work, we define jargon as any term or span of words that may be hard to understand for lay readers who are not in the medical domain, such as technical abbreviations or complex terminology requiring simplification. However, jargon detection models are often evaluated on isolated datasets, leaving significant questions about their generalization capabilities.

Recent resources like MedReadMe (Jiang and Xu, 2024) and PLABA (Attal et al., 2023; Ondov et al., 2025) provide valuable benchmarks for jargon-centric research, but they employ different annotation schemes, MedReadMe categorizes term difficulty for lay readers, while PLABA identifies

PLABA Dataset

We studied 36 drop seizures in 5 patients with myoclonic astatic epilepsy of early childhood (MAEE) with simultaneous split-screen video recording and polygraph. Sixteen were falling attacks and 20 were either less severe attacks exhibiting only deep head nodding or seizures equivalent to drop attacks in terms of ictal pattern but recorded in the supine position. All seizures except those that occurred in patients in the supine position showed sudden momentary head dropping or collapse of the whole body downward.

MedReadMe Dataset

The long-acting bronchodilator tiotropium and single-inhaler combination therapy of inhaled corticosteroids and long-acting beta 2-agonists (ICS/LABA) are commonly used for maintenance treatment of patients with chronic obstructive pulmonary disease (COPD). Combining these treatments, which have different mechanisms of action, may be more effective than administering the individual components.

Figure 1: Example annotations from PLABA and MedReadMe datasets. Underlined terms indicate identified jargon.

terms requiring simplification via expert adaptations (see Figure 1). This discrepancy creates a fundamental barrier to cross-dataset evaluation and generalization.

To address this, we first reconstruct the MedReadMe experimental setup and extend evaluation to PLABA to probe cross-dataset generalization. Second, we investigate whether explicitly surfacing detected jargon in prompts improves LLM-based simplification of medical abstracts.

Our contributions are:

- We replicate MedReadMe’s jargon identifica-

tion and release our implementation, establishing baselines on PLABA and a relabeled subset for cross-dataset evaluation.

- We assess cross-dataset generalization, showing transfer learning is limited by annotation mismatches and that aligning schemes improves performance.
- We introduce and evaluate jargon-aware prompting strategies for simplification, finding benefits are model-dependent and often trade off against readability.

We release code and data to support reproducibility and future work on jargon-aware medical text simplification: https://github.com/taikilazos/thesis_codebase.

2 Related Work

Text simplification aims to make specialized content accessible without sacrificing meaning, a particular challenge in medicine where technical terminology is dense (Agrawal and Carpuat, 2024). Our work connects four areas: lexical complexity detection, biomedical simplification resources, LLM prompting strategies, and evaluation. Lexical complexity and jargon detection have evolved from surface heuristics to contextual models like BERT (Devlin et al., 2019), with MedReadMe providing fine-grained jargon categories for lay reader difficulty analysis (Jiang and Xu, 2024).

Work on biomedical text simplification leverages datasets such as PLABA, which offers expert-authored adaptations with span-level links to technical terms (Attal et al., 2023; Ondov et al., 2025), though Bakker and Kamps (2024) and others highlight challenges in sentence-level alignment (Devaraj et al., 2021; Goldsack et al., 2022; Guo et al., 2024). There has been limited exploration of LLM prompting strategies that explicitly surface jargon to control simplification (Xia et al., 2025). For evaluation, we assess how automatic metrics like FKGL and SARI (Kincaid et al., 1975; Xu et al., 2016) align with human judgments in this jargon-aware setting.

The PLABA dataset has enabled research into more controlled simplification approaches. Notably, Xia et al. (2025) conducted a study on jargon-aware simplification by using detected jargon spans to structure prompts for large language models. Their findings suggest that while explicitly surfacing jargon can be beneficial, its effectiveness is not

Dataset	# Sentences	# Jargon
PLABA Training	1,602	2,586
PLABA Validation	178	296
PLABA Testing	4,500	9,126
MedReadMe Training	2,587	5,207
MedReadMe Validation	784	1,789
MedReadMe Testing	1,140	2,112

Table 1: Number of examples and total jargon terms in the PLABA and MedReadMe datasets.

Metric	PLABA	MedReadMe
FKGL	10.73	14.08
Jargon / sent	1.92	1.76
Jargon Length	2.98	3.35

Table 2: Comparison of metrics between the PLABA and MedReadMe datasets.

consistent across models and often comes at the cost of readability, highlighting the complexity of integrating detection with generation.

3 Methodology

3.1 Dataset Analysis

We study two biomedical datasets with distinct objectives and annotation schemes. MedReadMe comprises 4,520 sentences from 180 complex-simple article pairs sampled from 15 medical simplification resources (Guo et al., 2024; Goldsack et al., 2022; Devaraj et al., 2021) and provides a hierarchical jargon taxonomy (binary/3-class/7-class) annotated by non-experts to approximate lay comprehension (Jiang and Xu, 2024). PLABA consists of PubMed abstracts paired with expert-authored plain-language adaptations and marks spans that require simplification (Attal et al., 2023; Ondov et al., 2025). As shown in Table 1, the datasets differ in the number of examples and annotated jargon terms; we preprocess PLABA to a sentence-level format to match MedReadMe. In difficulty characteristics (Table 2), MedReadMe exhibits higher lexical and grammatical complexity (FKGL 14.08 vs. 10.73). PLABA shows slightly higher jargon density but shorter jargon terms. Only 276 jargon terms exactly overlap across datasets, underscoring divergent annotation goals and target audiences. For examples of the differing annotation focus, see Figure 1.

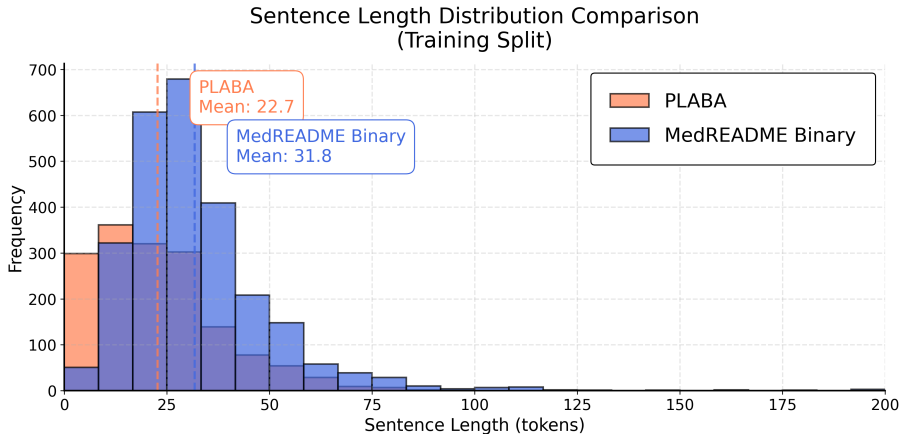


Figure 2: Sentence length distribution for train set: the mean value for MedReadMe is 31.8 and for PLABA 22.7

We also analyzed quantitative metrics to compare the two datasets. Figure 2 illustrates the sentence length distribution in the training splits, with MedReadMe sentences averaging 31.8 tokens compared to PLABA’s 22.7 tokens.

3.2 Model Selection and Training

We use standard BIO tagging for span detection: MedReadMe is labeled at binary/3-class/7-class levels, while PLABA is binary-only. Subword tokenization is handled via tokenizer `word_ids()`, assigning B- to the first subword and I- to subsequent subwords; special tokens ([CLS], [SEP], [PAD]) are masked with -100 in the loss. Both datasets are processed at the sentence level with a maximum sequence length of 250 and attention masks to ignore padding.

We reproduced the MedReadMe experiment using BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), BioBERT (Lee et al., 2019), and PubMedBERT (Tinn et al., 2023), testing both base and large variants of each model. Since the original paper only referenced the Hugging Face API without specifying model versions for BioBERT and PubMedBERT, our specific choices are documented in Table 3.

We standardized fine-tuning across models: optimizer AdamW (Loshchilov and Hutter, 2019), learning rate $2e-5$ (MedReadMe) and $1e-5$ (PLABA), batch size 32, up to 20 epochs with early stopping (patience 3) on validation entity-level F1.

3.3 Transfer Learning

We evaluated transferability via two settings: (1) direct transfer (train on MedReadMe \rightarrow evaluate on PLABA; train on PLABA \rightarrow evaluate

Family	Type	Model
BERT	Base	bert-base-uncased
RoBERTa	Base	roberta-base
BERT	Large	bert-large-uncased
RoBERTa	Large	roberta-large
BioBERT	Base	biobert-base-v1.1 [†]
PubMedBERT	Base	biomed-base-uncased [‡]
BioBERT	Large	biobert-large-v1.1 [†]
PubMedBERT	Large	biomed-large-uncased [‡]

Table 3: HuggingFace models used in experiments: generic model architectures (top half) and biomedical variants (bottom half). [†]From dmis-lab, [‡]From microsoft/BiomedNLP.

on MedReadMe) and (2) sequential fine-tuning (PLABA \rightarrow MedReadMe evaluated on MedReadMe and PLABA; MedReadMe \rightarrow PLABA evaluated on MedReadMe and PLABA).

All models were based on RoBERTa-large, fine-tuned with a learning rate of 1×10^{-5} , batch size of 16, and early stopping with max epoch of 20. Since MedReadMe contains multi-class annotations, we standardized both datasets to a binary classification setting (jargon vs. non-jargon) for consistency. Performance was primarily measured using entity-level F1.

3.4 Manual Annotation of PLABA Sentences Using the MedReadMe Scheme

To ensure that any observed differences in model performance when transferring between MedReadMe and PLABA are not solely due to mismatches in annotation schemes, we manually re-annotated 100 PLABA sentences using the

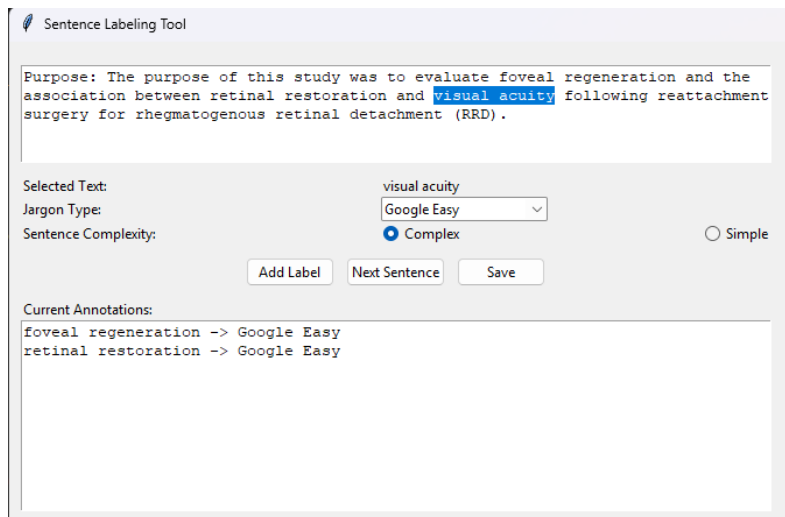


Figure 3: Screenshot of the custom sentence labeling tool. The tool allows the annotator to highlight spans corresponding to jargon terms and assign one of seven MedReadMe classes (e.g., Google Easy, Google Hard, Medical Name, etc.). The annotator can also specify whether a sentence is considered complex or simple, with the latter defined as sentences containing no jargon or only a single Google Easy term.

MedReadMe seven-class taxonomy: Google-Easy, Google-Hard, Medical Named Entity, Medical Abbreviation, General Abbreviation, General Complex Term, and Multi-sense Word (Jiang and Xu, 2024). This approach allows us to directly assess the impact of annotation scheme alignment on model performance.

A custom annotation tool (see Figure 3) was developed to facilitate this process, allowing the annotator to highlight jargon spans and assign the appropriate class. Sentences were also labeled as "complex" or "simple," with "simple" defined as containing no jargon or only a single Google Easy term, and all the other cases as "complex."

All annotations were performed by a single annotator (the main author), following MedReadMe guidelines (Jiang and Xu, 2024).

This relabeled subset allows for a fairer evaluation of model transfer: if model performance improves on the MedReadMe-labeled PLABA data, it suggests that the original drop in transfer performance was primarily due to annotation scheme mismatch and label distribution differences, rather than a fundamental inability of the model to generalize. Sentences were classified as "simple" if they contained no jargon or only a single Google Easy term (aligned with MedReadMe’s lower CEFR levels); "complex" sentences included any with additional jargon (e.g., multiple Google Easy or Google Hard/Medical terms), though this feature was not used in classification. The class distribution of the relabeled data can be seen in the Table 4.

Class	Count
Google Easy	203
Google Hard	187
Medical Name	15
Medical Abbreviation	16
General Abbreviation	0
General Complex	16
Multisense	0

Table 4: True label distribution (token-level) for each class in the evaluation set.

3.5 PLABA Test Set for Simplification

The PLABA test set comprises 300 medical abstracts with 3,315 sentences, of which 3,041 (91.7%) contain at least one jargon term (Attal et al., 2023). Sentences contain between 1 and 18 jargon spans, with most (64.6%) containing 1–4. Each span is annotated with one or more recommended simplification actions (e.g., substitute, explain, generalize, omit, exemplify). On average, abstracts contain 26.57 jargon terms. Action distribution is skewed toward substitution (65.62%), followed by explanation (17.59%), omission (10.25%), generalization (6.12%), and exemplification (0.43%). Average jargon length is 1.79 words.

Reference simplifications were 40% shorter at the sentence level (26.18 to 15.94 words) and 6.5 grade levels easier (FKGL 13.55 to 7.04; Kincaid

Original Text:

The patient exhibited tachycardia and dyspnea during examination.

Base Instructions (applied to all prompts):

1) Write a clear sentence; 2) Preserve distinctions and numbers; 3) Replace medical terms only if meaning stays exact; 4) No notes or multiple versions.

1. Simple Prompt: Provide one simplified sentence for the input (focusing on lexical simplification of jargon).

2. Jargon-aware Prompt: Highlight detected terms (e.g., tachycardia, dyspnea) and simplify cautiously.

3. Ground Truth Jargons Prompt (GT): Use ground truth jargon terms as guidance.

4. Ground Truth Actions Prompt (GT action): Specify per-term actions (e.g., substitute, explain).

Figure 4: Prompting strategies for sentence-level medical simplification. See Appendix B for detailed information.

et al., 1975), yet they contained more sentences (19 vs. 11.05), indicating frequent sentence splitting.

3.6 Prompt Design

We evaluate four prompting strategies, from a simple instruction baseline to prompts that explicitly surface jargon terms and, in the most guided variant, specify actions per term. Jargon spans are obtained from our PLABA jargon detector. Simplification operates at the sentence level; simplified sentences are concatenated for abstract-level evaluation.

Ground-truth variants estimate the upper bound of jargon-aware prompting: if gold-guided prompts outperform detected-jargon prompts, the bottleneck lies in detection rather than prompting.

3.7 LLM Models for Simplification

We compare a general-purpose model, Llama-3.1-8B-Instruct¹, with a domain-specialized alternative, Medicine-Llama3-8B². We standardize output

¹<https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct>

²<https://huggingface.co/instruction-pretrain/medicine-Llama3-8B>

cleaning to remove prefixes and meta-commentary, retaining only the simplified sentence for evaluation.

3.8 Evaluation Methods for Simplification

For Jargon Detection tasks, we report F1 due to class imbalance in medical texts. Token-level F1 measures correct classification of individual tokens while ignoring padding tokens (-100) and non-entity (O) predictions. Entity-level F1 requires exact matches between predicted and gold entities in both span boundaries and type. We run generation with fixed decoding settings: temperature = 0.2, top_p = 0.9, repetition penalty = 1.3, no_repeat_ngram_size = 3, and max tokens = 512. Experiments use NVIDIA A100 GPUs; models are loaded from shared storage for throughput. Evaluation is computed at the abstract level by concatenating sentence-level outputs.

As for the simplification task, we report readability (FKGL) (Kincaid et al., 1975) and semantic similarity with BERTScore³, and we use SARI and BLEU (via EASSE)⁴ to assess add/keep/delete operations relative to original and reference.

To validate our automatic metrics, we conducted a human evaluation study⁵. We evaluated at the abstract level rather than sentence level to better reflect real-world reading, where users consume full abstracts; complex sentences aggregated at this level provide a fairer assessment of overall difficulty. Each original medical text was presented alongside five simplified versions: four generated by our models and one gold-standard reference (PLABA’s expert-authored simplifications), randomly ordered to prevent bias. We recruited N=5 annotators, all fluent English speakers with at least a graduate-level background in Computer Science-related fields from the University of Amsterdam. Each annotator evaluated three different medical abstracts. For each abstract, they rated the five simplified versions on a 1–5 scale along three aspects: meaning preservation (accuracy of medical information), simplicity (lexical accessibility to non-experts, focusing on jargon reduction), and fluency (natural and coherent writing).

For an action-based perspective, annotators also evaluated the model’s ability to perform specific

³<https://github.com/feralvam/easse/blob/master/easse/bertscore.py>

⁴<https://github.com/feralvam/easse>

⁵<https://qualitativexphthesis-biomed.streamlit.app/>

Model	Token-Level			Entity-Level			
	Bin	3Cls	7Cls	Bin	3Cls	7Cls	
Base	BERT	88.12 (85.4)	86.61 (80.4)	75.02 (66.3)	70.85 (77.0)	67.74 (72.5)	56.47 (63.3)
	RoBERTa	89.89 (86.2)	88.72 (81.7)	76.72 (66.7)	57.83 (79.7)	69.29 (75.2)	53.23 (66.6)
	BioBERT	87.83 (84.2)	87.39 (79.6)	76.10 (66.4)	68.51 (77.1)	67.12 (72.8)	58.22 (64.1)
	PubMedBERT	84.98 (85.2)	84.72 (81.2)	76.71 (67.7)	71.57 (75.8)	72.50 (74.8)	63.68 (66.3)
Large	BERT	88.05 (86.1)	87.18 (80.9)	76.25 (67.9)	67.70 (78.5)	68.93 (74.1)	58.71 (43.9)
	RoBERTa	89.73 (86.8)	88.72 (82.3)	78.65 (68.6)	73.42 (80.2)	68.87 (75.9)	62.63 (67.9)
	BioBERT	87.80 (85.3)	86.33 (80.7)	75.98 (67.0)	73.40 (78.4)	70.51 (72.2)	60.19 (64.9)
	PubMedBERT	86.39 (85.7)	85.67 (82.3)	75.31 (68.3)	72.32 (79.0)	69.70 (75.2)	61.67 (66.5)

Table 5: F1 scores (%) on the MedReadMe dataset. Our results are shown with original results in parentheses. The highest value per column is bolded.

text transformation operations (substitute, generalize) informed by PLABA action annotations (Attal et al., 2023; Ondov et al., 2025). In this setting, each annotator rated three action types across five randomly selected sentences, using the same 1–5 scale. Detailed examples are provided in Appendix B.

4 Results and Analysis

4.1 Jargon Detection Performance

We successfully replicated the MedReadMe (MRM) experiment, though with notable differences. As shown in Table 5, our implementation achieved higher token-level F1 scores (e.g., 89.89% vs. 86.8% for RoBERTa-base) but lower entity-level F1 scores (e.g., 73.42% vs. 80.2% for RoBERTa-large) compared to the original study. This suggests our models were better at classifying individual tokens but worse at identifying exact span boundaries, potentially due to differences in the evaluation pipeline.

Performance varied significantly across jargon categories (Table 6). The RoBERTa model excelled at identifying medical abbreviations (F1=0.869) but struggled with nuanced distinctions, such as differentiating Google-Hard from Google-Easy terms (F1=0.514). It failed completely on rare classes like multisense terms, highlighting the impact of severe class imbalance.

On the PLABA dataset, all models performed worse than on MRM, with RoBERTa-large achieving the highest entity-level F1 of 46.70% (Table 7). Surprisingly, domain-specific models like BioBERT showed no clear advantage. This performance gap is likely due to PLABA’s smaller size

Class	Prec.	Rec.	F1	Supp.
G_EASY	0.697	0.828	0.756	3,939
G_HARD	0.748	0.391	0.514	1,178
MED_ABBR	0.831	0.910	0.869	933
MED_NAME	0.506	0.701	0.588	455
GEN_CPLX	0.695	0.628	0.660	489
GEN_ABBR	0.866	0.792	0.827	130
MULTI	0.000	0.000	0.000	28

Table 6: 7-class performance for RoBERTa-large on MedReadMe (MRM) dataset. G: Google, MED: Medical, GEN: General, ABBR: Abbreviation, NAME: Name Entity, CPLX: Complex, MULTI: Multisense.

Model	F1	Precision	Recall
BERT	44.17	39.74	49.70
RoBERTa	46.70	46.06	47.36
BioBERT	43.42	46.99	40.35
PubMedBERT	45.43	43.19	47.92

Table 7: Entity-level performance metrics across different language models (large version) on the PLABA dataset.

and, crucially, its different annotation objective.

4.2 Transfer Learning and the Impact of Annotation Schema

Cross-dataset transfer learning yielded only modest gains, underscoring the challenge of generalizing across differently annotated resources (Table 8). For instance, a model trained on MedReadMe achieved only 33.71% entity F1 when evaluated directly on PLABA.

To test if this was due to annotation mismatch,

Experiment	Token F1	Entity F1
MRM → PLABA	61.22	33.71
PLABA+MRM → PLABA	62.94	37.01
MRM+PLABA → PLABA	66.84	37.71
PLABA → MRM	59.01	25.03
PLABA+MRM → MRM	89.80	73.84
MRM+PLABA → MRM	73.96	46.64

Table 8: F1 scores (%) for transfer learning experiments. Sequential transfer refers to fine-tuning on a second dataset after initial training. MRM stands for MedReadMe.

Setting	SARI↑	BERTScore↑	FKGL↓	BLEU↑
Simple	29.87	19.91	13.53	2.34
Jargon	29.92	19.07	14.16	2.40
GT	30.62	20.51	14.16	3.03
GT action	32.26	11.55	15.36	4.06

Table 9: Performance metrics for Llama-3.1-8B-Instruct across different prompts.

we manually relabeled a 100-sentence PLABA subset with the MedReadMe schema. When evaluated on this aligned data, the MedReadMe-trained model’s performance improved markedly from 33.71% to 42.00% entity F1. This confirms that the performance drop was primarily due to divergent annotation schemes rather than a model limitation. When labels are aligned, models generalize effectively.

4.3 Jargon-Aware Text Simplification

We next investigated whether explicitly highlighting jargon in prompts improves LLM-based simplification. We evaluated four prompting strategies of increasing complexity on both a general-purpose (Llama-3.1, Table 9) and a domain-specialized (Medicine-Llama3, Table 10) model.

The results were model-dependent and revealed a consistent trade-off. For Llama-3.1, more explicit guidance (e.g., providing ground-truth actions) led to the best performance on operation-based metrics like SARI (32.26) but at the cost of readability, yielding the highest FKGL (15.36). In contrast, the simple prompt achieved the best readability (FKGL=13.53).

Contrary to expectations, the Medicine-Llama3 model performed best across all metrics with the simple prompt and its performance degraded with

Setting	SARI↑	BERTScore↑	FKGL↓	BLEU↑
Simple	28.81	13.59	12.69	1.69
Jargon	28.64	12.66	12.76	1.65
GT	28.70	11.00	13.67	1.60
GT action	28.57	8.90	13.87	1.42

Table 10: Performance metrics for Medicine-Llama3-8B across different prompts.

Version/Task Type	Mean Rating	SD
Ground Truth	5.00	0.00
Simple	3.33	1.41
Jargon	2.34	0.94
GT	2.84	1.18
GT action	2.67	0.94
Action-based Tasks	4.60	0.57

Table 11: Average Ratings and Standard Deviations by Version and Task Type

more complex, jargon-aware instructions. This suggests that domain-specific pre-training does not automatically translate into an ability to effectively leverage explicit jargon instructions.

4.4 Human Evaluation

A qualitative human evaluation (Table 11) revealed that while the reference simplifications received perfect scores, all model outputs were perceived as lower quality. The simple prompt was competitive (Mean=3.33), while jargon-aware prompts did not reliably improve perceived quality. Notably, the high standard deviations indicate substantial disagreement among raters. A key observation was that sentence-level simplification often led to a loss of context and information across the abstract, limiting overall coherence. See Appendix C for examples of generated outputs for different prompting strategies.

5 Discussion and Conclusions

In this work, we thoroughly evaluated the automatic jargon detection methods for biomedical texts. We reproduced MedReadMe experiments, established PLABA baselines, and showed that cross-dataset transfer is limited primarily by annotation mismatches. We then experimented with jargon-aware prompting strategies for the automatic simplification of these texts.

On jargon detection, our replications achieved

higher token-level but lower entity-level F1 than the original report, highlighting remaining challenges in precise span boundary modeling. Category-wise analyses showed strong performance on frequent, well-formed classes (e.g., medical abbreviations) and weaknesses on rarer or nuanced classes (e.g., Google-Hard, multisense), reflecting severe class imbalance. Importantly, evaluating on a PLABA subset re-annotated with the MedReadMe scheme (100 sentences) improved entity-level F1 from 33.71% to 42.00%, demonstrating that schema alignment substantially boosts transferability.

Turning to simplification, our experiments show that the effect of jargon-aware prompting is model-dependent rather than uniformly beneficial. The general-purpose Llama-3.1-8B-Instruct benefited from more explicit guidance (best SARI/BLEU with ground-truth actions), but with reduced readability (higher FKGL). In contrast, the domain-specialized Medicine-Llama3-8B performed best with simple prompts, suggesting that domain pre-training does not automatically translate into better handling of explicit jargon instructions. This dependency may arise from how models process prompts: general models require explicit jargon surfacing to prioritize medical terms, while specialized models implicitly handle them, making simple instructions sufficient.

Across models, we observed a consistent trade-off: more detailed prompting can improve operation- and overlap-based metrics (SARI, BLEU) while harming readability (FKGL). Qualitative judgments echoed this tension: references set a clear upper bound; simple prompts were competitive, whereas jargon-aware prompts did not reliably improve perceived quality, and sentence-level processing likely contributed to information loss across abstracts.

Thus, explicitly including identified jargon in prompts does not consistently improve LLM medical text simplification. Jargon matters, but surfacing terms alone is insufficient; benefits depend on the model and come with readability trade-offs.

Future work should (i) improve span boundary modeling and mitigate class imbalance in detection, (ii) explore schema-aware or multi-task training for cross-dataset robustness, and (iii) couple detection with controllable, document-level generation and evaluation that jointly captures medical fidelity and accessibility. We release code and data to support further research.

6 Limitations

The lack of multiple annotators for manual re-annotation is a limitation and should be addressed in future work to improve reliability. The small-scale human evaluation (only 5 annotators from a computer science background, evaluating just 3 abstracts each) and resulting high rater disagreement may limit generalizability of perceived quality. Automatic metrics like SARI may not fully capture jargon-specific changes, and the low BLEU scores in simplification experiments indicate challenges in generating high-quality outputs. While the evaluation was at the document level, generation was at the sentence level, losing the global context. Future work could explore better prompt engineering, larger-scale evaluations with diverse annotators, and document-level generation to address these issues.

7 Lay Summary

Medical texts are full of complex terms that can confuse people without a scientific background. This makes it hard for patients and the general public to understand health information. Our research focuses on two key areas: identifying these difficult terms (called "jargon") and simplifying medical texts so they're easier to read.

First, we studied how well computer models can spot jargon in medical writing. We compared two datasets: MedReadMe, which labels terms by how hard they are for lay people to understand, and PLABA, which marks terms that experts think need simplifying. We found that models trained on one dataset don't work as well on the other because the datasets have different goals. But when we manually relabeled some PLABA data to match MedReadMe's style, the models improved a lot, showing that aligning how we define jargon helps cross-dataset learning.

Second, we tested ways to make large language models (like AI chatbots) simplify medical texts. We tried simple prompts and more complex ones that highlight detected jargon. Surprisingly, the simple prompts often worked just as well or better than the jargon-focused ones. Results depended on the model—general-purpose models liked more guidance, but specialized medical models did better with basics. This suggests that just telling an AI to simplify might be enough, without needing to point out every jargon term.

Our work shows that making medical info ac-

cessible is tricky, but better data alignment and smarter prompting can help. We hope this leads to tools that make health communication clearer for everyone, improving patient understanding and outcomes. All our code and data are publicly available to support future research.

Acknowledgments

Experiments in this paper were carried out on the National Supercomputer Snellius, supported by SURF and the HPC Board of the University of Amsterdam. Jan Bakker and Jaap Kamps are partly funded by the Netherlands Organization for Scientific Research (NWO NWA # 1518.22.105). Jaap Kamps is also partly funded by the University of Amsterdam (AI4FinTech program), and ICAI (AI for Open Government Lab). We thank the annotators for their contributions to the human evaluation study. Views expressed in this paper are not necessarily shared or endorsed by those funding the research.

References

- Sweta Agrawal and Marine Carpuat. 2024. [Do text simplification systems preserve meaning? a human evaluation via reading comprehension](#). *Transactions of the Association for Computational Linguistics*, 12:432–448.
- Kush Attal, Brian Ondov, and Dina Demner-Fushman. 2023. [A dataset for plain language adaptation of biomedical abstracts](#). *Scientific Data*, 10(1):8.
- Jan Bakker and Jaap Kamps. 2024. [Cochrane-auto: An aligned dataset for the simplification of biomedical abstracts](#). In *Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024)*, pages 41–51, Miami, Florida, USA. Association for Computational Linguistics.
- Jan Bakker, Taiki Papandreou-Lazos, and Jaap Kamps. 2024. [Biomedical text simplification models trained on aligned abstracts and lay summaries](#). In *The Thirty-Third Text REtrieval Conference Proceedings (TREC 2024)*, Gaithersburg, MD, USA, November 15-18, 2024, volume 1329 of *NIST Special Publication*. National Institute of Standards and Technology (NIST).
- Ashwin Devaraj, Iain Marshall, Byron Wallace, and Junyi Jessy Li. 2021. [Paragraph-level simplification of medical texts](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4972–4984, Online. Association for Computational Linguistics.
- 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.
- Tomas Goldsack, Zhihao Zhang, Chenghua Lin, and Carolina Scarton. 2022. [Making science simple: Corpora for the lay summarisation of scientific literature](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10589–10604, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Yue Guo, Wei Qiu, Gondy Leroy, Sheng Wang, and Trevor Cohen. 2024. [Retrieval augmentation of large language models for lay language generation](#). *Journal of Biomedical Informatics*, 149:104580.
- Chao Jiang and Wei Xu. 2024. [MedReadMe: A systematic study for fine-grained sentence readability in medical domain](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17293–17319, Miami, Florida, USA. 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, Naval Technical Training Command Millington TN Research Branch.
- Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. [Biobert: a pre-trained biomedical language representation model for biomedical text mining](#). *Bioinformatics*, 36(4):1234–1240.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [Roberta: A robustly optimized bert pretraining approach](#). *arXiv preprint arXiv:1907.11692*.
- Ilya Loshchilov and Frank Hutter. 2019. [Decoupled weight decay regularization](#). In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.
- Brian Ondov, William Xia, Kush Attal, Ishita Unde, Jerry He, and Dina Demner-Fushman. 2025. [Lessons from the trec plain language adaptation of biomedical abstracts \(plaba\) track](#). *arXiv preprint arXiv:2507.14096*.
- Taiki Papandreou, Jan Bakker, and Jaap Kamps. 2025. [University of Amsterdam at the CLEF 2025 SimpleText Track](#). In *Working Notes of CLEF 2025*:

Conference and Labs of the Evaluation Forum, volume 4038 of *CEUR Workshop Proceedings*, pages 4356–4362. CEUR-WS.org.

Panagiotis Taiki Papandreou-Lazos. 2025. [Medical text simplification: From jargon detection to automated simplification](#). Master’s thesis, University of Amsterdam.

Robert Tinn, Hao Cheng, Yu Gu, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2023. [Fine-tuning large neural language models for biomedical natural language processing](#). *Patterns*, 4(4):100729.

William Xia, Ishita Unde, Brian David Ondov, and Dina Demner-Fushman. 2025. [JEBS: A fine-grained biomedical lexical simplification task](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 17654–17666, Vienna, Austria. Association for Computational Linguistics.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. [Optimizing statistical machine translation for text simplification](#). *Transactions of the Association for Computational Linguistics*, 4:401–415.

A Data, code, and models

We release code and data to support reproducibility and future work on jargon-aware medical text simplification: https://github.com/taikilazos/thesis_codebase.

Extensive further documentation can be found in (Papandreou-Lazos, 2025).

Related experiments were reported at the TREC 2024 PLABA track (Bakker et al., 2024) and at the CLEF 2025 SimpleText Track (Papandreou et al., 2025).

B Prompt Design

Original Text

The patient exhibited tachycardia and dyspnea
↪ during examination.

Base Instructions (applied to all prompts)

IMPORTANT: Follow these rules exactly:

1. Write a clear sentence
2. Keep ALL medical distinctions and patterns
3. Keep exact numbers and measurements
4. Replace medical terms with plain words ONLY if
↪ meaning stays exactly the same
5. Keep medical terms if simplifying would lose
↪ precision
6. No explanations or notes
7. No multiple versions

1. Simple Prompt

Text to simplify: The patient exhibited
↪ tachycardia and dyspnea during examination.
Write one simplified sentence.

2. Jargon-aware Prompt

Replace these terms ONLY if you can keep their
↪ exact medical meaning:
- tachycardia
- dyspnea
Text to simplify: The patient exhibited
↪ tachycardia and dyspnea during examination.
Write one simplified sentence.

3. Ground Truth Jargons Prompt

You must simplify the following terms in the
↪ sentence below (if present):
- tachycardia
- dyspnea
Text to simplify: The patient exhibited
↪ tachycardia and dyspnea during examination.
Write one simplified sentence.

4. Ground Truth Actions Prompt

For each marked term in the sentence below, take
↪ the specified action:
- tachycardia: substitute
- dyspnea: explain
Text to simplify: The patient exhibited
↪ tachycardia and dyspnea during examination.
Write one simplified sentence:

C Generated Text Examples

To illustrate the differences in simplification quality across prompting strategies, we provide excerpts from one test abstract. We selected two key sentences with medical jargon and show how each version handles them. Figure 5 shows the source, reference (Ground truth human simplification), and prediction based on either the Simple prompt, the Jargon-aware prompt, the Ground truth jargons prompt, and the Ground truth actions prompt

These examples highlight how simpler prompts can be fluent but sometimes inaccurate, while jargon-aware prompts vary in readability and fidelity.

Original Sentence 1

Background: The purpose of the study was to examine whether early repeated infections due to *Trichomonas vaginalis* among human immunodeficiency virus (HIV)-positive and HIV-negative women are reinfections, new infections, or cases of treatment failure.

Ground Truth human simplification

For the study's background, the study's aim was to find whether early repeated infections of *Trichomonas vaginalis* (a parasite that can cause a sexually transmitted infection or STI) among women with human immunodeficiency virus (HIV) and without HIV are reinfections, new infections, or cases of treatment failing.

Simple prompt

The goal of this research is to figure out what happens when many people get infected again by an STI called trichomoniasis in both those who have AIDS and do not have it.

Jargon-aware prompt

The goal is to figure out what happens when people get infected again by this type of bacteria called trichomonas's vaginalis, which can happen even after they have gotten better because either it came back through their sexual partner as a re-infection from them, or got another infection on its own, or maybe the cure didn't work for some reason so now we need to find out why.

Ground Truth Jargons Prompt

The goal of this research is to determine why some people infected by *Trichomonas* get reinfected again after being treated for it.

Ground Truth Actions Prompt

The goal of this research is to determine what causes frequent repeat occurrences of an infection caused by *Trichomonas vaginalis* among both HIV-positive and negative females - either they get it again after being treated previously, contract it for the first time despite previous treatments, or their current condition does not respond well to medication.

Original Sentence 1

Methods: Women attending an HIV outpatient clinic and a family planning clinic in New Orleans, Louisiana, who had culture results positive for *T. vaginalis* were treated with 2 g of metronidazole under directly observed therapy.

Ground Truth human simplification

For the study's methods, women attending an HIV clinic and family planning clinic in New Orleans, Louisiana, who had tests detecting *T. vaginalis*, were treated with 2 grams (g) of metronidazole (an antibiotic for infections) under directly observed treatment.

Simple prompt

Women from two clinics received treatment at this place because they tested positive for *Trichomonas vaginalis* using cultures; each woman was given 2000 milligrams of Metronidazole while being watched by healthcare workers during their visit.

Jargon-aware prompt

Women from clinics attended by people living with AIDS and those seeking birth control services received treatment after testing positive for a parasite causing STDs; this included taking two grams of antibiotics while being watched as you swallow them.

Ground Truth Jargons Prompt

Women from two clinics received treatment after testing positive for *Trichomonas* infection; they took 2000 milligrams of antibiotic medication while being watched by healthcare workers during their visits.

Ground Truth Actions Prompt

Women from two clinics in New York received treatment after testing positive for *Trichomonas* infection; they took 2000 milligrams of antibiotic medication while being closely monitored by healthcare workers during their visits.

Figure 5: Examples: Effect of prompt variation on the output.

Readability Reconsidered: A Cross-Dataset Analysis of Reference-Free Metrics

Catarina G. Belem¹, Parker Glenn², Alfy Samuel², Anoop Kumar², Daben Liu²,

¹University of California Irvine, ²Capital One

Correspondence: cbelem@uci.edu

Abstract

Automatic readability assessment plays a key role in ensuring effective and accessible written communication. Despite significant progress, the field is hindered by inconsistent definitions of readability and measurements that rely on surface-level text properties. In this work, we investigate the factors shaping human perceptions of readability through the analysis of 897 judgments, finding that, beyond surface-level cues, information content and topic strongly shape text comprehensibility. Furthermore, we evaluate 15 popular readability metrics across five English datasets, contrasting them with six more nuanced, model-based metrics. Our results show that four model-based metrics consistently place among the top four in rank correlations with human judgments, while the best performing traditional metric achieves an average rank of 8.6. These findings highlight a mismatch between current readability metrics and human perceptions, pointing to model-based approaches as a more promising direction.

1 Introduction

Readability assessment can be used to determine the level of comprehension of a piece of text (DuBay, 2004; Collins-Thompson, 2014). In domains such as science communication (Kerwer et al., 2021; August et al., 2023), health (Friedman and Hoffman-Goetz, 2006; Hershenhouse et al., 2024), law (Curtotti et al., 2015; Cheong et al., 2024), and education (Vajjala and Lučić, 2018), readability assessment plays a key role in making information accessible to individuals regardless of their background or cognitive needs (Collins-Thompson, 2014). It is important for highly-specialized fields characterized by dense jargon and complex language (Friedman and Hoffman-Goetz, 2006; Han et al., 2024), as well as for applications

engaging with users of varied familiarity with the domain (Joshi et al., 2025; Puech et al., 2025).

One challenge in advancing automatic readability assessment is that *readability* is an overloaded term, measured in different ways by prior work. Some studies treat readability as *text difficulty*, using surface-level properties such as word length, word frequency, and various word type counts (Flesch, 1948; Kincaid et al., 1975; Leroy et al., 2008). Others broaden the definition of readability to consider syntactic and discourse-level organization, including cohesion and coherence properties (Graesser et al., 2004; Petersen, 2007; Pitler and Nenkova, 2008; Feng et al., 2010; Es-lami, 2014; Zhuang et al., 2025). A third line of work views readability as a combination of text characteristics and information content (Xia et al., 2016; August et al., 2024).

Taken together, the diversity of interpretations highlight the difficulty of pinning down readability, and have led to the continued use of proxy metrics that may not fit the task, domain, or are misaligned with human comprehension judgments (Ahmed, 2023; Liu and Lee, 2023; Han et al., 2024).

2 Related Work

Readability Datasets. Despite growing interest in readability assessment, high-quality datasets remain scarce (Xia et al., 2016). Existing document-level datasets can be subdivided into *parallel* corpora (Vajjala and Lučić, 2018; August et al., 2024; Joshi et al., 2025) and *non-parallel* corpora (Lu et al., 2022; Crossley et al., 2024) and span various tasks and content type, including literary and informational (Crossley et al., 2024), academic (August et al., 2024), or information-seeking content (Lu et al., 2022; Joshi et al., 2025). Recently, sentence-level datasets have also been introduced (Arase

et al., 2022; Naous et al., 2024).

Readability Metrics. While human judgments remain the gold standard for readability evaluation, their collection is often time-consuming and expensive (Rooein et al., 2024). Automated metrics have emerged as a cheaper and quicker alternative. Examples include metrics relying on basic linguistic features, including sentences, words, and syllables counts, average reading time (Demberg and Keller, 2008), language model perplexity (Collins-Thompson, 2014; Pitler and Nenkova, 2008), and fraction of functional (Leroy et al., 2008, 2010) or uncommon words (August et al., 2024). Surface-form features have been further combined to form *readability tests*, such as the Automatic Reading Index (Senter and Smith, 1967), Dale-Chall Readability Score (Dale and Chall, 1948), Flesch-Kincaid Reading Ease (Flesch, 1948), and Linsear Write Formula (Klare, 1974). Despite critiques of brittleness (Rooein et al., 2024; Collins-Thompson, 2014) and limited domain suitability (Leroy et al., 2010), these formulas continue to be used. Recently, both fine-tuning (Arase et al., 2022; Naous et al., 2024) and LLM-as-a-judge approaches (Rooein et al., 2024; Trott and Rivière, 2024) have been proposed to capture more abstract and nuanced aspects of readability. However, since these methods rely on implicitly learned representations, they are regarded as less interpretable than those grounded in surface-level textual features.

3 How Do Humans Perceive Readability?

Given the divergent definitions of readability and continued reliance on surface-form metrics, we take a human-centric perspective, asking: *What guides human perceptions of readability?* To address this question, we analyze a subset of the ELI-WHY (GPT-4) (Joshi et al., 2025) dataset, designed to study whether LLMs can generate explanations tailored to various readability levels. The dataset comprises GPT-4-generated explanations for 299 “Why” questions, each annotated by humans into three readability levels—Elementary, High School, and Graduate—along with accompanying rationales justifying their judgments. Each question–explanation pair was independently rated by three annotators, and final labels were determined via majority vote. For additional de-

tails, see the original paper. Table 11 (in Appendix) shows randomly selected examples of human rationales for each readability level.

Exploring Human Rationales. Although Joshi et al. (2025) collected human rationales supporting readability judgments, their analysis primarily focuses on the labels themselves, offering limited insight into the factors shaping human perceptions. We complement their study by providing a quantitative perspective on the key factors driving human text comprehension through the analysis of human rationales. Two authors of this paper annotated the human-provided readability rationales for 90 ELI-Why question–answer pairs, balanced evenly across classes. Building on the original human annotation instructions, each rationale was labeled with one or more of the following categories:

- *Wording/Terminology*: presence of scientific words, abbreviations, or complex synonyms;
- *Sentence Structure*: comments on sentence length or the number of concepts;
- *Examples/Analogies*: mentions of examples or analogies as key factors;
- *Details and Depth*: mentions of the presence or absence of details;
- *Curriculum-based*: links the information content or topic to a specific education level.

Figure 1 shows the consensus vote across readability classes. The average sample-level Jaccard index for the obtained annotations is 0.91, indicating high agreement between the two annotators. *Wording/Terminology* emerges as the predominant rationale for readability judgments, with annotators’ distinctions in lexical complexity (e.g., “Words like adherence are too advanced for elementary school”) or simplicity (“uses basic words”). The *Curriculum-based* category is invoked far more often to justify High School and Graduate judgments than Elementary, with annotators noting that “The scientific terms... require an introductory background or some foundational knowledge” or that “a concept that will be brought up in chemistry classes in undergrad.” Conversely, *Examples/Analogies* is disproportionately used to support Elementary judgments, with comments such

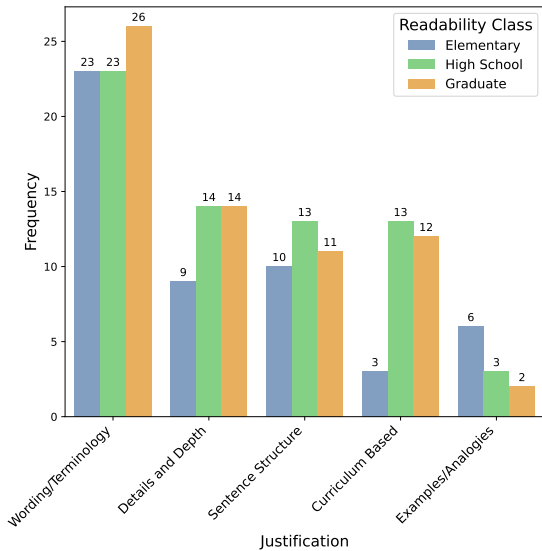


Figure 1: **Distribution of justification reasons across 90 examples in ELI-WHY (GPT-4)**. Counts are based on the consensus over 2-way annotations.

as “Examples are... what you’d say to a toddler” or “The analogies used... make it more accessible to elementary level.”. Notably, both categories rely on comprehension and common-sense reasoning that go beyond surface-level textual properties.

4 Re-evaluating Readability Metrics

Motivated by the gap between surface-form textual cues and human perceptions of readability, we investigate how well existing readability metrics correlate with human judgments across five diverse datasets (see statistics in Table 1).¹

4.1 Datasets

SCIENTIFIC PAPERS (August et al., 2024) consists of 180 total query-focused summaries about 10 different academic papers (*e.g.*, What did the paper do?) and cover topics from public policy to nanotechnology. Summaries were carefully curated by an expert science writer to reflect three levels of complexity: **Low**, **Medium**, and **High**.

CLEAR (Crossley et al., 2024) contains 4.7k text excerpts sourced from open digital libraries including Project Gutenberg and Wikipedia. The texts are self-contained and cover both literary and informational content. Approximately 111k pairwise readability judgments from 1.1k annotators were aggregated under a Bradley-Terry

¹Additional dataset details are available in Appendix A.

model to obtain continuous readability scores.

ELI-WHY (GPT-4) (Joshi et al., 2025) includes 897 GPT-4-generated explanations tailored to three readability levels—**Elementary**, **High School**, and **Graduate**—each annotated with human-assigned labels and rationales. Likewise, **ELI-WHY (HUMAN)** is a smaller dataset with 123 answers that were manually curated.

SCIENCEQA (Lu et al., 2022) is a multi-modal science reasoning dataset consisting of 21k multiple choice questions sourced from K-12 curriculum, covering various subjects (*e.g.*, natural science, language science, and social science). Each example is associated with a reference solution (or *explanation*) and reference knowledge (or *lecture*), both of which are written at the readability level of the intended student audience. We randomly sample 200 text-only examples per grade for our analysis.

4.2 Metrics ²

Surface-form metrics consist of direct counts of properties of the text, such as characters, syllables, monosyllables, polysyllables, words, and sentences. These also include other specialized variants such as estimated reading time in seconds, number of difficult words, and functional words.

Psycholinguistic metrics, known as *readability tests*, are typically formulated as weighted sums of ratios involving surface-form properties. For instance, *Automatic Readability Index* is based on characters-to-words and words-to-sentences ratio (Senter and Smith, 1967), the *Flesch Kincaid Reading Ease* on words-to-sentences and syllables-to-words (Flesch, 1948), and *Dale-Chall Readability* on the fraction of difficult words and words-to-sentences ratio (Dale and Chall, 1948). An exception is the *Linsear Write Formula*, which distinguishes easy from hard words using syllable counts and computes their frequencies in a text sample (Klare, 1974). We additionally report values for other popular metrics (Coleman and Liau, 1975; Gunning, 1952; Harry and Laughlin, 1969).

Model-based metrics are categorized into two main classes: *fine-tuned metrics* (Zhuang et al., 2025) and *LLM-as-a-judge metrics* (Zheng et al., 2023). In this work, we use two fine-tuned metrics based on ModernBERT (Warner et al., 2024) –

²We refer readers to Appendix B for additional details.

Dataset	Size	Label Type	Labels	Avg. #WORDS	Avg. #SENTS
SCIENTIFIC PAPERS (August et al., 2024)	180	categorical	Low < Medium < High	65.93	2.22
CLEAR (Crossley et al., 2024)	1000	continuous	N/A	199.23	9.45
ELI-WHY (GPT-4) (Joshi et al., 2025)	897	categorical	Elementary < High School < Graduate	144.21	6.97
ELI-WHY (HUMAN) (Joshi et al., 2025)	117	categorical	Elementary < High School < Graduate	99.03	4.22
SCIENCEQA (Lu et al., 2022)	2295	categorical	Grade 1 < Grade 2 < ... < Grade 12	183.08	13.26

Table 1: Dataset statistics, including dataset size, readability label type (continuous vs categorical), average number of words and sentences across examples.

META RATER (READABILITY) and META RATER (PROFESSIONALISM), which were recently introduced to evaluate texts along readability and professionalism dimensions, respectively (Zhuang et al., 2025). The former considers factors such as clarity, coherence, vocabulary complexity, and sentence structure with the goal of assessing whether a reader can understand a written text, whereas the latter relies on the depth and content accessibility to determine the degree of expertise or knowledge required to comprehend a text. Additionally, we include a complementary BERT-based metric—README++ (Naous et al., 2024)—which predicts readability in terms of language learning capabilities through the use of the 6-point Common European Framework of Reference for Languages scale.

We test three different LLM-as-a-judge approaches, including the zero-shot continuous score approach by Trott and Rivière (2024) (dubbed LLM-AS-A-JUDGE CONTINUOUS 0-100). We also test a categorical setting, in which a model is tasked with predicting one of three readability labels - Elementary, High School or Graduate. We prompt the model with the same instructions provided to human annotators in Joshi et al. (2025) and, in the 5-shot setting, include the five example annotations (two Elementary, two Graduate, one High School). All LLM-as-a-judge approaches are performed using Llama-3.3-70B-Instruct with greedy decoding (temperature=0).

4.3 Results & Discussion

An ideal metric should correlate strongly with human judgments of readability. To operationalize this, and given that readability labels are ordinal,

we map the discrete labels to monotonically increasing numeric values ranging from 0 to $k - 1$. We apply a similar transformation to the outputs of model-based metrics to obtain numerical values and then compute the correlation between metric outputs and human annotations using the Kendall Tau-b coefficient (Kendall, 1938).³⁴ To assess overall performance, we report the average rank order across all datasets (Avg. Rank).

Table 2 shows that *model-based metrics systematically achieve stronger correlations with human judgments*, surpassing surface-form and psycholinguistic metrics by up to 0.24 absolute points. Notably, all three LLM-as-a-judge metrics consistently rank in the top three (average ranks 2.4–3.2), followed closely by the fine-tuned META RATER (PROFESSIONALISM) and README++ models. Looking at the disagreements between metrics, we find LLM-as-a-judge metrics to be more sensitive to specialized terminology and sentence structure, whereas fine-tuned models like README++ are more sensitive to information density and presence of connectors and cohesive devices. Comparing META RATER (PROFESSIONALISM) with META RATER (READABILITY), the latter shows an average correlation rank of 21.0, falling below psycholinguistic and surface-form metrics, where the best traditional metric achieves 8.6. This may be because examples are generally clear, grammatically correct, and coherent, leading the model to systematically assign the same readability class. Conversely, because META RATER (PROFESSIONALISM) reflects the depth and expertise demanded

³We use the implementation available in `scipy.stats`.

⁴See Appendix B for details on the categorical-to-numerical mappings used for each metric.

Type	Metric	SCIENTIFIC PAPERS (August et al., 2024)	CLEAR (Crossley et al., 2024)	ELI-WHY (GPT-4) (Joshi et al., 2025)	ELI-WHY (HUMAN) (Joshi et al., 2025)	SCIENCEQA (Lu et al., 2022)	Avg. Rank
Surface-form	# Words	0.16*	-0.06*	0.46*	0.15	0.28*	17.0
	# Sentences	0.25*	0.23*	0.38*	-0.07	0.09*	17.0
	Avg. Sentence Length	-0.15	-0.25*	0.21*	0.40*	0.39*	16.4
	Avg. Reading Time (s)	0.20*	-0.23*	0.47*	0.25*	0.32*	14.8
	# Syllables	0.22*	-0.28*	0.47*	0.28*	0.33*	13.2
	# Monosyllables	0.08	0.16*	0.39*	0.01	0.22*	18.8
	# Polysyllables	0.31*	-0.33*	0.46*	0.47*	0.41*	9.6
	# Difficult Words	0.26*	-0.40*	0.45*	0.46*	0.48*	8.6
	TE Score	0.35*	-0.18*	0.34*	0.34*	0.06*	17.2
Psycholinguistics	Automatic Readability Index	0.07	-0.33*	0.36*	0.56*	0.40*	11.0
	Coleman Liau Index	0.30*	-0.32*	0.31*	0.54*	0.35*	16.8
	Dalle Chall Readability Score	0.37*	-0.37*	0.37*	0.52*	0.22*	12.4
	Flesch Reading Grade	0.15	-0.36*	0.37*	0.58*	0.40*	11.6
	Flesch-Kincaid Reading Ease	-0.32*	0.37*	-0.35*	-0.58*	-0.36*	11.8
	Gunning Fog	0.15*	-0.37*	0.39*	0.57*	0.37*	14.0
	Linsear Write Formula	-0.06	-0.31*	0.24*	0.45*	0.40*	14.2
	SMOG Index	0.14	-0.38*	0.37*	0.59*	0.37*	12.2
Model-based	README++	0.40*	-0.45*	0.50*	0.50*	0.44*	6.2
	Meta Rater (readability)	-0.17	0.14*	0.00	0.00	0.09*	21.0
	Meta Rater (professionalism)	0.49*	-0.40*	0.51*	0.67*	0.44*	4.2
	LLM-as-a-judge (0-shot)	0.57*	-0.50*	0.49*	0.73*	0.60*	2.4
	LLM-as-a-judge (5-shot)	0.61*	-0.55*	0.43*	0.71*	0.61*	3.2
LLM-as-a-judge (continuous 0-100)	-0.56*	0.59*	-0.53*	-0.68*	-0.52*	2.4	

Table 2: **Rank correlations between readability metrics and human judgments of correctness across 5 datasets.** We report the Kendall Tau coefficient and boldface the four metrics exhibiting strongest correlations with human judgments. * indicates correlation coefficients with p-value < 0.01.

by each input, we hypothesize it better aligns with human perceptions of readability which go beyond lexical and syntactic cues (see Section 3).

Together these results demonstrate the strong performance of LLM-as-a-judge metrics. However, we highlight the trade-off with inference cost, as LLM-based evaluations typically require generating text for each instance, making them slower and more resource-intensive approaches than fine-tuned models. We also note that despite achieving the strongest correlations with human judgments (up to 0.73), **model-based metrics remain far from perfect alignment**, suggesting room for improvement.

Overall, **no single model-based metric consistently dominates**: while the continuous LLM-as-a-judge metric achieves the highest correlations on three datasets, it underperforms relative to LLM-AS-A-JUDGE (0-SHOT) on ELI-WHY (HUMAN) and SCIENCEQA. The two metrics differ considerably: the continuous variant penalizes texts containing numbers and named entities (e.g., “The Barber of Seville”), whereas the discriminative one is more sensitive to scientific terminology (e.g., “hydrophobic effect”, “endergonicity”), complex sentence structures, and equations. Despite its finer granularity, the continuous approach shows marked score saturation in SCIENCEQA (Li et al., 2025), with 81.30% of scores confined to three values.

Surface-form metrics outperform psycholin-

guistic metrics on 4 (out of 5) datasets. With the exception of ELI-WHY (HUMAN) dataset, Table 2 shows that there is always a simpler surface-level metric (e.g., # DIFFICULT WORDS, or #SYLLABLES) that is on par or outperforms popular metrics, such as the Automatic Readability Index or the Flesch Kincaid Reading Ease. Upon further analysis, we find that the stronger correlation observed for average sentence length in the ELI-WHY (GPT-4) can be attributed to length bias in the generations, where perceived readability is linked to the explanation’s length (see Figure 3).

5 Conclusion

This work tackles the inconsistency of readability definitions (and metrics) in the literature by showing that human perceptions of readability go beyond lexical and syntactic features, also considering topic and information content. Furthermore, we benchmark 20+ reference-less metrics—including LLM-as-a-judge and fine-tuned models—across five datasets. Our results show that model-based metrics correlate more strongly with human judgments than popular readability metrics, suggesting they capture more nuanced features. Together, these findings call for clearer definitions of readability and more rigorous validation of metrics, paving the way for assessments that better reflect how humans understand text.

Limitations

The analysis conducted in this paper is limited to the available datasets in the English language, therefore providing limited generalization to other languages. While we are partially motivated by the lack of high quality labeled data in other languages, a few exceptions exist namely in the French language (François and Fairon, 2012). Future work may consider expanding on this work through the creation of additional readability datasets in other languages or by expanding our analysis to other languages.

Section 3 concerns the investigation of the main factors shaping human readability judgments. While our findings are intuitive and generally aligned with prior discussion in the literature (August et al., 2024; Klare, 1974), they are based on information extracted from a single dataset in QA, potentially leading to concerns about their generalizability. However, reasoning judgments are not widely available in readability datasets, making it non-trivial to extend this analysis to other datasets. Future work could include building additional datasets, therefore, facilitating the expansion of this analysis to other domains and tasks.

Lay Summary

Readability assessment helps ensure that information can be understood by people with different backgrounds and abilities. A key goal is to automate this process and reduce the need for human evaluation.

Many datasets and methods have been developed for automatic readability assessment, but they often rely on different definitions of what makes text readable. Even today, most approaches still use basic measures, like the number of words, syllables, or sentences, to estimate readability.

In this work, we show that people’s perceptions of readability depend on more than simple text features—they are strongly influenced by the content and topic of the text. We compare traditional readability measures with more advanced model-based metrics across five datasets and find that conventional measures often fail to capture what humans consider readable. Our results emphasize the need for clearer, standardized definitions of readability and for moving beyond simple, surface-level met-

rics.

Acknowledgments

We thank the anonymous reviewers, the members of the Capital One research team for their helpful feedback.

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, and 1 others. 2024. *GPT4 Technical Report*. Preprint, arXiv:2303.08774.
- Arif Ahmed. 2023. *Beyond vocabulary: Capturing readability from children’s difficulty*. In *Proceedings of the Second Workshop on Text Simplification, Accessibility and Readability*, pages 134–141, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.
- Yuki Arase, Satoru Uchida, and Tomoyuki Kajiwara. 2022. *CEFR-based sentence difficulty annotation and assessment*. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6206–6219, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Tal August, Kyle Lo, Noah A. Smith, and Katharina Reinecke. 2024. *Know your audience: The benefits and pitfalls of generating plain language summaries beyond the "general" audience*. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, CHI '24*, New York, NY, USA. Association for Computing Machinery.
- Tal August, Lucy Lu Wang, Jonathan Bragg, Marti A. Hearst, Andrew Head, and Kyle Lo. 2023. *Paper plain: Making medical research papers approachable to healthcare consumers with natural language processing*. *ACM Trans. Comput.-Hum. Interact.*, 30(5).
- Inyoung Cheong, King Xia, K. J. Kevin Feng, Quan Ze Chen, and Amy X. Zhang. 2024. *(a)i am not a lawyer, but...: Engaging legal experts towards responsible llm policies for legal advice*. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, FAccT '24*, page 2454–2469, New York, NY, USA. Association for Computing Machinery.
- Meri Coleman and Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283.
- Kevyn Collins-Thompson. 2014. *Computational assessment of text readability: A survey of current and future research* running title: *Computational assessment of text readability*.

- S.A. Crossley, Y. Tian, P. Baffour, A. Franklin, M. Benner, and U. Boser. 2024. [A large-scale corpus for assessing written argumentation: Persuade 2.0](#). *Assessing Writing*, 61:100865.
- Michael Curtotti, Eric McCreath, Tom Bruce, Sara Frug, Wayne Weibel, and Nicolas Ceynowa. 2015. [Machine learning for readability of legislative sentences](#). In *Proceedings of the 15th International Conference on Artificial Intelligence and Law, ICAIL '15*, page 53–62, New York, NY, USA. Association for Computing Machinery.
- Edgar Dale and Jeanne S. Chall. 1948. [A formula for predicting readability](#). *Educational Research Bulletin*, 27(1):11–28.
- Vera Demberg and Frank Keller. 2008. [Data from eye-tracking corpora as evidence for theories of syntactic processing complexity](#). *Cognition*, 109(2):193–210.
- William H. DuBay. 2004. [The principles of readability](#).
- Hedayat Eslami. 2014. [The effect of syntactic simplicity and complexity on the readability of the text](#). *Journal of Language Teaching and Research*, 5(5).
- Lijun Feng, Martin Jansche, Matt Huenerfauth, and Noémie Elhadad. 2010. [A comparison of features for automatic readability assessment](#). In *Coling 2010: Posters*, pages 276–284, Beijing, China. Coling 2010 Organizing Committee.
- Rudolf Franz Flesch. 1948. [A new readability yardstick](#). *The Journal of applied psychology*, 32 3:221–33.
- Thomas François and Cédric Fairon. 2012. [An “AI readability” formula for French as a foreign language](#). In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 466–477, Jeju Island, Korea. Association for Computational Linguistics.
- Daniela B. Friedman and Laurie Hoffman-Goetz. 2006. [A systematic review of readability and comprehension instruments used for print and web-based cancer information](#). *Health Education amp; Behavior*, 33(3):352–373.
- Arthur C. Graesser, Danielle S. McNamara, Max M. Louwerse, and Zhiqiang Cai. 2004. [Coh-matrix: Analysis of text on cohesion and language](#). *Behavior Research Methods, Instruments, amp; Computers*, 36(2):193–202.
- Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. [A survey on llm-as-a-judge](#). *Preprint*, arXiv:2411.15594.
- Robert Gunning. 1952. [The technique of clear writing](#). (*No Title*).
- Yu Han, Aaron Ceross, and Jeroen HM Bergmann. 2024. [The use of readability metrics in legal text: A systematic literature review](#). *arXiv preprint arXiv:2411.09497*.
- G Harry and Mc Laughlin. 1969. [Smog grading - a new readability formula](#). *The Journal of Reading*.
- Jacob S. Hershenhouse, Daniel Mokhtar, Michael B. Eppler, Severin Rodler, Lorenzo Storino Ramacciotti, Conner Ganjavi, Brian Hom, Ryan J. Davis, John Tran, Giorgio Ivan Russo, Andrea Cocci, Andre Abreu, Inderbir Gill, Mihir Desai, and Giovanni E. Cacciamani. 2024. [Accuracy, readability, and understandability of large language models for prostate cancer information to the public](#). *Prostate Cancer and Prostatic Diseases*, 28(2):394–399.
- Brihi Joshi, Keyu He, Sahana Ramnath, Sadra Sabouri, Kaitlyn Zhou, Souti Chattopadhyay, Swabha Swayamdipta, and Xiang Ren. 2025. [Eli-why: Evaluating the pedagogical utility of language model explanations](#). *Preprint*, arXiv:2506.14200.
- M. G. Kendall. 1938. [A new measure of rank correlation](#). *Biometrika*, 30:81–93.
- Martin Kerwer, Anita Chasiotis, Johannes Stricker, Armin Günther, and Tom Rosman. 2021. [Straight from the scientist’s mouth—plain language summaries promote laypeople’s comprehension and knowledge acquisition when reading about individual research findings in psychology](#). *Collabra: Psychology*, 7(1).
- Peter Kincaid, Robert P. Fishburne, 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](#).
- George R. Klare. 1974. [Assessing readability](#). *Reading Research Quarterly*, 10(1):62.
- Gondy Leroy, Stephen Helmreich, and James R. Cowie. 2010. [The influence of text characteristics on perceived and actual difficulty of health information](#). *International Journal of Medical Informatics*, 79(6):438–449. Special Issue: Information Technology in Health Care: Socio-technical Approaches.
- Gondy Leroy, Stephen Helmreich, {James R.} Cowie, Trudi Miller, and Wei Zheng. 2008. [Evaluating online health information: beyond readability formulas](#). *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, pages 394–398.
- Qingquan Li, Shaoyu Dou, Kailai Shao, Chao Chen, and Haixiang Hu. 2025. [Evaluating scoring bias in llm-as-a-judge](#). *Preprint*, arXiv:2506.22316.

- Fengkai Liu and John Lee. 2023. [Hybrid models for sentence readability assessment](#). In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 448–454, Toronto, Canada. Association for Computational Linguistics.
- Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.
- Glenda M. McClure. 1987. [Readability formulas: Useful or useless?](#) *IEEE Transactions on Professional Communication*, PC-30(1):12–15.
- Tarek Naous, Michael J Ryan, Anton Lavrouk, Mohit Chandra, and Wei Xu. 2024. [ReadMe++: Benchmarking multilingual language models for multi-domain readability assessment](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12230–12266, Miami, Florida, USA. Association for Computational Linguistics.
- Sarah E. Petersen. 2007. *Natural language processing tools for reading level assessment and text simplification for bilingual education*. Ph.D. thesis, USA. AAI3275902.
- Emily Pitler and Ani Nenkova. 2008. [Revisiting readability: A unified framework for predicting text quality](#). In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 186–195, Honolulu, Hawaii. Association for Computational Linguistics.
- Romain Puech, Jakub Macina, Julia Chatain, Mrinmaya Sachan, and Manu Kapur. 2025. [Towards the pedagogical steering of large language models for tutoring: A case study with modeling productive failure](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26291–26311, Vienna, Austria. Association for Computational Linguistics.
- Donya Rooein, Paul Röttger, Anastassia Shaitarova, and Dirk Hovy. 2024. [Beyond flesch-kincaid: Prompt-based metrics improve difficulty classification of educational texts](#). In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 54–67, Mexico City, Mexico. Association for Computational Linguistics.
- RJ Senter and Edgar A Smith. 1967. Automated readability index. Technical report.
- Sean Trott and Pamela Rivière. 2024. [Measuring and modifying the readability of English texts with GPT-4](#). In *Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024)*, pages 126–134, Miami, Florida, USA. Association for Computational Linguistics.
- Sowmya Vajjala and Ivana Lučić. 2018. [On-eStopEnglish corpus: A new corpus for automatic readability assessment and text simplification](#). In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 297–304, New Orleans, Louisiana. Association for Computational Linguistics.
- Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, and 1 others. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *arXiv preprint arXiv:2412.13663*.
- Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. 2016. [Text readability assessment for second language learners](#). In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 12–22, San Diego, CA. Association for Computational Linguistics.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS ’23*, Red Hook, NY, USA. Curran Associates Inc.
- Xinlin Zhuang, Jiahui Peng, Ren Ma, Yinfan Wang, Tianyi Bai, Xingjian Wei, Qiu Jiantao, Chi Zhang, Ying Qian, and Conghui He. 2025. [Meta-rater: A multi-dimensional data selection method for pre-training language models](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10856–10896, Vienna, Austria. Association for Computational Linguistics.

A Additional Details: Datasets

In this section, we provide additional details about the datasets. Table 1 summarizes the general statistics about the five datasets considered in this study, including the readability label type, the size of the dataset, but also the average example length in terms of word count and sentence count.

A.1 SCIENCEQA (Lu et al., 2022)

SCIENCEQA is collected from elementary and high school science curricula sourced from IXL learning⁵ and with topics ranging from natural, social,

⁵<https://www.ixl.com>

SCIENCEQA readability example

```
Lecture: {{lecture}}  
Explanation: {{explanation}}
```

Figure 2: Formatting of each SCIENCEQA example. Whenever examples miss the corresponding `{{lecture}}` or `{{explanation}}` fields, we omit them from the template above.

and natural sciences. To ensure coverage across grades 1–12, we sample from the full dataset. We draw 200 examples per grade, except for 1st grade where only 95 are available, yielding 2295 examples overall. Although primarily a multiple-choice QA dataset, it also includes a *lecture* covering the knowledge needed to answer each question and a solution outlining how to use it to derive the answer. For every question, we compute the readability by concatenating the two fields as demonstrated in Figure 2. For some qualitative examples, see Table 3. To compute the correlation with human judgments, we use grades 1-12 as the readability judgments (12-way classification), where a higher grade implies added difficulty in comprehending a text.

A.2 CLEAR (Crossley et al., 2024)

CLEAR consists of 4.7k text excerpts sampled from online digital libraries. Each example is curated to ensure the text is self-contained and composed of full sentences. Unlike the other datasets, the readability score in CLEAR is continuous and represents the easiness of comprehension of a given text (*BT_easiness*). We refer to the original paper for additional details regarding the dataset. Table 5 illustrates a few examples from this dataset and corresponding readability score. To balance efficiency with generalization, we randomly sample 1k examples without replacement from the original dataset and use them for our correlation analysis. Table 6

A.3 SCIENTIFIC PAPERS (August et al., 2023)

SCIENTIFIC PAPERS dataset is a parallel corpus for readability, comprising 3 human-edited variants of the same summary for each example. Table 7 shows three human-curated versions of the question “What did the paper find?” at different complexity levels. The correlation analysis considers we all examples and map the ordinal classes—Low <

Medium < High —onto a 0–2 scale.

A.4 ELI-WHY (GPT-4) (Joshi et al., 2025)

Our analyses reveal the presence of length bias, where there seems to exist a correlation between the length of GPT4-generated explanations and human perceived readability (see Figure 3). In fact, we observe a propensity for responses deemed higher readability to be longer, which can be explained by the added detail and specificity often emphasized by human experts. Future work could explore ways of mitigating this bias by enforcing strict generation lengths or, if a reference document with relevant information is available by controlling the information content within each generation (August et al., 2024).

A.5 ELI-WHY (HUMAN) (Joshi et al., 2025)

Table 8 illustrates a few randomly selected examples for the ELI-WHY (HUMAN) datasets. These explanations were manually curated by two authors of the paper.

B Automated Metrics

This section discusses the implementation details of the metrics evaluated in the main paper. All experiments are implemented in Python: `textstat`⁶ is used to compute surface-form and psycholinguistic metrics; `transformers` is used to implement the model-based metrics, including fine-tuned and LLM-as-a-judge approaches.

B.1 Psycholinguistics Metrics

The metrics listed below are commonly referred to as *readability tests* and commonly used to gauge the difficulty that human readers may have in understanding a given text.

Automatic Readability Index (ARI) (Senter and Smith, 1967) estimates the US grade level needed to comprehend a text. To do so, it uses the ratio of characters-to-words and words-to-sentences. Intuitively, these ratios capture the idea that longer words and longer sentences are more difficult to grasp. The character counts include both numbers and letters. A score of 1 and 14 would match that of a Kindergarten and a College student,

⁶<https://pypi.org/project/textstat/>

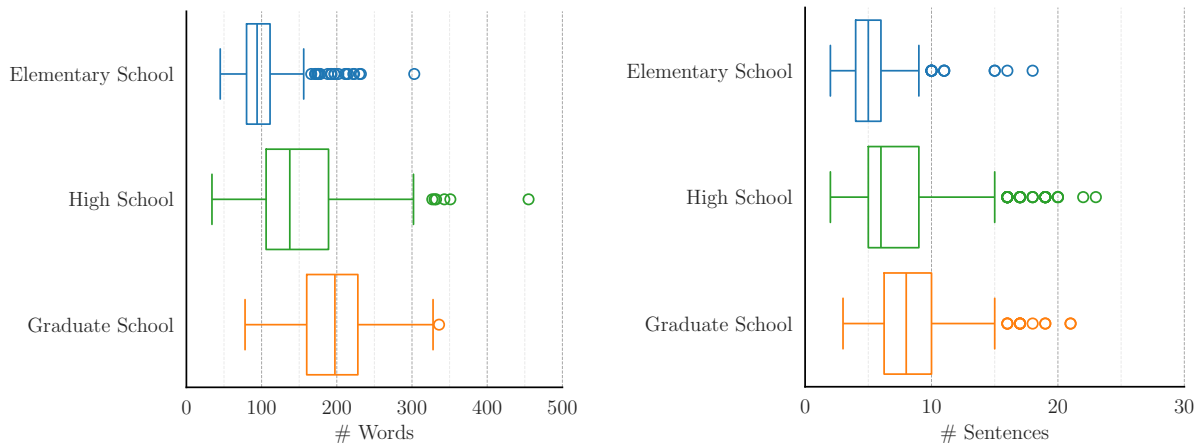


Figure 3: Distribution of number of words (# WORDS) and sentences (# SENTENCES) per readability label in the ELI-WHY (GPT-4) dataset.

respectively.

$$\left[4.71 \left(\frac{\#chars}{\#words} \right) + 0.5 \left(\frac{\#words}{\#sentences} \right) - 21.43 \right]$$

Coleman Liau Index (CLI) (Coleman and Liau, 1975) similarly to ARI, it also yields an estimate of the minimum US grade level necessary to understand a piece of text. It is defined in terms of the average counts of letters and sentences per 100 words in a text sample.

$$0.0588 \cdot \#letters - 0.296 \cdot \#sentences - 15.8$$

Dale Chall Readability (Dale and Chall, 1948) leverages the fraction of difficult words in the document, as well as the average word-to-sentence count ratio to gauge the difficulty of a given text. By design, the metric relies on a pre-defined subset of 3k words that is empirically expected to be familiar to the majority of 4th graders. The formula is designed such that scores ≤ 4.9 match grade 4 and below, and scores ≥ 10 match grades 16 and above. Below we write the new Dale-Chall Formula:

$$\left[64 - 0.95 \left(\frac{\#difficult_words}{\#words} \right) - 0.69 \left(\frac{\#words}{\#sentences} \right) \right]$$

Flesch-Kincaid Reading Ease (FKRE) and **Flesch-Kincaid Grade Level (FKGL)** (Flesch, 1948) rely on the same core properties of language, such as average word length and average sentence length, differing only in the coefficients. The formulas were defined by the US Navy to gauge the readability of the technical material and later adopted by a few US states to impose readability

requirements on various legal documents (*e.g.*, insurance policies) (McClure, 1987). The FKRE is defined in as follows:

$$206.835 - 1.015 \left(\frac{\#words}{\#sentences} \right) - 84.6 \left(\frac{\#syllables}{\#words} \right)$$

whereas the FKGL is defined as:

$$0.39 \left(\frac{\#words}{\#sentences} \right) + 11.8 \left(\frac{\#syllables}{\#words} \right) - 15.59$$

Gunning Fog Index (GFI) (Gunning, 1952) provides an estimate of the number of formal education required to understand the text on a first reading. It works by first computing the average sentence length, *i.e.*, word-to-sentence ratio of a passage and then computing the ratio of complex words in the passage. In this formula, complex words are defined as words with 3+ syllables that are not proper nouns, familiar words, or compound words. Conventionally, scores range between 6 and 17 which indicate that 6th grade and College Graduate are necessary to be able to understand a piece of text, respectively.

$$0.4 \left[\left(\frac{\#words}{\#sentences} \right) + 100 \left(\frac{\#complex_words}{\#words} \right) \right]$$

Linsear Write Formula (LWF) (Klare, 1974) counts the number of easy and hard words in a 100-word sample. To distinguish easy from hard words, it utilizes the number of syllables in each word: polysyllable words are considered hard words, whilst words with less than 3 syllables are considered easy. It was originally designed to gauge

the readability of the technical manuals used in the US Air Force.

$$r = \frac{3 \cdot \# \text{hard_words} + 1 \cdot \# \text{easy_words}}{\# \text{words}}$$

where the final linear write score is given by

$$\text{LWF} = \begin{cases} r/2 & \text{if } r > 20 \\ r/2 - 1 & \text{else} \end{cases}$$

SMOG grade (Harry and Laughlin, 1969) was proposed as a more accurate and easier to compute alternative to Gunning Fog Index. It is defined in terms of polysyllable counts (words with 3+ syllables) across three 10-sentence long texts.

$$1.043 \sqrt{\# \text{polysyllables} \cdot \frac{30}{\# \text{sentences}}} + 3.1291$$

B.2 Model-based Metrics

META RATER (PROFESSIONALISM) and **META RATER (READABILITY)** (Zhuang et al., 2025) are two fine-tuned based metrics, both operationalized using a ModernBERT-base model. The models are designed to evaluate the *degree of required expertise* and *ease of understanding* in a 0-5 point scale, respectively. To obtain the metric score associated with a given text, each text is fed through the model and the class with maximum probability is selected (*i.e.*, greedy prediction). This score is then used to compute the correlation with human judgments.

README++ (Naous et al., 2024) is a model-based metric that grounds readability assessment in the capabilities of second-language learners. Specifically, we use tareknaous/readabert-en, a BERT-based model fine-tuned on the English portion of the README++ corpus—a sentence-level readability dataset spanning multiple domains (*e.g.*, finance, economics, poetry, agriculture). Readability scores are provided on a six-point scale aligned with the Common European Framework of Reference for Languages (CEFR), where higher values indicate greater language proficiency.

Since README++ was originally trained on single sentences, we hypothesize that it may not generalize well to multi-sentence inputs, such as those in SCIENCEQA or ELI-WHY (GPT-4). To address this limitation, we adopt a bottom-up approach: for each document, we first compute the README++ score for each sentence, then average them to obtain a document-level score (README++ (AVG)).

We also evaluate another variant, README++ (MAX), which reflects the hypothesis that advanced readers can understand simpler texts, but not vice versa. Table 9 summarizes the results. While both README++ and README++ (MAX) exhibit the same average rank (1.8), we observe that README++ exhibits stronger correlations with human judgments in 3 (out of 5) evaluated datasets. Notably, README++ (AVG) exhibits an average rank of 2.4, suggesting that this variant systematically underperforms the other two variants in terms of correlating with human judgments. For brevity, and because of its superior performance, we restrict the analysis in the main paper to the original method—README++.

LLM-AS-A-JUDGE (0-SHOT) and **LLM-AS-A-JUDGE (5-SHOT)** are prompt-based strategies to extract the readability level from any text. Specifically, we use Llama-3.3-70B-Instruct⁷, a popular open-source instruction-following model. To ensure that models’ predictions strongly align with human readability judgments, we re-use instructions previously provided to humans (Joshi et al., 2025). The prompt is discriminative in nature, being designed to extract 3-way readability labels – Elementary, High School, Graduate. Originally, the prompt includes 5 examples of readability judgments spanning the three classes, which we refer to as LLM-AS-A-JUDGE (5-SHOT). Although these examples improve alignment with human judgments and help constrain output structure, they add runtime overhead. We therefore evaluate a 0-shot version that replaces examples with explicit format instructions. Since our goal is to compute correlations with human judgments, which can be expressed as either categorical or continuous, we map textual labels to numbers. Treating the labels as ordinal (Elementary < High School < Graduate), we assign them a 0–2 scale for correlation analysis.

LLM-AS-A-JUDGE (CONTINUOUS 0-100), first proposed in Trott and Rivière (2024), elicits its continuous 0-100 readability scores from GPT-4-Turbo and GPT-4o-mini (Achiam et al., 2024), with higher values denoting easier texts to understand. We follow the same prompt as in the original paper (Figure 4), but replace the model

⁷<https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct>

LLM-AS-A-JUDGE (CONTINUOUS 0-100)

```
{
  "content": "You are an experienced teacher
skilled at identifying the readability of
different texts.",
  "role": "system"
}, {
  "content": "Read the text below. Then
indicate the readability of the text, on a
scale from 1 (extremely challenging to
understand) to 100 (very easy to read and
understand). In your assessment, consider
factors such as sentence structure,
vocabulary complexity, and overall
clarity.\n<Text>{{text}}</Text>\nOn a scale
from 1 (extremely challenging to
understand) to 100 (very easy to read and
understand), how readable is this text?
Please answer with a single number.",
  "role": "user"
}
```

Figure 4: Prompt used to extract a 0-100 continuous score associated with the ease of readability of a given text. The placeholder `{{text}}` is either the explanation to a question or the text excerpts depending on the dataset being evaluated.

with Llama-3.3-70B-Instruct to ensure comparability among LLM-as-a-judge metrics.⁸

In the main paper, we ensure the reproducibility of LLM-as-a-judge evaluations by reporting correlations obtained from greedy generations (temperature=0).⁹ This decoding strategy is not only deterministic but also commonly adopted in prior work (Trott and Rivière, 2024; Gu et al., 2025), being representative of the most likely (or modal) behavior of the LLM.

C Human Perceptions of Readability

In the main paper, we examine the reasons driving the human’s annotations of various perceived readability levels. To this end, we employ various automatic pattern extraction techniques, including frequency-based analysis (represented in the form of wordclouds) and n-gram feature importance. The following sections provide additional details about each of these experiments.

⁸Llama-3.3-70B-Instruct consistently generates a number between 1–100.

⁹Continuous LLM-as-a-judge approaches (LLM-AS-A-JUDGE (CONTINUOUS 0-100)) are configured to generate at most 3 tokens, whereas the discriminative approaches (LLM-AS-A-JUDGE (0-SHOT) and LLM-AS-A-JUDGE (5-SHOT)) are configured to generate at most 20 tokens. We then extract the corresponding readability label through the use of regular expressions.

C.1 Frequency-based Analysis

As part of our analysis, we conduct a frequency-based analysis of the rationales behind the readability judgments provided by the human annotators in the ELI-WHY (GPT-4) dataset.

Methodology. We conduct our analysis by first separating the dataset into three subsets according to the perceived readability level of the GPT4-generated explanations. In doing so, we obtain a total of 324, 694, and 182 examples corresponding to the **Elementary**, **High School**, and **Graduate**, respectively. Subsequently, we merge the annotators justification field for each subset, remove the English stopwords (as provided by the NLTK library). To aggregate words with similar meanings, we further lemmatize each word using the WORDNETLEMMATIZER¹⁰.

C.2 Predictive Analysis

We also conduct a model-based approach to determine the discriminative power of different phrases for each readability class. In this analysis, each annotator’s justification is considered to be an individual document and both term and document frequencies are used to determine the readability class of a annotators’ justifications.

Methodology. Similarly to the frequency-based analysis, we first decompose the ELI-WHY (GPT-4) dataset into three exclusive subsets based on the human perceived readability label. Additionally, we expand the justification field into individual documents, resulting in 707, 1665, and 416 total documents for **Elementary**, **High School**, and **Graduate**, respectively. As preprocessing steps, we remove the English stopwords using the NLTK default list, lemmatize the text using the WORDNETLEMMATIZER, and lowercase the text. Finally, we compute the term-to-document frequency matrix using SKLEARN’s TFIDFVECTORIZER. To ensure that we capture complex phrases and not just individual words, we consider n-grams where $n \in \{1, 2, 3, 4\}$ and, to avoid overfitting to terms that appear in a single document, set `MIN_DF=2`.

Having the term-to-document frequency matrix, we adopt a one-vs-all approach, where we itera-

¹⁰<https://www.nltk.org/api/nltk.stem.WordNetLemmatizer.html>

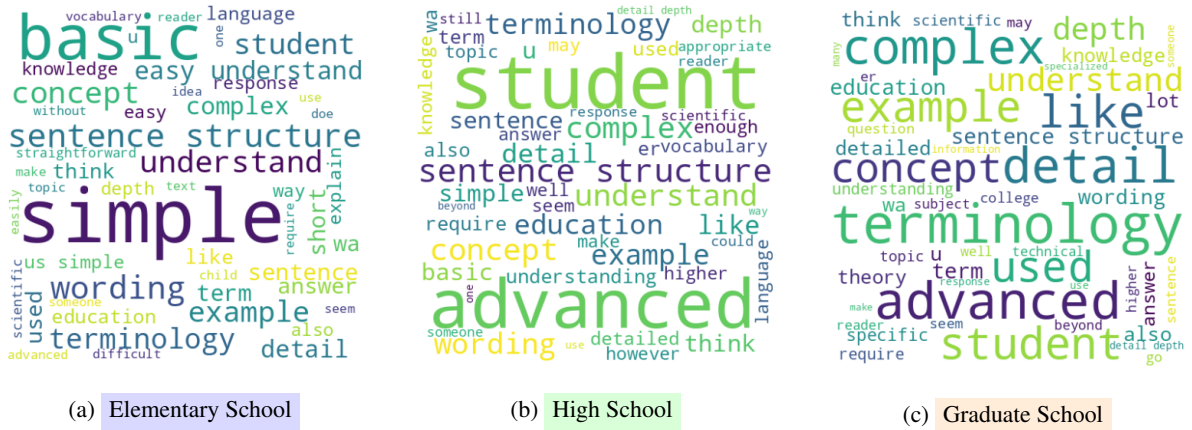


Figure 5: Frequency-based analysis of the language expressions used by human annotators when judging the perceived readability of various GPT4-generated explanations in ELI-WHY (GPT-4). These word clouds are collected over 324, 694, and 182 examples annotated for `Elementary`, `High School`, and `Graduate`, respectively.

tively fit a linear model to discriminate justifications of one class (e.g., `Elementary`) from justifications outside of this class (e.g., `High School` and `Graduate`). While focusing on linear models such as logistic regression allow us to directly examine the predictive importance of different n-grams, it pre-assumes that the most is a strong predictor. With the intent of selecting a good predictive model, we perform hyperparameter optimization using 10-fold cross-validation while using predictive accuracy as the evaluation criteria. We consider the following hyperparameters and employ grid search:

- estimator = `LogisticRegression()`
- max_iter = `{100, 300}`
- C = `{0.01, 0.1, 1, 10, 100, 500}`
- penalty = `{l1, l2, elasticnet}`
- solver = `{liblinear, saga}`

We list the best obtained models for each readability class in Table 10. Across all readability classes, we find that the fitted logistic regression outperforms a simple baseline that predicts the majority class (MAJORITY ACCURACY) by at least 3% and up to 15% absolute points.

D Related Work

In this section, we extend the discussion of readability metrics provided in the main paper. Specifically,

we elaborate on the limitations of the previously proposed LLM-as-a-judge approaches and remaining challenges.

Readability Assessment using LLMs. Rooein et al. (2024) show that combining yes/no prompts with conventional metrics yields stronger correlations with human judgments than using either set of metrics alone. Trott and Rivière (2024) use 0-shot prompts to extract continuous readability scores which correlate strongly with human judgments. In spite of promising results, these approaches have seen little adoption in practice. Their reliance on repeated prompting introduces significant inference overhead, making them costly for large-scale evaluation or use as reward functions. They also require allocating part of the already limited readability data to calibrate combinations or thresholds, further limiting their practicality. Finally, although prior work has explored continuous readability assessments with LMs, to our knowledge their ability to distinguish coarse-grained readability classes remains unexplored.

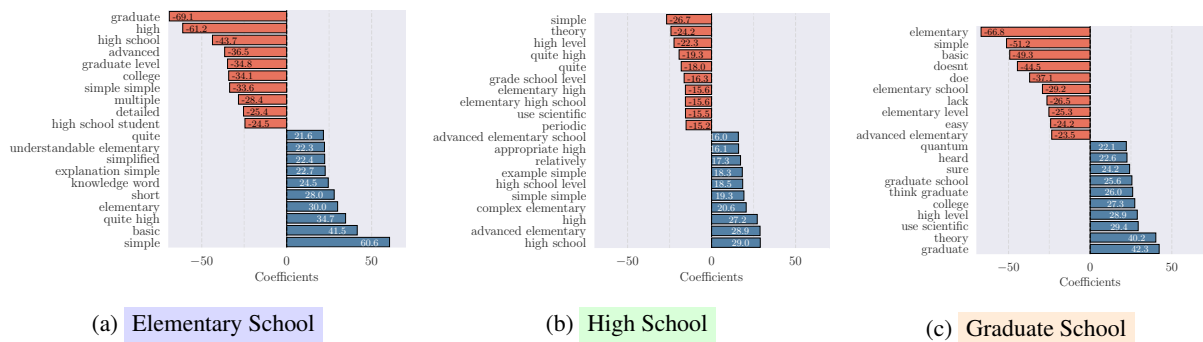


Figure 6: Regression analysis of the language expressions used by human annotators when judging the perceived readability of various GPT4-generated explanations in ELI-WHY (GPT-4). These values clouds are collected over 324, 694, and 182 examples annotated for **Elementary**, **High School**, and **Graduate**, respectively.

Table 3: Randomly selected ScienceQA examples across 6 different readability classes (**grade**).

Grade	Subject (Category)	Formatted Example
1	language science (comprehension strategies)	Explanation: A book is made of paper.\nA book tells a story.\nA teacher may read a book out loud.
3	natural science (weather and climate)	Lecture: The atmosphere is the layer of air that surrounds Earth. Both weather and climate tell you about the atmosphere.\nWeather is what the atmosphere is like at a certain place and time. Weather can change quickly. For example, the temperature outside your house might get higher throughout the day.\nClimate is the pattern of weather in a certain place. For example, summer temperatures in New York are usually higher than winter temperatures.\n\nExplanation: Read the text carefully.\nWhere Sarah lives, winter is the rainiest season of the year.\nThis passage tells you about the usual precipitation where Sarah lives. It does not describe what the weather is like on a particular day. So, this passage describes the climate.
5	natural science (traits and heredity)	Lecture: Organisms, including people, have both inherited and acquired traits. Inherited and acquired traits are gained in different ways.\nInherited traits are passed down through families. Children gain these traits from their parents. Inherited traits do not need to be learned.\nAcquired traits are gained during a person's life. Some acquired traits, such as riding a bicycle, are gained by learning. Other acquired traits, such as scars, are caused by the environment. Children do not inherit their parents' acquired traits.\n\nExplanation: People are not born knowing how to cook. Instead, many people learn how to cook. So, cooking is an acquired trait.
7	natural science (designing experiments)	Lecture: Experiments can be designed to answer specific questions. When designing an experiment, you must identify the supplies that are necessary to answer your question. In order to do this, you need to figure out what will be tested and what will be measured during the experiment.\nImagine that you are wondering if plants grow to different heights when planted in different types of soil. How might you decide what supplies are necessary to conduct this experiment?\nFirst, you need to identify the part of the experiment that will be tested, which is the independent variable. This is usually the part of the experiment that is different or changed. In this case, you would like to know how plants grow in different types of soil. So, you must have different types of soil available.\nNext, you need to identify the part of the experiment that will be measured or observed, which is the dependent variable. In this experiment, you would like to know if some plants grow taller than others. So, you must be able to compare the plants' heights. To do this, you can observe which plants are taller by looking at them, or you can measure their exact heights with a meterstick.\nSo, if you have different types of soil and can observe or measure the heights of your plants, then you have the supplies you need to investigate your question with an experiment!
9	language science (literary devices)	Lecture: Figures of speech are words or phrases that use language in a nonliteral or unusual way. They can make writing more expressive.\nA euphemism is a polite or indirect expression that is used to de-emphasize an unpleasant topic.\nThe head of Human Resources would never refer to firing people, only to laying them off.\nHyperbole is an obvious exaggeration that is not meant to be taken literally.\nI ate so much that I think I might explode!\nAn oxymoron is a joining of two seemingly contradictory terms.\nSome reviewers are calling this book a new classic.\nA paradox is a statement that might at first appear to be contradictory, but that may in fact contain some truth.\nAlways expect the unexpected.\n\nExplanation: The text uses an oxymoron, a joining of two seemingly contradictory terms.\nOpen secret is a contradiction, because open describes something that is freely or publicly known, and a secret is hidden.

Table 3: Randomly selected ScienceQA examples across 6 different readability classes (**grade**). (continued)

Grade	Subject (Category)	Formatted Example
11	language science (word usage and nuance)	<p>Lecture: Words change in meaning when speakers begin using them in new ways. For example, the word peruse once only meant to examine in detail, but it's now also commonly used to mean to look through in a casual manner.</p> <p>When a word changes in meaning, its correct usage is often debated. Although a newer sense of the word may be more commonly used, many people consider a word's traditional definition to be the correct usage. Being able to distinguish the different uses of a word can help you use it appropriately for different audiences.</p> <p>Britney perused her notes, carefully preparing for her exam.</p> <p>The traditional usage above is considered more standard.</p> <p>David perused the magazine, absentmindedly flipping through the pages.</p> <p>The nontraditional usage above is now commonly used, but traditional style guides generally advise against it.</p> <p>Explanation: The first text uses travesty in its traditional sense: a ridiculous imitation; a parody.</p> <p>Doug's ill-researched essay about the Space Race received a poor grade because it presented such a travesty of the actual historical events.</p> <p>The second text uses travesty in its nontraditional sense: a disappointment or a tragedy.</p> <p>Doug realized that his essay about the Space Race was a bit inaccurate, but he still thought it a travesty that such an entertaining essay should receive a poor grade.</p> <p>Most style guides recommend to use the traditional sense of the word travesty because it is considered more standard.</p>

Table 4: Randomly selected examples from the CLEAR dataset. In contrast to other datasets, each example is associated with a continuous readability score obtained by fitting a Bradley–Terry model on pairwise human judgments of reading ease.

Readability Score	Category	Text
-2.91	Info	It must not be supposed that, in setting forth the memories of this half-hour between the moment my uncle left my room till we met again at dinner, I am losing sight of "Almayer's Folly." Having confessed that my first novel was begun in idleness—a holiday task—I think I have also given the impression that it was a much-delayed book. It was never dismissed from my mind, even when the hope of ever finishing it was very faint. Many things came in its way: daily duties, new impressions, old memories. It was not the outcome of a need—the famous need of self-expression which artists find in their search for motives. The necessity which impelled me was a hidden, obscure necessity, a completely masked and unaccountable phenomenon. Or perhaps some idle and frivolous magician (there must be magicians in London) had cast a spell over me through his parlour window as I explored the maze of streets east and west in solitary leisurely walks without chart and compass. Till I began to write that novel I had written nothing but letters, and not very many of these.
-1.44	Info	In the second place, the Emperor is an exceedingly intelligent and highly cultivated man. His mental processes are swift, but they go also very deep. He is a searching inquirer, and questions and listens more than he talks. His fund of knowledge is immense and sometimes astonishing. He manifests interest in everything, even to the smallest detail, which can have any bearing upon human improvement. I remember a half hour's conversation with him once over a cupping glass, which he had gotten from an excavation in the Roman ruin called the Saalburg, near Homburg. He always appeared to me most deeply concerned with the arts of peace. I have never heard him speak much of war, and then always with abhorrence, nor much of military matters, but improved agriculture, invention, and manufacture, and especially commerce and education in all their ramifications, were the chief subjects of his thought and conversation. I have had the privilege of association with many highly intelligent and profoundly learned men, but I have never acquired as much knowledge, in the same time, from any man whom I have ever met, as from the German Emperor.
-1.21	Literary	Moreover Grandmother Grant always dressed in one fashion; she had a calico dress for morning and a black silk for the afternoon, made with an old-fashioned surplice waist, with a thick plaited ruff about her throat; she sometimes tied a large white apron on, but only when she went into the kitchen; and she wore a pocket as big as three of yours, Matilda, tied on underneath and reached through a slit in her gown. Therein she kept her keys, her smelling-bottle, her pocket-book, her handkerchief and her spectacles, a bit of flagroot and some liquorice stick. I mean when I say this, that all these things belonged in her pocket, and she meant to keep them there; but it was one peculiarity of the dear old lady, that she always lost her necessary conveniences, and lost them every day.\n"Maria!" she would call out to her daughter in the next room, "have you seen my spectacles?"\n"No, mother; when did you have them?"\n"Five minutes ago, darnin' Harry's stockings; but never mind, there's another pair in the basket."
-0.37	Literary	The others were watching him closely. They guessed something of the nature of what must be passing through Ned's mind, for both Jack and Teddy followed his gaze up the uneven shore. Jimmy had the glasses again, and was busily engaged in scrutinizing the blur on the distant horizon, which all of them had agreed must be smoke hovering close to the water. Perhaps he half-believed the fanciful suggestion made by Teddy, with reference to Captain Kidd, and was wildly hoping to discover some positive sign that would stamp this fairy story with truth. All the previous adventures that had befallen himself and chums would sink into utter insignificance, could they go back home and show evidences of having made such a romantic discovery up there in the Hudson Bay country.\n"See the feather they say he always wore in his hat, Jimmy?" asked Frank.
0.06	Literary	The other day, as I was walking through a side street in one of our large cities, I heard these words ringing out from a room so crowded with people that I could but just see the auctioneer's face and uplifted hammer above the heads of the crowd.\n"Going! Going! Going! Gone!" and down came the hammer with a sharp rap.\nI do not know how or why it was, but the words struck me with a new force and significance. I had heard them hundreds of times before, with only a sense of amusement. This time they sounded solemn.\n"Going! Going! Gone!"\n"That is the way it is with life," I said to myself - "with time." This world is a sort of auction room; we do not know that we are buyers: we are, in fact, more like beggars; we have brought no money to exchange for precious minutes, hours, days, or years; they are given to us. There is no calling out of terms, no noisy auctioneer, no hammer; but nevertheless, the time is "going! going! gone!"

Table 5: Randomly selected examples from the CLEAR dataset. In contrast to other datasets, each example is associated with a continuous readability score obtained by fitting a Bradley–Terry model on pairwise human judgments of reading ease. (continued)

Readability Score	Category	Text
0.19	Info	There are various kinds of pitcher-plants. Some are shorter and broader than others; but they are all green like true leaves, and hold water as securely as a jug or glass. They grow in Borneo and Sumatra, hot islands in the East. The one shown in the drawing grows in Ceylon. Some grow in America; but they are altogether different from those in Borneo and Ceylon. One beautiful little pitcher-plant grows in Australia: but this is also very different from all the rest; for the pitchers, instead of being at the end of the leaves, are clustered round the bottom of the plant, close to the ground. All these pitcher-plants, though very beautiful to look at, are very cruel enemies to insects: for the pitchers nearly always have water in them; and flies and small insects are constantly falling into them, and getting drowned.

Table 6: Comparison of readability scores between the original CLEAR (Original) and the 1k subsample used to conduct the correlation analysis (Subsample).

Statistic	Original	Subsample
Count	4724	1000
Mean	-0.96	-0.97
Std	1.03	1.06
Min	-3.68	-3.68
25%	-1.70	-1.74
50%	-0.91	-0.89
75%	-0.20	-0.20
Max	1.71	1.71

Table 7: Randomly selected examples from the SCIENTIFIC PAPERS dataset, spanning all three readability classes.

Complexity Level	Text
Low	The researchers found that women who lived in countries that received less US foreign aid during the policy used less contraceptives and had both more pregnancies and more abortions during the years that the policy was in place. They also noted that the effects of the policy reversed once it had been rescinded, further strengthening the researchers' hypothesis that the Mexico City Policy has an effect on a nation's observed patterns of reproductive behavior.
Medium	The researchers found that abortions and pregnancies increased when the Mexico City Policy was in effect, which they correlate to a decreased availability in contraception during those years. They also found that the effects varied by exposure to the policy, as women in high exposure countries were more likely to experience abortion when the policy was enacted and less likely when it wasn't in effect. The alternating patterns of reproductive behavior depending on whether the policy was enacted also strengthens the researchers' hypothesis that it has a not insubstantial effect on abortion rates in sub-Saharan Africa.
High	When US support for international family planning organizations was conditioned on the policy, coverage of modern contraception fell and the proportion of women reporting pregnancy and abortions increased, in relative terms, among women in countries more reliant on US funding. Although the degree to which abortions increase when contraceptive supply is curtailed is poorly characterized, one analysis estimated that, depending on the total fertility of the population, a 10% decline in contraceptive use would lead to a 20-90% increase in abortions. The researchers posit that the observed changes in abortion could be due to changing availability of modern contraception, and that a change in the use of modern\contraception would be expected to result in a change in pregnancy rates. Women in high-exposure countries experienced a relative increase in abortion (and decrease in modern contraceptive use) when the policy was enacted and a relative decrease in abortion (and increase in modern contraceptive use) when the policy was rescinded.

Low	Study looks at pushup capacity and heart health, finding that those who could do the most (over 40) push ups had the lowest risk of heart disease.
Medium	Study examines the relationship between a person's push up ability and their physical health, finding that push ups are a good indicator of a person's cardiovascular fitness.
High	Association Between Pushup Exercise Capacity and Future Cardiovascular Events Among Active Adult Men

Table 8: Examples of different explanations for the ELI-WHY (HUMAN) for the questions “Why do we enjoy horror movies or stories?” and “Why does DNA have a double helix structure?”. Each set of three examples refers to the same question.

Readability	Topic	Formatted Explanation
Elementary	Psychology	All the same reasons people like sad songs, Halloween, war documentaries, apocalyptic fiction, etc. etc. It’s like any other film genre. Horror movies can be artistic; the performances can be entertaining; the movies can be well-constructed or conceived; they can be relatable or provide personal insight.
High School	Psychology	According to these researchers, stimulation is one of the driving forces behind the consumption of horror. Exposure to terrifying acts like stories of demonic possession or alien infestation can be stimulating both mentally and physically. These experiences can give rise to both negative feelings, such as fear or anxiety, and positive feelings, such as excitement or joy. And we tend to feel the most positive emotions when something makes us feel the most negative ones.
Graduate	Psychology	“The horror film occupies in popular culture roughly comparable to that of horror literature. That is to say, it is generally ignored, sometimes acknowledged with bemused tolerance, and viewed with alarm when it irritates authority - rather like a child too spirited to follow the rules that rendition has deemed acceptable” (p. ix), a view that is echoed elsewhere. For example, Tudor (1997) noted that “a taste of horror is a taste for something seemingly abnormal and is therefore deemed to require special attention” (p. 446). Part of the reason for the disdain, apart from the broad and base nature of the content, may be the relative cheapness of horror film: these are often much less expensive to create than are other genre films such as westerns, comedies, or science fiction.
Elementary	Biology	DNA is made up of small components called nucleotides. A nucleotide is made up of 3 parts: a phosphate group, a sugar, and a base. The base can be 1 of 4 varieties: A, T, C, or G. Simply put, because of their structures, A and T bond nicely with one another and C and G bond nicely as well. As a result, a single strand of DNA will bond to a strand with another strand with a "complimentary" sequence of bases. In other words, there will be two strands with opposite, for lack of a better word, sequences of bases. The production of strands that complement one another is a result of how DNA copies itself (known as DNA replication). The second strand’s base sequence is ordered based on the first strand’s sequence. So that answers the question of why DNA is a double helix instead of a single helix. But why is it a helix at all? Why not a shaped like a ladder? In a cell, most of the material present is water. Water is shaped in such a way where positive and negative charges aren’t spread evenly throughout the water’s molecules. This is known as being a polar molecule. Polar molecules "like" being around other polar molecules. Non-polar molecules don’t "like" being around polar molecules like water. In DNA, the bases are non-polar, but the phosphate groups are polar. As a result, the preferred shape puts the phosphates in contact with the water and the bases are covered by being on the inside. The twisting shape of DNA reduces the extent to which the bases are exposed to the water in the outside environment. TL;DR: DNA is made of 3 components: bases, sugars, and phosphates. The sugars bond with one another explaining the double-configuration of DNA. To be stable, the bases need to stay in the inside of the molecule and phosphates need to make up the molecule’s exterior, and a helical shape allows this to happen.
High School	Biology	The three-dimensional structure of DNA—the double helix—arises from the chemical and structural features of its two polynucleotide chains. Because these two chains are held together by hydrogen bonding between the bases on the different strands, all the bases are on the inside of the double helix, and the sugar-phosphate backbones are on the outside. To maximize the efficiency of base-pair packing, the two sugar-phosphate backbones wind around each other to form a double helix, with one complete turn every ten base pairs.
Graduate	Biology	The stability of double-stranded DNA (dsDNA) at physiological temperature is due to the hydrogen bonding between complementary bases and the stacking between neighboring bases. However, these base-stacking interactions are of the order of magnitude of a few $k_B T$ thermal energy and the thermal fluctuations can lead (even at physiological temperature) to local and transient unzipping of the double helix.

Metric	SCIENTIFIC PAPERS (August et al., 2024)	CLEAR (Crossley et al., 2024)	ELI-WHY (GPT-4) (Joshi et al., 2025)	ELI-WHY (HUMAN) (Joshi et al., 2025)	SCIENCEQA (Lu et al., 2022)	Avg. Rank
README++	0.40	-0.45	0.50	0.50	0.44	1.8
README++ (AVG)	0.23	-0.49	0.26	0.68	0.38	2.4
README++ (MAX)	0.35	-0.51	0.43	0.57	0.42	1.8

Table 9: Rank correlations between variants of the README++ metric and human judgments of correctness across 5 datasets. We boldface the variant exhibiting strongest correlation with human judgments. We report the Kendall Tau coefficient. All correlation coefficients are statistically significant with p-value < 0.01.

Table 10: Hyperparameter configurations of the Logistic Regression models fit for each readability class. We use a grid search to find the optimal combination over the hyperparameters C, PENALTY, and SOLVER. The best configuration is defined as the best achieving accuracy determined using 10-fold cross-validation.

Readability Class	Hyperparameters	Majority Accuracy (%)	Best Accuracy (%)
Elementary	C = 100 max_iter = 300 penalty = 11 solver = saga	74.64	88.05
High School	C = 500 max_iter = 100 penalty = 11 solver = saga	59.72	75.11
Graduate	C = 100 max_iter = 300 penalty = 11 solver = saga	85.08	88.34

Table 11: Human rationales underlying readability judgments across 3 different readability classes: Elementary , High School , Graduate . Each row refers to the analysis of the same “Why” question but different GPT-4 explanation, being sourced from ELI-WHY (GPT-4) (Joshi et al., 2025).

Elementary	High School	Graduate
<ul style="list-style-type: none"> - It’s probably too verbose for elementary levels, but I think people reading at that level could understand this explanation. The words are short enough. - The explanation uses basic English language to interpret why humans are inclined towards social interactions. There are not many technical or professional terminologies, making it easy to understand. The sentence structures are simple, making it easy to follow. 	<ul style="list-style-type: none"> - Pretty easy and straight forward to understand. Not using complex words or scientific words. - The sentences are short in length and easy to digest. It uses terms like “elements” and “conductivity and ductility” which require deeper understanding of elements and reactions. - The explanation is written in a way that is easy to understand, but the details and some of the words used such as “corrosion” would make it difficult for an elementary reader to comprehend. However, the material is not so specialized that you would learn it on the graduate level, meaning this falls into the high school reading level. 	<ul style="list-style-type: none"> - The terminology seems higher level and more complicated than elementary or high school; - This is borderline HS/GS to me. But the terms “parasocial” and “existential fears” are a bit much for a typical high school student. It should be simplified a bit for an HS student.
<ul style="list-style-type: none"> - The details are very surface level and it uses simple wording. - Simple sentence structure with simple and short explanations. Not detailed or in depth. - They used simple wording and examples to make their point. - It uses simple words like electricity, and can be easily understood - It gives clear examples like copper being easy to stretch and not rusting, the sentences are short and straightforward. It gives enough detail to understand why copper is used in wires. 	<ul style="list-style-type: none"> - The wording/terminology, examples, and details suggest high school-level engagement. It lacks the technicality of graduate school while being too advanced for elementary school; - Using terminology like “ritual”, “theological” and “philosophical” which requires basic knowledge of these terms. Depth and detail are also moderate levels but not quite a graduate level understanding; - Wording Terminology, Sentence Structure, Details and depth 	<ul style="list-style-type: none"> - No way most high school students could follow this; - The details and depth show of a graduate school person answering this.
<ul style="list-style-type: none"> - Simple wording, a concept that most students of elementary school age should be able to grasp. Also not too many details. - The explanation uses simple and direct language without complex terminology, making it accessible to children or adults with basic education. - I think this text’s wording, examples, sentence structure, and amount of detail are simple enough for an elementary-age student to comprehend. 	<ul style="list-style-type: none"> - This response includes references to Alzheimer’s, which I think would be outside the understanding of a typical 4th grader. It also references brain waves, which I think is covered in high school-level science courses. - It uses more elevated vocabulary than Elementary School, however the lack of citations and more complex concepts and narrative structure make it less than Graduate School. 	<ul style="list-style-type: none"> - The language is more advanced and mentions more specific scientific theories. - The amount of detail and specific terminology make me think it is a graduate level.

Evaluating Health Question Answering Under Readability-Controlled Style Perturbations

Md Mushfiqur Rahman and Kevin Lybarger

George Mason University

Fairfax, VA

{mrahma45, klybarga}@gmu.edu

Abstract

Patients often ask semantically similar medical questions in linguistically diverse ways that vary in readability, tone, and background knowledge. A robust question answering (QA) system should both provide semantically consistent answers across stylistic differences and adapt its response style to match the user’s input; however, existing QA evaluations rarely test this capability, creating critical gaps in QA evaluation that undermine accessibility and health literacy. We introduce SPQA, an evaluation framework and benchmark that applies controlled stylistic perturbations to consumer health questions while preserving semantic intent, then measures how model answers change across correctness, completeness, coherence, fluency, and linguistic adaptability using a human-validated LLM-based judge. The style axes include reading level, formality, and patient background knowledge; all perturbations are grounded in human annotations to ensure fidelity and alignment with human judgments. Our contributions include a readability-aware evaluation methodology, a style-diverse benchmark with human-grounded perturbations, and an automated evaluation pipeline validated against expert judgments. Evaluation results across multiple health QA models indicate that stylistic perturbations lead to measurable performance degradation, even when semantic intent is preserved during perturbation. The largest performance drops occur in answer correctness and completeness, while models also show limited ability to adapt their style to match the input. These findings underscore the risk of inequitable information delivery and highlight the need for accessibility-aware QA evaluation.

1 Introduction

Large Language Models (LLMs) have rapidly become central to consumer-facing question-answering (QA) systems, offering users quick and

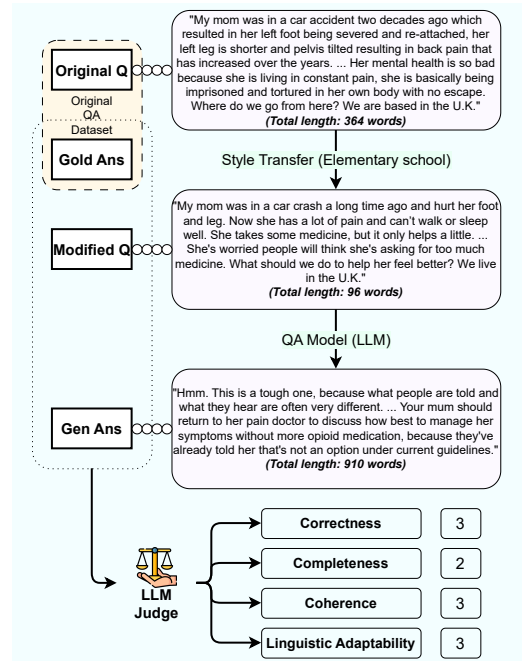


Figure 1: Example of the Style Perturbed Question Answering (SPQA) task

interactive access to information across a wide range of domains (Yu et al., 2024; Chiang et al., 2024; He et al., 2025). With this growing adoption, these systems are increasingly relied upon in critical areas such as healthcare, where users expect accurate and actionable guidance. However, as the user base becomes more diverse, linguistic variability in queries arising from differences in literacy, tone, and background knowledge presents a substantial challenge (Epner and Baile, 2012; Vela et al., 2022). Such diversity often affects the accessibility and reliability of responses, creating disparities in the quality of information retrieved. Despite its practical significance, this issue remains largely understudied, and existing evaluations rarely assess whether QA models can adapt to stylistic and readability differences, even when the underlying information need is unchanged.

While prior research has documented that demographic and stylistic factors influence model outputs (Qu and Wang, 2024; Gosavi et al., 2024), most evaluations have focused on narrow dimensions such as adversarial perturbations or typographical noise (Gan et al., 2024; Li et al., 2024; Wang et al., 2021). These approaches do not address real-world variability in question phrasing that affects user comprehension and system accessibility. Current QA assessments typically emphasize correctness and completeness but overlook whether responses maintain quality and align with the style of diverse user queries. We term this as *linguistic adaptability*, and it is essential for ensuring accessible information delivery and mitigating inequities in high-stakes domains like health communication.

To address this gap, we introduce Style Perturbed Question Answering (SPQA), an evaluation framework and benchmark (Figure 1). SPQA systematically perturbs user questions along predefined stylistic axes - A) reading level, B) formality, and C) domain knowledge, while preserving semantic intent. For each perturbed question, responses are evaluated against four comprehensive criteria: correctness, completeness, coherence and fluency, and linguistic adaptability. All perturbations are grounded in human annotations to ensure realism, and evaluations are conducted using a human-validated LLM-as-Judge for scalability and reliability. SPQA serves as a general framework for evaluating QA robustness under stylistic variation and provides a structured protocol for stress-testing QA systems under stylistic diversity, operationalizing accessibility as an evaluation dimension. In this work, we apply SPQA to consumer health question answering, where the selected stylistic axes capture key factors influencing comprehension and accessibility. Our key contributions are summarized below:

1. Readability-Aware Evaluation Framework: We introduce SPQA ¹, a framework for evaluating QA performance under readability- and style-controlled perturbations, addressing an underexplored dimension of accessibility in QA.

2. Human-Grounded Perturbations with Automated Evaluation: We generate stylistic question variants informed by human annotations and evaluate responses using GPT-4o as an automated judge, validated against expert annotations.

3. Comprehensive LLM Benchmarking: We benchmark major LLMs (Llama, DeepSeekR1, Qwen, and Phi) across multiple configurations, revealing their performance sensitivities to linguistic perturbations.

4. Focus on Consumer Health: We apply SPQA specifically to consumer health QA, emphasizing implications for health literacy, accessibility, and equity in medical information provision.

SPQA provides a systematic approach to evaluating QA systems under stylistic and linguistic variations, extending efforts in text simplification and readability assessment. Our experiments across ten state-of-the-art LLMs reveal that stylistic perturbations lead to measurable and statistically significant performance degradation, even when the underlying question intent remains unchanged. The largest declines occur in correctness and completeness, while linguistic adaptability remains a persistent challenge, indicating that models often fail to align their response style with user phrasing. Performance drops are particularly pronounced for graduate-level and expert-style variants, underscoring risks for both low-literacy and highly specialized users. These findings highlight the urgency of accessibility-aware evaluations to ensure equitable health information delivery.

2 Related Work

2.1 Open-ended QA Benchmarks for LLMs

LLMs are evaluated using a range of benchmarks that assess language understanding (Hendrycks et al., 2020; Bommasani et al., 2023), factual knowledge (Lin et al., 2021; Kwiatkowski et al., 2019; Thorne et al., 2018), reasoning (Zellers et al., 2019; Ghazal et al., 2017), and question answering (Abacha et al., 2017). While QA models frequently use multiple-choice question datasets like ARC (Clark et al., 2018), benchmarks targeting open-ended QA for practical, real-world applications remain limited because of the difficulty in evaluation. The prominent open-ended QA works (Yen et al., 2023; Prabhu and Anand, 2024; Shah et al., 2024), have used benchmarks like MT-Bench (Bai et al., 2024) for dialogue coherence and Chatbot Arena (Chiang et al., 2024) for pairwise response ranking. Medical QA benchmarks prioritize accuracy and clinical reliability. Notable datasets include MedQA (Jin et al., 2020), and PubMedQA (Jin et al., 2019). MedRedQA (Nguyen et al., 2023), the QA dataset we used in experimentation, evaluates

¹Code: github.com/mushfiqu11/spqa

responses to consumer-driven medical inquiries from Reddit, making it highly relevant for studying linguistic and stylistic variability in real-world health queries.

To address robustness, literature use techniques like, adversarial attacks (Huang et al., 2024; Singh et al., 2024), and specialized frameworks like RIT-FIS (Walsh et al., 2024). However, existing benchmarks rarely assess whether QA models maintain performance under stylistic variation, an essential dimension of accessibility and linguistic robustness.

2.2 Evaluation Criteria and Accessibility

Evaluation in QA traditionally emphasizes correctness, completeness, and coherence (Yalaman-chili et al., 2024; Liu et al., 2023a), while medical QA additionally incorporates trustworthiness (Zhu et al., 2020). Literature has explored simplifying biomedical text for lay readers (Shardlow et al., 2024; Ondov et al., 2022; Rahman et al., 2024; Štanjner et al., 2022) and transferring domain-specific language into more comprehensible forms. Such work advances text generation for readability but leaves unanswered how QA models respond to inputs that vary in readability and style, which SPQA explicitly evaluates.

2.2.1 Automated Metrics and LLM-Judge

Traditional QA metrics like BLEU (Papineni et al., 2002) and ROUGE-L (Lin, 2004) rely on n-gram overlap, limiting their ability to capture deeper semantic nuances or stylistic alignment. Embedding-based measures, like BERTScore (Zhang* et al., 2020), incorporate contextual embeddings but primarily measure semantic similarities in topics and themes rather than information accuracy. Recently, LLM-based evaluators have shown strong alignment with human judgments (Chiang et al., 2024; Bai et al., 2024; Dubois et al., 2024). Chatbot Arena (Chiang et al., 2024), MT-Bench (Bai et al., 2024), and AlpacaEval (Dubois et al., 2024) utilize LLM-based ranking systems for dialogue evaluation. GPT-4 has demonstrated moderate to strong correlation with human ratings in natural language generation tasks, with Spearman coefficients around 0.51–0.65 (Liu et al., 2023b; Sot-tana et al., 2023) and high interrater reliability with intraclass correlation (ICC) scores between 0.94 and 0.99 (Hackl et al., 2023). These findings suggest that LLM-Judge setups can serve as practical and scalable proxies for human assessment. How-

ever, applying such evaluators in readability-aware and health-sensitive QA contexts remains underexplored. Unlike prior LLM-based evaluation frameworks that focus primarily on general response quality or user preference, SPQA extends the LLM-as-Judge paradigm to explicitly address readability- and accessibility-aware QA performance using a human-validated setup that combines scalability with rigor informed by domain expertise and health question answering.

3 Methods

3.1 Dataset

For dataset preparation, we utilized MedRedQA (Nguyen et al., 2023), a large QA dataset comprising 51,000 consumer questions and their corresponding expert answers. During initial inspection, we identified a small number of incomplete questions or entries missing answers. To ensure data quality, we dropped the entries with incomplete questions or ill-formatted or incomplete expert answers. We randomly sampled 470 data points from 1000+ such clean QA pairs. Since the answers in the original dataset are expert verified or expert generated, we used these answers as the gold standard in our experiments.

We split our filtered dataset (of 470 samples) into two parts: SYSTEM-VAL (N=120) and QA-BENCH (N=350). In the SYSTEM-VAL subset, each of the 120 samples was assigned one of the eight perturbation types, resulting in 15 instances per perturbation type. These samples were used to validate the style transfer process and LLM-Judge (see §3.4.1). The QA-BENCH subset includes 350 unique original questions, each transformed into all eight stylistic variations, alongside the original version, totaling 3,150 QA pairs.

3.2 Task Formulation

The primary objective of QA systems is to generate accurate, informative, and contextually appropriate responses to user questions. Formally, this QA task is represented as the mapping function:

$$f : Q \rightarrow A' \quad (1)$$

where f denotes an LLM-based QA model that generates an answer A' given an input question Q . The quality of the generated answer is evaluated via a scoring function g , which compares the model-generated answer A' against a gold-standard, expert-validated answer A_{gold} :

Criteria	Definition (This Work)	Prior Work and Their Definition
Correctness	Measures the factual correctness and accuracy of the LLM generated response considering the gold answer as factually correct.	Literature defines correctness as the factual alignment of generated responses with ground-truth data in QA tasks (Adlakha et al., 2024; Yalamanchili et al., 2024; Scialom et al., 2021).
Completeness	Evaluates what portion of the question is fully answered by the LLM-generated response.	Literature examines the comprehensiveness of long-form answers, analyzing whether the responses fully address the posed questions without omitting essential information (Yalamanchili et al., 2024; Xu et al., 2023; Scialom et al., 2021).
Coherence and Fluency	Assesses the grammatical correctness and logical coherence of the generated response.	In literature, coherence is defined as response consistency, while fluency is defined as grammatical correctness and naturalness (Zhong et al., 2022).
Linguistic Adaptability	Measures how well an LLM adjusts its response based on variations in tone, and user expertise while preserving factuality.	No prior works systematically define this; our study introduces this criterion to assess LLM robustness to stylistic perturbations.

Table 1: Evaluation criteria used in this study for the perturbed QA task (See §A for details)

$$g(Q, A_{gold}, A') \quad (2)$$

To systematically evaluate how linguistic variations affect QA performance, we formulate a modified QA task by linguistically perturbing the original question Q , generating a transformed question Q^* . The new task now becomes:

$$st : Q \rightarrow Q^* \implies f^* : Q^* \rightarrow A' \quad (3)$$

consequently, the evaluation function is adjusted accordingly:

$$g(Q^*, A_{gold}, A') \quad (4)$$

Importantly, while Q^* differs from the original question in phrasing, tone, complexity, or style, the semantic intent remains constant. The gold-standard answer A_{gold} is based on the original question Q , emphasizing the necessity to verify the model-generated answer remains accurate and complete, while appropriately adapting its linguistic style to the perturbed input.

3.3 Automated Style Transfer (AST)

The SPQA framework is broadly applicable across various QA domains, with the specific linguistic styles requiring careful selection based on the target task and domain context. Because relevant linguistic styles vary significantly by domain, each application of SPQA must identify style dimensions critical to effective communication within that context.

In this study, we specifically apply SPQA to consumer health QA, given the critical importance of providing medically accurate, reliable, and easily understandable health information to diverse

user populations. To systematically assess QA robustness within this domain, we selected three linguistic dimensions, for which we identified eight distinct style variations: *reading level*, *formality spectrum*, and *domain-knowledge level* (see Table 2). These dimensions were specifically selected for their relevance to the consumer health context and their known influence on information accessibility and health literacy.

For the reading level dimension, we used four commonly employed sub-categories spanning a broad range of text complexity levels: elementary, middle school, high school, and graduate school (Petersen and Ostendorf, 2007; Balyan et al., 2020). Variations in formality (*formal* vs. *informal*) and domain knowledge (*domain expert* vs. *layperson*) were similarly incorporated to reflect the realistic range of ways consumers engage with health information, from casual and accessible to highly specialized and formal. Additional or alternative stylistic dimensions can be integrated based on the specific QA task or domain context.

The linguistic perturbations were generated via a zero-shot prompting approach utilizing GPT-4o-2024-08-06. Given an original question Q , the model produced transformed versions Q^* that preserved the semantic intent while varying linguistically according to the specified stylistic criteria.

3.3.1 Validating AST Framework

We validated each perturbation through a rigorous human validation process involving five health-informatics from George Mason University’s College of Public Health (Appendix §B.1). Each perturbed question Q^* in the SYSTEM-VAL subset was

Domain	Category	Definition
Grade levels	elementary	Text written with very basic vocabulary and simple sentence structures, as used by an elementary school student.
	middle	Text written with basic but varied vocabulary and slightly longer sentences, reflecting a middle school student’s style.
	high	Text featuring advanced vocabulary and sentence structures typical of a high school student.
	graduate	Text employing specialized terminology and dense, academic sentences characteristic of a graduate student.
Formality spectrum	formal	Text using precise grammar and elevated word choice appropriate for a professional report.
	informal	Text using casual phrasing and contractions common in everyday conversation.
Domain knowledge levels	domain-expert	Text incorporating field-specific terms and explanations suited for subject-matter experts.
	layperson	Text using everyday vocabulary and clear explanations geared toward a general audience.

Table 2: Definitions of each style transfer category

at first doubly annotated and then independently adjudicated for evaluation on a 3-point Likert scale using two criteria:

- **Style Transfer Success:** The degree to which the intended linguistic transformation (e.g., adjusting formality or reading level) was successfully implemented.
- **Meaning Preservation:** The extent to which the original medical meaning and intent of the question were preserved after perturbation.

During annotation, the annotators were not told what specific stylistic perturbation was performed on a given sample. This quality-control step ensured that observed performance differences across perturbations genuinely reflected model sensitivity to linguistic variations rather than unintended semantic changes.

3.4 LLM-Judge

A comprehensive and scalable evaluation of LLM-based QA systems using the SPQA framework requires an automated evaluation approach closely aligned with human judgments. To achieve this, we implemented an automated evaluation mechanism using GPT-4o as an *LLM-Judge*. Each generated answer was compared with the gold answer associated with the original question and assessed based on four criteria: *correctness*, *completeness*, *coherence and fluency*, and *linguistic adaptability*. Table 1 provides detailed definitions of these criteria. *Correctness* assesses the factual accuracy of the model-generated response, using the gold-standard answer as the reference. *Completeness* assesses the extent to which the generated answer fully addresses the information needs expressed

in the question. *Coherence and fluency* assess the grammatical quality, clarity, and logical flow of the generated answer. These three criteria are widely used in literature. *Linguistic adaptability*, a new criterion introduced in this study, evaluates how effectively a system adjusts the tone, formality, and style of its responses to align with the linguistic style of the input questions. Within health contexts, including patient-facing applications and educational tools, misaligned tone or style can undermine comprehension and negatively impact user experience (Okoso et al., 2025). Incorporating linguistic adaptability into our evaluation extends conventional QA assessment beyond factual and structural quality to include responsiveness to user style and context, thereby advancing the accessibility and inclusiveness of QA systems.

Each criterion is scored using a standardized 3-point Likert scale (1–3). Figure A presents the final zero-shot prompt used in the system. This prompt was refined based on 20 selected samples from the SYSTEM-VAL subset. Using these criteria and the LLM-as-Judge setup, we evaluated the outputs of 10 different LLMs from four model families as consumer health QA systems.

3.4.1 Validating LLM-Judge

We evaluated the reliability of the automated LLM-as-Judge through an annotation study conducted with three medical student annotators (Appendix §)². Annotators evaluated 120 selected QA pairs, each comprising a stylistically perturbed question (Q^*), the original expert answer (A_{gold}), and the model-generated answer (A'), using the same four evaluation criteria and Likert scale as the LLM-

²This is different from the annotation described in §3.3.1

Judge. Annotation occurred in four rounds: an initial calibration round, where each annotator evaluated eight samples followed by a training session to align scoring practices, and three subsequent rounds. The resulting 120 annotated samples were randomly split into two subsets, with 20 samples reserved for refining the LLM-Judge prompt (see §4.2 for results). To ensure the reliability of the automated evaluation, the LLM-Judge was tested using the held-out validation set (100 samples) that was never seen during prompt tuning. The judge operated under blinded conditions, where it was not informed whether a question was the original or style-transferred version. This procedure minimized bias and potential information leakage. Evaluation prompts were designed to independently compare generated and gold answers for each question instance. This structured process ensures rigorous assessment of the automated evaluation mechanism, enabling reliable identification of LLM strengths and weaknesses across realistic linguistic variations.

3.5 QA Benchmarking and Exp. Setup

Using our SPQA framework, we evaluated ten state-of-the-art LLM variants from four LLM families: Phi-4, Llama3, Qwen3, and DeepSeek-R1-Distilled³. Each model generated answers for the same set of 350 consumer health questions in their original forms and across eight stylistically transformed variants, resulting in 3,150 total generated answers per model. Each model’s generated response was compared against the gold-standard expert answer associated with the original question, regardless of stylistic perturbation. This ensured that all evaluations measured factual consistency and completeness relative to the same ground truth rather than stylistic similarity alone. Responses were evaluated using GPT-4o as an automated judge, scoring each answer on four criteria, *correctness*, *completeness*, *coherence*, and *linguistic adaptability*, using a 3-point Likert scale. These 3-point Likert scores were scaled and normalized to a 0-1 scale for ease of comparison.

All models were evaluated in a zero-shot setting via HuggingFace without additional fine-tuning. Inference was performed on an A100 GPU (80 GB VRAM), with average runtime per variant of

³For the DeepSeek model, we exclusively utilized locally downloaded pretrained weights without employing any external API, in compliance with institutional and state requirements.

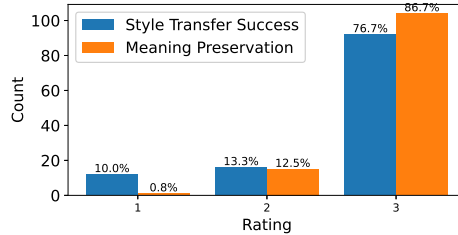


Figure 2: Distribution of ratings for Question Style Transfer Validation where 3 indicates successful, 2 indicates somewhat successful and 1 indicates failure

approximately five hours for larger models and two hours for smaller ones.

4 Results

4.1 AST Validation Results

Figure 2 presents the final adjudicated results from validating the stylistic transformations applied specifically to the questions. The results demonstrate that only 10.0% of the style-transferred questions did not fully achieve the desired stylistic modifications, and just 0.8% failed to retain the original meaning of the question. The high validation success indicates that the style transfer process reliably preserves meaning while effectively applying the intended stylistic modifications.

4.2 LLM-Judge Validation Results

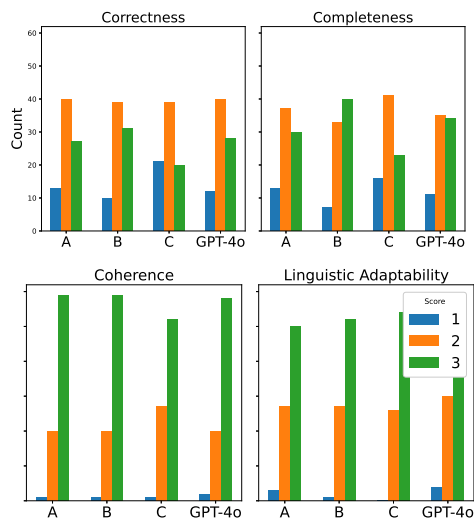


Figure 3: Score distribution for the human annotators (marked as A, B, and C) and the LLM-Judge (GPT-4o) across the four evaluation criteria, indicating similar scoring patterns between humans and the LLM-Judge

The moderate agreement score (0.47 Pearson correlation) among the human annotators (see Ta-

Agreement Type	Pearson Correlation (r)	Cohen’s Kappa (κ)
Human vs. Human (avg)	0.47	0.39
Human vs. GPT-4o (LLM-Judge)	0.36	0.33
Human vs. Llama3-70B-Inst.	0.23	0.18

Table 3: The agreement scores between human experts and the LLM-Judge are moderate. Human vs human agreement and human vs LLM-Judge agreement are quite similar indicating reliability of performance from the LLM-Judge

ble 3) indicate the inherent complexity and subjectivity involved in evaluating nuanced linguistic adaptations in open-ended QA and medical QA contexts. So, GPT-4o’s score of 0.36 when compared to human judgment (which is also considered as a moderate correlation (Kuckartz et al., 2013)) makes it a decent choice for LLM-Judge. The Cohen’s Kappa among humans (0.39) and between GPT-4o and humans (0.33) are not too far as well which further strengthens the claim for GPT-4o as an LLM-Judge.

Inter-annotator agreement among human annotators, as well as alignment between human annotators and the automated LLM-Judge, was assessed using Pearson correlation coefficients and Cohen’s Kappa scores.

Figure 3 presents the distribution of Likert scores for human annotators and the LLM-Judge across each evaluation criterion. This comparative analysis supports the reliability and suitability of the LLM-Judge for automated evaluation in nuanced medical QA tasks.

4.3 QA Benchmarking

Overall Degradation Across Styles

Table 4 provides results from the best performing models from each LLM family (full table in §6). The table shows the normalized scores for the original questions and the performance change for each stylistic variant compared to the original scores. To assess the significance of this performance drop, we performed a paired t-test with the null hypothesis of no performance degradation. Fields marked with * indicate statistically significant decreases ($p < 0.05$).

Across all models and metrics, the quality of the answers generated for stylistically altered ques-

tions significantly decreased (with statistical significance) compared to answers generated for original questions. These declines were most prominent for correctness and completeness, suggesting that models either misinterpreted the question or failed to provide adequate information. Linguistic adaptability, a criterion introduced in our SPQA framework to assess how well answer style matches question style, also showed substantial drops, suggesting models often fail to adjust their response style when question phrasing shifts. In contrast, coherence remained relatively stable, consistent with the known ability of LLMs to produce fluent text even when misinterpreting question intent.

Impact of Linguistic Axes

We further analyzed these performance drops to identify patterns. Figure 4 presents the average performance change across models, computed as the difference between the mean score on original questions and the mean score on stylistically altered variants. The results are grouped into two broader variants: (1) a simplified and informal style, averaging elementary, informal, and layperson variants; and (2) a formal and specialized style, averaging graduate, formal, and expert variants, representing advanced and specialized language usage.

As represented in the figure, the overall degradation in performance is higher in formal and specialized styles compared to simple and informal styles. This result was consistent for all ten LLM variants that we used in our experimentation.

Comparative Model Performance

All ten models demonstrated susceptibility to style-induced performance degradation, although the degree varied by model size and training approach. Larger models achieved higher baseline accuracy but were also more sensitive to stylistic perturbations. For example, DeepSeek-R1-Distilled-Llama3-70B achieved the highest baseline scores on original questions but experienced disproportionately greater performance drops under stylistic perturbations. Similarly, DeepSeek-R1-Distilled-Qwen experienced marked losses under formal and specialized styles, indicating brittleness despite its size. In comparison, Llama3-70B-Instruct and Qwen-3, though similar in size to their R1 counterparts, performed marginally better on linguistic adaptability.

Mid-sized models like Phi-4 exhibited more stable performance across styles, albeit with lower baseline performance. Qwen3-0.6B, the smallest

Model	Metric	Original	Drop in performance compared to original							
			Grade Level				Formality Spectrum		Domain-knowledge	
			Elementary	Middle	High	Graduate	Informal	Formal	Layperson	Expert
DS-Llama3-70B†	Coherence	0.71	-0.06*	-0.04*	-0.05*	-0.08*	-0.03*	-0.08*	-0.06*	-0.12*
	Completeness	0.5	-0.04*	-0.05*	-0.05*	-0.07*	-0.04*	-0.05*	-0.03*	-0.11*
	Correctness	0.62	-0.04*	-0.04*	-0.06*	-0.07*	-0.04*	-0.06*	-0.03*	-0.11*
	Linguistic Ad.	0.63	-0.06*	-0.03*	-0.07*	-0.13*	-0.03*	-0.1*	-0.05*	-0.14*
DS-Qwen3-32B†	Coherence	0.73	-0.06*	-0.07*	-0.05*	-0.1*	-0.05*	-0.09*	-0.07*	-0.12*
	Completeness	0.48	-0.03*	-0.03*	-0.04*	-0.06*	-0.04*	-0.04*	-0.04*	-0.07*
	Correctness	0.61	-0.05*	-0.04*	-0.02*	-0.07*	-0.03*	-0.07*	-0.03*	-0.09*
	Linguistic Ad.	0.64	-0.07*	-0.06*	-0.03*	-0.13*	0.0	-0.11*	-0.05*	-0.11*
Phi4	Coherence	0.69	-0.03*	-0.03*	-0.05*	-0.06*	-0.02*	-0.07*	-0.01*	-0.08*
	Completeness	0.44	-0.02*	-0.01	-0.01	-0.05*	-0.03*	-0.04*	-0.01	-0.05*
	Correctness	0.56	-0.02	-0.01	-0.01	-0.04*	-0.02	-0.04*	-0.02	-0.05*
	Linguistic Ad.	0.66	-0.04*	-0.03*	-0.02	-0.08*	-0.01	-0.07*	-0.05*	-0.08*
Qwen3-32B†	Coherence	0.7	-0.06*	-0.05*	-0.03*	-0.09*	-0.04*	-0.08*	-0.04*	-0.09*
	Completeness	0.5	-0.02	-0.03*	-0.03*	-0.05*	-0.03*	-0.04*	-0.03*	-0.08*
	Correctness	0.61	-0.04*	-0.01	-0.02*	-0.05*	-0.03*	-0.04*	-0.04*	-0.07*
	Linguistic Ad.	0.64	-0.09*	-0.06*	-0.06*	-0.13*	-0.02	-0.08*	-0.05*	-0.09*

Table 4: Normalized mean scores of the best performing models from each family (Rounded to 2 Decimal Places). Except for a few cases, all models have performed worse in case of the linguistic variants compared to the original. († indicates 8-bit quantization). * indicates statistical significance with $p < 0.05$. (See Figure 6 for full results and Figures 5, 6, and 7 for significance test results)

model, had the smallest absolute drop but also the lowest original performance. Interestingly, its resilience to informal and layperson styles may reflect its reduced specialization, leading to more consistent outputs (Yang et al., 2025).

These observations suggest that model scale and advanced training techniques (like Reinforcement Learning with Human Feedback (RLHF)), although beneficial for original phrasing, may amplify sensitivity to stylistic shifts. Instruction tuning may reinforce specific interaction norms that break down under atypical inputs.

Implications for Equity and Robustness

These results raise pressing concerns regarding QA robustness in real-world deployments. While the largest performance drops occurred with formal and expert-style queries, there was still notable degradation for simplified and informal styles. Users with low literacy or non-native speakers may frame queries in simplified or unconventional ways. Our findings show that such phrasing, though semantically equivalent, often results in lower answer quality. Conversely, expert users posing technically precise questions also receive degraded responses, an especially problematic outcome in clinical settings. The performance degradation likely stems from the models’ reliance on surface-level linguistic patterns during fine-tuning, which reduces their

ability to generalize across stylistically distinct but semantically equivalent inputs.

This dual vulnerability suggests that current LLMs may be more proficient with specific styles, likely shaped by standard web-based corpora and fine-tuning data that emphasize neutral, well-formed text (Cao et al., 2025). As a result, models fail to generalize across diverse communication styles, reducing their utility for a broad population.

5 Conclusion and Future Work

This study introduces SPQA, a framework and benchmark for evaluating linguistic robustness in question-answering systems under controlled stylistic perturbations. SPQA systematically assesses how stylistic variations in questions impact QA model performance across multiple evaluation dimensions. Unlike prior evaluations focused primarily on accuracy under standard inputs, SPQA captures a critical but overlooked dimension: the ability of models to deliver consistent, accessible answers across diverse linguistic contexts.

While broadly applicable, we applied SPQA to consumer health QA, revealing vulnerabilities in current LLMs when processing stylistic variations reflecting real-world linguistic diversity. Our experiments across ten state-of-the-art LLMs demonstrate that stylistic changes, even when semantic meaning is preserved, result in measurable per-

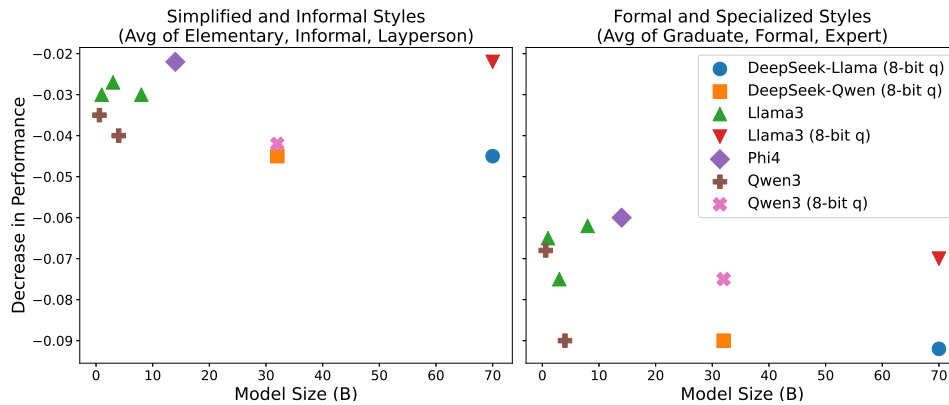


Figure 4: Average performance drop (across 4 metrics) for evaluated LLMs, indicating that larger models are more susceptible to performance degradation. Performance decline is more pronounced for formal and specialized stylistic variants compared to simplified styles

formance degradation in answer generation. The largest declines occur in correctness and completeness, and models frequently fail to align response style with question style. These findings reveal systematic risks to health information accessibility, affecting both users with limited literacy and expert users with specialized language needs, thereby reinforcing the urgency of equity-aware QA evaluations. SPQA provides a pathway toward addressing this gap by offering a systematic, human-grounded, and readability-aware evaluation protocol validated through LLM-as-Judge. Future research should extend SPQA to additional domains, including multimodal inputs, spoken interactions, and low-resource languages. Performance improvements may be achieved through adaptive prompting, style-diverse data augmentation, and patient-centered metrics. This work underscores the need for robust evaluation frameworks to ensure equitable access to reliable information for all. Future extensions should also incorporate readability-based user modeling to guide real-world deployment strategies.

Lay Summary

People ask medical questions in many different ways. Some use plain and simple words. Others use more formal wording or include medical terminology. This variability is important because an AI system might work well with one style but struggle with another. If an AI gives good answers to formal questions but fails for simple ones (or vice versa), then it may not serve all users equally.

To explore this issue, we create a formal process that can be followed step-by-step to evaluate how AI systems respond to changes in language style.

We start with real medical questions and rewrite each into several versions that differ in reading level, tone, and technical detail, while keeping the original meaning. We call this benchmark SPQA. By doing this, we can test whether AI models remain accurate and helpful no matter how a question is phrased.

We evaluate ten leading language models on every rewritten version and compare their responses across four dimensions: correctness, completeness, fluency, and how well they match the style of the question. Our results show that style variation has a clear impact on answer quality. Models frequently lose accuracy and completeness when questions are highly formal or contain dense medical jargon, and the same question can produce different answers depending on how it is phrased. Some models are more sensitive to these shifts than others, and most do not adjust their response style to match the user’s expression. These findings show that current systems handle writing styles unevenly, which may disadvantage some users.

Our work emphasizes the need for health question-answering systems that give reliable, inclusive, and understandable answers for everyone, regardless of how they phrase their questions.

Our work highlights the importance of developing health question-answering systems that provide reliable, fair, and easy-to-understand information for all users, regardless of how they write their questions.

Limitations

This study has several limitations. Methodologically, errors introduced during the style-transfer step could propagate through subsequent stages,

although validation showed that meaning was preserved in over 99% of cases. Occasional deviations from the intended style may still influence downstream outcomes. In addition, evaluating generated answers through human and LLM-based scoring introduces subjectivity. While the automated LLM-Judge achieved performance comparable to human annotators, moderate agreement levels suggest residual bias or inconsistency that could affect result validity. Future work should explore hybrid or multi-judge evaluation strategies to improve reliability.

Beyond these methodological constraints, the validation group consisted of medical students and domain experts, limiting sociolinguistic diversity and potentially affecting generalization across populations. The current evaluation also focuses on a single consumer health QA dataset; broader experimentation across datasets and domains is needed to establish the generalizability of the SPQA framework. Finally, reliance on GPT-4o for both style transfer and evaluation may introduce model-specific biases, underscoring the importance of future replication with diverse model architectures.

References

- Asma Ben Abacha, Eugene Agichtein, Yuval Pinter, and Dina Demner-Fushman. 2017. Overview of the medical question answering task at trec 2017 liveqa. In *TREC*, pages 1–12.
- Vaibhav Adlakha, Parishad BehnamGhader, Xing Han Lu, Nicholas Meade, and Siva Reddy. 2024. [Evaluating correctness and faithfulness of instruction-following models for question answering](#). *Transactions of the Association for Computational Linguistics*, 12:681–699.
- Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, and Wanli Ouyang. 2024. [MT-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7421–7454, Bangkok, Thailand. Association for Computational Linguistics.
- Renu Balyan, Kathryn S McCarthy, and Danielle S McNamara. 2020. Applying natural language processing and hierarchical machine learning approaches to text difficulty classification. *International Journal of Artificial Intelligence in Education*, 30(3):337–370.
- Rishi Bommasani, Percy Liang, and Tony Lee. 2023. Holistic evaluation of language models. *Annals of the New York Academy of Sciences*, 1525(1):140–146.
- Tianyu Cao, Neel Bhandari, Akhila Yerukola, Akari Asai, and Maarten Sap. 2025. Out of style: Rag’s fragility to linguistic variation. *arXiv preprint arXiv:2504.08231*.
- Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E Gonzalez, et al. 2024. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning*.
- Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457v1*.
- Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. 2024. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*.
- Daniel E Epner and Walter F Baile. 2012. Patient-centered care: the key to cultural competence. *Annals of oncology*, 23:iii33–iii42.
- Esther Gan, Yiran Zhao, Liying Cheng, Mao Yancan, Anirudh Goyal, Kenji Kawaguchi, Min-Yen Kan, and Michael Shieh. 2024. [Reasoning robustness of LLMs to adversarial typographical errors](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10449–10459, Miami, Florida, USA. Association for Computational Linguistics.
- Ahmad Ghazal, Todor Ivanov, Pekka Kostamaa, Alain Crolotte, Ryan Voong, Mohammed Al-Kateb, Waleed Ghazal, and Roberto V. Zicari. 2017. [Big-bench v2: The new and improved bigbench](#). *2017 IEEE 33rd International Conference on Data Engineering (ICDE)*, pages 1225–1236.
- Purva Prasad Gosavi, Vaishnavi Murlidhar Kulkarni, and Alan F Smeaton. 2024. Capturing bias diversity in llms. In *2024 2nd International Conference on Foundation and Large Language Models (FLLM)*, pages 593–598. IEEE.
- Veronika Hackl, Alexandra Elena Müller, Michael Granitzer, and Maximilian Sailer. 2023. Is gpt-4 a reliable rater? evaluating consistency in gpt-4’s text ratings. In *Frontiers in Education*, volume 8, page 1272229. Frontiers Media SA.
- Kai He, Rui Mao, Qika Lin, Yucheng Ruan, Xiang Lan, Mengling Feng, and Erik Cambria. 2025. A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics. *Information Fusion*, page 102963.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

- Guanhua Huang, Yuchen Zhang, Zhe Li, Yongjian You, Mingze Wang, and Zhouwang Yang. 2024. [Are AI-generated text detectors robust to adversarial perturbations?](#) In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6005–6024, Bangkok, Thailand. Association for Computational Linguistics.
- Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. [What disease does this patient have? a large-scale open domain question answering dataset from medical exams.](#) *Preprint*, arXiv:2009.13081.
- Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. [PubMedQA: A dataset for biomedical research question answering.](#) 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 2567–2577, Hong Kong, China. Association for Computational Linguistics.
- Udo Kuckartz, Stefan Rädiker, Thomas Ebert, and Julia Schehl. 2013. *Statistik: eine verständliche Einführung*. Springer-Verlag.
- Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. [Natural questions: A benchmark for question answering research.](#) *Transactions of the Association for Computational Linguistics*, 7:452–466.
- Zekun Li, Baolin Peng, Pengcheng He, and Xifeng Yan. 2024. [Evaluating the instruction-following robustness of large language models to prompt injection.](#) In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 557–568, Miami, Florida, USA. Association for Computational Linguistics.
- 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.
- Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. [Truthfulqa: Measuring how models mimic human falsehoods.](#) *arXiv preprint arXiv:2109.07958*.
- Nelson Liu, Tianyi Zhang, and Percy Liang. 2023a. [Evaluating verifiability in generative search engines.](#) In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7001–7025, Singapore. Association for Computational Linguistics.
- Yang Liu, Dan Iter, Yichong Xu, Shuhang Wang, Ruo Chen Xu, and Chenguang Zhu. 2023b. [G-eval: NLG evaluation using gpt-4 with better human alignment.](#) In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.
- Vincent Nguyen, Sarvnaz Karimi, Maciej Rybinski, and Zhenchang Xing. 2023. [MedRedQA for medical consumer question answering: Dataset, tasks, and neural baselines.](#) In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 629–648, Nusa Dua, Bali. Association for Computational Linguistics.
- Ayano Okoso, Keisuke Otaki, Satoshi Koide, and Yukino Baba. 2025. [Impact of tone-aware explanations in recommender systems.](#) *ACM Trans. Recomm. Syst.*, 3(4).
- Brian Ondov, Kush Attal, and Dina Demner-Fushman. 2022. A survey of automated methods for biomedical text simplification. *Journal of the American Medical Informatics Association*, 29(11):1976–1988.
- 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.
- Sarah Elizabeth Petersen and Mari Ostendorf. 2007. *Natural Language Processing Tools for Reading Level Assessment and Text Simplification for Bilingual Education*. Citeseer.
- Venkatesh V Deepali Prabhu and Avishek Anand. 2024. [Dexter: A benchmark for open-domain complex question answering using llms.](#) *arXiv preprint arXiv:2406.17158*.
- Yao Qu and Jue Wang. 2024. Performance and biases of large language models in public opinion simulation. *Humanities and Social Sciences Communications*, 11(1):1–13.
- Md Mushfiqur Rahman, Mohammad Sabik Irbaz, Kai North, Michelle S Williams, Marcos Zampieri, and Kevin Lybarger. 2024. Health text simplification: An annotated corpus for digestive cancer education and novel strategies for reinforcement learning. *Journal of Biomedical Informatics*, 158:104727.
- Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021. [QuestEval: Summarization asks for fact-based evaluation.](#) In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6594–6604, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Shalin Shah, Srikanth Ryali, and Ramasubbu Venkatesh. 2024. [Multi-document financial question answering using llms.](#) *arXiv preprint arXiv:2411.07264*.

- Matthew Shardlow, Horacio Saggion, Fernando Alva-Manchego, Marcos Zampieri, Kai North, Sanja Štajner, and Regina Stodden, editors. 2024. *Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024)*. Association for Computational Linguistics, Miami, Florida, USA.
- Ayush Singh, Navpreet Singh, and Shubham Vatsal. 2024. Robustness of llms to perturbations in text. *arXiv preprint arXiv:2407.08989*.
- Andrea Sottana, Bin Liang, Kai Zou, and Zheng Yuan. 2023. [Evaluation metrics in the era of GPT-4: Reliably evaluating large language models on sequence to sequence tasks](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8776–8788, Singapore. Association for Computational Linguistics.
- Sanja Štajner, Horacio Saggion, Daniel Ferrés, Matthew Shardlow, Kim Cheng Sheang, Kai North, Marcos Zampieri, and Wei Xu, editors. 2022. *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Virtual).
- James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. [FEVER: a large-scale dataset for fact extraction and VERification](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.
- Monica B Vela, Amarachi I Erondu, Nichole A Smith, Monica E Peek, James N Woodruff, and Marshall H Chin. 2022. Eliminating explicit and implicit biases in health care: evidence and research needs. *Annual review of public health*, 43(1):477–501.
- Matthew Walsh, David Schulker, and Shing hon Lau. 2024. [Beyond Capable: Accuracy, Calibration, and Robustness in Large Language Models](#).
- Boxin Wang, Chejian Xu, Shuohang Wang, Zhe Gan, Yu Cheng, Jianfeng Gao, Ahmed Hassan Awadallah, and Bo Li. 2021. Adversarial glue: A multi-task benchmark for robustness evaluation of language models. *arXiv preprint arXiv:2111.02840*.
- Fangyuan Xu, Yixiao Song, Mohit Iyyer, and Eunsol Choi. 2023. [A critical evaluation of evaluations for long-form question answering](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3225–3245, Toronto, Canada. Association for Computational Linguistics.
- Amulya Yalamanchili, Bishwambhar Sengupta, Joshua Song, Sara Lim, Tarita O. Thomas, Bharat B. Mittal, Mohamed E. Abazeed, and P. Troy Teo. 2024. [Quality of large language model responses to radiation oncology patient care questions](#). *JAMA Network Open*, 7(4):e2444630–e2444630.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- Howard Yen, Tianyu Gao, Jinhyuk Lee, and Danqi Chen. 2023. [MoQA: Benchmarking multi-type open-domain question answering](#). In *Proceedings of the Third DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pages 8–29, Toronto, Canada. Association for Computational Linguistics.
- Haoran Yu, Chang Yu, Zihan Wang, Dongxian Zou, and Hao Qin. 2024. Enhancing healthcare through large language models: A study on medical question answering. In *2024 IEEE 6th International Conference on Power, Intelligent Computing and Systems (ICPICS)*, pages 895–900. IEEE.
- Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. [Hellaswag: Can a machine really finish your sentence?](#) *Proceedings of the 57th Annual Meeting of the 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*.
- Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. [Towards a unified multi-dimensional evaluator for text generation](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Ming Zhu, Aman Ahuja, Da-Cheng Juan, Wei Wei, and Chandan K. Reddy. 2020. [Question answering with long multiple-span answers](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3840–3849, Online. Association for Computational Linguistics.

A LLM Judge Prompts

Since we used a zero-shot LLM-Judge, it was essential to have a rigorously engineered prompt for different phases of our workflow. Appendix A represents the user-prompt provided to the LLMs to generate the answers to the questions. Appendix A represents the system prompt with all the necessary definitions provided to the LLM-Judge. These were also used as the base instructions for the annotators validating the LLM-Judge. Keeping the instructions same, we ensured fair ground for the LLM-Judge and human experts.

User-prompt for LLM-Judge

```
1 """
2 Evaluate the following QA sample:
3 Modified Question (Q_mod): [SEP] {question} [SEP]
4 Generated Answer (A_gen): [SEP] {answer} [SEP]
5 Gold Standard Answer (A_gold): [SEP] {gold} [SEP]
6 """
```

System-prompt for LLM-Judge

```
1 """
2 You are a helpful assistant that evaluates medical QA samples.
3 For each sample, you must evaluate the generated answer (A_gen) as a response to
4 the modified question (Q_mod). Use the gold standard answer (A_gold) as the
5 medically accurate information regarding the topic.
6 In this QA task, the generated answer (A_gen) and the gold standard answer (
7 A_gold) may come from different linguistic distribution. Do not penalize A_gen
8 for being linguistically different from A_gold.
9
10 Use the following four criteria. For each criterion, assign a score from 0 (
11 lowest) to 2 (highest).
12
13 1. Correctness: Considering A_gold as medically correct, what portion of A_gen is
14 accurate in answering the Q_mod? A_gen can be significantly different from
15 A_gold.
16 - Score Definitions:
17 CR0: A_gen is completely incorrect. It does not have any medically accurate
18 information.
19 CR1: A_gen is mostly incorrect. It has very little medically correct advice
20 or information.
21 CR2: The generated answer is mostly correct.
22
23 2. Completeness: What portion of the queries made in Q_mod are answered by A_gen?
24 - Score Definitions:
25 CM0: A_gen is absolutely irrelevant and does not answer any of the queries
26 made in Q_mod.
27 CM1: A_gen is somewhat incomplete, missing the essential details required to
28 address Q_mod.
29 CM2: A_gen answers Q_mod sufficiently. Important parts of the question in
30 Q_mod is addressed by A_gen.
31
32 3. Fluency and Coherence: How well is A_gen written in terms of language fluency
33 and logical structure?
34 - Score Definitions:
35 FC0: A_gen is poorly written, with significant grammatical and structural
36 issues.
37 FC1: A_gen is somewhat fluent but contains noticeable issues and grammatical
38 errors.
39 FC2: A_gen is fluent and logically structured.
40
41 4. Linguistic Adaptability: How effectively does A_gen adopt the style and tone
42 of the modified question (Q_mod)?
43 - Score Definitions:
44 LA0: A_gen fails to match the tone or style of Q_mod and would be totally
45 unreadable for the user asking Q_mod.
46 LA1: A_gen somewhat matches the tone or style but would not be fully legible
47 for the user asking Q_mod.
48 LA2: A_gen is appropriate and easy to read for someone who asked the question
49 Q_mod.
50 """
```



```

31
32 Return your evaluation in JSON format as follows:
33
34 {
35     "correctness": <rating as an integer>,
36     "completeness": <rating as an integer>,
37     "fluency_and_coherence": <rating as an integer>,
38     "linguistic_adaptability": <rating as an integer>
39 }
40
41 Ensure that your output contains only the JSON object.
42 """

```

B Human Annotation Details

This appendix provides details of the two independent annotation streams that supported this work. Both streams involved the same set of 120 question–answer pairs but differed in purpose, annotator expertise, and evaluation criteria.

B.1 Annotation Stream 1: AST Validation

Objective: Assess whether stylistic perturbations (generated via GPT-4o) successfully applied the intended style changes (reading level, formality, patient background knowledge) while preserving original meaning.

Annotator Profile: Five health informatics graduate students with training in health data interpretation.

Procedure:

- Each question–variant pair was evaluated independently by two annotators and adjudicated by a third for disagreements.
- Annotations were based on two criteria using a 3-point Likert scale:
 1. *Style Transfer Success* – Did the variant reflect the assigned stylistic dimension?
 2. *Meaning Preservation* – Was the original medical intent maintained?
- Random audits (15% of samples) were conducted for quality control.

Outcome: High validation accuracy was achieved (Style Transfer Success: 76.7%, Meaning Preservation: 86.7%), confirming fidelity of stylistic transformations.

B.2 Annotation Stream 2: LLM-Judge Validation

Objective: Validate the reliability of GPT-4o as an automated judge by comparing its ratings to human annotations for QA responses.

Annotator Profile: Three medical students (clinical track) with prior training in patient communication.

Procedure:

- Annotators rated the same 120 question–answer pairs on four evaluation criteria:
 1. Correctness
 2. Completeness
 3. Coherence and Fluency
 4. Linguistic Adaptability
- Process included a calibration round (8 samples), followed by three main annotation rounds after a training session.
- Agreement metrics were computed using Pearson correlation and Cohen’s Kappa.

Outcome: Human–LLM alignment showed moderate agreement (Pearson’s $r = 0.36$, Cohen’s $\kappa = 0.33$), supporting the use of LLM-as-Judge for scalable evaluation.

B.3 Summary of Annotation Resources

- Total samples annotated: 120 (used for both streams).
- Annotators: 5 health informatics students (style validation) and 3 medical students (evaluation validation).

Table 5: Summary of Human Annotation Streams

Stream	Annotator Group	Goal	Criteria
Style Transfer Validation	5 Health Informatics Students	Validate stylistic perturbations	Style Transfer Success, Meaning Preservation
Evaluation Rubric Validation	3 Medical Students	Validate automated evaluation rubric	Correctness, Completeness, Coherence and Fluency, Linguistic Adaptability

C Additional results

C.1 Significance Test

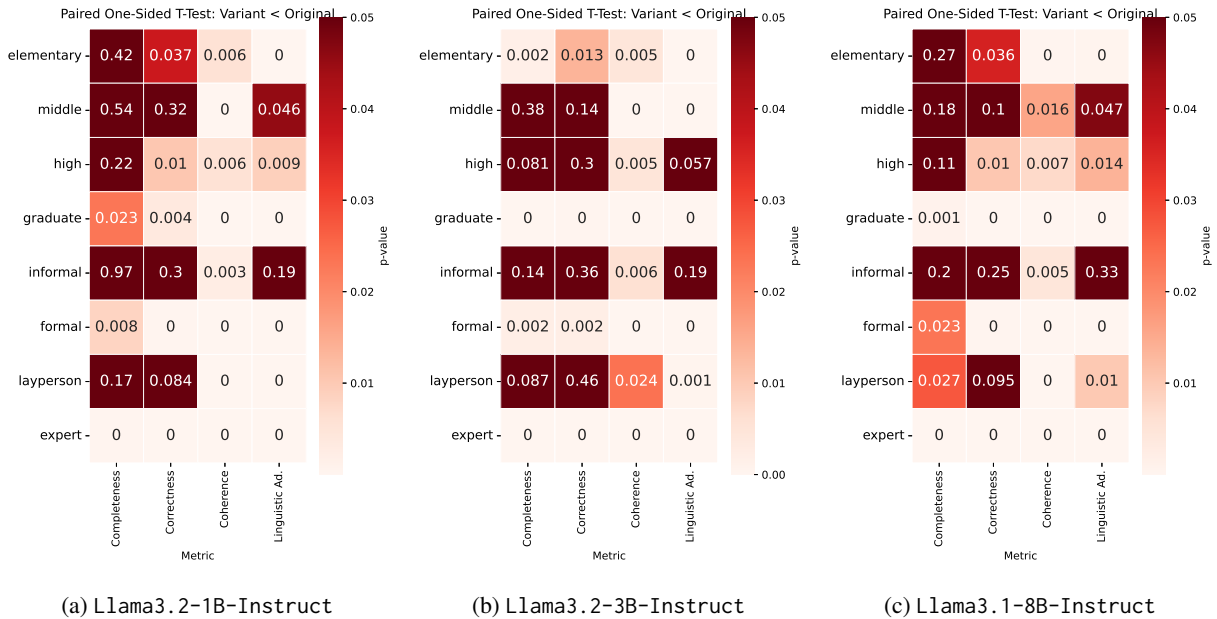


Figure 5: Paired One-Sided T-Test: Style Variant < Original (Rounded to nearest third decimal place)

Section 4 mentions that a significance test was performed. Figures 5, 6, and 7 represent heatmaps of the detailed results from the significance test.

C.2 Full Result

Table 6 represents the complete results table with all the models we have used in our experimentation. A shorter and more concise version of this table has been presented in the main paper.

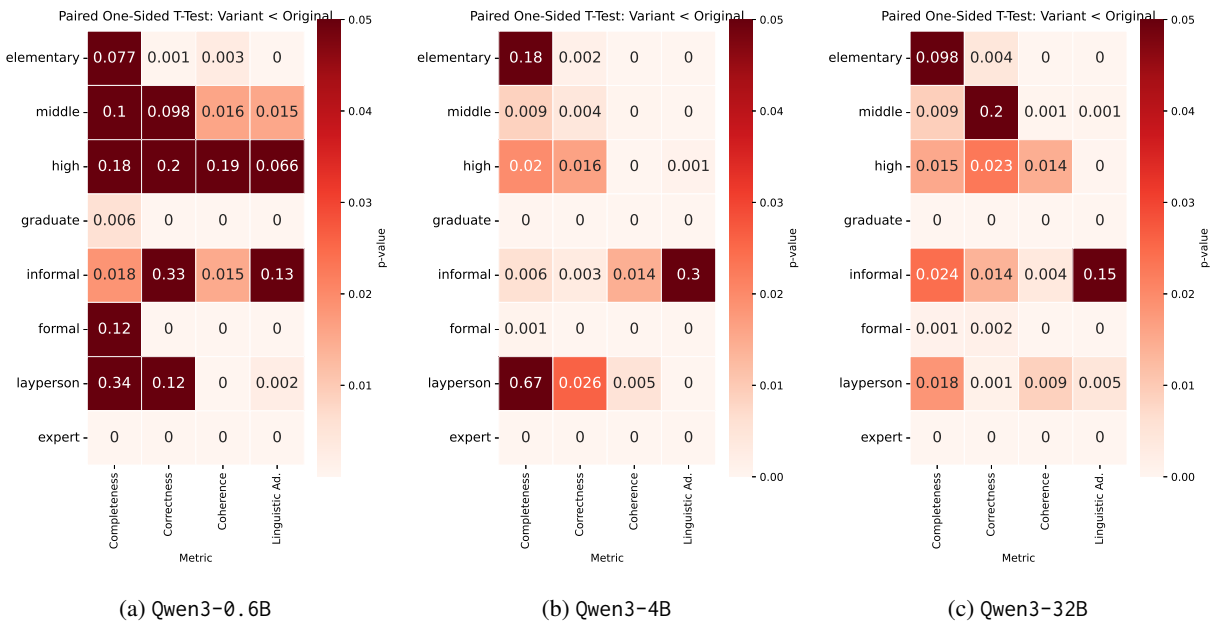


Figure 6: Paired One-Sided T-Test: Style Variant < Original (Rounded to nearest third decimal place)

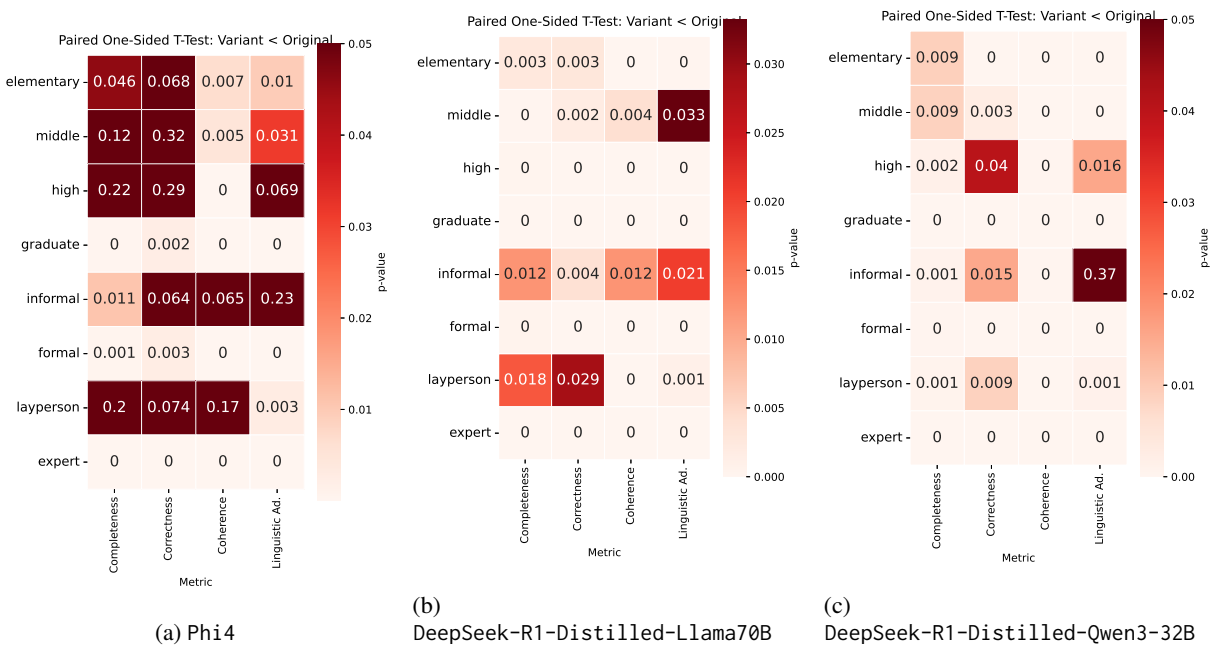


Figure 7: Paired One-Sided T-Test: Style Variant < Original (Rounded to nearest third decimal place)

		Drop in performance compared to original								
Model	Metric	Original	Grade Level				Formality Spectrum		Domain-knowledge	
			Elementary	Middle	High	Graduate	Informal	Formal	Layperson	Expert
DS-Llama-70B†	Coherence	0.71	-0.06	-0.04	-0.05	-0.08	-0.03	-0.08	-0.06	-0.12
DS-Llama-70B†	Completeness	0.5	-0.04	-0.05	-0.05	-0.07	-0.04	-0.05	-0.03	-0.11
DS-Llama-70B†	Correctness	0.62	-0.04	-0.04	-0.06	-0.07	-0.04	-0.06	-0.03	-0.11
DS-Llama-70B†	Linguistic Ad.	0.63	-0.06	-0.03	-0.07	-0.13	-0.03	-0.1	-0.05	-0.14
DS-Qwen-32B†	Coherence	0.73	-0.06	-0.07	-0.05	-0.1	-0.05	-0.09	-0.07	-0.12
DS-Qwen-32B†	Completeness	0.48	-0.03	-0.03	-0.04	-0.06	-0.04	-0.04	-0.04	-0.07
DS-Qwen-32B†	Correctness	0.61	-0.05	-0.04	-0.02	-0.07	-0.03	-0.07	-0.03	-0.09
DS-Qwen-32B†	Linguistic Ad.	0.64	-0.07	-0.06	-0.03	-0.13	0.0	-0.11	-0.05	-0.11
Llama3-1B	Coherence	0.71	-0.04	-0.05	-0.03	-0.08	-0.04	-0.06	-0.06	-0.09
Llama3-1B	Completeness	0.41	0.0	0.0	-0.01	-0.02	0.03	-0.03	-0.01	-0.05
Llama3-1B	Correctness	0.54	-0.03	-0.01	-0.03	-0.04	-0.01	-0.05	-0.02	-0.06
Llama3-1B	Linguistic Ad.	0.67	-0.08	-0.03	-0.04	-0.11	-0.02	-0.07	-0.07	-0.11
Llama3-3B	Coherence	0.71	-0.04	-0.05	-0.04	-0.1	-0.04	-0.08	-0.03	-0.11
Llama3-3B	Completeness	0.43	-0.04	0.0	-0.02	-0.05	-0.01	-0.04	-0.01	-0.06
Llama3-3B	Correctness	0.54	-0.03	-0.01	-0.01	-0.05	0.0	-0.04	0.0	-0.06
Llama3-3B	Linguistic Ad.	0.68	-0.07	-0.06	-0.03	-0.1	-0.01	-0.09	-0.05	-0.12
Llama3-8B	Coherence	0.71	-0.05	-0.03	-0.03	-0.08	-0.04	-0.07	-0.06	-0.1
Llama3-8B	Completeness	0.43	-0.01	-0.01	-0.02	-0.04	-0.01	-0.03	-0.03	-0.05
Llama3-8B	Correctness	0.56	-0.03	-0.02	-0.03	-0.06	-0.01	-0.05	-0.02	-0.07
Llama3-8B	Linguistic Ad.	0.66	-0.05	-0.03	-0.04	-0.07	-0.01	-0.07	-0.04	-0.07
Llama3-70B†	Coherence	0.69	-0.04	-0.04	-0.01	-0.06	-0.02	-0.06	-0.03	-0.08
Llama3-70B†	Completeness	0.45	-0.01	-0.01	-0.01	-0.05	0.0	-0.05	0.0	-0.07
Llama3-70B†	Correctness	0.57	-0.03	-0.03	-0.02	-0.06	-0.01	-0.06	-0.02	-0.07
Llama3-70B†	Linguistic Ad.	0.67	-0.07	-0.03	-0.03	-0.08	-0.01	-0.08	-0.04	-0.11
Phi4	Coherence	0.69	-0.03	-0.03	-0.05	-0.06	-0.02	-0.07	-0.01	-0.08
Phi4	Completeness	0.44	-0.02	-0.01	-0.01	-0.05	-0.03	-0.04	-0.01	-0.05
Phi4	Correctness	0.56	-0.02	-0.01	-0.01	-0.04	-0.02	-0.04	-0.02	-0.05
Phi4	Linguistic Ad.	0.66	-0.04	-0.03	-0.02	-0.08	-0.01	-0.07	-0.05	-0.08
Qwen3-0.6B	Coherence	0.69	-0.04	-0.03	-0.01	-0.07	-0.03	-0.07	-0.06	-0.09
Qwen3-0.6B	Completeness	0.46	-0.02	-0.02	-0.01	-0.03	-0.03	-0.01	0.0	-0.07
Qwen3-0.6B	Correctness	0.59	-0.05	-0.02	-0.02	-0.06	-0.01	-0.05	-0.02	-0.08
Qwen3-0.6B	Linguistic Ad.	0.63	-0.08	-0.04	-0.03	-0.09	-0.02	-0.07	-0.06	-0.11
Qwen3-4B	Coherence	0.72	-0.07	-0.05	-0.06	-0.1	-0.03	-0.1	-0.04	-0.12
Qwen3-4B	Completeness	0.48	-0.02	-0.04	-0.03	-0.05	-0.04	-0.05	0.0	-0.07
Qwen3-4B	Correctness	0.62	-0.04	-0.04	-0.03	-0.08	-0.04	-0.06	-0.03	-0.1
Qwen3-4B	Linguistic Ad.	0.65	-0.09	-0.07	-0.05	-0.11	-0.01	-0.09	-0.06	-0.13
Qwen3-32B†	Coherence	0.7	-0.06	-0.05	-0.03	-0.09	-0.04	-0.08	-0.04	-0.09
Qwen3-32B†	Completeness	0.5	-0.02	-0.03	-0.03	-0.05	-0.03	-0.04	-0.03	-0.08
Qwen3-32B†	Correctness	0.61	-0.04	-0.01	-0.02	-0.05	-0.03	-0.04	-0.04	-0.07
Qwen3-32B†	Linguistic Ad.	0.64	-0.09	-0.06	-0.06	-0.13	-0.02	-0.08	-0.05	-0.09

Table 6: Full results table. † indicates models with 8-bit quantization.

D Declaration of use of Generative AI

During the preparation of this manuscript, the authors used ChatGPT to obtain editorial assistance focused on writing clarity and proofreading. All scientific content, including analyses and interpretations, was developed independently by the authors. The authors carefully reviewed and revised the text following the use of these tools and assume full responsibility for the integrity and accuracy of the final manuscript.

A Multi-Agent Framework with Diagnostic Feedback for Iterative Plain Language Summary Generation from Cochrane Medical Abstracts

Felipe Arias Russi^{1,2}, Carolina Salazar Lara³, Rubén Manrique¹

¹Systems and Computing Engineering Department, Universidad de los Andes, Bogotá D.C.

²Department of Mathematics, Universidad de los Andes, Bogotá D.C.

³Department of Biomedical Engineering, Universidad de los Andes, Bogotá D.C.

{af.ariasr, c.salazar499, rf.manrique}@uniandes.edu.co

Abstract

Plain Language Summaries (PLS) improve health literacy and enable informed healthcare decisions, but writing them requires domain expertise and is time-consuming. Automated methods often prioritize efficiency over comprehension, and medical documents' unique simplification requirements challenge generic solutions. We present a multi-agent system for generating PLS using Cochrane PLS as proof of concept. The system uses specialized agents for information extraction, writing, diagnosis, and evaluation, integrating a medical glossary and statistical analyzer to guide revisions. We evaluated three architectural configurations on 100 Cochrane abstracts using six LLMs (both proprietary and open-source). Results reveal model-dependent trade-offs between factuality and readability, with the multi-agent approach showing improvements for smaller models and providing operational advantages in control and interpretability.

1 Introduction

Health literacy is the ability of an individual to access, understand, and apply health information. This is essential for informed decision-making and effective navigation of healthcare systems. Inadequate health literacy remains a global challenge, contributing to poor treatment adherence, higher hospitalization rates, and health disparities (Berkman et al., 2011; Sørensen et al., 2015; Bahador et al., 2020). Plain Language Summaries (PLS) is a way to reduce health literacy gaps by translating medical texts into clear, accurate, and accessible language for non-technical audiences (Bahador et al., 2020). However, producing high-quality PLS manually is resource-intensive and requires expertise in both medical content and health communication.

Recent advances in LLMs offer new opportunities to automate PLS generation. While early efforts showed that LLMs can produce readable and

semantically faithful summaries, most approaches relied on single-pass generation and lacked systematic guardrails for factuality, readability, and regulatory compliance (Turbitt et al., 2023; Van Veen et al., 2024). The increasing complexity of biomedical content and the need for domain-specific standards underscore the importance of structured, multi-step workflows over monolithic generation.

This work builds on our previous research in LLM-based PLS generation (Arias-Russi et al., 2025), which showed the potential of single-prompt models to translate Cochrane abstracts into PLS and Clinical Trials into Protocol Plain Language Summaries (PPLS). However, generating these kind of structured PLS that meet professional standards differs from generic text simplification; it needs adherence to specific templates and guidelines, diagnostic feedback, and systematic quality control. Unlike general simplification tasks that focus only on reducing complexity, structured PLS generation requires writing documents with well-defined structures that balance accessibility with medical accuracy.

Current LLMs struggle to balance simplicity with factual accuracy, often oversimplifying complex medical content or preserving meaning at the cost of readability (Li et al., 2024). Our prior work revealed similar domain-specific challenges, requiring distinct prompts for different document types (Cochrane PLS and PPLS). Also, this approach did not provide mechanisms to identify specific problems in the generated PLS drafts or provide targeted corrections.

Based on these limitations, this research aims to: (1) develop tools that allow a better understanding of what makes a text non-compliant with PLS standards and how to systematically address these issues, and (2) create a multi-agent framework supported by diagnostic tools that can both generate structured PLS and evaluate their quality through some iterative refinement.

We propose a framework that decomposes PLS generation into specialized subtasks, each handled by dedicated agents: information extraction, writing, diagnostic, and evaluation (Figure 1). The key component is a diagnostic feedback loop where evaluation agents identify specific complexity issues and guide targeted revisions using verifiable criteria. We instantiate this framework for Cochrane PLS generation, as their guidelines (Pitcher et al., 2022) provide an useful template ideal for testing structured document generation.

2 Related Work

2.1 Plain Language Summaries

Recent work in biomedical text simplification explores different approaches. The BioLaySumm shared task (Xiao et al., 2025) focuses on generating lay summaries from biomedical abstracts. Participants showed various strategies: supervised fine-tuning of T5 and LLaMA models (Zhang et al., 2025); extract-then-summarize pipelines with persona-based prompts and UMLS definitions (Gupta and Krishnamurthy, 2025); structured prompting with dynamic few-shot selection and RAG (Lossio-Ventura et al., 2025); and Tree-of-Thought prompting with hybrid methods (Sivagnanam et al., 2025). Fine-tuning approaches include QLoRA adaptation with iterative refinement (Binti Moriasi and Sung, 2025) and readability-controlled instruction tuning (Tran et al., 2025). Others focus on preprocessing (Dehkordi et al., 2025) or evaluation metrics (Lyu and Pergola, 2024a; Scholz and Wenzel, 2025). More related work can be found in our previous work (Arias-Russi et al., 2025).

Our work addresses a complementary task: generating structured PLS following Cochrane’s established template. Unlike lay summaries that prioritize readability alone, structured PLS must adhere to specific section requirements (Title, Key Messages, Background, Methods, Results, Limitations, Currency), maintain professional standards, and balance accessibility with regulatory compliance. We used some evaluation metrics from BioLaySumm and related work to assess both readability and structural conformance.

2.2 Multi-agent Systems for Text Simplification

Multi-agent systems have emerged as a promising approach for text processing. The Society of Medi-

cal Simplifiers (Lyu and Pergola, 2024b) simplifies biomedical literature into general plain language text, using five agents in three interaction loops—a Layperson Agent identifies technical terms, a Medical Expert provides clarifications, and a Simplifier Agent edits text, focusing on making content accessible without following specific templates or guidelines. ExpertEase (Mo and Hu, 2024) generates grade-specific simplified documents for educational purposes, using Expert, Teacher, and Student agents that calibrate text complexity for target reading levels like 2nd-3rd grade. For diagnostic applications, MedAgent-Pro (Wang et al., 2025) produces evidence-based medical diagnoses with supporting visual evidence rather than simplified text, employing RAG, Planner, and Tool agents to integrate clinical guidelines for diseases like glaucoma. AgentSimp (Fang et al., 2025) creates general simplified documents focusing on coherence and metaphor handling, using nine agents including a Metaphorical Analyst and Terminology Interpreter, but without adherence to medical communication standards or structured templates.

Rather than generating general simplified text, we aim to create structured PLS that facilitate the work of medical writers, helping to automate the process to get high-quality PLS drafts. Multi-agent systems are particularly suited for this task because structured documents can be decomposed into separate subtasks (extraction, integration, evaluation, and refinement) that align naturally with specialized agents (see the conceptual framework in Figure 1). From this abstract idea of generating structured PLS, we designed a multi-agent approach specifically for Cochrane PLS, as their detailed guidelines provide a well-defined template that serves as an ideal test case for our framework (Pitcher et al., 2022). Our primary approach uses an on-demand evaluator tool that the editor agent calls when needed. Inspired by Self-Refine (Madaan et al., 2023), we also tested an alternative iterative approach where the evaluator runs multiple refinement cycles independently, instead of being invoked by the editor agent.¹

3 Methodology

We present the methodology for developing and evaluating a multi-agent system for automatic gen-

¹All materials including agent prompts, datasets, evaluation scripts, and workflow implementation are available at: <https://github.com/feliperussi/tsar-2025-medical-writing-agent-cochrane>

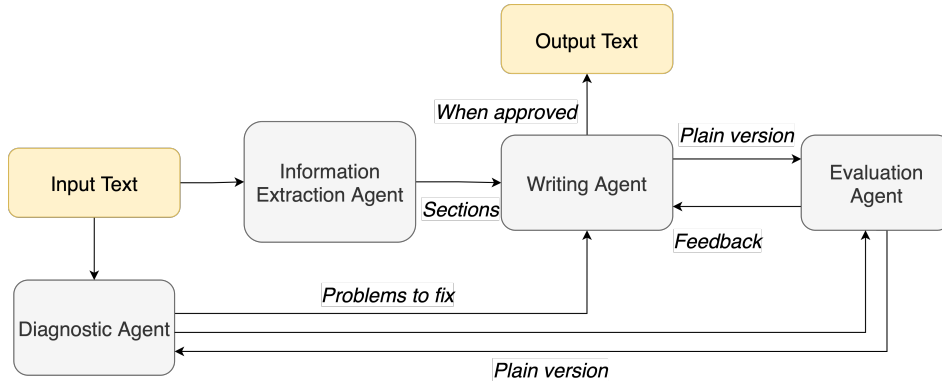


Figure 1: Conceptual framework for generating PLS. The system orchestrates specialized agents for information extraction, writing, terminology simplification, and evaluation. A diagnostic feedback loop enables evaluation agents to identify complexity issues and guide targeted revisions. This modular architecture supports structured, accurate, and readable PLS aligned with Cochrane standards.

eration of PLS from Cochrane medical abstracts. The approach leverages linguistic analysis, medical glossaries, domain-specific glossaries, and advanced language models to produce accessible medical communication. The methodology follows six main steps: (1) Data Collection and Processing, (2) PLS Linguistic Feature Extraction, (3) Percentiles for PLS Diagnosis, (4) Develop Diagnostic Tools, (5) Define Specialization, and (6) Define Architecture for the multi-agent system (see Appendix A for the complete workflow diagram).

3.1 Data Collection and Processing

3.1.1 Data Source

We collected 9,469 Abstract-PLS pairs (18,938 documents) extracted from the Cochrane Database of Systematic Reviews (1996-2025), spanning nearly 30 years of evolving medical communication practices (Cochrane Library, 2025). Prior work has highlighted significant content misalignment between abstracts and their corresponding PLS, where summaries often incorporate information from full-text articles (Bakker and Kamps, 2024). To address this issue, the authors proposed a new dataset (Cochrane-auto) that ensures better alignment between abstracts and PLS. Although we recognize this problem, we preferred to use the original Cochrane data to maintain the integrity of the dataset, and ensure easier evaluations and comparison between different strategies.

3.1.2 Data Processing Pipeline

We filtered and split the dataset into reference and test sets, applying minimum length thresholds (200 words for abstracts, 150 for PLS). After filtering, we retained 16,308 documents from the original

18,938 documents. For our experiments, we used only PLS texts from the reference corpus to compute statistical thresholds for the evaluation agent (ignoring their paired abstracts), and the test set pairs for generation and evaluation. The test set comprised recent publications (2023-2025) to align with the 2022 Cochrane PLS guidelines. Table 1 shows the data used in our study (complete dataset splits are available in the repository).

Data	PLS	Abstracts	Total
Reference	6,754	–	6,754
Test	100	100	200
Total	6,854	100	6,954

Table 1: Dataset distribution. Reference corpus: PLS texts for computing percentile thresholds. Test set: Abstract-PLS pairs for evaluation.

3.2 PLS Linguistic Feature Analysis

3.2.1 Feature Extraction

We extracted 18 linguistic features from each document (see Appendix B.1) comprising 9 readability indices, 4 structural metrics, 3 vocabulary measures, and 2 content density indicators. These features enable the multi-agent system to compare any text against typical PLS patterns using percentile distributions. The Cochrane PLS guidelines (Pitcher et al., 2022) recommend specific criteria: maximum 850 words, active voice, personal pronouns, and 20 words per sentence average.

These metrics answer concrete diagnostic questions: "Where does this text's passive voice usage fall compared to typical PLS?" or "Is this sentence

length in the common range?" Using the percentile thresholds from Section 3.2.2, the evaluator agent identifies specific deviations and provides feedback to the editor agent for improving the draft.

3.2.2 Percentiles for PLS Diagnosis

We computed statistical thresholds from our reference corpus of PLS texts based on percentiles to create a reference baseline for evaluation. These thresholds function as an interpretable diagnostic tool for the multi-agent system (or a human evaluator), providing explicit information about where generated text falls within the distribution of each linguistic feature. This approach enables specific improvements based on concrete positional feedback rather than abstract quality scores. We chose percentiles over machine learning approaches (e.g., gradient boosting, decision trees with feature importance) because when communicating diagnostic feedback to an evaluator, percentile distributions provide the most natural and interpretable way to identify which features have atypical values and where they fall relative to typical patterns.

We use a dual percentile system that adapts to each metric’s direction. For metrics where lower values are preferred, we label the ranges as P25, P50, P75, and P90, corresponding to the actual percentiles. For metrics where higher values are preferred, we maintain the same labels but apply them to the inverse percentiles (P75, P50, P25, and P10 respectively). This ensures P25 and P75 consistently identify the best quartile regardless of metric direction.

The tool provides the evaluator agent with specific positional information (e.g., "passive voice falls in P90 range") that directly translates to actionable feedback. While deviation from typical patterns (beyond P10 or P90) suggests that a revision may be warranted, such deviations do not automatically indicate poor quality. For instance, a text scoring in the P90 range for complex vocabulary may still be considered plain language if those terms are medically necessary and properly defined. Complex medical procedures may require precise technical terminology that cannot be simplified without losing essential meaning (a limitation discussed in Section 7). The percentile ranges serve as diagnostic indicators rather than absolute quality judgments, guiding targeted improvements while preserving content accuracy. Appendix B.2 presents the complete thresholds used as the diagnostic baseline. These thresholds are

then integrated into the PLS Evaluation Tool (Section 3.3.2) to enable automated quality assessment.

3.3 Diagnostic Tools Development

We developed two deterministic tools that emulate the resources and decision-making process of professional medical writers: a medical glossary service and a PLS evaluation tool. These tools provide the editor agent with the same type of guidance a human medical writer would use, including professionally-recommended terminology simplifications and rapid interpretable indicators to identify atypical text patterns. By grounding our tools in professional practices mentioned in the Cochrane PLS guidelines, we enable systematic evaluation and improvement of generated text.

3.3.1 Medical Glossary Tool

We collected 20,637 medical terms with plain language alternatives from 11 professional dictionaries recommended by the Williams (2025) and Cochrane Plain Language Summary Guidelines (Pitcher et al., 2022, page 29, Appendix 1). Table 2 shows the distribution of terms across sources, spanning cancer terminology, public health, diabetes, genetics, clinical trials, and other healthcare domains (see Appendix C for detailed source descriptions). The tool uses a longest-match regex algorithm to identify medical terms in submitted text and returns structured JSON with the term, its plain language alternative, and source; mirroring how a medical writer would consult reference materials during revision.

Source	Focus Area	Terms
NCI-C	Cancer terminology	9,416
NCI-D	Cancer drugs	9,144
CDC-T	Public health	891
ADA-D	Diabetes	247
NCI-G	Genetics	242
UIowa	General	242
MRCT	Clinical trials	187
WA-PH	Immunization	104
WebMD-A	Asthma	75
CCHIO	Insurance	59
Cochrane	Systematic reviews	30
Total		20,637

Table 2: Medical glossary sources with term counts and focus areas.

3.3.2 PLS Evaluation Tool

Using the percentile thresholds computed in Section 3.2.2, the evaluation tool provides rapid, inter-

pretable assessments that a medical writer would typically perform manually. Given a text, it computes 18 linguistic features using standard computational linguistics algorithms (no LLMs involved) and maps each to its percentile range based on the thresholds. The tool generates evaluation reports showing word count compliance, metric-by-metric analysis, percentile assignments, and revision suggestions for atypical patterns. This human-interpretable output allows the editor agent to make informed decisions about which deviations warrant revision and which are contextually justified (see Appendix B.3 for example output).

3.4 Multi-Agent System Architecture

We designed 14 specialized agents and their corresponding prompts organized into four functional groups: Information Extraction, Writer, Diagnostic, and Evaluation. Each prompt was developed based on the Cochrane PLS template (Pitcher et al., 2022), iteratively refined through a combination of Gemini 2.5 Pro and Claude Opus 4.1 generations with human revision to ensure alignment with Cochrane guidelines (see the repository for all the prompts). Figure 2 illustrates the complete multi-agent architecture with all components and their interactions. We first describe the core agent functionalities below; the architectural variants (Baseline, V₁, V₂) are presented in Section 3.4.4.

3.4.1 Information Extraction Agents

These agents work (mostly) in parallel to extract different parts of the abstract simultaneously, with each agent based on a specific section of the Cochrane PLS template (Pitcher et al., 2022).

To begin, the **Plain Titles Agent** reformulates technical review titles into patient-friendly questions, following Cochrane’s recommendation to use question-based titles that directly address patient concerns. When complex medical terms appear in titles, they are either replaced with plain language alternatives or clearly defined for patient understanding. Similarly, the **Key Messages Agent** extracts 2–3 main findings as bullet points, ensuring technical terms are either avoided or explained.

For the introductory content, the **Background Agent** creates 2–3 subsections with question-based headings that explain what the health condition is (“What is [condition]?”), why it matters, and what the researchers wanted to find out (“What did we want to find out?”). This output includes the review aims, which are then referenced by the

Methods Agent. Building on these aims, the Methods Agent writes “What did we do?” in 1–2 sentences, ensuring direct connection to the research goals. It describes three key actions—searching for studies, combining results, and rating confidence in evidence—using standardized phrases like “We searched for studies that compared. . .” while avoiding specific study design mentions unless essential.

The **Results Agent** generates “What did we find?” by coordinating two specialized tool agents: the **Characteristics Agent** extracts study details (number of studies, participants, duration, countries), while the **Findings Agent** translates technical findings into plain language, simplifying narratives and avoiding technical statistical terms.

To complete the extraction pipeline, the **Limitations Agent** identifies constraints from the review findings, and the **Date Extraction Agent** standardizes when the evidence was collected.

3.4.2 Writer Agents

The **Assembly Agent** takes all the pieces from the extraction agents and combines them into one complete summary. It follows the exact order required by Cochrane: title, key messages, background sections, “What did we want to find out?”, “What did we do?”, “What did we find?”, and so on.

The **Editor Agent** improves the assembled draft through revision. It checks for problems like unexplained medical terms, complicated sentences, or forbidden elements (like acronyms or statistical data), working in coordination with evaluation mechanisms to ensure quality standards are met.

3.4.3 Diagnostic and Evaluation Agents

These agents provide specialized diagnostic support and quality assessment throughout the writing process. The **Technical Terms Recognizer Agent** identifies remaining medical terms that require explanation in plain language context.

The **Evaluator Agent** is a hybrid agent with dual functionality. In its diagnostic capacity, it identifies specific issues by leveraging the diagnostic tools developed in the previous section. As an evaluator, it performs comprehensive quality checks by verifying factual accuracy through comparing drafts against original extraction outputs to detect hallucinations, ensuring all required sections are present, and using the PLS Evaluation Tool to assess readability metrics against the thresholds from Section 3.2.2.

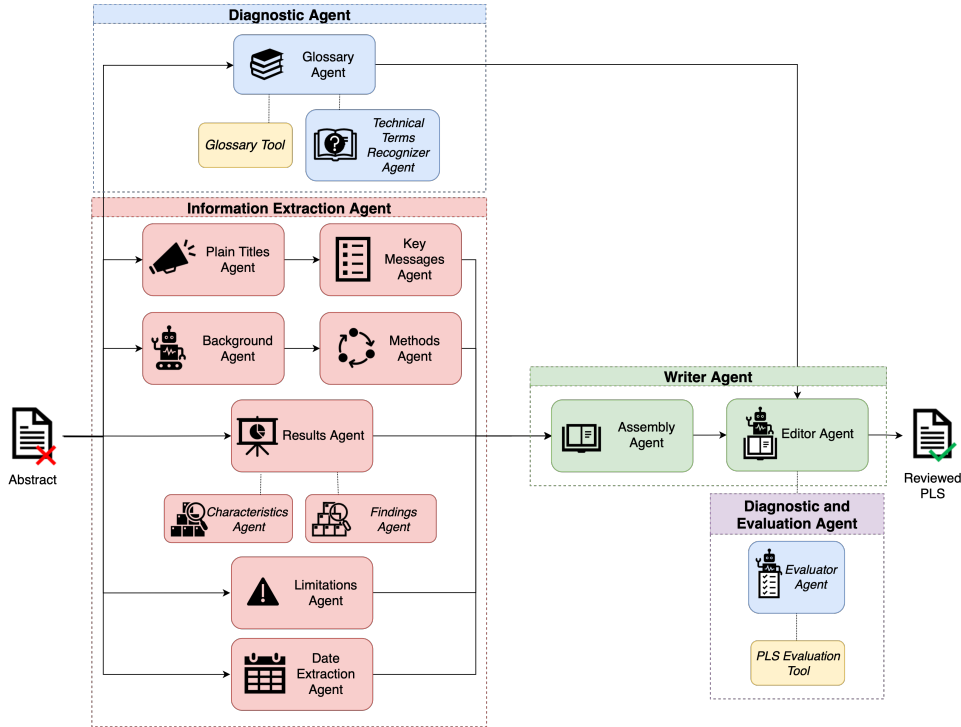


Figure 2: Multi-agent system architecture for PLS generation (V_1). Information Extraction Agents (red) process the abstract in parallel to extract different components. Writer Agents (green) assemble and refine content sequentially. Diagnostic Agents (blue) provide specialized terminology support using deterministic tools (yellow). The hybrid Diagnostic and Evaluation Agent (purple) performs both diagnostic analysis and final quality assessment. The pipeline flows from extraction through assembly and iterative refinement until the PLS is approved.

3.4.4 Architectural Variants

We evaluated three system configurations to assess different approaches to task decomposition and quality control:

- **Baseline:** This configuration consolidates the entire workflow of the agents viewed into a single, complete prompt. This allows the model to generate the complete PLS in a single step, without the need for iterative refinement or specialized agent roles.
- **Version 1 (V_1):** The primary multi-agent architecture shown in Figure 2 that integrates all the agents and tools described in the previous sections. In this configuration, specialized agents handle different subtasks as designed, the Evaluator Agent operates as an on-demand tool that the Editor can invoke when needed, and the Medical Glossary Tool provides terminology support through the Glossary Agent.
- **Version 2 (V_2):** An iterative variant where the Evaluator operates as a standalone agent that systematically evaluates each draft and provides feedback to the Editor. The pro-

cess terminates when either (1) the Evaluator approves the draft based on quality criteria, or (2) the maximum of 10 iterations is reached. Additionally, V_2 removes the Medical Glossary Tool to assess whether explicit medical dictionaries are necessary. For more details, V_2 iterative approach is presented in Appendix A.1.

These variants allow us to compare monolithic versus multi-agent approaches, tool-based versus iterative evaluation strategies, and assess the impact of explicit glossaries on generation quality.

4 Evaluation and Results

We evaluated the three architectural configurations (Baseline, V_1 , V_2) described in Section 3.4.4 across multiple language models to assess the impact of multi-agent decomposition and iterative refinement on PLS generation quality.

4.1 Experimental Setup

We implemented the system using *n8n* (2025) workflow automation platform and evaluated on 100 Abstract-PLS pairs from 2023-2025. We tested

six language models: Gemini 2.5 Pro (Google AI for Developers, 2025), GPT-5 (OpenAI, 2025b) via Azure OpenAI, GPT-OSS-120B and GPT-OSS-20B (OpenAI, 2025a) via Together AI (2025), Llama 3.3 70B, and Llama 3.2 3B.

All models were evaluated with the Baseline configuration. V_1 was tested with Gemini 2.5 Pro, GPT-5, and GPT-OSS-120B. V_2 was tested with Gemini 2.5 Pro, GPT-OSS-120B, GPT-OSS-20B, Llama 3.3 70B, and Llama 3.2 3B (GPT-5 could not be run with V_2 due to implementation constraints). Additionally, we tested a hybrid variant where Llama 3.3 70B (V_2) used Gemini 2.5 Flash as the Evaluator Agent. All models used temperature 0.0 except GPT-5 (temperature 1.0 due to API constraints). For each run, all agents within a configuration used the same underlying model.

4.2 Evaluation Metrics

Generated PLS texts were evaluated across three dimensions:

1. Relevance: Measuring the semantic similarity between the LLM-generated summaries and human-written reference PLS using BERTScore (Zhang et al., 2020), which computes token-level similarity through contextual embeddings. We also calculated similarity against original abstracts to assess information retention.

2. Factuality: Evaluating the consistency of generated content with source abstracts (ensuring no contradictory information is introduced) using AlignScore (Zha et al., 2023) and MeaningBERT (Beauchemin et al., 2023), which measure factual alignment and semantic equivalence respectively.

3. Readability: Assessing grammaticality and ease of comprehension through computational metrics. Additionally, we computed percentile distributions across all 18 linguistic features to measure conformity with typical PLS patterns, with “Best 25%” representing the percentage of features in the optimal quartile.

4.3 Results and Analysis

Our evaluation reveals inconclusive results with mixed patterns across models and configurations (Tables 3 and 4). No single architectural approach consistently outperforms others across all quality dimensions and model types, with results suggesting fundamental trade-offs between factuality and readability that manifest differently depending on base model characteristics.

Gemini 2.5 Pro is the only model where we can validly compare all three configurations (GPT-OSS-120B V_1 had implementation limitations where the evaluator tool could only be invoked once, making it unsuitable for valid comparison). Gemini shows a clear trade-off pattern: the V_2 configuration achieves the best semantic similarity and factuality scores to the reference corpus, but the baseline produces more readable text across most readability indices. The V_1 configuration (with medical glossary and evaluator as tool) achieves readability metrics similar to baseline and better than V_2 , suggesting that the medical glossary may help balance factuality and simplicity in agentic workflows, though we cannot confirm this conclusively. This pattern reflects a general trend where agentic configurations tend to improve relevance and factuality metrics, while baseline configurations often produce more readable outputs, though this varies across models.

GPT-5 agentic (V_1) underperforms its baseline across most metrics, with the exception of AlignScore to the original abstract. For GPT-OSS-120B, the baseline outperforms both V_1 and V_2 variants in factuality to reference and readability metrics, though V_2 shows improvements over V_1 .

Smaller models exhibit distinct behaviors. GPT-OSS-20B demonstrates considerable improvements with the V_2 architecture in semantic similarity and factuality compared to baseline, with modest impact on readability. Llama 3.2 3B shows an interesting pattern where both baseline and agentic configurations achieve the highest AlignScore to the original abstract among all tested models (with baseline being globally highest), yet both produce the least readable outputs, with the agentic version particularly affected. This suggests smaller models may compensate for limited capabilities by maintaining strict alignment to source material while struggling with linguistic transformation. The hybrid configuration (Llama 3.3 70B with Gemini 2.5 Flash as evaluator) achieves competitive semantic quality while substantially improving readability compared to standard Llama 3.3 70B, demonstrating that evaluator quality impacts generation quality.

Examining conformity to typical PLS patterns (Best 25% in Table 4), results are inconclusive. Gemini 2.5 Pro improves with agentic configurations, but this does not generalize to other models. Most configurations achieve conformity levels comparable to human reference patterns. Ap-

Model	Approach	VS. ORIGINAL ABSTRACT			VS. REFERENCE PLS		
		BERTScore	MeaningBERT	AlignScore	BERTScore	MeaningBERT	AlignScore
Reference (human)		0.8482	0.6825	0.7551	–	–	–
Gemini 2.5 Pro	Baseline	<u>0.8352</u>	0.5957	<u>0.7820</u>	0.8701	0.7162	<u>0.7002</u>
	Agentic (V1)	0.8420	<u>0.5928</u>	0.7909	0.8708	<u>0.6907</u>	0.7157
	Agentic (V2)	0.8469	0.6136	0.7992	0.8736	0.7153	0.7219
GPT-5	Baseline	0.8342	0.6075	<u>0.7692</u>	0.8619	0.6873	0.6598
	Agentic (V1)	<u>0.8278</u>	<u>0.5863</u>	0.7887	<u>0.8499</u>	<u>0.6718</u>	<u>0.6522</u>
GPT-OSS-120B	Baseline	0.8407	0.6477	<u>0.7696</u>	0.8650	0.7346	0.6878
	Agentic (V1)	<u>0.8321</u>	<u>0.6468</u>	0.7995	<u>0.8562</u>	<u>0.7304</u>	0.6708
	Agentic (V2)	0.8464	0.6557	0.7953	0.8651	0.7393	0.6595
GPT-OSS-20B	Baseline	<u>0.8327</u>	<u>0.5857</u>	<u>0.7396</u>	<u>0.8565</u>	<u>0.6697</u>	<u>0.6480</u>
	Agentic (V2)	0.8422	0.6680	0.8002	0.8615	0.7534	0.6590
Llama 3.3 70B	Baseline	0.8514	0.6985	<u>0.7536</u>	<u>0.8679</u>	<u>0.7158</u>	0.7076
	Agentic (V2)	0.8549	0.6818	0.7823	0.8708	0.7446	0.7140
	Agentic (V2 + Gemini 2.5 Flash)	<u>0.8485</u>	<u>0.6514</u>	<u>0.7644</u>	0.8711	0.7325	<u>0.6982</u>
Llama 3.2 3B	Baseline	0.8477	<u>0.6566</u>	<u>0.7644</u>	0.8467	0.6302	0.6982
	Agentic (V2)	0.8551	0.6952	<u>0.8403</u>	0.8532	0.6672	<u>0.6706</u>

Table 3: Semantic similarity and factuality metrics for all tested models and approaches. **Bold** indicates best performance within each model, underlined indicates worst within each model. Gray shading highlights best global performance, red shading highlights worst global performance. Human reference excluded from comparisons. All metrics averaged across 100 test samples.

Model	Approach	Words	FKGL↓	ARI↓	CLI↓	FRE↑	GFI↓	LIX↓	SMOG↓	RIX↓	DCRS↓	Best 25%	P25%	P50%	P75%	P90%	P10%
Original Abstract		868	13.85	14.07	11.11	42.12	20.39	59.26	17.37	8.60	8.75	27.61	7.39	11.89	20.22	15.89	1.00
Reference PLS (human)		655	11.38	11.46	11.21	49.23	16.21	50.35	14.23	6.07	7.37	52.50	25.89	17.89	26.61	14.78	0.22
Gemini 2.5 Pro	Baseline	661	8.26	8.01	9.07	64.45	12.38	41.21	12.74	4.02	6.53	<u>78.44</u>	45.33	5.28	33.11	12.89	0.00
	Agentic (V1)	591	8.56	8.49	9.42	63.16	12.51	42.02	12.16	4.21	6.39	82.00	46.56	9.22	35.44	7.94	0.00
	Agentic (V2)	567	<u>9.40</u>	<u>9.47</u>	<u>10.16</u>	<u>58.78</u>	<u>13.54</u>	<u>44.64</u>	<u>12.42</u>	<u>4.74</u>	<u>6.65</u>	83.22	44.67	9.11	38.56	7.06	0.00
GPT-5	Baseline	866	9.62	9.89	10.46	57.97	13.49	45.08	14.62	4.87	6.74	63.83	41.33	5.56	22.50	9.61	0.00
	Agentic (V1)	879	<u>10.97</u>	<u>11.59</u>	<u>11.71</u>	<u>51.14</u>	<u>15.12</u>	<u>48.34</u>	<u>15.50</u>	<u>5.61</u>	<u>7.71</u>	<u>49.33</u>	26.67	14.00	22.67	8.22	0.00
GPT-OSS-120B	Baseline	623	9.82	9.94	9.74	59.10	14.76	45.98	12.96	5.15	6.77	73.44	39.39	10.56	34.06	13.22	0.00
	Agentic (V1)	743	11.05	11.25	11.12	<u>51.08</u>	<u>16.55</u>	<u>49.62</u>	<u>15.03</u>	<u>5.91</u>	<u>8.01</u>	<u>46.72</u>	22.50	17.94	24.22	12.11	0.00
	Agentic (V2)	623	<u>10.83</u>	<u>11.00</u>	<u>10.82</u>	52.66	16.32	49.21	13.87	5.82	7.75	83.22	44.67	9.11	38.56	7.06	0.00
GPT-OSS-20B	Baseline	541	10.17	8.89	7.08	58.28	13.85	44.24	11.84	5.04	6.32	71.88	40.11	15.80	31.76	8.85	0.11
	Agentic (V2)	632	<u>10.50</u>	<u>10.27</u>	<u>10.14</u>	<u>53.76</u>	<u>15.78</u>	<u>47.78</u>	<u>13.84</u>	<u>5.51</u>	<u>7.74</u>	<u>59.32</u>	31.00	13.52	28.32	16.97	0.17
Llama 3.3 70B	Baseline	477	12.01	12.31	11.05	48.55	16.89	51.89	12.65	6.57	6.87	61.56	29.17	23.33	32.39	9.61	0.44
	Agentic (V2)	497	12.23	12.46	<u>12.59</u>	<u>42.72</u>	16.91	53.08	13.41	<u>6.60</u>	<u>7.54</u>	<u>57.28</u>	26.17	17.61	31.11	16.78	0.00
	Agentic (V2 + Gemini 2.5)	514	<u>9.97</u>	<u>10.13</u>	10.91	54.97	14.24	46.69	<u>12.48</u>	5.11	6.84	58.89	20.78	24.22	38.11	10.39	0.17
Llama 3.2 3B	Baseline	477	9.58	8.95	10.60	53.56	14.32	45.81	12.88	4.68	6.82	69.67	45.11	19.33	24.56	6.06	0.28
	Agentic (V2)	536	<u>12.93</u>	<u>12.72</u>	<u>12.65</u>	<u>37.95</u>	<u>17.57</u>	<u>54.41</u>	<u>14.06</u>	<u>6.95</u>	<u>7.96</u>	<u>45.69</u>	19.98	21.93	25.71	14.30	1.50

Table 4: Readability metrics and percentile distribution for all tested models. Left: average readability scores (arrows: ↓ lower is better, ↑ higher is better). Right: percentage of linguistic features in each percentile range across 18 selected features. Best 25% represents percentage in optimal quartile. **Bold** indicates best performance within each model, underlined indicates worst within each model. Gray shading highlights best global performance, red shading highlights worst global performance. Original Abstract and Reference PLS (human) excluded from comparisons. Averages computed across 100 test samples.

pendix A.2 presents a detailed example of the V_2 iterative refinement process, illustrating how the Evaluator Agent provides structured feedback that guides draft improvements from 83.33% to 94.44% best quartile conformity.

5 Discussion

Our evaluation reveals inconclusive results regarding which approach is superior, with both baseline and multi-agent configurations showing distinct advantages depending on use case requirements. The baseline proves remarkably effective when properly designed with comprehensive instructions based on Cochrane guidelines, demonstrating that systematic prompt engineering grounded in domain standards can produce high-quality PLS. The multi-agent architecture, while not completely superior, provides specific benefits in certain contexts.

For smaller models, the multi-agent approach shows improvements in relevance and factuality metrics. GPT-OSS-20B improves in semantic similarity and factual alignment when using the V_2 configuration, though with a slight deterioration in readability. Notably, smaller models achieve competitive or higher factuality scores compared to larger models, something we noticed in our prior work (Arias-Russi et al., 2025), suggesting these models may be more conservative in adhering to source material. A potential strategy to address readability limitations would involve using a smaller model for content extraction followed by a larger model for final linguistic refinement, potentially offering cost-effective generation while preserving factual accuracy.

The multi-agent architecture provides operational advantages in terms of control and inter-

pretability. Separating extraction, assembly, and evaluation into distinct agents makes each component transparent and independently modifiable. While the baseline consolidates all instructions into a single comprehensive prompt, the decomposed approach allows for targeted refinement of specific subtasks without affecting the entire pipeline. This modularity also produces intermediate outputs for each section, which are stored separately in our [repository](#) and can be inspected individually for diagnostic purposes.

However, the multi-agent approach incurs higher token costs due to multiple agent invocations and memory preservation through context repetition across agents. The V_1 configuration, which includes the medical glossary tool, is particularly token-intensive, with processing costs reaching approximately 10M tokens (input + output combined) per 100-abstract batch. Most tokens are input tokens, which are typically cheaper than output tokens, though iterative configurations may generate similar text multiple times, increasing output costs. The baseline, requiring only a single model invocation per abstract, proves more cost-effective in terms of API usage. This cost-benefit trade-off must be considered when selecting an approach for production deployment.

The effectiveness of smaller models like GPT-OSS-20B with the multi-agent approach could enable local deployment in medical contexts where data privacy is critical. While the current task of simplifying published Cochrane reviews does not involve sensitive information, other medical text simplification scenarios could benefit from local processing. For instance, healthcare providers might need to simplify patient-specific medical reports or treatment explanations without transmitting sensitive data to external APIs. In such contexts, the ability to run smaller models locally while maintaining reasonable quality through multi-agent decomposition could provide a viable solution.

Beyond technical performance metrics, this work addresses the practical need of facilitating medical writers' work and improving health information accessibility. Our framework provides diagnostic tools and automated first drafts that meet professional PLS standards, potentially reducing the manual effort required to produce accessible health communication materials.

6 Future Work

Future work could explore alternative evaluation approaches beyond percentile-based diagnostics, including different metrics and quality assessment methods for iterative refinement. Multi-agent hybrid systems where different agents use specialized models could balance cost and quality. Extending the architecture to other plain language formats such as PLS or documents following the Federal Plain Language Guidelines ([Williams, 2025](#)) would test its generalizability. Additionally, incorporating full-text papers as input sources rather than abstracts alone could address content coverage limitations, potentially leveraging datasets like Cochrane-auto for improved alignment between source and simplified text.

7 Limitations

Our system may perpetuate suboptimal information prioritization patterns from the training data ([Bakker and Kamps, 2024](#)). Computational constraints and API rate limits restricted experimentation scope, and we did not conduct formal statistical hypothesis testing. GPT-5 could not be tested with V_2 due to API rate limits, only supports temperature 1.0² preventing deterministic generation, and exhibited inconsistent behavior (sometimes the model did not use the evaluation tool for V_1). GPT-OSS-120B V_1 had implementation issues with the evaluator tool. V_1 intermediate outputs were lost due to storage issues; V_2 outputs are available in the [repository](#). The V_2 example in [Appendix A.2](#) represents a single cherry-picked case. Our percentile-based evaluation framework represents statistical conformity rather than absolute quality, and strict percentile ranges could penalize innovative plain language strategies.

8 Lay Summary

Medical research papers often contain complex language that makes them difficult for patients and the general public to understand. Plain Language Summaries help solve this problem by explaining research findings using everyday words. Organizations like Cochrane create these summaries for their systematic reviews, which combine results from many studies to answer health questions. However,

²<https://web.archive.org/web/20250903093505/https://community.openai.com/t/temperature-in-gpt-5-models/1337133/20>

writing plain language summaries requires medical expertise and takes considerable time. This creates a bottleneck in making health information accessible to everyone.

We wanted to find out whether computer systems using AI could automatically generate high-quality plain language summaries. Specifically, we tested whether breaking down the writing task into smaller steps handled by specialized AI agents would work better than using a single comprehensive instruction. We also wanted to know if this approach would be more helpful for some AI models than others.

We built a system that divides summary writing into four stages: extracting information from the medical abstract, assembling it into a draft, checking for medical terms that need simplification, and evaluating readability. The system uses a medical dictionary with over 20,000 terms and their plain language alternatives. It also includes a statistical analyzer that compares the generated text against patterns found in human-written summaries. We tested this system using 100 Cochrane medical abstracts and six different AI models, ranging from large commercial models to smaller open-source ones.

Our results are mixed and there is no clear winner among the designed architectures. The baseline approach, which uses a single well-designed instruction, performed surprisingly well. The multi-agent system did not consistently outperform the baseline, but it showed specific advantages. Smaller AI models improved notably when using the multi-agent approach, achieving better accuracy in preserving medical facts, though sometimes at the cost of readability. We also found that the multi-agent system provides greater control and less black-box effect, allowing users to inspect and modify individual steps separately.

Healthcare organizations and research institutions working with limited computational resources could benefit from these findings. The results suggest that smaller, locally-run AI models combined with the multi-agent approach could generate reasonably accurate summaries while maintaining data privacy. The modular design also makes it easier to adapt the system for different types of medical writing beyond Cochrane summaries. However, more work is needed to improve readability when using smaller models and to reduce the computational costs of the multi-agent approach.

References

- American Diabetes Association. 2024. [Common Diabetes Terms](#). Accessed: August 23, 2025.
- Jonathan Anderson. 1983. [Lix and Rix: Variations on a Little-known Readability Index](#). *Journal of Reading*, 26(6):490–496. Publisher: [Wiley, International Reading Association].
- Andrés Arias-Russi, Carolina Salazar-Lara, and Rubén Manrique. 2025. [Bridging the gap in health literacy: Harnessing the power of large language models to generate plain language summaries from biomedical texts](#). In *Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health)*, pages 269–284, Albuquerque, New Mexico. Association for Computational Linguistics.
- B. Bahador, S. Baedorf Kassis, H. Gawrylewski, and et al. 2020. [Promoting equity in understanding: A cross-organizational plain language glossary for clinical research](#). *Medical Writing*, 29(4):10–15.
- Jan Bakker and Jaap Kamps. 2024. [Cochrane-auto: An Aligned Dataset for the Simplification of Biomedical Abstracts](#). In *Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024)*, pages 41–51, Miami, Florida, USA. Association for Computational Linguistics.
- David Beauchemin, Horacio Saggion, and Richard Khoury. 2023. [MeaningBERT: assessing meaning preservation between sentences](#). *Frontiers in Artificial Intelligence*, 6. Publisher: Frontiers.
- N. D. Berkman, S. L. Sheridan, K. E. Donahue, D. J. Halpern, and K. Crotty. 2011. [Low health literacy and health outcomes: an updated systematic review](#). *Annals of Internal Medicine*, 155(2):97–107.
- Nur Alya Dania Binti Moriasi and Mujeen Sung. 2025. [KHU_LDI at BioLaySumm2025: Fine-tuning and refinement for lay radiology report generation](#). In *Proceedings of the 24th Workshop on Biomedical Language Processing (Shared Tasks)*, pages 256–268, Vienna, Austria. Association for Computational Linguistics.
- Center for Consumer Information & Insurance Oversight. 2024. [Uniform glossary of health coverage and medical terms](#). Updated for plan or policy years beginning on or after January 1, 2024. Accessed: August 23, 2025.
- Centers for Disease Control and Prevention. 2011. [Plain Language Thesaurus for Health Communications](#). Accessed: August 23, 2025.
- Jeanne Sternlicht Chall and Edgar Dale. 1995. *Readability Revisited: The New Dale-Chall Readability Formula*. Brookline Books. Google-Books-ID: 2nbuAAAAMAAJ.
- Cochrane Library. 2025. Cochrane Database of Systematic Reviews. <https://www.cochranelibrary.com/cdsr/reviews>.

- Meri Coleman and T. L. Liau. 1975. [A Computer Readability Formula Designed for Machine Scoring](#). *Journal of Applied Psychology*, 60(2):283–284. Place: US Publisher: American Psychological Association.
- Mahshad Koochi Habibi Dehkordi, Yehoshua Perl, Fadi P. Deek, Zhe He, Vipina K. Keloth, Hao Liu, Gai Elhanan, and Andrew J. Einstein. 2025. [Improving Large Language Models’ Summarization Accuracy by Adding Highlights to Discharge Notes: Comparative Evaluation](#). *JMIR Medical Informatics*, 13(1):e66476. Company: JMIR Medical Informatics Distributor: JMIR Medical Informatics Institution: JMIR Medical Informatics Label: JMIR Medical Informatics Publisher: JMIR Publications Inc., Toronto, Canada.
- Dengzhao Fang, Jipeng Qiang, Xiaoye Ouyang, Yi Zhu, Yunhao Yuan, and Yun Li. 2025. [Collaborative Document Simplification Using Multi-Agent Systems](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 897–912, Abu Dhabi, UAE. Association for Computational Linguistics.
- Rudolph Flesch. 1948. [A New Readability Yardstick](#). *Journal of Applied Psychology*, 32(3):221–233. Place: US Publisher: American Psychological Association.
- Google AI for Developers. 2025. [Gemini models: Gemini API | Google AI for Developers](#). Latest update: August 2025; archived at <https://web.archive.org/web/20250825013047/https://ai.google.dev/gemini-api/docs/models#previous-experimental-models>.
- Robert Gunning. 1952. *The Technique of Clear Writing*. McGraw-Hill. Google-Books-ID: ofI0AAAAMAAJ.
- Aaradhya Gupta and Parameswari Krishnamurthy. 2025. [Shared task at biolaysumm2025 : Extract then summarize approach augmented with UMLS based definition retrieval for lay summary generation](#). In *Proceedings of the 24th Workshop on Biomedical Language Processing (Shared Tasks)*, pages 185–189, Vienna, Austria. Association for Computational Linguistics.
- Human Subjects Office, University of Iowa. 2021. [Medical terms in lay language](#). Last updated: 05/03/2021. Accessed: August 23, 2025.
- Peter Kincaid, Robert P. Fishburne, 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](#). In *Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel*.
- Zihao Li, Samuel Belkadi, Nicolo Micheletti, Lifeng Han, Matthew Shardlow, and Goran Nenadic. 2024. [Large Language Models for Biomedical Text Simplification: Promising But Not There Yet](#). *arXiv preprint*. ArXiv:2408.03871 [cs].
- Juan Antonio Lossio-Ventura, Callum Chan, Arshitha Basavaraj, Hugo Alatrística-Salas, Francisco Pereira, and Diana Inkpen. 2025. [5cNLP at BioLay-Summ2025: Prompts, retrieval, and multimodal fusion](#). In *Proceedings of the 24th Workshop on Biomedical Language Processing (Shared Tasks)*, pages 215–231, Vienna, Austria. Association for Computational Linguistics.
- Chen Lyu and Gabriele Pergola. 2024a. [SciGisPy: a novel metric for biomedical text simplification via gist inference score](#). In *Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024)*, pages 95–106, Miami, Florida, USA. Association for Computational Linguistics.
- Chen Lyu and Gabriele Pergola. 2024b. [Society of Medical Simplifiers](#). In *Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024)*, pages 61–68, Miami, Florida, USA. Association for Computational Linguistics.
- Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegrefe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. [Self-refine: Iterative refinement with self-feedback](#). *Preprint*, arXiv:2303.17651.
- G. Harry Mc Laughlin. 1969. [SMOG Grading: A new readability formula](#). *Journal of Reading*, 12(8):639–646. Publisher: [Wiley, International Reading Association].
- Kaijie Mo and Renfen Hu. 2024. [ExpertEase: A multi-agent framework for grade-specific document simplification with large language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9080–9099, Miami, Florida, USA. Association for Computational Linguistics.
- n8n. 2025. [n8n: Workflow automation platform \(v1.109.0\)](#). <https://github.com/n8n-io/n8n/releases/tag/n8n@1.108.0>. Version 1.109.0, released August 25, 2025.
- National Cancer Institute. 2025a. [NCI Dictionary of Cancer Terms](#). Accessed: August 23, 2025.
- National Cancer Institute. 2025b. [NCI Dictionary of Genetics Terms](#). Accessed: August 23, 2025.
- OpenAI. 2025a. [gpt-oss-120b & gpt-oss-20b model card](#). *Preprint*, arXiv:2508.10925.
- OpenAI. 2025b. [OpenAI models: Open AI API](#). Latest update: August 2025; archived at <http://web.archive.org/web/20250817212345/https://platform.openai.com/docs/models/gpt-5>.

- Nicole Pitcher, Denise Mitchell, and Carolyn Hughes. 2022. *Template and guidance for writing a Cochrane Plain language summary*. Cochrane. Archived at <https://web.archive.org/web/20250824000540/https://www.cochrane.org/authors/handbooks-and-manuals/handbook/current/guidance-writing-cochrane-plain-language-summary.pdf> (Accessed 2025-08-23).
- Readability. 2019. *Readability 0.3.1*. Accessed August 2025.
- Karen Scholz and Markus Wenzel. 2025. *Evaluating readability metrics for German medical text simplification*. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6049–6062, Abu Dhabi, UAE. Association for Computational Linguistics.
- R. J. Senter and E. A. Smith. 1967. Automated readability index. Technical Report AMRL-TR-6620, Wright-Patterson Air Force Base, Ohio, USA.
- Bhuvaneshwari Sivagnanam, Rivo Krishnu C H, Princi Chauhan, and Saranya Rajiakodi. 2025. *CUTN_Bio at BioLaySumm: Multi-task prompt tuning with external knowledge and readability adaptation for layman summarization*. In *Proceedings of the 24th Workshop on Biomedical Language Processing (Shared Tasks)*, pages 269–274, Vienna, Austria. Association for Computational Linguistics.
- SpaCy. 2025. *SpaCy 3.8.7*. Accessed August 2025.
- Kristine Sørensen, Jürgen M. Pelikan, Florian Röthlin, Kristin Ganahl, Zofia Slonska, Gerardine Doyle, James Fullam, Barbara Kondilis, Demosthenes Agrafiotis, Ellen Uiters, Maria Falcon, Monika Mensing, Kancho Tchamov, Stephan van den Broucke, and on behalf of the HLS-EU Consortium Brand, Helmut. 2015. *Health literacy in Europe: comparative results of the European health literacy survey (HLS-EU)*. *European Journal of Public Health*, 25(6):1053–1058.
- The Multi-Regional Clinical Trials Center of Brigham and Women’s Hospital and Harvard. 2025. *Clinical Research Glossary*. Accessed: August 23, 2025.
- Together AI. 2025. *gpt-oss-120B API*. Archived at <https://web.archive.org/web/20250821140846/https://www.together.ai/models/gpt-oss-120b>.
- Hieu Tran, Zonghai Yao, Won Seok Jang, Sharmin Sultana, Allen Chang, Yuan Zhang, and Hong Yu. 2025. *MedReadCtrl: Personalizing medical text generation with readability-controlled instruction learning*. *arXiv preprint*. ArXiv:2507.07419 [cs].
- Oisín Turbitt, Robert Bevan, and Mouhamad Aboshokor. 2023. *MDC at BioLaySumm task 1: Evaluating GPT models for biomedical lay summarization*. In *Proceedings of the 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 611–619, Toronto, Canada. Association for Computational Linguistics.
- Dave Van Veen, Cara Van Uden, Louis Blanke-meier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, Nidhi Rohatgi, Poonam Hosamani, William Collins, Neera Ahuja, Curtis P. Langlotz, Jason Hom, Sergios Gavidis, John Pauly, and Akshay S. Chaudhari. 2024. *Adapted large language models can outperform medical experts in clinical text summarization*. *Nature Medicine*, 30(4):1134–1142.
- Ziyue Wang, Junde Wu, Linghan Cai, Chang Han Low, Xihong Yang, Qiakuan Li, and Yueming Jin. 2025. *Medagent-pro: Towards evidence-based multi-modal medical diagnosis via reasoning agentic workflow*. *Preprint*, arXiv:2503.18968.
- Washington State Department of Health. 2023. *Glossary of immunization and public health terms*. Updated: June 2023. Accessed: August 23, 2025.
- WebMD. 2022. *Asthma Glossary*. Medically Reviewed on September 22, 2022. Accessed: August 23, 2025.
- Mary Ann Williams. 2025. *Guides: Health Literacy Resources: Plain Language Resources*. <https://guides.hshsl.umaryland.edu/c.php?g=94026&p=7981462>. Accessed: August 23, 2025.
- Chenghao Xiao, Kun Zhao, Xiao Wang, Siwei Wu, Sixing Yan, Tomas Goldsack, Sophia Ananiadou, Noura Al Moubayed, Liang Zhan, William Cheung, and Chenghua Lin. 2025. *Overview of the biolaysumm 2025 shared task on lay summarization of biomedical research articles and radiology reports*. In *The 24th Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, Vienna, Austria. Association for Computational Linguistics.
- Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. *AlignScore: Evaluating Factual Consistency with a Unified Alignment Function*. *arXiv preprint*. ArXiv:2305.16739 [cs].
- 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*.
- Wenjun Zhang, Shekhar Chandra, Bevan Koopman, Jason Dowling, and Aaron Nicolson. 2025. *AEHRC at BioLaySumm 2025: Leveraging t5 for lay summarisation of radiology reports*. In *Proceedings of the 24th Workshop on Biomedical Language Processing (Shared Tasks)*, pages 171–178, Vienna, Austria. Association for Computational Linguistics.

A System Architecture and Methodology Workflow

Figure A1 illustrates the complete methodology workflow for developing and evaluating the multi-agent system.

A.1 Iterative Architecture

Initial experiments revealed that the baseline configuration considerably outperformed the V_1 architecture in several metrics. This motivated the development of V_2 as an alternative approach to test different architectural strategies and increase result diversity. The V_2 configuration (Figure A2) addresses two key observations from preliminary testing: (1) smaller models struggle to properly utilize tools, performing significantly better with iterative workflows rather than on-demand tool invocation, and (2) the Medical Glossary Tool, while conceptually useful, consumes substantial tokens without proportional quality gains.

The V_2 architecture maintains the same specialized agents as V_1 but reorganizes their interaction pattern. Instead of the Editor invoking the Evaluator as a tool when needed, the Evaluator operates as an independent agent that systematically reviews each draft and provides structured feedback to the Editor. This iterative loop continues for up to 10 cycles or until the Evaluator approves the output based on quality criteria. When the maximum iteration limit is reached without approval, we select the iteration that achieves the highest percentage of linguistic features in the optimal quartile. For each generated text, the PLS Evaluation Tool computes 18 linguistic features and maps them to percentile ranges (P25, P50, P75, P90).

The optimal quartile corresponds to P25 for lower-is-better metrics and P75 for higher-is-better metrics. We calculate what percentage of all features fall within this optimal range and select the iteration with the highest percentage. While metrics like AlignScore might provide more reliable quality assessment, we opted for this percentile-based approach due to computational efficiency constraints, as computing AlignScore for each iteration would significantly extend processing time. The removal of the Medical Glossary Tool reduces token consumption while testing whether explicit medical dictionaries are necessary for quality PLS generation.

This architectural variant allowed us to compare monolithic (Baseline), tool-based (V_1), and iter-

ative (V_2) approaches using identical underlying models and agents, providing insights into how interaction patterns affect generation quality across different model sizes.

A.2 Example V_2 Iterative Refinement Process

This section illustrates the V_2 iterative refinement process using Cochrane review CD002120.PUB4 (“Combined oral contraceptive pill for primary dysmenorrhoea”) as a representative example.

The Information Extraction Agents processed the abstract and produced structured outputs for each required PLS component:

Plain Title: What are the benefits and harms of the combined birth control pill for treating painful periods (primary dysmenorrhoea)?

Key Messages:

- For women with painful periods, the combined birth control pill reduces pain more than a placebo. However, it may cause unwanted effects such as irregular bleeding, headaches, and nausea.
- Taking the pill continuously may be better at reducing pain than taking it in the standard way.
- Because the studies were short, the long-term effects are not known. It is also unclear if the pill is more effective than other pain medicines due to a lack of studies directly comparing them.

Background: Many women experience painful menstrual cramps, a condition known as dysmenorrhoea. When these painful cramps are not caused by another underlying medical condition in the pelvis, it is called ‘primary dysmenorrhoea’. Combined oral contraceptive pills are a common treatment for this pain.

Methods: We searched for studies that compared combined oral contraceptive pills with a placebo, other contraceptive pills, or anti-inflammatory drugs for treating painful periods. We then combined the results and rated our confidence in the evidence.

Results: This review included 21 randomised controlled trials involving 3723 women. For women with painful menstrual cramps, OCPs reduce pain more effectively than placebo. However, OCPs increase the risk of irregular bleeding. Continuous use may reduce pain more effectively than the standard schedule. . .

Limitations: Our confidence in the evidence is only moderate because there are not enough studies to be certain about all outcomes. The evidence does not cover long-term effects.

Currency: The evidence is up to date to March 2023.

The Assembly Agent constructed the initial draft of 617 words, achieving 83.33% of features in the best quartile. The Evaluator identified three

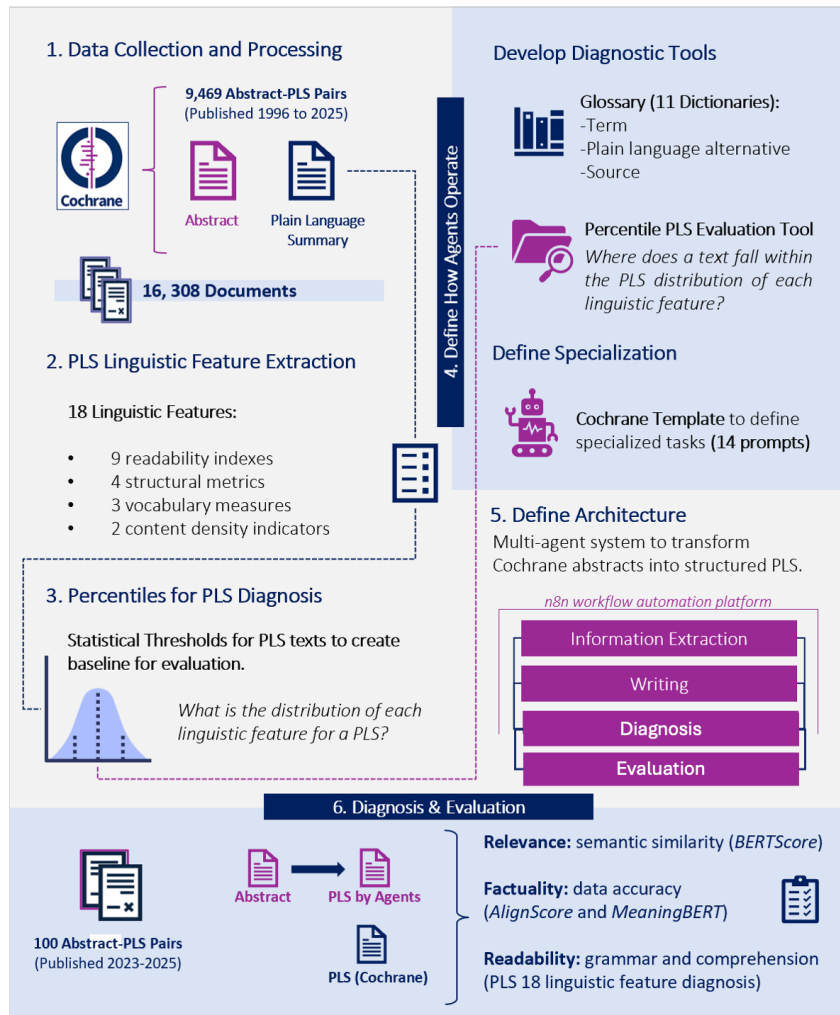


Figure A1: Methodology for developing and evaluating a multi-agent system for automatic generation of PLS from Cochrane medical abstracts.

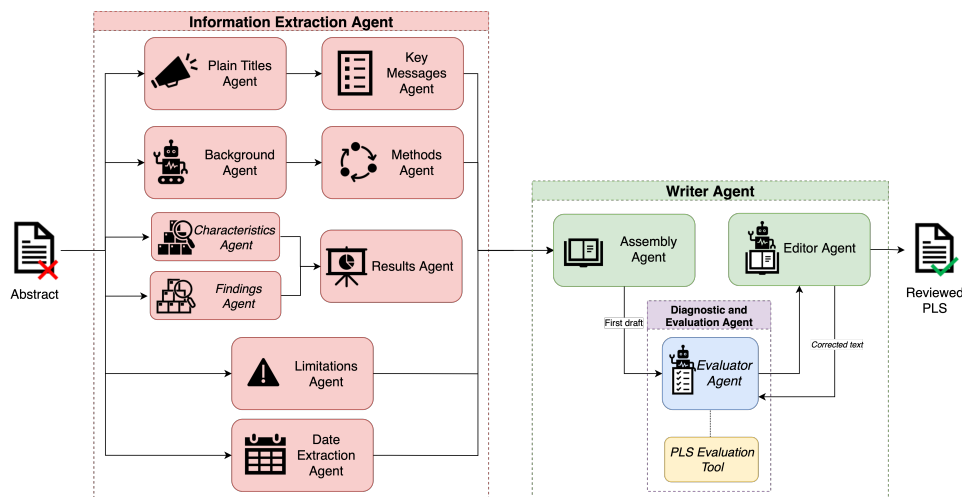


Figure A2: V_2 iterative architecture for PLS generation. Unlike V_1 where the Editor invokes the Evaluator as a tool on-demand, V_2 establishes an independent iterative loop where the Evaluator systematically reviews drafts and provides feedback to the Editor. The process continues for up to 10 iterations or until approval. This architecture removes the Medical Glossary Tool and proves particularly effective for smaller models that struggle with tool invocation.

key issues: (1) a non-question subheading (“Main results: oral contraceptives for painful menstrual cramps”), (2) complex phrasing such as “When these painful cramps are not caused by another underlying medical condition in the pelvis. . .”, and (3) excessive passive voice constructions. Three metrics fell in the P90 range: nouns (158), complex words (241), and long words (165). The draft was marked NOT_APPROVED.

What are the benefits and harms of the combined birth control pill for treating painful periods. . .

Key messages [same as extracted]

What is primary dysmenorrhoea and what did this review study?

What is primary dysmenorrhoea? Many women experience painful menstrual cramps, a condition known as dysmenorrhoea. It is a common and major cause of pain. When these painful cramps are not caused by another underlying medical condition. . .

What did we want to find out? We wanted to find out about the benefits and harms of using combined oral contraceptive pills. . .

How was the review conducted? We searched for studies that compared combined oral contraceptive pills with a placebo, other contraceptive pills, or anti-inflammatory drugs. . .

Main results: oral contraceptives for painful menstrual cramps For women with painful menstrual cramps, the pill reduces pain more effectively than a placebo. However, the pill comes with unwanted effects. . .

How reliable is this evidence? Our confidence in the evidence is only moderate because there are not enough studies to be certain. . .

How up to date is this evidence? The evidence is up to date to March 2023.

The Editor incorporated this feedback in Draft 2, reducing word count to 561 and improving the best quartile percentage to 94.44%. The Editor converted the non-question subheading to “What are the main results of the review?”, simplified phrasing (e.g., “This is called ‘primary dysmenorrhoea’ when the pain is not caused by another medical problem” instead of the more complex original wording), and reduced passive voice from 17 to 13 instances. All linguistic metrics moved within P75 or better, with no metrics remaining in the P90 range. The draft was marked PASS.

What are the benefits and harms of the combined birth control pill for treating painful periods. . .

Key messages [same as extracted]

What is primary dysmenorrhoea? Many women experience painful menstrual cramps, a condition known as dysmenorrhoea. This is called ‘primary dysmenorrhoea’ when the pain is not caused by another medical problem in the pelvis. . .

What did we want to find out? We wanted to find out about the benefits and harms of using combined oral contraceptive pills. . .

How did we conduct this review? We searched for studies that compared combined oral contraceptive pills with a placebo. We found 21 studies, known as randomised controlled trials, with a total of 3723 women. In these studies, researchers randomly put people into one of 2 or more treatment groups. . .

What are the main results of the review? For women with painful menstrual cramps, the pill reduces pain more effectively

than a placebo. However, the pill comes with unwanted effects. . .

How reliable is this evidence? Our confidence in the evidence is only moderate because there are not enough studies to be certain about all health effects. . .

How up to date is this evidence? The evidence is up to date to March 2023.

Table A1 quantifies the improvements between iterations.

Metric	Draft 1	Draft 2
Word count	617	561
FKGL	9.68	9.54
Passive voice	17	13
Nouns	158	142
Complex words (DC)	241	213
Long words	165	147
Best quartile (%)	83.33	94.44
P90 metrics	3	0
Decision	FAIL	PASS

Table A1: Metric improvements from Draft 1 to Draft 2 for CD002120.PUB4.

B Linguistic Analysis Framework

B.1 Linguistic Features

We computed 20 linguistic features for each document using the [Readability \(2019\)](#) and [SpaCy \(2025\)](#) libraries. Here we describe the 20 features selected for quality assessment (see Table B1 for the percentile thresholds):

- Words:** Total word count in the text.
- Sentences:** Total sentence count in the text.
- Flesch Reading Ease (FRE):** Produces a score where higher values indicate easier readability ([Flesch, 1948](#); [Kincaid et al., 1975](#)).
- Flesch-Kincaid Grade Level (FKGL):** Estimates the U.S. school grade level needed to comprehend the text ([Flesch, 1948](#); [Kincaid et al., 1975](#)).
- Gunning Fog Index (GFI):** Estimates the number of years of formal education needed to understand the text ([Gunning, 1952](#)).
- SMOG Readability Formula (SMOG):** Estimates readability by counting polysyllabic words ([Mc Laughlin, 1969](#)).
- Dale-Chall Readability Score (DCRS):** Assesses readability by comparing text words against a list of familiar words ([Chall and Dale, 1995](#)).

8. **Coleman-Liau Index (CLI):** Measures readability based on letter and word counts per sentence (Coleman and Liau, 1975).
9. **Automated Readability Index (ARI):** Computes readability using characters, words, and sentences (Senter and Smith, 1967).
10. **LIX:** Calculates readability by analyzing the proportion of long words in the text (Anderson, 1983).
11. **RIX:** Computes readability from the number of long words per sentence (Anderson, 1983).
12. **Words per Sentence:** Average number of words per sentence, computed as total words divided by total sentences.
13. **Passive Voice:** Frequency of passive voice constructions, determined via verb forms tagged as VBN (e.g., “was given”).
14. **Active Voice:** Frequency of active voice constructions, counted as verbs not tagged as VBN (e.g., “ran”, “decided”).
15. **Nominalization:** Count of nominalizations, where verbs or adjectives are transformed into nouns (e.g., “development” from “develop”).
16. **Complex Words (DC):** Count of complex words according to the Dale-Chall method (unknown polysyllabic words from a list of basic words).
17. **Long Words:** Count of words exceeding 7 letters in length.
18. **Complex Words:** Count of words with three or more syllables (e.g., “inconceivable”).
19. **Pronouns:** Count of pronouns in the text, determined by tokens with the part-of-speech PRON (e.g., “him”, “she”).
20. **Nouns:** Count of nouns in the text, determined by tokens with the part-of-speech NOUN (e.g., “book”, “concept”).

B.2 Percentile-Based Reference Thresholds

We derived these thresholds from the training corpus of 6,754 Plain Language Summaries. The labeling system adapts to metric directionality as described in Section 3.2.2: for lower-is-better metrics (\downarrow), the labels correspond to actual percentiles; for

higher-is-better metrics (\uparrow), the same labels represent inverse percentiles. This ensures P25 and P75 consistently identify what we could consider the “best” quartile across all metrics (having in mind that not being in this quartile does not necessarily mean that the text is bad/not plain).

B.3 Example PLS Evaluation Tool Output

The PLS Evaluation Tool generates structured text output (both JSON and human-readable format) that is directly provided to the LLM agents as input. Table B2 illustrates the tool’s analysis of an abstract-PLS pair from the same publication. The original abstract deviates significantly from typical PLS patterns, while its professionally written PLS achieves better conformity. For each metric deviating from typical patterns (P90 or beyond), the tool automatically generates specific feedback suggesting reduction to median values. The actual tool produces formatted text, but we present it here in tabular form for clarity.

C Medical Glossary Sources

We compiled medical glossaries from eleven authoritative sources to support plain language translation (Table 2 in the Methodology section presents the source distribution and term counts). Most of these resources are compiled in the University of Maryland’s Williams (2025), which provides comprehensive plain language resources for health communication. Each source provides specialized terminology translations for different healthcare domains:

- **NCI-C (National Cancer Institute, 2025a):** National Cancer Institute’s comprehensive cancer dictionary covering types, treatments, procedures, and side effects for patient education.
- **NCI-D (National Cancer Institute, 2025a):** National Cancer Institute’s drug database with chemotherapy agents, targeted therapies, and immunotherapy medications.
- **CDC-T (Centers for Disease Control and Prevention, 2011):** CDC’s thesaurus providing plain language alternatives for epidemiological and public health terminology.
- **ADA-D (American Diabetes Association, 2024):** American Diabetes Association’s glossary covering diabetes types, management, complications, and monitoring terms.

READABILITY INDICES				
Feature	P25/P75*	P50	P75/P25*	P90/P10*
FRE ↑	≥ 48.17	≥ 40.48	≥ 32.68	< 25.27
FKGL ↓	≤ 11.77	≤ 13.16	≤ 14.59	> 16.05
GFI ↓	≤ 16.15	≤ 17.79	≤ 19.39	> 21.05
SMOG ↓	≤ 10.68	≤ 12.11	≤ 13.58	> 14.87
DCRS ↓	≤ 7.19	≤ 7.65	≤ 8.16	> 8.63
CLI ↓	≤ 11.36	≤ 12.66	≤ 13.96	> 15.16
ARI ↓	≤ 12.02	≤ 13.60	≤ 15.29	> 16.95
LIX ↓	≤ 50.25	≤ 54.32	≤ 58.40	> 62.38
RIX ↓	≤ 6.04	≤ 7.04	≤ 8.14	> 9.36
STRUCTURAL COMPLEXITY				
Words/Sent. ↓	≤ 19.81	≤ 22.13	≤ 24.76	> 27.60
Passive Voice ↓	≤ 9	≤ 13	≤ 18	> 23
Active Voice ↑	≥ 41	≥ 29	≥ 20	< 14
Nominalization ↓	≤ 8	≤ 13	≤ 20	> 27
VOCABULARY COMPLEXITY				
Complex Words (DC) ↓	≤ 115	≤ 160	≤ 213	> 277
Long Words ↓	≤ 88	≤ 122	≤ 164	> 208
Complex Words ↓	≤ 60	≤ 84	≤ 115	> 145
CONTENT DENSITY				
Pronouns ↑	≥ 21	≥ 13	≥ 8	< 5
Nouns ↓	≤ 83	≤ 116	≤ 157	> 202

Table B1: Percentile-based reference ranges for 18 linguistic features derived from 6,754 PLS texts. Column headers show actual percentiles for lower-is-better metrics (↓) and with asterisk (*) for higher-is-better metrics (↑). For example, P25/P75* means 25th percentile for ↓ metrics and 75th percentile for ↑ metrics, both representing the best quartile.

- **NCI-G** (National Cancer Institute, 2025b): National Cancer Institute’s genetics dictionary explaining hereditary conditions, genetic testing, and molecular biology concepts.
- **UIowa** (Human Subjects Office, University of Iowa, 2021): University of Iowa’s general medical term translations designed for informed consent documents and patient communication.
- **MRCT** (The Multi-Regional Clinical Trials Center of Brigham and Women’s Hospital and Harvard, 2025): Multi-Regional Clinical Trials Center’s glossary for clinical research protocols, trial phases, and regulatory terminology.
- **WA-PH** (Washington State Department of Health, 2023): Washington State’s glossary focused on vaccine types, immunization schedules, and disease prevention terminology.
- **WebMD-A** (WebMD, 2022): WebMD’s asthma-specific dictionary covering triggers, medications, devices, and respiratory symptoms.
- **CCIO** (Center for Consumer Information & Insurance Oversight, 2024): Health insurance glossary explaining coverage terms, benefits, deductibles, and healthcare plan types.
- **Cochrane** (Pitcher et al., 2022): Cochrane’s guideline for writing systematic review summaries with standardized plain language templates.

D Complete Linguistic Feature Analysis

Table D1 presents a comprehensive analysis of all 20 linguistic features across all tested models, organized into three categories: Structure (word/sentence metrics), Style (voice and pronoun usage), and Readability & Complexity (traditional readability indices and vocabulary measures).

(a) Original abstract: 1019 words				(b) Corresponding PLS: 672 words			
Metric	Value	Target	Rating	Metric	Value	Target	Rating
<i>Within typical ranges</i>				<i>Within typical ranges</i>			
FRE	38.55	> 40.48	P25	Words/Sent.	19.76	< 22.13	P25
Active Voice	83	> 29	P75	FKGL	11.68	< 13.16	P25
Pronouns	30	> 13	P75	GFI	15.47	< 17.79	P25
CLI	11.95	< 12.66	P50	Active Voice	52	> 29	P75
				Pronouns	36	> 13	P75
				CLI	12.75	< 13.96	P75
				FRE	46.53	> 40.48	P50
				ARI	12.52	< 13.60	P50
				LIX	52.06	< 54.32	P50
				RIX	6.38	< 7.04	P50
				DCRS	7.30	< 7.65	P50
<i>Deviating from typical patterns</i>				<i>Deviating from typical patterns</i>			
FKGL	15.11	< 13.16	P90	Nominalization	22	< 13	P90
ARI	16.17	< 13.60	P90	Nouns	198	< 116	P90
Words/Sent.	29.11	< 22.13	Beyond P90	SMOG	14.18	< 12.11	P90
Passive Voice	42	< 13	Beyond P90	Complex Words (DC)	269	< 160	P90
Nominalization	44	< 13	Beyond P90	Complex Words	127	< 84	P90
Nouns	326	< 116	Beyond P90	Passive Voice	31	< 13	Beyond P90
GFI	22.72	< 17.79	Beyond P90	Long Words	217	< 122	Beyond P90
LIX	64.15	< 54.32	Beyond P90				
RIX	10.20	< 7.04	Beyond P90				
SMOG	18.97	< 12.11	Beyond P90				
DCRS	9.30	< 7.65	Beyond P90				
Complex Words (DC)	507	< 160	Beyond P90				
Complex Words	282	< 84	Beyond P90				
Long Words	357	< 122	Beyond P90				

Table B2: Example of PLS Evaluation Tool output comparing an abstract-PLS pair. The original abstract (a) exceeds the word limit and shows poor conformity with 14 of 18 metrics deviating from typical patterns. Its corresponding PLS (b) meets the word limit and achieves moderate conformity with only 7 metrics deviating. The tool automatically generates feedback suggesting median target values for all metrics at P90 or beyond.

Model	Approach	Structure			Lexical Features							
		Words	Sent.	WPS	Pass.	Act.	Pron.	Nom.	Nouns	CW-DC	CW	LW
	Original Abstract	868	33.82	26.04	24.56	49.85	20.42	43.35	278.62	411.69	215.47	285.83
	Reference PLS (human)	655	33.08	20.08	19.64	57.95	35.01	25.02	187.50	263.36	133.12	197.00
Gemini 2.5 Pro	<i>Baseline</i>	661	41.64	16.05	14.75	66.43	45.52	19.05	171.80	240.15	98.59	166.07
	<i>Agentic (V1)</i>	591	36.10	16.53	11.49	63.20	40.19	17.68	159.35	208.00	86.35	149.57
	<i>Agentic (V2)</i>	567	32.73	17.48	12.05	59.57	35.07	17.92	153.56	207.12	92.02	152.80
GPT-5	<i>Baseline</i>	866	48.69	17.90	18.59	90.29	46.74	22.09	249.45	321.40	136.71	235.15
	<i>Agentic (V1)</i>	879	45.60	19.49	21.67	92.90	44.73	25.89	278.74	375.02	160.88	253.35
GPT-OSS-120B	<i>Baseline</i>	623	32.47	19.34	14.17	63.93	35.52	19.92	160.30	229.30	109.29	165.78
	<i>Agentic (V1)</i>	743	37.70	19.80	18.73	75.59	39.10	25.68	218.46	330.97	160.50	221.94
	<i>Agentic (V2)</i>	623	31.62	19.80	14.37	64.15	33.24	21.29	177.80	266.55	130.59	182.99
GPT-OSS-20B	<i>Baseline</i>	541	27.52	19.31	12.18	57.55	32.31	16.08	141.26	193.89	87.96	141.78
	<i>Agentic (V2)</i>	632	33.41	18.88	15.47	65.61	32.76	20.92	179.55	275.15	131.01	183.65
Llama 3.3 70B	<i>Baseline</i>	477	21.65	22.23	9.51	44.43	29.26	19.56	130.86	173.90	95.13	141.06
	<i>Agentic (V2)</i>	497	25.26	19.85	10.89	46.97	23.49	19.93	148.66	207.16	112.08	165.90
	<i>Agentic (V2 + Gemini 2.5 Flash)</i>	514	29.37	17.62	9.57	54.74	32.64	17.86	143.63	194.28	92.17	148.95
Llama 3.2 3B	<i>Baseline</i>	477	31.40	15.28	11.94	38.27	25.36	19.09	127.76	185.00	98.34	145.72
	<i>Agentic (V2)</i>	536	27.83	19.79	15.73	40.84	17.42	25.17	162.41	239.07	130.42	187.85

Table D1: Linguistic feature analysis (Structure and Lexical Features) across all tested models and approaches. Gray shading highlights original abstract and human reference baseline. Abbreviations: Words Per Sentence (WPS), Passive voice (Pass.), Active voice (Act.), Pronouns (Pron.), Nominalization (Nom.), Complex Words Dale-Chall (CW-DC), Complex Words (CW), Long Words (LW). Averages computed across 100 test samples.

Efficient On-Device Text Simplification for Firefox with Synthetic Data Fine-Tuning

Pablo Romero¹, Zihao Li², Matthew Shardlow¹,

¹Manchester Metropolitan University, ²University of Manchester,

Correspondence: pablo2004romero@gmail.com, jeremy.li@manchester.ac.uk, m.shardlow@mmu.com

Abstract

This work presents a system for on-device text simplification that enables users to process sensitive text without relying on cloud-based services. Through the use of quantization techniques and a novel approach to controllable text simplification, we reduce model size by up to 75% with minimal performance degradation. Our models demonstrate efficient state-of-the-art results using a synthetic dataset of 2,909 examples, outperforming prior work trained on 300K examples. This efficiency stems from: (1) a single control token strategy that precisely targets specific reading levels, (2) a multi-level training approach that exposes models to transformations from multiple source complexity levels, and (3) individual models that dedicate full parameter capacity to specific reading level transformations. Our best models achieve up to 82.18 BLEU (at the Advanced level) and 46.12 SARI (at the Elementary level) on standard benchmarks, with performance preserved even after aggressive quantization. This work is implemented as a collaboration with the Mozilla AI team to process text entirely locally, ensuring sensitive information never leaves the user’s device. We have a demonstration video¹ and a web demo available at: <https://pablrom2004.github.io/Simplification-Web-Demo/>

1 Introduction

Text simplification aims to reduce textual complexity while preserving essential meaning, thereby improving accessibility for a broad range of readers (Alva-Manchego et al., 2021). Today, sequence-to-sequence neural models provide state-of-the-art results, but many existing solutions require server-side processing, raising concerns about data privacy and latency when processing sensitive content like medical information or legal documents.

Our work addresses these issues by introducing an on-device text simplification system with two core innovations. First, we implement a single control token strategy instead of relying on multiple complexity metrics. This approach departs from previous work (Li et al., 2022) that employed multiple tokens to represent various linguistic features.

Second, we create a high-quality synthetic dataset comprising only 2,909 examples generated by carefully prompting large language models. Despite its relatively small size, this dataset outperforms the 300K-example WikiLarge corpus (Zhang and Lapata, 2017) on standard benchmarks, highlighting that data quality can far outweigh quantity for efficient model training.

We deploy our models locally in the browser via *transformers.js*², ensuring that all data processing happens directly on the user’s device. Experiments on standard benchmarks show that our models rival or outperform previous approaches, and maintain their quality even when quantized for efficient on-device operation.

2 Related Work

Text simplification has progressed from rule-based approaches (Elhadad and Sutaria, 2007; Yatskar et al., 2010; Biran et al., 2011) to data-driven methods leveraging parallel corpora (Surya et al., 2019; Martin et al., 2020; Omelianchuk et al., 2021; Martin et al., 2022). For comprehensive overviews of recent developments, see (North et al., 2025). Pre-trained language models like BART (Lewis et al., 2019) have enabled more fluent and faithful simplifications.

Control tokens for controllable text simplification were first introduced by (Scarton and Specia, 2018), with (Martin et al., 2020) and (Spring et al., 2021) further expanding on controllable simplification. (Li et al., 2022) explored different control

¹<https://youtu.be/TzmaxnARMzg>

²<https://huggingface.co/docs/transformers.js>

token configurations and found that tokens representing distinct aspects of complexity (e.g., dependency tree depth, word rank, and length ratio) could effectively guide generation. However, their approach used multiple tokens simultaneously, potentially creating competition for model attention and parameter space. Our work builds on this line of research by using a single, level-based control token.

Synthetic data generation using large language models has emerged as a promising direction for low-resource NLP tasks (Wang et al., 2023). Recent work has demonstrated that high-quality synthetic data can match or exceed naturally collected data for various applications (Yang et al., 2023). Our work emphasizes quality and diversity over quantity, showing that careful prompt engineering can produce highly effective training examples.

On-device NLP has gained traction with growing privacy concerns and the need for offline capability. Mozilla’s Firefox Translations project (Mozilla NLP Team, 2023) pioneered browser-based machine translation using ONNX format (Foundation, 2017) (Open Neural Network Exchange, an open standard for machine learning deployment) and WebAssembly. Model compression techniques like distillation and quantization (Jain, 2022) have been essential for deploying models in resource-constrained environments. Our work extends these approaches to text simplification, demonstrating successful deployment with minimal performance loss.

3 Model

3.1 Synthetic Dataset Creation

We created a synthetic dataset of 2,909 examples, each with three levels of simplification: Elementary, Secondary, and Advanced. The dataset generation involved a two-stage approach:

First, we developed a detailed prompt³ using Claude 3.5 Sonnet (Anthropic, 2024), specifically describing task requirements, formatting, and quality expectations. This meta-prompting approach, asking one language model to create prompts for another, represents a valuable technique for data creation. Our prompt included specific instructions for creating sentences at different reading levels, with clear definitions for each simplification level:

³<https://github.com/pabloRom2004/Simple-Synthetic-Dataset/blob/main/Prompt1.txt>

- **ELEMENTARY**: Uses very simple words and straightforward structure. Suitable for 5th-7th grade.
- **SECONDARY**: Simple vocabulary. Suitable for 8th-12th grade.
- **ADVANCED**: Keep as one sentence but slightly simpler than complex.

The resulting prompt was then used with OpenAI’s o1 model (OpenAI, 2024) (a large multi-modal model with strong reasoning capabilities) to generate 2,284 synthetic examples (an example is the complex sentence and the three levels of simplification). We chose the o1 model due to its large context window and reasoning capabilities, which helped ensure high-quality, diverse examples. The prompt instructed the model to generate content across various topics (news, technology, health, education) that Firefox users might encounter during web browsing. We enhance model training by including multiple examples (e.g., training the Elementary model on Complex → Elementary, Advanced → Elementary, and Secondary → Elementary inputs).

To ground our dataset in established benchmarks, we supplemented the synthetic data with 625 examples derived from the WikiLarge training set, where we extracted complex sentences and used our o1 prompt⁴ to generate three simplified versions. This hybrid approach balances novel generation with grounding in established datasets. The dataset can be found here: <https://github.com/pabloRom2004/Simple-Synthetic-Dataset>

3.2 Control Token vs. Separate Models Approaches

We explored two distinct approaches to modeling reading level control:

3.2.1 Control Token Model

For the Control Token Model approach, we trained a single BART-base model with a prepended control token (e.g., <LEVEL_ELEMENTARY>, <LEVEL_SECONDARY>, or <LEVEL_ADVANCED>) indicating the desired reading level. Unlike previous approaches that used multiple control tokens for different aspects of simplification (Li et al., 2022), our approach uses a single token for the entire simplification level.

⁴<https://github.com/pabloRom2004/Simple-Synthetic-Dataset/blob/main/Prompt2WikiLarge.txt>

This single token strategy simplifies the control mechanism and allows the model to learn a more direct mapping between the token and the desired output style. For example, rather than specifying multiple complexity dimensions such as *dependency tree depth*, *word rank*, *levenshtein distance* and *length ratio*, we simply use `<LEVEL_ELEMENTARY>` to indicate the overall desired simplification level, in line with the main takeaway from the bitter lesson⁵, a principle in AI research suggesting that general methods leveraging computation ultimately outperform human-engineered approaches.

3.2.2 Individual Models

For the Individual Models approach, we trained three separate models, each dedicated to one target reading level. This approach exposes models to transformations from multiple source complexity levels during training. The Elementary model is trained using Complex \rightarrow Elementary, Secondary \rightarrow Elementary, and Advanced \rightarrow Elementary transformation pairs. Similarly, the Secondary model incorporates Complex \rightarrow Secondary and Advanced \rightarrow Secondary examples, while the Advanced model focuses only on Complex \rightarrow Advanced transformation pairs. This multi-level training strategy provides each model with a richer set of transformation patterns, helping it learn more robust simplification strategies by observing how text at different complexity levels can be transformed to its target level.

From a theoretical perspective, individual models allocate the entire parameter space to learning one specific level of transformation, avoiding the parameter competition that might occur in a single model trying to learn multiple transformation levels simultaneously. This parameter efficiency becomes particularly important in smaller models like BART-base, where the capacity to represent multiple complex transformations may be limited.

3.3 Training Configuration

We used a BART-base sequence-to-sequence model (Lewis et al., 2019) as our foundation, with the following hyperparameters: 8 epochs, batch size of 8, learning rate of $1e-4$ with Adam optimizer, weight decay of 0.01, and 10% warmup steps. The dataset was split using a 99/0.8/0.2 ratio for train/test/validation. This high train percentage was chosen because we evaluated on the standard ASSET Test Set as our benchmark rather than using

⁵<http://www.incompleteideas.net/IncIdeas/BitterLesson.html>

our own test set, making it possible to allocate more data for training.

Because of the relatively small dataset size, training on an NVIDIA 4080 Super took only approximately 30 minutes, highlighting the computational efficiency of our approach. This rapid training time stands in stark contrast to models trained on large datasets like WikiLarge, which can take many hours to train. The efficiency gain comes from both the smaller dataset size and the clear learning signal provided by our high-quality synthetic examples.

3.4 Deployment via *transformers.js*

We deploy our models on-device using HuggingFace’s *transformers.js* framework directly adhering to the Firefox documentation⁶, which allows running models hosted on HuggingFace⁷ with the ONNX format directly in web browsers. Our deployment pipeline involves two key steps:

We convert our trained models to ONNX format for compatibility with a script provided by *transformers.js*. This conversion is essential for allowing the models to run in standard web browsers without specialized hardware acceleration. In this same script, quantization is applied to reduce model size and memory footprint. We experimented with both INT8 and 4-bit quantization, which reduced file sizes by up to 75% compared to the full-precision FP-32 model with minimal performance degradation.

The resulting system stores models in the browser’s local storage, enabling persistent availability across sessions without repeated downloads. Once downloaded, models can be used completely offline, ensuring privacy and reliability even in disconnected environments.

4 Results

4.1 Evaluation Setting

We evaluate on two widely-used text simplification benchmarks: The **ASSET Test Set** (Alva-Manchego et al., 2020) and **TurkCorpus** (Xu et al., 2016), these two benchmarks feature human made simplifications of 359 English sentences from Wikipedia, focusing on fluency, meaning preservation, and simplicity. As a relatively recent benchmark, it offers multiple reference simplifications

⁶<https://firefox-source-docs.mozilla.org/toolkit/components/ml/>

⁷<https://huggingface.co/collections/pabRomero/firefox-simplification-67d70f0d3dcb47939026303f>

per sentence, created through crowdsourcing with detailed guidelines.

For evaluation, we employ multiple complementary metrics:

- **BLEU** (Papineni et al., 2002): Measures n-gram overlap with reference texts, capturing fluency and preservation of meaning
- **SARI** (Xu et al., 2016): Focuses on evaluating Add, Delete, and Keep operations compared to references, specifically designed for text simplification evaluation
- **BERTScore** (Zhang et al., 2020): Captures semantic similarity using contextual embeddings, offering a more nuanced measure of meaning preservation
- **LENS** (Maddela et al., 2023) a learnt evaluation metric for simplicity assessment
- **SALSA** (Heineman et al., 2023): An edit-level simplification evaluation metric

4.2 Distillation and Quantization

Table 1 shows that quantized models (INT8 or BNB-4) lose minimal performance compared to FP-32 while reducing size by up to 75%. Notably, the INT8 model (136MB) achieves the highest readability scores (LENS, SALSA) despite its smaller size, while BNB-4 outperforms on semantic preservation metrics (BERT-P/R/F1). These results suggest quantization may act as beneficial regularization for certain simplification aspects, making these compressed models ideal for resource-constrained environments without sacrificing quality. The quantization of the models was performed using the framework provided by transformers.js.

4.3 Model Performance on ASSET Test Set

Table 2 compares our approaches with a baseline BART model (Li et al., 2022) trained on WikiLarge.

Both our control token model and individual models outperform the baseline trained on WikiLarge across all metrics, despite our synthetic dataset containing 100 times fewer examples (300,000 vs. 2,909). This striking result challenges conventional wisdom, suggesting that a small set of high-quality synthetic examples can be more effective than a large corpus of lower-quality or less focused examples. The individually trained models

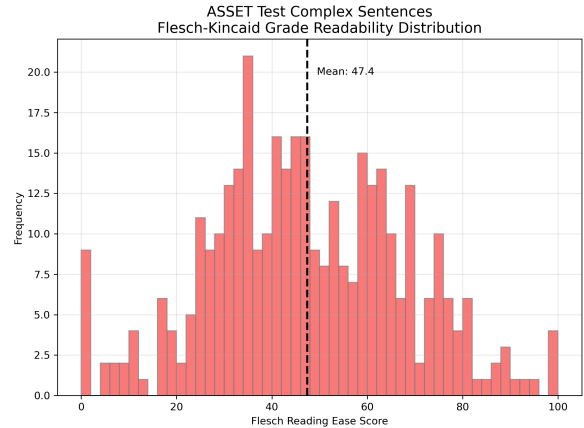


Figure 1: Distribution of Flesch-Kincaid readability scores for original ASSET Test Set sentences. Mean score is around 47.4, reflecting moderate complexity.

outperform the control token model and the baseline, showing the efficacy of specialization into a single, well-defined task for these small language models like BART.

4.4 Readability Analysis

Figure 1 shows the distribution of Flesch-Kincaid readability scores for the original ASSET Test Set sentences. These scores range from 0-100, with higher scores indicating easier readability. Figure 2 shows the distribution of Flesch-Kincaid readability scores of the individual models after simplification.

Figure 3 visualizes how our Individual Models transform text readability according to the Flesch-Kincaid score. Each point represents a sentence, with readability score of the original text on the x-axis and the readability of the simplified text on the y-axis. Points above the diagonal line ($y=x$) indicate simplification; points below it show increased complexity.

This visualization confirms that our three models effectively target distinct reading levels, with clear separation in their simplification behaviours. The Advanced level model stays around the $y=x$ line, which means that the model is generally just re-writing the sentences in a similar level of complexity just with different structure, potentially helping a user understand the sentence once it is re-worded. The Elementary level model shows clear simplification from input to output, nearly all of the examples show clear improvements in their Flesch-Kincaid scores which shows great simplification ability from the model. While we acknowledge that Flesch-Kincaid and other automatic readabil-

Dataset	Model	BLEU	SARI	BERT-P	BERT-R	BERT-F1	LENS	SALSA	Size
ASSET	FP-32	51.17	42.07	0.682	0.667	0.657	58.80	68.88	540MB
	INT8	49.96	42.06	0.677	0.668	0.657	59.19	69.86	136MB
	BNB-4	51.07	42.12	0.688	0.674	0.665	58.66	68.86	212MB

Table 1: Performance comparison of BART model variants with different quantization levels on text simplification tasks.

Model	Level	BLEU	SARI	BERT-P	BERT-R	BERT-F1	LENS	SALSA
Baseline	–	51.17	42.07	0.682	0.667	0.657	58.80	68.88
Control	Elem	57.25	43.21	0.728	0.701	0.701	68.97	75.42
Control	Sec	62.98	41.76	0.766	0.750	0.745	61.21	69.83
Control	Adv	52.96	40.32	0.680	0.699	0.678	50.49	65.03
Indiv.	Elem	58.41	46.12	0.754	0.747	0.737	71.39	77.54
Indiv.	Sec	72.81	42.19	0.828	0.828	0.817	62.46	68.91
Indiv.	Adv	82.18	35.82	0.870	0.877	0.866	59.42	65.32

Table 2: Model performance on the ASSET Test Set. Individual Models outperform both the baseline and the Control Token model.

ity metrics have known limitations in evaluating text simplification quality (Alva-Manchego et al., 2021), we present these scores as exploratory indicators of relative complexity changes across our models rather than definitive measures of simplification success.

4.5 Example Simplifications

Table 4 in the Appendix presents selected examples from our models, highlighting successes and challenges across different reading levels.

These examples highlight both the strengths of our models (effective simplification at appropriate levels) and areas for improvement (maintaining factual accuracy and avoiding unnecessary transformations).

5 Prototype Implementation

We developed a web-based prototype that demonstrates our text simplification models operating directly in the browser. The implementation uses *transformers.js*, for all processing locally, ensuring privacy by keeping sensitive text on the user’s device. The source code is available at <https://github.com/pabloRom2004/Simplification-Web-Demo>.

5.1 Technical Architecture

The prototype follows a fully client-side architecture, operating entirely within the browser without

server-side processing. The core components include:

- **Model Management:** Handles downloading, storing, and loading of quantized ONNX models
- **Text Processing:** Implements sentence splitting for input paragraphs, tokenization, and recombination
- **Inference Pipeline:** Configures and executes the simplification models
- **Readability Analysis:** Calculates Flesch-Kincaid scores for original and simplified text
- **Visualization:** Provides interactive display of simplification results with sentence mapping

When a user first visits the application, they select a quantization level (FP32, INT8, or BNB-4) based on their device capabilities and memory constraints. They can then download one or more models at their chosen reading levels. Once downloaded, models persist in the browser’s local storage, eliminating the need for re-downloading in future sessions.

For longer texts, we implement a sentence-splitting algorithm that identifies sentence boundaries while accounting for common abbreviations

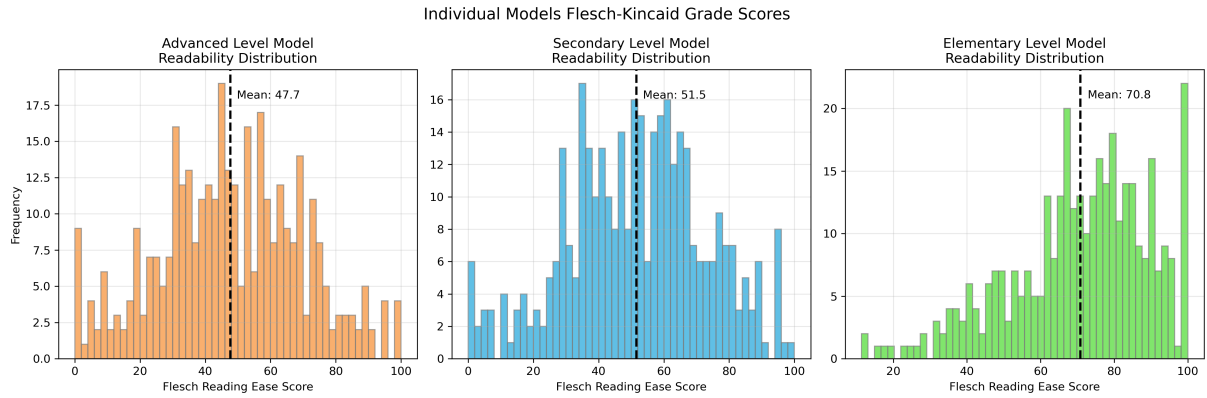


Figure 2: Distribution of Flesch-Kincaid readability scores for individual models on the ASSET Test Set sentences.

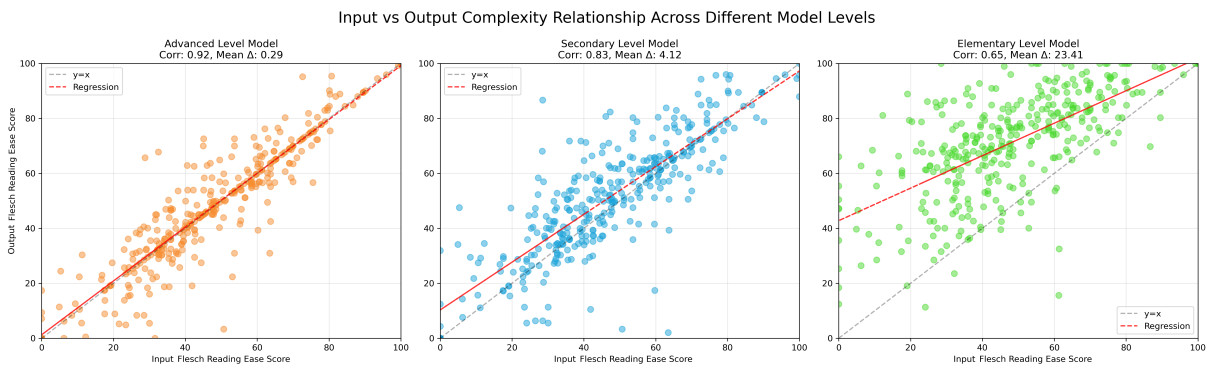


Figure 3: Scatter plot of original (x-axis) vs. simplified (y-axis) readability for the Individual Models. Points above the diagonal line represent simplification, while points below it indicate increased complexity.

and special cases. Each sentence is processed independently before being recombined, allowing efficient handling of paragraphs without exceeding browser memory constraints.

6 Discussion

Our experiments reveal several key insights with broader implications. First, our small but high-quality synthetic dataset (2,909 examples) outperforms the WikiLarge corpus (300K examples), challenging conventional wisdom about data requirements for fine-tuning. This finding suggests that pre-trained models already possess substantial linguistic knowledge and primarily need clear, unambiguous direction rather than extensive examples. Our synthetic data succeeds by precisely pointing the model toward the desired task through concise, well-crafted examples that demonstrate the exact transformation patterns required. The clarity and quality of this directional signal prove significantly more valuable than quantity, indicating that carefully engineered LLM prompts can create highly effective training data for a wide range of downstream NLP tasks.

Second, our single control token strategy demonstrates that simplicity can outperform complexity in control mechanisms. By using a single token that directly indicates the target reading level rather than multiple tokens representing different complexity features, we reduce potential parameter competition and make training more efficient for smaller models.

Third, our finding that individually trained models outperform the control token approach highlights the importance of parameter efficiency in smaller architectures. By dedicating the entire parameter space to learning one specific transformation, these models develop more robust simplification strategies for their target reading levels.

Finally, the impressive stability of performance across quantization levels (with size reductions up to 75%) indicates that many NLP tasks may not require full floating-point precision. The INT8 model’s superior performance on readability metrics despite its smaller size suggests that quantization may actually function as beneficial regularization for certain aspects of text simplification.

While effective for sentence-level simplification,

we found extending to longer contexts or domain-specific text challenging for our BART-base models, suggesting larger architectures may be needed for these scenarios.

7 Conclusion

This work presents an on-device text simplification approach using synthetic data and model quantization that processes text locally in Firefox browsers. Our contributions include: (1) demonstrating synthetic LLM-generated data can outperform much larger human-annotated datasets, (2) showing specialized models outperform control token approaches for smaller architectures, and (3) providing a privacy-preserving implementation with state-of-the-art quantization techniques that together enable efficient and private language technologies.

Limitations

Our approach has several limitations. Our reliance on synthetic data, while effective, may miss certain nuances of human-authored simplifications. Our current implementation emphasizes sentence-level simplification rather than document-level coherence, potentially creating local optimizations that do not maintain global coherence in longer texts. We have also not conducted extensive human evaluation, which would be valuable for assessing subjective aspects of simplification quality that automatic metrics may not capture.

Future work will incorporate structured human evaluation with university students to validate our findings beyond automatic metrics, and explore extending the approach to longer contexts and additional languages.

Lay Summary

Reading complex text online can sometimes be too hard to read for some users, especially when encountering technical articles, legal documents, or medical information. This work presents a system that simplifies difficult text directly in your web browser without sending your data to external servers, protecting your privacy.

Our approach makes two key innovations. First, instead of training our models on hundreds of thousands of examples like previous work, we created just 2,909 high-quality examples by carefully prompting advanced AI systems to generate simplified versions of sentences at three reading levels:

Elementary (suitable for middle school), Secondary (suitable for high school), and Advanced (slightly simplified but maintaining sophistication). Surprisingly, this small, carefully crafted dataset outperformed much larger datasets, demonstrating that quality matters more than quantity.

Second, we made these models small enough to run in a web browser by compressing them to 25% of their original size while maintaining performance. This means users can simplify sensitive text like medical records or legal documents without that information ever leaving their device.

The system offers three simplification levels, allowing users to choose how much simplification they need. For example, a medical article about "dextromethorphan occurring as a white powder in its pure form" might become "Dextromethorphan is a white powder" at the Elementary level, while maintaining more detail at higher levels.

We built this as a working web demo and collaborated with Mozilla's Firefox team to integrate it into the browser. The models work entirely offline once downloaded, making simplified reading accessible even without an internet connection. This work shows that privacy-preserving, accessible language technology can be both practical and powerful.

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, Carolina Scarton, and Lucia Specia. 2021. *The (un)suitability of automatic evaluation metrics for text simplification*. *Computational Linguistics*, 47(4):861–889.
- Anthropic. 2024. *Introducing claude 3.5 sonnet*. Accessed on March 14, 2025.
- Or Biran, Samuel Brody, and Noémie Elhadad. 2011. *Putting it simply: a context-aware approach to lexical simplification*. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 496–501.
- Noemie Elhadad and Komal Sutaria. 2007. *Mining a lexicon of technical terms and lay equivalents*. In *Biological, translational, and clinical language processing*, pages 49–56.

- Linux Foundation. 2017. ONNX | onnx.ai. <https://onnx.ai/>. [Accessed 14-03-2025].
- David Heineman, Yao Dou, Mounica Maddela, and Wei Xu. 2023. **Dancing between success and failure: Edit-level simplification evaluation using SALSA**. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3466–3495, Singapore. Association for Computational Linguistics.
- Shashank Mohan Jain. 2022. Hugging face. In *Introduction to transformers for NLP: With the hugging face library and models to solve problems*, pages 51–67. Springer.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.
- Zihao Li, Matthew Shardlow, and Saeed Hassan. 2022. **An investigation into the effect of control tokens on text simplification**. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 154–165, Abu Dhabi, United Arab Emirates (Virtual). Association for Computational Linguistics.
- Mounica Maddela, Yao Dou, David Heineman, and Wei Xu. 2023. **LENS: A learnable evaluation metric for text simplification**. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16383–16408, Toronto, Canada. Association for Computational Linguistics.
- Louis Martin, Éric de la Clergerie, Benoît Sagot, and Antoine Bordes. 2020. **Controllable sentence simplification**. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4689–4698, Marseille, France. European Language Resources Association.
- Louis Martin, Angela Fan, Éric de la Clergerie, Antoine Bordes, and Benoît Sagot. 2022. **MUSS: Multilingual unsupervised sentence simplification by mining paraphrases**. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1651–1664, Marseille, France. European Language Resources Association.
- Mozilla NLP Team. 2023. **Mozilla translations: Open-source neural translation in the browser**. In *Proceedings of Machine Translation Summit*.
- Kai North, Tharindu Ranasinghe, Matthew Shardlow, and Marcos Zampieri. 2025. **Deep learning approaches to lexical simplification: A survey**. *Journal of Intelligent Information Systems*, 63:111–134.
- Kostiantyn Omelianchuk, Vipul Raheja, and Oleksandr Skurzhanyskiy. 2021. **Text Simplification by Tagging**. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 11–25, Online. Association for Computational Linguistics.
- OpenAI. 2024. **Learning to reason with LLMs**. <https://openai.com/index/learning-to-reason-with-llms/>. [Accessed 14-03-2025].
- 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.
- Carolina Scarton and Lucia Specia. 2018. **Learning simplifications for specific target audiences**. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 712–718, Melbourne, Australia. Association for Computational Linguistics.
- Nicolas Spring, Annette Rios, and Sarah Ebling. 2021. **Exploring German multi-level text simplification**. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1339–1349, Held Online. INCOMA Ltd.
- Sai Surya, Abhijit Mishra, Anirban Laha, Parag Jain, and Karthik Sankaranarayanan. 2019. **Unsupervised neural text simplification**. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2058–2068, Florence, Italy. Association for Computational Linguistics.
- Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khoshabi, and Hannaneh Hajishirzi. 2023. **Self-instruct: Aligning language models with self-generated instructions**. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*.
- Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. **Optimizing statistical machine translation for text simplification**. *Transactions of the Association for Computational Linguistics*, 4:401–415.
- Rui Yang, Haotian Lin, Cheng Wang, and Hao Qian. 2023. **Gpt4tools: Teaching large language models to use tools via self-instruction**. In *Proceedings of the 40th International Conference on Machine Learning*.
- Mark Yatskar, Bo Pang, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2010. **For the sake of simplicity: Unsupervised extraction of lexical simplifications from wikipedia**. *arXiv preprint arXiv:1008.1986*.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. **Bertscore: Evaluating text generation with bert**. *Preprint*, arXiv:1904.09675.

Xingxing Zhang and Mirella Lapata. 2017. Sentence simplification with deep reinforcement learning. *arXiv preprint arXiv:1703.10931*.

A Additional Results and Examples

Model	Level	BLEU	SARI	BERT-P	BERT-R	BERT-F1	LENS	SALSA
Baseline	–	46.94	36.87	0.678	0.613	0.636	58.65	68.88
Control	Elem	51.25	37.63	0.720	0.652	0.679	68.67	75.42
Control	Sec	59.87	38.28	0.778	0.726	0.745	61.62	69.83
Control	Adv	52.75	36.76	0.693	0.682	0.681	51.22	65.03
Indiv.	Elem	50.60	39.48	0.733	0.693	0.704	70.89	77.54
Indiv.	Sec	71.45	40.37	0.846	0.814	0.825	63.00	68.91
Indiv.	Adv	84.92	37.63	0.912	0.894	0.900	61.10	65.32

Table 3: Model performance on TurkCorpus.

B Dataset Creation Prompts

The prompts used to generate our synthetic dataset (including the WikiLarge-based simplifications) are available at: <https://github.com/pabloRom2004/Simple-Synthetic-Dataset>

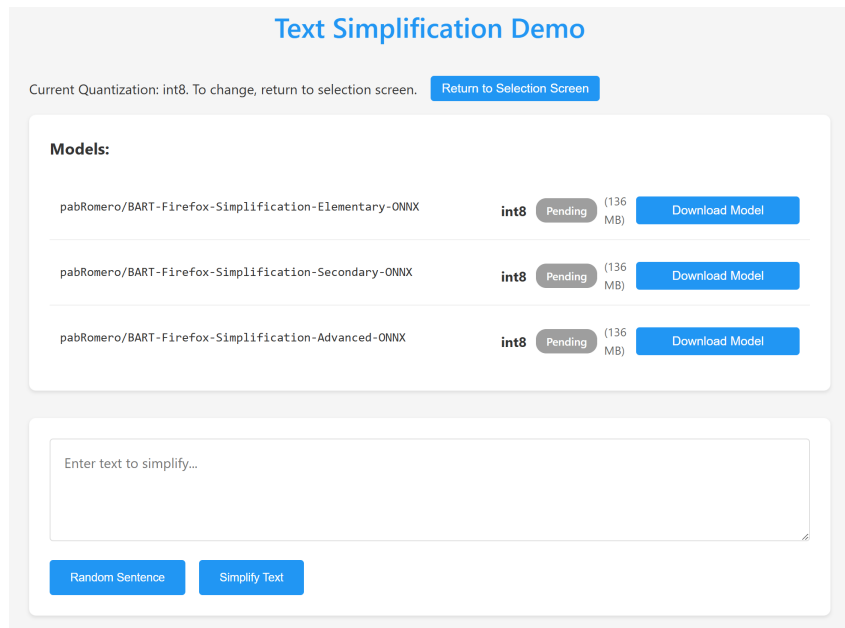


Figure 4: Screenshot of the web-based prototype interface showing model selection, text input, and results display with color-coded sentence mapping.

Type	Example
<i>Elementary (Success)</i>	
Original	The SAT Reasoning Test (formerly Scholastic Aptitude Test and Scholastic Assessment Test) is a standardized test for college admissions in the United States.
Simplified	The SAT is a test that students can take to get a good grade. It's used to help them find colleges.
<i>Elementary (Failure)</i>	
Original	In its pure form, dextromethorphan occurs as a white powder.
Simplified	Dextromethorphan is a white powder.
<i>Secondary (Success)</i>	
Original	Below are some useful links to facilitate your involvement.
Simplified	Here are some helpful resources to help you get involved.
<i>Secondary (Failure)</i>	
Original	Protests across the nation were suppressed.
Simplified	Demonstrations across the nation were peaceful.
<i>Advanced (Success)</i>	
Original	Minor planet designations are number-name combinations overseen by the Minor Planet Center, a branch of the IAU.
Simplified	Minor planets are classified by the Minor Planet Center, a branch of the IAU.
<i>Advanced (Failure)</i>	
Original	About 95 species are currently accepted.
Simplified	Approximately 95 species are currently accepted.

Table 4: Example simplifications from different models and reading levels. "Failures" are cases where the model either fails to maintain factual accuracy (e.g., changing "suppressed" to "peaceful"), performs minimal transformation when simplification is needed, or unnecessarily complicates simple text (e.g., "about" to "approximately").

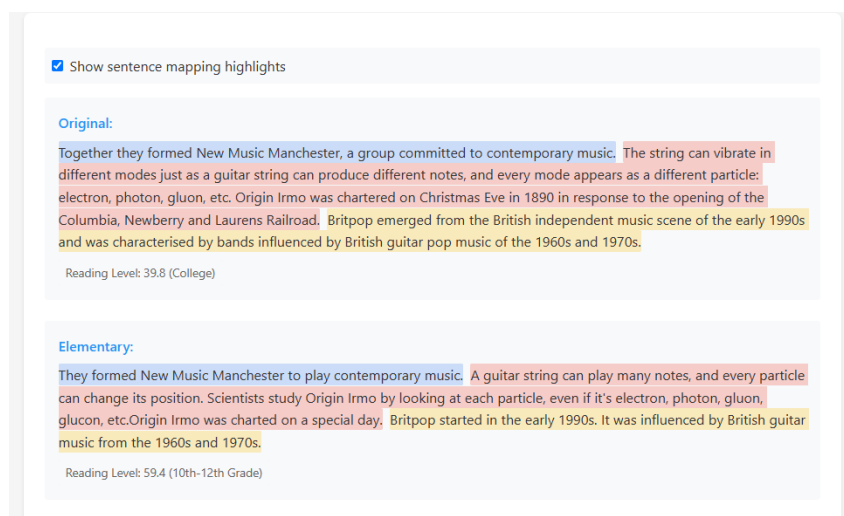


Figure 5: Screenshot of the web-based prototype interface showing sentence splitting from a paragraph, each sentence is individually processed by the model, then re-constructed.

Findings of the TSAR 2025 Shared Task on Readability-Controlled Text Simplification

Fernando Alva-Manchego¹, Regina Stodden², Joseph Marvin Imperial^{3,4},
Abdullah Barayan^{1,5}, Kai North⁶, Harish Tayyar Madabushi³

¹Cardiff University, ²Bielefeld University, ³University of Bath,

⁴National University Philippines, ⁵King Abdulaziz University, ⁶Cambium Assessment

CORRESPONDENCE: alvamanchegof@cardiff.ac.uk

Abstract

This paper presents the findings of the first Shared Task on Readability-Controlled Text Simplification at TSAR 2025. The task required systems to simplify English texts to specific target readability levels of the Common European Framework of Reference for Languages (CEFR). We received 48 submissions from 20 participating teams, with approaches predominantly based on large language models (LLMs), which included iterative refinement, multi-agent setups, and LLM-as-a-judge pipelines. For this shared task, we developed a new dataset of pedagogical texts and evaluated submissions using a weighted combination of semantic similarity and CEFR-level accuracy. The results of the participating teams demonstrate that while LLMs can perform substantially well on this task, dependable and controlled simplification often requires complex, multi-iterative processes. Our findings also suggest that the capabilities of current systems are beginning to saturate existing automatic evaluation metrics, underscoring the need for re-evaluation and practicality.

1 Introduction

Text simplification consists of automatically rewriting a text to make it easier to read and understand while preserving meaning, supporting applications in education, accessibility, and second-language learning (Alva-Manchego et al., 2020). Many previous shared tasks have focused on lexical simplification (Specia et al., 2012; Saggion et al., 2022; Shardlow et al., 2024) or on complexity prediction (Paetzold and Specia, 2016; Yimam et al., 2018; Shardlow et al., 2021). However, these tasks typically do not require control over output readability level, which is essential if simplification is to be adaptive to learner needs.

In this work, we introduce a new shared task for readability-controlled text simplification in English, in which systems must simplify a source text

to a specified CEFR level (e.g., A2 or B1). In this way, the target complexity explicitly aligns with educational and pedagogical goals. Our task builds on, but also departs from, earlier shared tasks in lexical simplification, in that participating systems must simplify short passages under a CEFR constraint, rather than single words or phrases based on a target audience.

To support this shared task, we curated a **new CEFR-based reference dataset of 100 paragraph-level English texts** drawn from pedagogical reading materials for language learners. Each source text was manually simplified by experienced English-language teachers to two lower target levels, resulting in a total of 200 reference simplifications. In addition, we trained a **CEFR evaluator model** to estimate the readability level of system outputs automatically. The model was trained on CEFR-annotated texts and fine-tuned to classify English texts into CEFR levels with high reliability. Both resources are publicly released to support further research.

The shared task challenged participants to generate simplified versions of the same source texts at specified CEFR targets, requiring systems to demonstrate both readability control and semantic fidelity. Submissions were assessed using the CEFR evaluator model to measure compliance with the target level and MeaningBERT (Beauchemin et al., 2023) to assess source and reference-wise meaning preservation. The resulting metric scores were combined into a single ranking using AUTORANK (Kocmi et al., 2025), an aggregation method that normalizes metric scales and mitigates the effect of outliers.

2 Related Work

The task of readability-controlled text simplification (RCTS) aims to generate simplified text aligned with specific difficulty levels, often us-

ing frameworks such as the Common Core Standards or the CEFR for reference (Xu et al., 2015; Uchida et al., 2018; Scarton et al., 2018). The main challenge with RCTS is that it requires finer-grained generation control across multiple versions of the exact text, tailored to different audiences. Early approaches treated RCTS as a supervised sequence-to-sequence problem requiring a parallel corpus annotated with readability (Scarton and Specia, 2018), which then followed by more advanced techniques including the addition of low-level control tokens (Nishihara et al., 2019; Agrawal and Carpuat, 2023), lexical-based constrained decoding for better handling of complex words (Zetsu et al., 2022), and reinforcement learning to reward readability-aligned texts (Yanamoto et al., 2022; Ribeiro et al., 2023; Malik et al., 2024).

Despite the substantial progress, current methods remain heavily dependent on gold-standard parallel data. To address this, more recent works have explored techniques that take advantage of instruction-tuned LLMs’ few-shot learning capabilities for shifting the readability levels of texts without the need for additional training data (Kew et al., 2023; Imperial and Tayyar Madabushi, 2023; Imperial et al., 2024; Farajidizaji et al., 2024; Malik et al., 2024; Barayan et al., 2025). However, achieving precise readability control while maintaining the quality of generated simplifications remains challenging, motivating further research into fine-grained level alignment.

3 CEFR Evaluator Model

We detail the steps we followed to train the CEFR-based evaluator model we will use to evaluate the CEFR alignment of system submissions with gold-standard reference simplifications.

3.1 Data

We used the English subset of the UNIVERSAL-CEFR (Imperial et al., 2025) dataset, which contains gold-standard CEFR-annotated texts at multiple granularities (sentence-, paragraph-, and document-level). We constructed three separate training sets with variations of granularities and language coverage:

TRAIN_DOC_EN This split contains 650 English documents from CAMBRIDGEEXAMS (Xia et al., 2016) and ELG-CEFR-EN (Breuker, 2022).¹

¹AbdullahBarayan/ModernBERT-base-doc_en-Cefr

TRAIN_DOC_SENT_EN This split contains 13,476 English sentence- and document-level texts, combining TRAIN_DOC_EN with additional data from CEFR-SP (Arase et al., 2022) and README++ (EN) (Naous et al., 2024).²

REFERENCE_ALLLANG This is the largest split with 56,963 multilingual instances at the sentence, paragraph, and document levels, integrating data from multiple languages together with TRAIN_DOC_EN.³ See Table 7 in the Appendix for the list of datasets and languages.

For our validation and test sets, we use the corresponding test splits of the English document-level subsets CAMBRIDGEEXAMS and ELG-CEFR-EN. Stratified sampling was applied to maintain a proportional representation across CEFR levels, with 15% of the data allocated to validation and 15% to testing. The final distribution of instances across splits and CEFR levels is shown in Table 6 (Appendix).

3.2 Base Model Architecture

We fine-tuned MODERNBERT-BASE (Warner et al., 2024), a 395M-parameter LLM on each of the three training sets previously discussed. Using this training regime, we produce the following: (1) a document-level English evaluator model using the TRAIN_DOC_EN split, (2) a combination of sentence and document multi-level English evaluator model using the TRAIN_DOC_SENT_EN split, and (3) a multilingual sentence, paragraph, and document evaluator model using the REFERENCE_ALLLANG split. All evaluator model variations were trained for 10 epochs, with the best checkpoint selected based on the highest weighted F1 score on the validation set. Additional training details can be found in Table 4 of the Appendix.

In addition to training MODERNBERT-BASE, we explored two ensemble-based strategies for resolving the final CEFR prediction:

- **Majority Vote:** Labels were assigned based on agreement among at least two models. In cases without a majority, the median CEFR level was chosen.
- **Confidence-Based:** Predictions were taken from the model with the highest confidence score for the given instance.

²AbdullahBarayan/ModernBERT-base-doc_sent_en-Cefr

³AbdullahBarayan/ModernBERT-base-reference_AllLang2-Cefr2

Model Setup	A1	A2	B1	B2	C1	C2	Avg	AdjAcc	RMSE
TRAIN_DOC_EN	0.80	0.90	0.84	0.84	0.74	0.83	0.83	0.97	0.50
TRAIN_DOC_SENT_EN	0.00	0.83	0.90	0.94	0.85	0.86	0.86	0.99	0.38
REFERENCE_ALLLANG	0.50	0.86	0.89	0.97	0.88	0.88	0.89	1.00	0.32
MAJORITY VOTE	0.50	0.87	0.92	0.95	0.85	0.86	0.89	0.99	0.35
CONFIDENCE-BASED	0.00	0.89	0.94	0.94	0.87	0.89	0.89	0.99	0.34

Table 1: Performance of various training data and model prediction setups integrated with ModernBERT-Base on the **test set**. We selected the CONFIDENCE-BASED setup for our final CEFR evaluator model due to its optimal performance (low RMSE and high averages).

3.3 Results

Table 1 reports the performance of the three fine-tuned models and the two ensemble strategies on the test set (Table 8 in the Appendix reports results in the validation set). Results are presented in terms of class-wise F1 scores, weighted average F1, adjusted accuracy, and RMSE. The results demonstrate that all three fine-tuned models achieve strong performance, with the confidence-based ensemble providing the most consistent accuracy and lowest error across both validation and test sets. We therefore adopt this ensemble to assess CEFR compliance of simplified outputs.

4 Shared Task Dataset

Our primary shared task dataset was designed to support readability-controlled simplification in English aligned with the Common European Framework of Reference for Languages (CEFR). The dataset is also aimed towards evaluating systems that can simplify texts to a target readability level while preserving meaning.

4.1 Data Source

All texts were extracted from The British Council’s LearnEnglish website⁴, a major UK-based open-access platform offering pedagogical content for learners of English. We were granted formal permission to use the materials for research and distribution as part of this shared task. The acquired material includes graded reading passages, each associated with a specific CEFR level. The content covers a range of everyday topics and was authored and reviewed by professional educators.

From the available materials, we selected a subset of texts originally labeled as C1 and B2 to serve as source texts for simplification. These upper-intermediate and advanced texts provide sufficient lexical and syntactic complexity to enable meaningful simplification toward lower CEFR levels.

⁴<https://learnenglish.britishcouncil.org/>

4.2 Data Annotation

The annotation process involved producing simplified versions of the original C1 and B2 source texts at lower CEFR levels. Each selected text was simplified to B1 and A2 target levels, resulting in two simplified versions per source. The dataset was divided into two parts: (1) **trial data** (20 instances), simplified by one annotator and released for system development; and (2) **test data** (80 instances), simplified by two annotators and used for official evaluation. All annotators were teachers of English as a foreign language and familiar with the CEFR framework, although with varied experience levels.

Before the main annotation phase, all annotators completed a qualification task to ensure a consistent understanding of CEFR levels and simplification principles. During annotation, each annotator received: (1) the original C1 or B2 text; (2) the target CEFR level (A2 or B1); and (3) a set of annotation guidelines describing the expected linguistic characteristics of the simplified output. These guidelines specified that simplifications should preserve meaning, reduce syntactic and lexical complexity in line with the target level, and maintain grammaticality and fluency. The full qualification task and annotation guidelines are included in the dataset release.

4.3 Quality Control

All simplifications were reviewed by the organisers for formatting consistency and completeness. No additional post-editing or filtering was applied to preserve the natural stylistic variation introduced by each annotator.

Inter-annotator agreement was not computed, as the annotators worked on disjoint subsets of the data. Instead, we assessed reliability by comparing the target CEFR levels assigned during annotation with the levels predicted by our automatic CEFR evaluator model (Sec. 3). The average RMSE was 0.6, suggesting moderate agreement between hu-

man and model estimates. While lower values would indicate stronger alignment, this level of divergence is expected given the subjectivity of CEFR judgments and the coarse step size between adjacent levels. Detailed metrics per annotator are provided in the Appendix.

5 Evaluation Setup

We describe the evaluation pipeline, combining normalization procedures, weighting decisions, and ranking methodology, used to assess the performance of system submissions using the shared task data and a trained CEFR evaluator model. Following the AUTORANK framework proposed in the WMT 2025 General Machine Translation Shared Task (Kocmi et al., 2025), we aggregate multiple evaluation metrics into a single overall ranking to increase robustness against outliers and improve interpretability. We present two forms of AUTORANK rankings: one for all submitted runs and another for the best run per team.

5.1 Metrics

We evaluated each system submission based on three variables: CEFR level compliance, meaning preservation, and gold-standard reference similarity. We describe the metrics chosen to measure each variable below:

1. **CEFR Level Compliance.** We use the Root Mean Squared Error (RMSE) from the trained CEFR model evaluator to assess CEFR level compliance. Lower RMSE values indicate better control of a submitted system with respect to the target CEFR readability level.
2. **Meaning Similarity.** We use MeaningBERT (Beauchemin et al., 2023) to measure the semantic similarity between the source text and the submitted system’s output.
3. **Reference Similarity.** Similar to Meaning Similarity, we also use MeaningBERT (Beauchemin et al., 2023) to measure the semantic similarity between the expert-written simplifications and a submitted system’s output.

We considered other computed metrics, including adjacent accuracy, weighted F1, and BERTScore, but they were ultimately not used in the official ranking for several reasons. Adjacent accuracy is less informative than RMSE because it does not account for the degree of mismatch. On

the other hand, RMSE penalizes predictions proportionally to their distance from the target level. Weighted F1 reflects categorical performance but does not capture the severity of misclassification. Lastly, while BERTScore (Zhang et al., 2020) is a popular general-purpose similarity metric, it was not trained for simplification and often overestimates similarity when there is lexical overlap without true semantic preservation. MeaningBERT, on the other hand, was trained on human annotations for preserving meaning during simplification and is a more task-appropriate choice.

5.2 Submission Filtering

We observed that some teams submitted model simplifications with fewer runs than the expected total of 200. Since missing outputs would bias the evaluation, all runs with fewer than 200 outputs were discarded before scoring and ranking.

5.3 Normalization of Metrics

Metrics operate on different scales and distributions. RMSE included very low outliers, including values close to 0.0, representing perfect or near-perfect CEFR compliance. MeaningBERT values were tightly clustered. If we combined these raw values directly, it is evident that RMSE would dominate because of its larger relative variance. To address this, following (Kocmi et al., 2025), we applied *median–interpercentile scaling* to each metric. This normalization method reduces the influence of outliers while making scores comparable across metrics. Unlike min–max scaling, which is highly sensitive to outliers, this approach ensures that mid-ranked systems remain fairly distinguished. For RMSE, since lower values are better, we invert the scaled scores so that for all metrics, higher is always better.

5.4 Weighting

To reflect the balance required to optimize readability control and meaning preservation, we assign equal global weights of 50% to these two variables. While both semantic similarities in the source text and reference simplifications are essential, the latter generally correlates more strongly with expert judgments of simplification quality. For this reason, we weight the reference-based score twice as much as the source text-based score. The final weights are 0.500 for RMSE, 0.167 for meaning similarity via MeaningBERT, and 0.333 for reference similarity via MeaningBERT.

5.5 Aggregation and AUTORANK Mapping

After normalization and weighting, we computed the system-level scores as weighted averages across the three metrics. These averages were linearly scaled to the range of $[1, N]$, where N is the number of valid runs. Following the WMT25 AUTORANK convention, we apply a final linear mapping where the best-performing system was assigned $\text{AUTORANK} = 1$, the worst-performing system was assigned $\text{AUTORANK} = N$, and the intermediate middle-ranking systems were spaced proportionally between these endpoints.

6 Participants and System Descriptions

6.1 Overview

We received overall 48 submissions by 20 participating teams. Each team was allowed to submit outputs of up to three systems or runs. The most dominant strategy of the submissions was prompting (28 submissions), including evaluations with a CEFR labeling system (14 submissions) or an LLM as a judge (6 submissions). Other strategies ranged from rule-based systems (4 submissions), agentic approaches (7), fine-tuning of LLMs (4), training of neuronal networks (3), and other approaches (2).

6.2 System Summaries

Archaeology (Roscan and Nisioi, 2025) submitted three submissions. For two of them, they prompt an LLM (Claude-Sonnet-4 vs. Llama-3.1-8B-Instruct) to generate simplifications iteratively. They added feedback on the enforced CEFR level to the prompt at each iteration until the level is reached or 5 runs have been completed. They select the best candidates of the model with Minimum Bayes Risk. Additionally, they fine-tuned a lightweight Llama model on synthetic data with CEFR levels A2 and B1 and repeated the process previously described, but achieved lower scores with this approach.

BU-IntelPA proposed two multi-agent systems with zero-shot simplification using either GPT-OSS-20B or Mistral-NeMo-12B. Unfortunately, we cannot provide more information as no system description paper has been submitted.

Cappuccino submitted system outputs based on zero-shot prompting using LLMs. Their prompts are focusing on simplification, meaning preservation, short sentences, controlled vocabulary, and a consistent style. Unfortunately, we cannot provide

more information as this team has not submitted a system description paper.

EasyJon (Barbu et al., 2025) proposed a system that selects the best simplification out of seven LLMs' outputs. They prompt seven different LLMs (Qwen3-235B-A22B, Llama-3.3-70B-Instruct, DeepSeek-R1, Gemma-3-27B, GPT-OSS-120B, Claude-Sonnet-4, Mistral-Medium-3.1) with either short, descriptive, and descriptive prompts with examples and use an LLM-as-a-judge approach to select the best candidates.

EhiMeNLP (Miyata et al., 2025) is ranked first overall with one of their three different systems. All three systems follow the same two-fold strategy of generating simplified candidates by prompting LLMs with four prompt types (fine-grained simplification, controlling CEFR level, US grade levels, and edit operations), and evaluating the candidates based on CEFR labelling and meaning preservation relative to the source text. In their best performing submission, they ensemble several LLMs (GPT-5, GPT-4.1, o3, GPT-OSS-20B, Qwen3-32B, Llama-3.3-70B-Instruct) and all prompting strategies. In the other approaches, they use GPT-5 and combine only two of the four prompting strategies.

GPLSI submitted three runs, including one that is based on a fine-tuned Flan-T5 model, whereas the others are based on prompting Llama-3.2-3B-Instruct. Unfortunately, we cannot provide more information as this team has not submitted a system description paper.

GRIPF (Alfter and Gombert, 2025) proposed three different systems variations. The simplifications of their first system are the outputs of a discussion between two LLMs (i.e., GPT-5 and Claude-Opus-4.1.) which generate, criticize, and revise each others simplified outputs. The submitted candidate is either selected by the two models or, if they do not agree, by a third LLM (Llama-3.2-3B-Instruct), which judges the best output. In their second approach, the authors also provide specific vocabulary for each CEFR level to the LLMs. The third approach uses GPT-4o to generate the simplified candidates. Afterwards, another LLM provides feedback based on the CEFR level of the generated texts and potentially rewrites the text, provided the level matches (or after three runs).

HIT-YOU (Shimada et al., 2025) explored readability-controlled simplification with a prompt-

ing and LLM-as-a-judge approach. For two of their three approaches, they generated multiple candidates with 4 LLMs (i.e., GPT-5, Gemini-2.5-Flash, Claude-4-Sonnet, and o1) and three prompting techniques, and used an LLM-as-a-judge (either Gemini or GPT-5). Their prompts are either zero-shot, one-shot, or include a round-trip translation. The third approach contains a self-refinement loop in which a CEFR labeling system provides continuous feedback until the generated text matches the expected level or the maximum number of iterations is reached.

HOPE (Maharjan and Shrestha, 2025) proposed two rule-based approaches and an additional multi-stage pipeline. Their first rule-based approach focuses only on lexical simplification, whereas the second approach combines lexical and syntactical simplification via word substitution and sentence splitting. In the multi-stage pipeline, they combined lexical and syntactical preprocessing with zero-shot prediction of a T5 model.

HULAT-UC3M (Sanchez-Gomez et al., 2025) submitted two systems to the shared task. Both approaches are based on prompting LLMs (Ettin Suite and LLama-3). Their reinforced prompts contain either detailed descriptions of CEFR levels or only brief descriptions.

ITU (Dinç et al., 2025) explored prompting GPT-4o in a three-step manner. First, they ask the model to simplify with respect to syntactic simplification. In the second round, they ask the model to consider lexical simplification during generation based on syntactically simplified text. Finally, they ask the model also to include elaborations in the simplification. Each prompt includes some rules for simplification as well as examples for it.

Know-AI (Wu et al., 2025) proposed two different submissions. Both submissions can be summarized as an iterative generation of simplifications using GPT-4o until the target CEFR level is reached. In the first submission, the CEFR level is estimated based on an alignment between the CEFR levels and the Flesch-Kincaid Grading Level. For the other approach, they evaluate the readability with the CEFR leveling system provided by the shared task organizers.

MMU_NLP tackled readability-controlled simplification via neural networks. They use existing parallel simplification corpora and enrich them

first with CEFR levels. Afterwards, they generate sentence embeddings for the additional data using SONAR and train a simple feed-forward neural network on them and the target level. This model is then used to generate simplifications of the sentences of the test set. In the different approaches of MMU_NLP, a) separate models are trained per CEFR level (run 1), b) the models are trained on all data at once (run 2), or c) are trained at once, but using 1-hot vectors (run 3). Unfortunately, this team has not submitted a system description paper, which would provide more insight into their approach.

OneNRC (Vajjala, 2025) proposed two agentic approaches using Gemini-2.5-Flash and Gemma-3-12B. For both submissions, they use zero-shot prompting and two tools to evaluate the generated simplifications. The first tool is a CEFR labeling system, and the second tool measures the preservation of meaning in the original text.

OUNLP (Huynh and Cao, 2025) submitted the output of two systems, which were built on top of each other. The first model is rule-based, whereas the second model uses the output of the first as part of an input prompt for an LLM, i.e., GPT-4o. The prompt is enriched with instructions for improving readability, e.g., synonym replacement, clause trimming, sentence splitting, and word limit restrictions.

SQUIREL (Sokova et al., 2025) used a fine-tuning approach for two of their three submissions. In the first approach, they fine-tune Llama-3.2-1B Instruct with two reward functions via the CEFR level of the generated text and sentence similarity between the generated text and the source texts. In the second approach, they focus on simplifications with a larger gap between the source and target CEFR levels. Therefore, they use a more relaxed CEFR reward function that reduces penalties for larger gaps between levels. In their third approach, they focus on lexical simplification. For the substitution, they iteratively select words that could be simplified based on the WordNet lexicon. Afterwards, they use an LLM for integrating the words well into the text.

STARLING (Przybyła, 2025) proposed three submissions to the shared task. All three submissions are prompting-based approaches based on the BLESS benchmark using Gemma-3-27B. Based on multiple generated candidates, the best

one is selected using the CEFR classifier provided by the shared task organizers. This team compared whether selecting the best candidates works better when providing 5 good candidates, 10 good candidates, or 5 random candidates.

taskGen (Oviedo et al., 2025) submitted three approaches using prompting strategies with the same LLM, i.e., Llama-3.1-8B. Each of their prompt lists the relevant CEFR descriptors and examples of appropriate vocabulary, morphological, and grammatical structures. In comparison, their first submission contains no cleaning, whereas the second includes some cleaning, and the third contains candidate selection based on CEFR levels and meaning preservation.

Uniandes (Russi et al., 2025) proposed a few-shot learning and agent-based approach using different LLMs for each of the three submissions. For all submissions, they use an LLM as a judge, which provides feedback for the next iteration in the simplification loop. In the first run, Gemini-2.5-Pro is prompted, Gemini-2.5-Flash in the second, and GPT-OSS-120B and Gemini-2.5-Pro in the third.

UoL-UPF (Hayakawa et al., 2025) addressed the task of readability-controlled simplification via prompting and candidate selection out of different model and prompt combinations. For candidate selection, they use Minimum Bayes Risk and the CEFR labeler provided by the shared task organizers. One of their approaches focuses on simplifying on the paragraph level, another on the sentence level, and the third one combines the sentence and paragraph levels.

7 Results

We report two separate rankings to provide a comprehensive view of system performance. Table 2 presents every valid submission, allowing us to compare the relative performance of all system variants submitted by each team. In contrast, Table 3 shows only the highest-scoring run per team to highlight each team’s most performant approach. Analyzing model variations across the submitted systems, we identified that 34 distinct LLMs were used for the shared task, of which 23 were open-source and the remaining were commercial.

Comparing approaches that used only one LLM, system submissions that used commercial ones like GPT-5 or Gemini 2.5 performed substantially better in terms of average performance (EhiMeNLP,

HIT-YOU, or Uniandes) than with system submissions that used open-weight models like Llama or Gemma (taskGen, Archaeology, and STARLING). The best model, which uses only open-weight models, is taskGen, with an AUTORANK score of 7.48 using Llama-3.1-8B. On the other hand, system submissions that used multiple LLMs achieved even better results, including the top 5 submissions, where one is from EhiMeNLP and two from UoL-UPF and HIT-YOU. All these top-scoring system submissions used at least four LLM variants, except for EasyJon, which has also used a collection of seven LLMs, but was only ranked 8.88. Thus, it is clear that the selection of LLMs and the prompting technique largely affect the performance. Likewise, we also observe a clear pattern from the system submissions where the use of commercial GPT-based models leads to a substantially stronger performance.

Based on the system submissions, prompting strategies achieved, on average, the best AUTORANK placement of 7.70, followed by agentic approaches with 8.48 and rule-based approaches with 13.00. Comparing the prompting strategies, the approaches with an LLM-as-a-judge performed slightly better, with an AUTORANK placement of 6.33, than the approaches using a CEFR labeling system for evaluation, with 6.70. Due to the lack of training data, some participants used external resources, such as related simplification corpora, lexicons, and more informative descriptions of CEFR levels to enhance their prompts. The teams also used reference-less metrics, such as readability-based formulas like Flesch-Kincaid and semantic similarity via MeaningBERT, to pre-evaluate their systems and feed the results back into their models in an iterative process.

8 Discussion

Are LLMs the Only Way Forward for Text Simplification?

The use of LLMs as a core resource across the majority of system submissions reflects a decisive shift away from traditional rule-based simplification methods to generative models. We observe a similar pattern with the proliferation of LLMs achieving state-of-the-art performance across general readability control benchmarks (Kew et al., 2023; Imperial et al., 2025). This transition has significant implications for evaluation, shifting from measuring output quality to assessing real-world impact (Reiter, 2025). As a

Team	Model	RMSE	MeaningBERT (Src)	MeaningBERT (Ref)	Avg	AUTORANK
EhiMeNLP ★	run1	0.000	0.902	0.845	0.636	1.000
UoL-UPF ★	uol-upf_submission3	0.000	0.856	0.857	0.603	1.410
UoL-UPF ★	uol-upf_submission1	0.000	0.849	0.856	0.590	1.580
HIT-YOU	run2_gpt_ensemble_4	0.158	0.852	0.835	0.429	3.610
HIT-YOU	run1_gemini_ensemble_4	0.187	0.863	0.833	0.424	3.670
EhiMeNLP	run3	0.234	0.847	0.840	0.390	4.100
EhiMeNLP	run2	0.200	0.838	0.816	0.322	4.970
HIT-YOU	run3_self_refine	0.245	0.822	0.820	0.282	5.460
Uniandes	run_1	0.212	0.817	0.814	0.275	5.560
Uniandes	run_2	0.200	0.825	0.803	0.260	5.740
Archaeology	claude_sonnet_4	0.122	0.779	0.804	0.238	6.010
Uniandes	run_3	0.510	0.847	0.813	0.138	7.280
taskGen	submit_3	0.628	0.856	0.826	0.122	7.480
ounlp	test_data_output_for_First_Program	0.755	0.855	0.849	0.121	7.500
Archaeology	llama_3.1_8b	0.265	0.782	0.789	0.109	7.640
BU-intelPA	run1	0.628	0.831	0.830	0.099	7.780
Cappuccino	Cappuccino_TSAR2025_Submission	0.718	0.826	0.843	0.077	8.050
GRIPF	tsar2025_ezscalar_lexical_gripf	0.689	0.857	0.820	0.070	8.130
GRIPF	tsar2025_ezscalar_nonlexical_gripf	0.721	0.856	0.824	0.060	8.270
Know-AI	run2	0.700	0.821	0.835	0.053	8.350
Know-AI	run1	0.659	0.801	0.832	0.036	8.560
EasyJon	run_1	0.822	0.838	0.836	0.011	8.880
SQUREL	SQUREL_Run3	1.153	0.979	0.819	-0.022	9.300
HULAT-UC3M	run2_llama3-8b_reinforced-prompt	0.608	0.793	0.806	-0.028	9.370
oneNRC	onenrc_google25flash_withtoolcall	0.534	0.772	0.800	-0.033	9.440
STARLING	starling_1_g5-best	0.621	0.811	0.791	-0.053	9.690
ITUNLP	itunlp	0.632	0.797	0.797	-0.063	9.820
oneNRC	onenrc_gemma312b_react_notool	0.579	0.761	0.803	-0.069	9.880
SQUREL	SQUREL_Run1	0.718	0.821	0.797	-0.076	9.980
taskGen	submit_1	0.592	0.791	0.786	-0.084	10.070
HULAT-UC3M	run1_llama3-8b_reinforced-prompt	0.682	0.790	0.791	-0.122	10.560
GRIPF	tsar2025_saga_gripf	0.831	0.827	0.796	-0.140	10.780
SQUREL	SQUREL_Run2	0.632	0.779	0.778	-0.153	10.950
STARLING	starling_2_g10-best	0.678	0.795	0.777	-0.160	11.040
Archaeology	llama_3.2_1b	0.212	0.706	0.731	-0.165	11.100
taskGen	submit_2	0.561	0.752	0.773	-0.169	11.150
STARLING	starling_3_g5-random	0.812	0.816	0.785	-0.180	11.280
HOPE	HOPE_run1	1.428	0.945	0.815	-0.255	12.230
GPLSI	run1_llama_knowledge	0.998	0.865	0.772	-0.258	12.270
GPLSI	run2_llama_zs	0.640	0.772	0.750	-0.258	12.270
ounlp	test_data_output_for_Second_Program	0.714	0.865	0.701	-0.313	12.960
HOPE	HOPE_EXPERT_run1	1.402	0.919	0.797	-0.337	13.260
MMU_NLP	mmu_tsar25_test_system2	1.005	0.845	0.754	-0.350	13.430
BU-intelPA	run2	0.612	0.715	0.739	-0.368	13.650
MMU_NLP	mmu_tsar25_test_system3	1.010	0.832	0.752	-0.381	13.830
MMU_NLP	mmu_tsar25_test_system_1	0.822	0.735	0.676	-0.664	17.390
HOPE	HOPE_SOTA_run1	1.600	0.841	0.730	-0.795	19.030
GPLSI	run3_flan_knowledge	0.883	0.221	0.182	-3.093	48.000

Table 2: Final ranked results for all submitted runs using AUTORANK with custom weighting.

Team	Model	RMSE	MeaningBERT (Src)	MeaningBERT (Ref)	Avg	AUTORANK
EhiMeNLP ★	run1	0.000	0.902	0.845	0.636	1.000
UoL-UPF ★	uol-upf_submission3	0.000	0.856	0.857	0.603	1.410
HIT-YOU ★	run2_gpt_ensemble_4	0.158	0.852	0.835	0.429	3.610
Uniandes	run_1	0.212	0.817	0.814	0.275	5.560
Archaeology	claude_sonnet_4	0.122	0.779	0.804	0.238	6.010
taskGen	submit_3	0.628	0.856	0.826	0.122	7.480
ounlp	test_data_output_for_First_Program	0.755	0.855	0.849	0.121	7.500
BU-intelPA	run1	0.628	0.831	0.830	0.099	7.780
Cappuccino	Cappuccino_TSAR2025_Submission	0.718	0.826	0.843	0.077	8.050
GRIPF	tsar2025_ezscalar_lexical_gripf	0.689	0.857	0.820	0.070	8.130
Know-AI	run2	0.700	0.821	0.835	0.053	8.350
EasyJon	run_1	0.822	0.838	0.836	0.011	8.880
SQUREL	SQUREL_Run3	1.153	0.979	0.819	-0.022	9.300
HULAT-UC3M	run2_llama3-8b_reinforced-prompt	0.608	0.793	0.806	-0.028	9.370
oneNRC	onenrc_google25flash_withtoolcall	0.534	0.772	0.800	-0.033	9.440
STARLING	starling_1_g5-best	0.621	0.811	0.791	-0.053	9.690
ITUNLP	itunlp	0.632	0.797	0.797	-0.063	9.820
HOPE	HOPE_run1	1.428	0.945	0.815	-0.255	12.230
GPLSI	run1_llama_knowledge	0.998	0.865	0.772	-0.258	12.270
MMU_NLP	mmu_tsar25_test_system2	1.005	0.845	0.754	-0.350	13.430

Table 3: Best run per team using AUTORANK with custom weighting.

result, the central question is no longer "*Can this model generate simplified texts?*" But rather, "*Are this model's outputs of sufficient quality to be used with CEFR-based learners?*" Hence, this shared task supports further research on pedagogical validations (e.g., expert-in-the-loop) to ensure that LLM-generated simplifications are aligned with CEFR-based learning objectives rather than producing superficial, simpler texts.

Where Do We Position Synthetic Data? While the shared task was ultimately successful in producing a new parallel CEFR-based reference dataset for validation and testing, it is worth noting the substantial effort and financial support required for resource development. In line with this, we ask the question "*Is it time to use LLM-generated CEFR data to complement expert-produced data?*" While adjacent NLP tasks like grammatical correction and essay scoring have benefited from the performance advantages of synthetic data, its practicality for readability and text simplification applications has only ever been explored recently (Stahlberg and Kumar, 2021; Klöser et al., 2024; Latouche et al., 2024; Qwaider et al., 2025). For text simplification that is anchored on a real-world language proficiency framework like CEFR, there are both opportunities and risks. On the one hand, LLMs are capable of generating fluent text guided by CEFR specifications, as evident in the top submissions of this shared task. This could potentially be valuable for low-resource languages across various domains, genres, and text types where expert-annotated resources are scarce. However, using synthetic data without careful validation checks risks producing noisy approximations of CEFR levels, which may reinforce undesirable simplification patterns. As such, our stance on this is that we should establish a community-accepted framework to integrate synthetic data, involving steps such as filtering and expert validation stages. For the next iteration of this shared task, we propose a direction exploring how the use of LLM-generated synthetic data can perform well on test and validation data generated by experts.

Cost-Performance Tradeoffs Our leaderboard results show that top-ranked systems achieved strong performance by leveraging ensembles of multiple LLMs. However, these submissions come with substantial computational and financial costs. For example, the rank-one system EhiMeNLP combined six LLMs, including GPT-5, GPT-4.1, o3,

GPT-OSS-20B, Qwen3-32B, and Llama-3.3-70B-Instruct to generate up to 120 candidate simplifications per input. While such methods clearly demonstrate the potential of ensemble-based techniques for producing precise readability-controlled simplifications, their heavy resource demands may limit their adoption in educational and resource-constrained contexts where computational budget to run more than one commercial model is unavailable. Future work should therefore explore approaches that strike a balance between achieving decent performance and computational efficiency. Likewise, combining *resource-aware evaluation procedures* with performance-based metrics may encourage participants to propose innovative and computationally viable methods that are effective in real-world CEFR-based text simplification settings.

9 Conclusion

The TSAR 2025 Shared Task introduced a benchmark for readability-controlled text simplification explicitly aligned with CEFR levels. Two main resources were developed as contributions to the community: a) a CEFR-aligned dataset of pedagogical paragraph-level English texts simplified to A2 and B1 levels, and b) a CEFR evaluator model fine-tuned to estimate text difficulty along the CEFR scale.

The results from participating teams indicate that although LLMs achieve strong performance on this task, achieving dependable and fine-grained control over simplification often relies on complex, iterative generation strategies. Moreover, our analysis suggests that current systems are approaching the limits of existing automatic evaluation metrics, underscoring the need to adapt these metrics for greater robustness and practical relevance.

Limitations

We acknowledge several limitations in the conduct of this shared task, which we believe can serve as a springboard for future iterations.

Dataset and Annotation Coverage. For our primary shared task dataset, we use a single-source English-only dataset from The British Council. This is the result of our prioritized search for a new gold-standard CEFR-based parallel dataset that has not been published before, as a contribution to the community from the Shared Task. The

dataset we acquired was provided in paragraph-level format due to the unavailability of resources in longer forms. One can argue for transforming the paragraph-level data into sentence-level data to account for variance; however, we did not pursue this option since sentence-level CEFR data are already available, such as CEFR-SP (Arase et al., 2022) and ReadMe (Naous et al., 2024). Likewise, due to budget constraints, we were only able to provide one expert-written simplification for each instance of the trial and test set. We acknowledge that expanding this, for example, by asking more language experts to produce separate annotations will allow better convergence of evaluation scores.

Automatic Metrics for System Evaluations.

Our main evaluation pipeline, which determined the system rankings, primarily relied on automatic metrics including a combination of weighted RMSE and ModernBERT scores, which compared system outputs to reference and target expert-written simplifications. While this setup is convenient from a shared task perspective, an additional round of expert validation from text simplifications of each system submission would be valuable in assessing linguistic and pedagogical appropriateness that automatic metrics do not capture. However, we were unable to conduct this due to time and funding constraints.

Ethics Statement

The dataset used for this shared task was acquired with the British Council’s formal permission. The collection of manual simplifications received a favourable opinion from the Ethics Committee of the School of Computer Science and Informatics at Cardiff University. All artifacts, including the dataset with expert reference simplifications and evaluation scripts, will be released to the research community to support future work. All participating teams are credited for their submissions. System description papers were included when available, ensuring transparency and proper attribution of their methods.

We acknowledge that the LLMs used by participants and in our evaluation tools may contain inherent biases reflecting their training data. This work is an analysis of system performance and does not constitute an endorsement of these models for direct pedagogical applications without further expert-in-the-loop validation. Our aim is to benchmark the current state of the art to encourage the

responsible development of text simplification technology.

Acknowledgements

We thank all participants who expressed interest in joining the shared task and worked towards submitting their proposed systems. We also thank the annotators for their valuable contributions to create the reference text simplification dataset. Regina Stodden was supported by the European Regional Development Fund within the project: LLM4KMU - Optimierter Einsatz von Open Source Large Language Modellen in KMU. Joseph Imperial is supported by the National University Philippines and the UKRI Centre for Doctoral Training in Accountable, Responsible, and Transparent AI [EP/S023437/1] of the University of Bath.

References

- Sweta Agrawal and Marine Carpuat. 2023. [Controlling pre-trained language models for grade-specific text simplification](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12807–12819, Singapore. Association for Computational Linguistics.
- David Alfter and Sebastian Gombert. 2025. GRIPF at TSAR 2025 Shared Task: Towards controlled CEFR level simplification with the help of inter-model interactions. In *Proceedings of the Fourth Workshop on Text Simplification, Accessibility, and Readability (TSAR 2025)*, Suzhou, China. Association for Computational Linguistics.
- Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2020. [Data-driven sentence simplification: Survey and benchmark](#). *Computational Linguistics*, 46(1):135–187.
- Yuki Arase, Satoru Uchida, and Tomoyuki Kajiwara. 2022. [CEFR-based sentence difficulty annotation and assessment](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6206–6219, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Abdullah Barayan, Jose Camacho-Collados, and Fernando Alva-Manchego. 2025. [Analysing zero-shot readability-controlled sentence simplification](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6762–6781, Abu Dhabi, UAE. Association for Computational Linguistics.
- Paul-Gerhard Barbu, Adrianna Lipska-Dieck, and Lena Lindner. 2025. EasyJon at TSAR 2025 Shared Task: Evaluation of Automated Text Simplification with

- LLM-as-a-Judge. In *Proceedings of the Fourth Workshop on Text Simplification, Accessibility, and Readability (TSAR 2025)*, Suzhou, China. Association for Computational Linguistics.
- David Beauchemin, Horacio Saggion, and Richard Khoury. 2023. [Meaningbert: assessing meaning preservation between sentences](#). *Frontiers in Artificial Intelligence*, 6.
- Mark Breuker. 2022. [CEFR labelling and assessment services](#). In *European Language Grid: A Language Technology Platform for Multilingual Europe*, pages 277–282. Springer International Publishing Cham.
- Kutay Arda Dinç, Fatih Bektaş, and Gülşen Eryiğit. 2025. ITU NLP at TSAR 2025 Shared Task: A Three-Stage Prompting Approach for CEFR-Oriented Text Simplification. In *Proceedings of the Fourth Workshop on Text Simplification, Accessibility, and Readability (TSAR 2025)*, Suzhou, China. Association for Computational Linguistics.
- Asma Farajidizaji, Vatsal Raina, and Mark Gales. 2024. [Is it possible to modify text to a target readability level? an initial investigation using zero-shot large language models](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9325–9339, Torino, Italia. ELRA and ICCL.
- Akio Hayakawa, Nouran Khallaf, and Horacio Saggion. 2025. UoL-UPF at TSAR 2025 Shared Task: A Generate-and-Select Approach for Readability-Controlled Text Simplification. In *Proceedings of the Fourth Workshop on Text Simplification, Accessibility, and Readability (TSAR 2025)*, Suzhou, China. Association for Computational Linguistics.
- Cuong Huynh and Jie Cao. 2025. OUNLP at TSAR 2025 Shared Task: AI-Generated Multi-Round Sentence Simplifier. In *Proceedings of the Fourth Workshop on Text Simplification, Accessibility, and Readability (TSAR 2025)*, Suzhou, China. Association for Computational Linguistics.
- Joseph Marvin Imperial, Abdullah Barayan, Regina Stodden, Rodrigo Wilkens, Ricardo Muñoz Sánchez, Lingyun Gao, Melissa Torgbi, Dawn Knight, Gail Forey, Reka R. Jablonkai, Ekaterina Kochmar, Robert Reynolds, Eugénio Ribeiro, Horacio Saggion, Elena Volodina, Sowmya Vajjala, Thomas François, Fernando Alva-Manchego, and Harish Tayyar Madabushi. 2025. [UniversalCEFR: Enabling Open Multilingual Research on Language Proficiency Assessment](#). *arXiv preprint arXiv:2506.01419*.
- Joseph Marvin Imperial, Gail Forey, and Harish Tayyar Madabushi. 2024. [Standardize: Aligning language models with expert-defined standards for content generation](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1573–1594, Miami, Florida, USA. Association for Computational Linguistics.
- Joseph Marvin Imperial and Harish Tayyar Madabushi. 2023. [Flesch or fumble? evaluating readability standard alignment of instruction-tuned language models](#). In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 205–223, Singapore. Association for Computational Linguistics.
- Tannon Kew, Alison Chi, Laura Vásquez-Rodríguez, Sweta Agrawal, Dennis Aumiller, Fernando Alva-Manchego, and Matthew Shardlow. 2023. [BLESS: Benchmarking large language models on sentence simplification](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13291–13309, Singapore. Association for Computational Linguistics.
- Lars Klöser, Mika Beele, Jan-Niklas Schagen, and Bodo Kraft. 2024. [German text simplification: Finetuning large language models with semi-synthetic data](#). In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 63–72, St. Julian’s, Malta. Association for Computational Linguistics.
- Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Natalia Fedorova, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakouga, Jessica Lundin, Kenton Murray, Masaaki Nagata, and 9 others. 2025. [Preliminary Ranking of WMT25 General Machine Translation Systems](#). *Preprint*, arXiv:2508.14909.
- Gaetan Lopez Latouche, Marc-André Carbonneau, and Benjamin Swanson. 2024. [Zero-shot cross-lingual transfer for synthetic data generation in grammatical error detection](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3002–3016, Miami, Florida, USA. Association for Computational Linguistics.
- Sujal Maharjan and Astha Shrestha. 2025. HOPE at TSAR 2025 Shared Task: Balancing Control and Complexity in Readability-Controlled Text Simplification. In *Proceedings of the Fourth Workshop on Text Simplification, Accessibility, and Readability (TSAR 2025)*, Suzhou, China. Association for Computational Linguistics.
- Ali Malik, Stephen Mayhew, Christopher Piech, and Klinton Bicknell. 2024. [From tarzan to Tolkien: Controlling the language proficiency level of LLMs for content generation](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15670–15693, Bangkok, Thailand. Association for Computational Linguistics.
- Rina Miyata, Koki Horiguchi, Risa Kondo, Yuki Fujiwara, and Tomoyuki Kajiwara. 2025. EhiMeNLP at TSAR 2025 Shared Task: Candidate Generation via Iterative Simplification and Reranking by Readability and Semantic Similarity. In *Proceedings of*

- the Fourth Workshop on Text Simplification, Accessibility, and Readability (TSAR 2025)*, Suzhou, China. Association for Computational Linguistics.
- Tarek Naous, Michael J Ryan, Anton Lavrouk, Mohit Chandra, and Wei Xu. 2024. [ReadMe++: Benchmarking multilingual language models for multi-domain readability assessment](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12230–12266, Miami, Florida, USA. Association for Computational Linguistics.
- Daiki Nishihara, Tomoyuki Kajiwara, and Yuki Arase. 2019. [Controllable text simplification with lexical constraint loss](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 260–266, Florence, Italy. Association for Computational Linguistics.
- Juan Cruz Oviedo, Elisabet Comelles Pujadas, Laura Alonso Alemany, and Jordi Atserias Batalla. 2025. [taskGen at TSAR 2025 Shared Task: Exploring prompt strategies with linguistic knowledge](#). In *Proceedings of the Fourth Workshop on Text Simplification, Accessibility, and Readability (TSAR 2025)*, Suzhou, China. Association for Computational Linguistics.
- Gustavo Paetzold and Lucia Specia. 2016. [SemEval 2016 task 11: Complex word identification](#). In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 560–569, San Diego, California. Association for Computational Linguistics.
- Piotr Przybyła. 2025. [STARLING at TSAR 2025 Shared Task: Leveraging Alternative Generations for Readability Level Adjustment in Text Simplification](#). In *Proceedings of the Fourth Workshop on Text Simplification, Accessibility, and Readability (TSAR 2025)*, Suzhou, China. Association for Computational Linguistics.
- Chatrine Qwaider, Bashar Alhafni, Kirill Chirkunov, Nizar Habash, and Ted Briscoe. 2025. [Enhancing Arabic automated essay scoring with synthetic data and error injection](#). In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 549–563, Vienna, Austria. Association for Computational Linguistics.
- Ehud Reiter. 2025. [We Should Evaluate Real-World Impact](#). *Computational Linguistics*, pages 1–13.
- Leonardo F. R. Ribeiro, Mohit Bansal, and Markus Dreyer. 2023. [Generating summaries with controllable readability levels](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11669–11687, Singapore. Association for Computational Linguistics.
- Rares-Alexandru Roscan and Sergiu Nisioi. 2025. [Archaeology at TSAR 2025 Shared Task: Teaching Small Models to do CEFR Simplifications](#). In *Proceedings of the Fourth Workshop on Text Simplification, Accessibility, and Readability (TSAR 2025)*, Suzhou, China. Association for Computational Linguistics.
- Felipe Arias Russi, Kevin Cohen Solano, and Ruben Manrique. 2025. [Uniandes at TSAR 2025 Shared Task: Multi-Agent CEFR Text Simplification with Automated Quality Assessment and Iterative Refinement](#). In *Proceedings of the Fourth Workshop on Text Simplification, Accessibility, and Readability (TSAR 2025)*, Suzhou, China. Association for Computational Linguistics.
- Horacio Saggion, Sanja Štajner, Daniel Ferrés, Kim Cheng Sheang, Matthew Shardlow, Kai North, and Marcos Zampieri. 2022. [Findings of the TSAR-2022 shared task on multilingual lexical simplification](#). In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 271–283, Abu Dhabi, United Arab Emirates (Virtual). Association for Computational Linguistics.
- Jesus M. Sanchez-Gomez, Lourdes Moreno, Paloma Martínez, and Marco Antonio Sanchez-Escudero. 2025. [HULAT-UC3M at TSAR 2025 Shared Task on Readability-Controlled Text Simplification](#). In *Proceedings of the Fourth Workshop on Text Simplification, Accessibility, and Readability (TSAR 2025)*, Suzhou, China. Association for Computational Linguistics.
- Carolina Scarton, Gustavo Paetzold, and Lucia Specia. 2018. [Text simplification from professionally produced corpora](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Carolina Scarton and Lucia Specia. 2018. [Learning simplifications for specific target audiences](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 712–718, Melbourne, Australia. Association for Computational Linguistics.
- Matthew Shardlow, Fernando Alva-Manchego, Riza Batista-Navarro, Stefan Bott, Saul Calderon Ramirez, Rémi Cardon, Thomas François, Akio Hayakawa, Andrea Horbach, Anna Hülsing, Yusuke Ide, Joseph Marvin Imperial, Adam Nohejl, Kai North, Laura Occhipinti, Nelson Pérez Rojas, Nishat Raihan, Tharindu Ranasinghe, Martin Solis Salazar, and 3 others. 2024. [The BEA 2024 shared task on the multilingual lexical simplification pipeline](#). In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 571–589, Mexico City, Mexico. Association for Computational Linguistics.
- Matthew Shardlow, Richard Evans, Gustavo Henrique Paetzold, and Marcos Zampieri. 2021. [SemEval-2021 task 1: Lexical complexity prediction](#). In *Proceedings of the 15th International Workshop on Se-*

- mantic Evaluation (SemEval-2021)*, pages 1–16, Online. Association for Computational Linguistics.
- Mao Shimada, Kexin Bian, Zhidong Ling, and Mamoru Komachi. 2025. HIT-YOU at TSAR 2025 Shared Task: Leveraging Similarity-Based Few-Shot Prompting, Round-Trip Translation, and Self-Refinement for Readability-Controlled Text Simplification. In *Proceedings of the Fourth Workshop on Text Simplification, Accessibility, and Readability (TSAR 2025)*, Suzhou, China. Association for Computational Linguistics.
- Daria Sokova, Anastasiia Bezobrazova, and Constantin Orasan. 2025. SUREL at TSAR 2025 Shared Task: CEFR-Controlled Text Simplification with Prompting and Reinforcement Fine-Tuning. In *Proceedings of the Fourth Workshop on Text Simplification, Accessibility, and Readability (TSAR 2025)*, Suzhou, China. Association for Computational Linguistics.
- Lucia Specia, Sujay Kumar Jauhar, and Rada Mihalcea. 2012. [SemEval-2012 task 1: English lexical simplification](#). In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 347–355, Montréal, Canada. Association for Computational Linguistics.
- Nicolas Spring, Annette Rios, and Sarah Ebling. 2021. [Exploring German multi-level text simplification](#). In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1339–1349, Held Online. INCOMA Ltd.
- Felix Stahlberg and Shankar Kumar. 2021. [Synthetic data generation for grammatical error correction with tagged corruption models](#). In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 37–47, Online. Association for Computational Linguistics.
- Regina Stodden, Omar Momen, and Laura Kallmeyer. 2023. [DEplain: A German parallel corpus with intralingual translations into plain language for sentence and document simplification](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16441–16463, Toronto, Canada. Association for Computational Linguistics.
- Satoru Uchida, Shohei Takada, and Yuki Arase. 2018. [CEFR-based lexical simplification dataset](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Sowmya Vajjala. 2025. OneNRC at TSAR2025 Shared Task: Small Models for Readability Controlled Text Simplification. In *Proceedings of the Fourth Workshop on Text Simplification, Accessibility, and Readability (TSAR 2025)*, Suzhou, China. Association for Computational Linguistics.
- Laura Vásquez-Rodríguez, Pedro-Manuel Cuenca-Jiménez, Sergio Morales-Esquivel, and Fernando Alva-Manchego. 2022. [A benchmark for neural readability assessment of texts in Spanish](#). In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 188–198, Abu Dhabi, United Arab Emirates (Virtual). Association for Computational Linguistics.
- Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. [Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference](#). *Preprint*, arXiv:2412.13663.
- Yiheng Wu, Anisia Katinskaia, Jue Hou, and Roman Yangarber. 2025. Know-AI at TSAR 2025 Shared Task: Difficulty-aware Text Simplification System. In *Proceedings of the Fourth Workshop on Text Simplification, Accessibility, and Readability (TSAR 2025)*, Suzhou, China. Association for Computational Linguistics.
- Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. 2016. [Text readability assessment for second language learners](#). In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 12–22, San Diego, CA. Association for Computational Linguistics.
- Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. [Problems in current text simplification research: New data can help](#). *Transactions of the Association for Computational Linguistics*, 3:283–297.
- Daiki Yanamoto, Tomoki Ikawa, Tomoyuki Kajiwara, Takashi Ninomiya, Satoru Uchida, and Yuki Arase. 2022. [Controllable text simplification with deep reinforcement learning](#). In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 398–404, Online only. Association for Computational Linguistics.
- Seid Muhie Yimam, Chris Biemann, Shervin Malmasi, Gustavo Paetzold, Lucia Specia, Sanja Štajner, Anaïs Tack, and Marcos Zampieri. 2018. [A report on the complex word identification shared task 2018](#). In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 66–78, New Orleans, Louisiana. Association for Computational Linguistics.
- Tatsuya Zetsu, Tomoyuki Kajiwara, and Yuki Arase. 2022. [Lexically constrained decoding with edit operation prediction for controllable text simplification](#).

In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 147–153, Abu Dhabi, United Arab Emirates (Virtual). 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 Appendix

A.1 Data Distribution

Table 6 shows the distribution of instances across CEFR levels in the training, validation, and test splits.

A.2 Hyperparameter Values

Table 4 reports the hyperparameters used to fine-tune the pre-trained MODERNBERT model.

Hyperparameter	Value
Learning rate	3.6×10^{-5}
Train batch size	3
Evaluation batch size	3
Random seed	42
Gradient accumulation steps	16
Total effective batch size	48
Optimizer	adamw_torch_fused
Betas	(0.9, 0.999)
Epsilon	10^{-8}
Learning-rate scheduler	linear
Warm-up ratio	0.1

Table 4: Hyperparameter values used to fine-tune the pre-trained MODERNBERT model.

A.3 Annotators’ Reliability

Table 5 shows different metrics between the target CEFR level assigned during annotation with the levels predicted by our automatic CEFR evaluator model in the shared task dataset.

Annotator	Data (N)	ρ	Acc	AdjAcc	RMSE
1	Trial (20)	0.68	0.63	1.00	0.61
2	Test (20)	0.37	0.65	1.00	0.59
3	Test (80)	0.43	0.61	1.00	0.62

Table 5: Reported reliability and accuracy scores of the annotators with respect to the data splits they produced the reference text simplifications.

A.4 Classifier Performance

Table 8 presents the performance of the CEFR classifier in the validation set.

Split	A1	A2	B1	B2	C1	C2	Total
TRAIN_DOC_EN	17	122	152	148	115	96	650
TRAIN_DOC_SENT_EN	323	2,066	4,080	4,374	2,236	397	13,476
REFERENCE_ALLLANG	2,318	19,838	22,270	7,257	3,802	1,478	56,963
VALIDATION	4	26	33	31	24	21	139
TEST	3	26	33	32	25	21	140

Table 6: Distribution of instances across CEFR levels in training, validation, and test splits.

Source Name	Language
cefr-sp (Arase et al., 2022)	en
apa-lha (Spring et al., 2021)	de
deplain-apa-doc (Stodden et al., 2023)	de
deplain-apa-sent (Stodden et al., 2023)	de
deplain-web-doc (Stodden et al., 2023)	de
elg-cefr-de (Breuker, 2022)	de
elg-cefr-nl (Breuker, 2022)	nl
hablacultura (Vásquez-Rodríguez et al., 2022)	es
kwiziq (Vásquez-Rodríguez et al., 2022)	es
kwiziq (Imperial et al., 2025)	fr
learn_welsh_cy (Imperial et al., 2025)	cy
readme (Naous et al., 2024)	en, ar, fr, hi, ru

Table 7: List of datasets and languages included in the REFERENCE_ALLLANG split

Model Setup	A1	A2	B1	B2	C1	C2	Avg	AdjAcc	RMSE
TRAIN_DOC_EN	0.57	0.85	0.77	0.81	0.78	0.91	0.81	0.97	0.50
TRAIN_DOC_SENT_EN	0.40	0.87	0.82	0.81	0.84	0.95	0.84	0.99	0.42
REFERENCE_ALLLANG	0.00	0.84	0.79	0.78	0.81	0.95	0.80	1.00	0.43
MAJORITY VOTE	0.40	0.87	0.84	0.82	0.84	0.95	0.84	0.99	0.42
CONFIDENCE-BASED	0.40	0.88	0.87	0.84	0.86	0.98	0.87	0.99	0.39

Table 8: Performance of various training data and model prediction setups integrated with ModernBERT-Base on the **validation set**. We selected the CONFIDENCE-BASED setup for our final CEFR evaluator model due to its optimal performance (low RMSE and high averages).

OneNRC@TSAR2025 Shared Task: Small Models for Readability Controlled Text Simplification

Sowmya Vajjala

National Research Council, Canada

sowmya.vajjala@nrc-cnrc.gc.ca

Abstract

In this system description paper, we describe the team OneNRC’s experiments on readability controlled text simplification, focused on using smaller, quantized language models ($< 20B$). We compare these with one large proprietary model and show that the smaller models offer comparable results in some experimental settings. The approach primarily comprises of an agentic workflow, and tool calling. The best results were achieved while using a CEFR proficiency classifier as a verification tool for the language model agent. In terms of comparison with other systems, our submission that used a quantized Gemma3:12B model that ran on a laptop achieved a rank of 9.88 among the submitted systems as per the AUTORANK framework used by the organizers. We hope these results will lead into further exploration on the usefulness of smaller models for text simplification.

1 Introduction

Automatic Text Simplification (ATS) is the task of translating a text written in a more advanced language into more accessible language. ATS research in NLP is over two decades old and most of the research focused on a single complex \rightarrow simple language text generation task, considering the unavailability of large scale graded simplification corpora. Siddharthan (2014) and Alva-Manchego et al. (2020) are two well-known surveys on the topic. The advent of Large Language Models (LLMs) made it possible to overcome the ATS dataset barrier to some extent with good zero-/few- shot performance on English text simplification (Kew et al., 2023) and some recent research explored zero-shot text simplification targeting specific reading levels (Farajidizaji et al., 2024; Barayan et al., 2025). Extending this strand of research, the TSAR 2025 Shared Task on Readability-Controlled Text Simplification (Alva-Manchego et al., 2025) aimed to

compare systems that can generate simplified versions of text that conform to a target CEFR reading level, while preserving the meaning. No training data was provided and the test data had 100 texts, each with targeted simplification levels of A2 and B1 (i.e., 200 in total). Reference simplifications were provided after the task completion. AUTORANK (Kocmi et al., 2025) was used to rank the submitted systems.

Approach Summary: Team OneNRC’s approach to this task focused on comparing how far can small language models go in zero-shot readability controlled ATS when they are supported with an agentic workflow and tool support. We used two tools: a) a CEFR prediction tool to help the language model agent verify its generated output and regenerate as needed, and b) an optional second tool to check the similarity between the generated text and the original text. For comparison, we also re-ran the same experiments with a larger proprietary model (Gemini-2.5-flash), and presented an analysis of the differences in the paper.

Motivation: Several small, open-weight language models were released with 1B-20B parameter range over the past year or two, and some of them can even be run for inference on consumer laptops, often in a quantized form. However, it is unclear how useful are they for many NLP tasks as they are not commonly compared with larger models in existing research. This gap motivates OneNRC’s submission to this shared task.

In terms of the performance, one of the runs we submitted (using a 4-bit quantized Gemma3:12B model on a laptop) achieved an AUTORANK of 9.88, among the 40 submitted runs (Alva-Manchego et al., 2025). In this paper, we describe our approach in detail (Section 2), discuss other experiments which achieve better results than the submitted runs, share some qualitative observations (Section 3) along with broad conclusions on learnings through this study. (Section 4).

2 Approach

Our approach primarily concentrated on exploring the power of small language models for this task, under the following zero-shot settings:

1. Plain prompting with local models
2. Prompting in a ReAct (Yao et al., 2022) agent setup using smolagents¹ with or without tool support. Two tools were used: a) a CEFR level prediction tool and b) a meaningBERT similarity score comparison tool - both provided by the organizers as a part of the shared task.
3. Prompting in a ReAct agent setup, but with structured prompt, input and output specification using dspy² (Khattab et al., 2024) instead of free-form prompting used in smolagents. The same two tools were explored in this setting too. Note that dspy supports further inbuilt automatic prompt optimization using few-shot examples, but we did not explore that in this paper.

These three settings focus on incrementally enhancing an LLM’s built in capabilities without fine-tuning. The first setting relies on a hand-crafted prompt, and is potentially subject to more variation in the output depending on even minor changes in prompts. The ReAct approach aims to interleave reasoning and actions performed by the LLM with or without additional tools, and is expected to provide more reliable outputs supported by a reasoning trace. Dspy focuses on structured specification of prompts, inputs, and outputs and emphasizes on a more programmatic interface to interacting with an LLM instead of handcrafted prompts, which further adds another layer of support to the base LLM. Thus, these three experimental settings can be viewed as a sequence of incremental adjustments to LLMs.

Compute and Costs: All experiments were run on a Macbook Pro M1 Pro laptop with a 32 GB RAM, and the costs for running the gemini-2.5-flash experiments amounted to 5 USD. All the implementation code including the used prompts are provided in the github for replication and reproduction.³

¹<https://smolagents.org/>

²<https://dspy.ai/>

³<https://github.com/nishkalavallabhi/tsarst2025-paper>

3 Results

We report results using the three official measures (RMSE, MeaningBERT-Orig, MeaningBERT-Ref) along with the weighted-F1 for CEFR level compliance (which our approach optimized for), and the following three models:

1. A 4-bit quantized Gemma3:12B (Team et al., 2025) model that does not natively offer tool support (which we nevertheless used with tool support).
2. GPT-oss-20B (Agarwal et al., 2025), a model post-trained with quantization of the model’s mixture-of-expert weights to MXFP4 format⁴, and can run natively without additional quantization on a laptop.
3. Gemini-2.5-Flash (Comanici et al., 2025), a large proprietary model as a comparison.

The first two models run on a consumer laptop using ollama⁵. The gemini-2.5-flash model was accessed through API calls with OpenRouter⁶. Our final results are summarized in Table 1.

Model	CEFR F1	RMSE	MB-Orig	MB-Ref
Basic Prompting with local models				
gemma	0.6765	0.5745	0.7514	0.7917
gptoss	0.5657	0.7517	0.8216	0.8308
ReAct Agent, No Tools				
gemma	0.6585	0.6042	0.7648	0.8055
gptoss	0.5612	0.7937	0.8178	0.826
gemini	0.6528	0.6124	0.7702	0.8152
ReAct Agent, CEFR Compliance Tool				
gemma	0.6559	0.6403	0.7304	0.7605
gptoss	0.7903	0.4796	0.795	0.8152
gemini	0.9494	0.2449	0.7689	0.8029
ReAct Agent, two tools				
gemma	0.6676	0.6671	0.7377	0.7616
gptoss	0.7969	0.4796	0.7938	0.8177
gemini	0.917	0.3	0.7749	0.8007
ReAct Agent in DSpy, No Tools				
gemma	0.622	0.6083	0.7466	0.7753
gptoss	0.5766	0.7616	0.8209	0.8308
gemini	0.631	0.6	0.7449	0.8027
ReAct Agent in DSpy, CEFR Compliance Tool				
gemma	0.7204	0.5292	0.7488	0.7844
gptoss	0.6071	0.7937	0.8144	0.8243
gemini	0.9646	0.2	0.7618	0.7997

Table 1: Comparison of results across different approaches

⁴<https://ollama.com/library/gpt-oss>

⁵<https://ollama.com/>

⁶<https://openrouter.ai/google/gemini-2.5-flash>

Discarded models: We initially experimented with a some of the small (quantized) models that support tool use e.g., LLama3.2:1B and 3B, Qwen3:4B and 8B, primarily to start with as small models as possible, but adding tool support or agentic setup resulted in a drastic decline in performance for all models under 12B that we explored. For example, a react agent regenerated the same text for 20 iterations in one case, with the LLama3.2:1B model. So, we did not perform additional experiments with the models under 12B.

3.1 Discussion

Without any tool calling, in a plain prompt based setup, the quantized Gemma3 model is surprisingly better than a larger gpt-oss model in terms of CEFR compliance and there is no substantial difference compared to even a much larger Gemini model when no tools are used in a react agent setup. It is also important to note that there is not much difference between using a ReAct agent without tools versus just prompting the model across different evaluation measures. The local Gemma3 model without ReAct agent even shows slightly better CEFR compliance than the Gemini model in the same setup. While using a tool did not benefit Gemma3 much in a prompt + react agent setup, it did result in a 5% boost in CEFR compliance compared to basic prompt setup, when used with dspy, which could be attributed to the differences in the react prompting setup between smolagents and dspy, and dspy’s focus on a more structured input approach to prompting. Considering that Gemma3 does not natively support tool use, it is interesting to note the performance improvement it could achieve with tool support and structured prompting. This could be further explored in future with few-shot prompt optimization. Gpt-oss and Gemini saw major improvements in overall CEFR compliance after tool support, but this came with a drop in both the meaningBERT similarity scores.

The usage of a second tool to verify meaning similarity did not seem to benefit any of the models, in any settings. Gpt-oss model consistently maintained better meaningBERT scores across all our experimental settings, and was consistently better than Gemma3 model even in CEFR compliance once tool calling was added to plain prompting. Clearly, adding tool calling benefited both in terms of CEFR compliance and meaning preservation.

Best results: Overall, the best result we achieved with a small model in terms of the official evaluation measures is with gpt-oss:20B (RMSE: 0.4796, MeaningBERT-Orig:0.7938, MeaningBERT-Ref:0.8177) in the ReAct+CEFR tool setup with smolagents, and the best result with Gemini-2.5-Flash is achieved in the ReAct+CEFR Tool+Dspy setting (RMSE: 0.2, MeaningBERT-Orig:0.7618, MeaningBERT-Ref:0.7997). For a comparison, the best submitted system amongst all the shared task systems (EHIMENLP) achieved the scores of 0 for RMSE, 0.902 and 0.845 for MeaningBERT-Orig and Ref respectively (Alva-Manchego et al., 2025). While we don’t know the details of this system yet, it is interesting to see that high scores on meaningBERT were maintained despite low RMSE, which clearly was not the case in our experiments. In our experiments, we consistently observed that lower RMSEs are accompanied by lowering of the meaningBERT scores. Additionally, all our approaches had a higher MeaningBERT-Ref score than MeaningBERT-Orig score, unlike the EHIMENLP system.

Qualitative Comparison: To get a feel for how much difference do these approaches bring into the outputs themselves, Table 2 shows an example text and the outputs with the some of the approaches, along with the reference simplification. In this example, interestingly, except one (Gemini+React+2 Tools), all the approaches left out the expansion of the abbreviation. Gemma3 based approach appears to focus on generating short sentences. There is a small amount of hallucination in the gpt-oss model’s output for this example, where "mother" in the original was transformed into "sister" in the simplified version. Only one model seems to miss the detail that that the asteroid is 800 meters wide, which is present in the original and reference texts. Going beyond these minor differences between models, the generated text generally seems to have undergone some form of simplification, and there are substantial similarities with the given reference text.

Table 3 shows an example text and a 4-bit quantized Gemma3 model’s output for the two target CEFR levels-A2 and B1 using plain prompting without any tool support or agentic setup. The output shows that even this basic model is able to distinguish between the two levels to some extent in terms of language use. That indicates that smaller

Original	The first is DART – the Double Asteroid Redirection Test. Plans are scheduled to test DART on the moon of an asteroid called Didymos. 'Didymoon' is 150 metres wide, orbiting its 800-metre mother, and hopefully the impact of DART will knock it out of its orbit enough for Earth-based telescopes to pick up.
Plain Gemma3	There is a project called DART. DART will test something on the moon of an asteroid. The asteroid is called Didymos. Didymoon is part of Didymos. It is about 150 metres wide. Didymoon goes around Didymos, which is bigger - about 800 metres wide. We hope DART will push Didymoon a little bit. Then, telescopes on Earth will be able to see the change.
Gemma3+ Re-Act + CEFR Tool	DART is a test. It will hit the moon of an asteroid called Didymos. The moon is called Didymoon. It is small, 150 meters. The asteroid is bigger, 800 meters. We want to see if DART can change Didymoon's path.
Gpt-oss + Re-Act +CEFR Tool	DART is a mission to change an asteroid. Scientists will test it on the moon of the asteroid Didymos. The moon is called Didymoon. It is 150 metres wide. It goes around its larger sister, an asteroid 800 metres wide. The plan is that DART will hit Didymoon and move it so that telescopes on Earth can see it.
Gemini + Re-Act + CEFR Tool	First, there's DART. It's a test to change an asteroid's path. We will test DART on Didymoon, a moon of an asteroid named Didymos. Didymoon is 150 meters wide. It moves around its parent asteroid. We hope DART will hit Didymoon and change its orbit a little. Then, telescopes on Earth can see this change.
Gemini + Re-Act + 2 Tools	The first thing is DART (Double Asteroid Redirection Test). People want to test DART on a moon of an asteroid called Didymos. This small moon is called 'Didymoon'. It is 150 meters wide. It goes around its mother asteroid, which is 800 meters. We hope DART's hit will move Didymoon a little. Then telescopes on Earth can see it.
Reference	The first project is DART- the Double Asteroid Redirection Test. It will be tested on the moon of an asteroid called Didymos. This moon is called "Didymoon." It is 150 metres wide and goes around its bigger "mother" asteroid, which is 800 metres wide. Scientists hope that when DART hits Didymoon, it will change its path enough for telescopes on Earth to see.

Table 2: Example outputs across different systems for Textid:60-A2)

Original	Cape Town can be a dangerous place for baboons. Some have been injured or killed in car accidents and others have been electrocuted as they played with electric cables. Baboons are an endangered species and to save them and minimise conflicts between humans and baboons, the city council has employed a team of Baboon Monitors. The monitors' job is to chase baboons out of urban areas or catch them and release them in the countryside. This policy has had some success, though it is proving hard to stop baboons from coming back to the city once they see it as a source of easy food.
Plain Gemma3-A2	Baboons can be in danger in Cape Town. Sometimes they get hurt or die in car accidents. Sometimes they get hurt by electricity. Baboons are rare, so we need to protect them. The city has people called Baboon Monitors. Their job is to move baboons away from the city. They take them to the countryside. This helps, but it's hard to stop baboons from coming back because they find food easily in the city.
Plain Gemma3-B1	Baboons can face dangers in Cape Town. Sometimes they are hurt or die in car accidents, or get hurt by electricity when they play near wires. Baboons are rare, and we need to protect them. To help the baboons and avoid problems with people, the city has a team of Baboon Monitors. These monitors try to move baboons away from towns or catch them and take them to the countryside. This has helped a little, but it's difficult to stop baboons from returning to the city if they think they can easily find food there.
Reference	Cape Town can be dangerous for baboons. Some are hurt or die in car accidents. Some die from playing with electric wires. Baboons are animals that are in danger of disappearing. To help them, the city now has a team called Baboon Monitors. These people try to move baboons out of the city and take them back to the countryside. This plan works a little, but the baboons often come back because they find food in the city.

Table 3: Textid:25 in the A2 and B1 versions generated by the 4-bit quantized Gemma3:12B model, just using a simple prompt

language models may hold potential in generating simplifications and perhaps as writing support tools for creating simplified versions of text. Looking at multiple model generations may also be more informative for human writers.

4 Conclusions

Overall, our experiments lead us to three main conclusions:

1. Small language models are capable of doing some readability controlled text simplification

even in a zero-shot setting.

2. Supplying appropriate tool calls is useful, especially for the larger models.
3. There appears to be a tradeoff between achieving the right CEFR level and preserving the meaning.

Evaluation: Evaluation in this shared task primarily relied on automated models, and what appears like an ad-hoc weighting scheme. While a smaller scale human evaluation may be needed to understand the generated texts better, explana-

tory measures using LLM-as-a-Judge approaches or checklist based evaluation approaches (Cook et al.; Mohammadkhani and Beigy, 2025) may provide more informative evaluation for text simplification. Small language models may hold a potential as judges too, providing low-cost, low-carbon footprint options. Finally, manual analysis revealed the possibility of hallucination even in these constrained generation scenarios. As the use of LLMs for text simplification increases, the evaluation measures need to also account for hallucination as a potential dimension.

Future Work: While a laptop environment is not conducive to conduct experiments on a larger scale, we believe future work should focus on exploring the power of smaller models in few-shot settings or fine-tuning using synthetic data generation. Dspy’s automatic prompt optimization capabilities in few shot settings should be better explored for this task, to understand the true potential of smaller models for text simplification. The current shared task focused only on English, but these experiments lead us to hypothesize that there may be some possibilities to replicate these results for at least some of the high resource languages, which needs to be explored. Finally, using language models as support tools to aid the authors of simplified texts also seems to be a useful direction to pursue.

Limitations

We did not venture into few-shot prompting or any other in-context learning strategies in this paper. Further, since everything was ran on a laptop, we only used 4-bit quantized versions of several small models, and their unquantized versions are potentially more powerful. This can be explored further in future, and the current results should be considered under these limitations.

Lay Summary

Automatic Text Simplification (ATS) is the task of translating a text written in a more advanced language into more accessible language. A common approach to solve this problem is to collect large number of examples of sentences and their manually simplified variants (e.g., from some form of sentence level alignment of Wikipedia and Simple Wikipedia articles) and use them as inputs to computational algorithms that are capable of learning patterns from such large amounts of data. The output of this process creates a simplification "model",

which can be used to simplify any given new text. One way to add more nuance to this process is to instruct the whole process to do graded simplification, based on some scale such as the CEFR scale, which is used to describe language proficiency at various levels. However, that would also mean collecting large amounts of examples for each level. Hence, this kind of research was largely restricted to languages where such data is already available or can be collected easily.

The arrival of general purpose Artificial Intelligence based models like ChatGPT gave some flexibility to this approach. We can now describe our problem, potentially give a few examples, and prompt a large language model to simplify a given text, targeting a specific CEFR level. This shared task (where different groups of researchers work on solving a given problem under the same data/evaluation conditions) focuses on CEFR targeted text simplification for English. The approach described in this paper explores whether we really need very large models for this task, asking the question: what can small language models that can run on laptops achieve, if they are given some support in the form of tools that can verify their output and tools that can automatically adapt the human written prompts to suit the language model’s requirements. It turns out, they can put up a strong competition to larger ones, sometimes.

References

- Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, and 1 others. 2025. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*.
- Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2020. Data-driven sentence simplification: Survey and benchmark. *Computational Linguistics*, 46(1):135–187.
- Fernando Alva-Manchego, Regina Stodden, Joseph Marvin Imperial, Abdullah Barayan, Kai North, and Harish Tayyar Madabushi. 2025. Findings of the TSAR 2025 shared task on readability-controlled text simplification. In *Proceedings of the Fourth Workshop on Text Simplification, Accessibility, and Readability (TSAR 2025)*, Suzhou, China. Association for Computational Linguistics.
- Abdullah Barayan, Jose Camacho-Collados, and Fernando Alva-Manchego. 2025. Analysing zero-shot readability-controlled sentence simplification. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6762–6781.

- Gheorghe Comanici, Eric Bieber, Mike Schaeckermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.
- Jonathan Cook, Tim Rocktäschel, Jakob Nicolaus Foerster, Dennis Aumiller, and Alex Wang. Ticking all the boxes: Generated checklists improve llm evaluation and generation. In *Language Gamification-NeurIPS 2024 Workshop*.
- Asma Farajidizaji, Vatsal Raina, and Mark Gales. 2024. Is it possible to modify text to a target readability level? an initial investigation using zero-shot large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9325–9339.
- Tannon Kew, Alison Chi, Laura Vásquez-Rodríguez, Sweta Agrawal, Dennis Aumiller, Fernando Alva-Manchego, and Matthew Shardlow. 2023. Bless: Benchmarking large language models on sentence simplification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13291–13309.
- Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Saiful Haq, Ashutosh Sharma, Thomas T Joshi, Hanna Moazam, Heather Miller, and 1 others. 2024. Dspy: Compiling declarative language model calls into state-of-the-art pipelines. In *The Twelfth International Conference on Learning Representations*.
- Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Natalia Fedorova, Mark Fishel, Markus Freitag, and 1 others. 2025. Preliminary ranking of wmt25 general machine translation systems. *arXiv preprint arXiv:2508.14909*.
- Mohammad Ghiasvand Mohammadkhani and Hamid Beigy. 2025. Checklist engineering empowers multilingual llm judges. *arXiv preprint arXiv:2507.06774*.
- Advaith Siddharthan. 2014. A survey of research on text simplification. *ITL-International Journal of Applied Linguistics*, 165(2):259–298.
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. In *NeurIPS 2022 Foundation Models for Decision Making Workshop*.

GRIPF at TSAR 2025 Shared Task: Towards controlled CEFR level simplification with the help of inter-model interactions

David Alfter

GRIDH

Literature, History of Ideas and Religion
University of Gothenburg, Sweden
david.alfter@gu.se

Sebastian Gombert

Educational Technologies

Information Centre for Education
DIPF, Germany
s.gombert@dipf.de

Abstract

In this contribution to the CEFR level simplification TSAR 2025 Shared Task, we propose two systems, *EZ-SCALAR* and *SAGA*, that implement two differing approaches to prompting LLMs for proficiency-adapted simplification. Our results place us in the middle of the participating teams, and reveal that using external lexical resources to guide simplification improves overall results.

1 Introduction

This paper presents the details of the GRIPF team in the TSAR 2025 Shared Task on CEFR level simplification (Alva-Manchego et al., 2025). The workshop website describes the task as follows: “The task targets English-language paragraphs written at upper-intermediate or advanced levels and requires participants to simplify them according to a specified target readability level, defined using the Common European Framework of Reference for Languages (CEFR). Specifically, participants will be asked to simplify texts originally at B2 level or above to target levels of A1, A2, or B1” (TSAR Workshop).¹

2 Related Work

As Bahrainian et al. (2024) put it, “[t]ext simplification is the process of rewriting a piece of text using simpler vocabulary and grammatical structure in order to make the text more accessible and understandable for a larger audience”.

Text simplification has evolved from rule-based approaches that rely on predefined linguistic transformations to neural methods that learn simplification patterns from data. Early work focused on lexical substitution and syntactic restructuring using hand-crafted rules, while recent advances have

leveraged transformer-based models for end-to-end simplification. Zetsu et al. (2024) demonstrate that lexically constrained decoding with edit operations can effectively guide neural simplification models, addressing the challenge of loose constraints in previous approaches.

Controllable text generation has emerged as a particularly relevant area, where models are guided to produce outputs meeting specific criteria such as reading level or target audience. Crippwell et al. (2023) propose document-level planning approaches that decompose simplification into discrete operations (copy, rephrase, split, or delete), showing that structured planning can guide generation more effectively than end-to-end approaches. Recent work by Maddela and Alva-Manchego (2025) addresses the challenge of evaluating document-level simplification, proposing methods to adapt sentence-level metrics for longer texts. Within this landscape, CEFR-level simplification represents a specialized challenge, as it requires not only reducing complexity but doing so in alignment with established language proficiency standards. Ryan et al. (2023) highlight the particular challenges of multilingual simplification, while Horiguchi et al. (2025) extend this work by developing parallel corpora for the simplification of medical texts in nine languages, demonstrating that multilingual training can improve the performance of simplification.

Ensemble methods in natural language processing have demonstrated consistent improvements over single-model approaches across various tasks. The principle of combining multiple models to leverage their complementary strengths has been successfully applied to machine translation, text generation, and classification tasks. Recent work by Valiev and Tutubalina (2024) shows that in-prompt ensemble methods, where multiple model predictions are integrated as separate expert solutions with trust scores, can achieve strong per-

¹While the webpage specifies the target level A1, the actual task only targets A2 and B1.

formance in specialized domains. Parfenova and Pfeffer (2025) demonstrate that smaller ensemble models with refined outputs can consistently outperform individual models and even large-scale LLMs, suggesting that ensemble approaches may be more effective than relying solely on large standalone models. Niess and Kern (2025) extend ensemble methods to watermarking applications, showing that multi-feature ensemble approaches achieve 98% detection rates and maintain robustness against paraphrasing attacks.

Our work extends this concept to controllable simplification, where inter-model critique and refinement can potentially address the limitations of individual models. This approach aligns with recent trends in self-correction and iterative improvement in language models, where systems refine their output through multiple generation cycles. Pan et al. (2024) provide a comprehensive survey of automated correction strategies, categorizing approaches into training-time, generation-time, and post-hoc methods, while Ferraz et al. (2024) demonstrate that decompose-critique-refine pipelines can significantly improve constraint following in language models. However, recent critical surveys suggest that the effectiveness of self-correction varies significantly across tasks and model types (Kamoi et al., 2024), highlighting the importance of ensemble approaches that can leverage external critique mechanisms rather than relying solely on self-evaluation.

3 Methodology

We use two similar approaches to tackle the problem of lexical simplification. The first system is called EZ-SCALAR (Ensemble Zero-shot Simplify, Criticize, Adapt with Lexical Assessment, and Referee). The second system is called SAGA (Self-Assessed Guided Adaptation). Both systems estimate complex vocabulary, either using external resources or by prompting the model. Both models adjust the simplification based on the identified complex vocabulary.

3.1 EZ-SCALAR

EZ-SCALAR uses two state-of-the-art LLMs: GPT-5 and Claude Opus 4.1. In the first step, both models are prompted to simplify the text to the target level. In the second step, each model receives the simplified text of the other model and is asked to critique the simplification given the target

level. The first stage generates independent simplifications, allowing each model to apply its learned patterns without bias from the other. The critique stage is crucial for identifying potential issues that individual models might miss, as different architectures often exhibit distinct biases. In the third step, each model receives the other model’s critique and is asked to refine its text based on the critique; the refinement stage incorporates these insights, potentially producing outputs superior to either model alone. In the last step, each model receives the original text, the two simplifications, and the target level, and is asked to choose the text that best fits the target level in terms of **meaning preservation**, **readability** and **clarity**. In case the models disagree, a third model – Llama-3.2-3B-Instruct – is used as judge.

3.2 EZ-SCALAR Lex

As an extension of the base system, we also experiment with using external lexical resources. The system is identical to EZ-SCALAR, but between the second and third step, a lexical assessment module is implemented. Figure 1 in the Appendix illustrates both architectures. This module uses the EFLLex vocabulary list (Dürlich and François, 2018) to estimate the complexity of vocabulary and addresses a key limitation of purely neural approaches: the lack of explicit control over vocabulary complexity. By integrating the EFLLex vocabulary list, the system bridges rule-based and neural methods, providing concrete guidance about problematic words. Since the EFLLex list contains words with their part-of-speech and frequency distributions over CEFR levels, we use the first-occurrence approach to map each (word, pos) tuple to its target level. This means that each word is linked to a distribution of frequencies over levels (how often was this word used at A1, A2, etc.), and the level at which the frequency first becomes non-zero is taken as target level. Since this method of level assignment is brittle when it comes to low frequencies, all levels above A1 are shifted downwards one level. The decision to shift CEFR levels downward reflects practical considerations about the reliability of frequency-based level assignments, acknowledging that conservative estimates are preferable to overly optimistic ones in educational contexts. The module uses stanza to lemmatize and pos-tag. Since the pos-tags are different, we map the lexical tags *nouns*, *verbs*, *adjectives* and *adverbs* to a common tag set (the tag

set used by EFLLex).² The module takes as input the simplified text and target level and identifies words which are above the target level. These words are then added to the prompt for the third step, along with a prompt to pay special attention to these words.

3.3 SAGA

SAGA uses two components, which we refer to as the *proposer* and the *reviewer*. Given a text and a desired CEFR level, the proposer generates an adjusted version of the text that supposedly satisfies that CEFR level. The reviewer then classifies whether the proposed version of the text actually meets the required level. In case the *reviewer* predicts that the text is indeed satisfying the level, we accept. Otherwise, the *proposer* needs to regenerate a text based on the previously generated version, which is then again judged by the *reviewer*. This process is repeated until an agreement is reached or a maximum number of iterations is reached. Figure 2 in the Appendix illustrates the SAGA architecture.

For our submission, the *proposer* is implemented with *GPT-4o* via the conversational API. Given, a custom system prompt (see appendix), the model is first prompted to identify words and phrases within the given input text that are not appropriate for the desired CEFR level and also to provide more appropriate alternatives along with the identified words and phrases. In the second step, the model is then prompted to reformulate the input text by changing the respective phrases and using the proposed alternatives instead.

The resulting text is then given to the *reviewer*, which, in our case is implemented in the form of a *ModernBERT-large* (Warner et al., 2025) classification model fine-tuned on a corpus of English texts labelled for their CEFR level³ (1x5 cross-validation resulted in a weighted F1 of 0.689). If the *reviewer* detects a different CEFR level than expected, the *proposer* is prompted to identify additional words and phrases that could have brought the reviewer to this judgement, again with suitable alternatives. Following this, it is again prompted to reformulate the text accordingly. For the shared task submission, we set the maximum iterations to 5.

²We surmise that marking interjections and other closed-class part-of-speech may be unnecessary and introduce unnecessary noise in the process.

³<https://www.kaggle.com/datasets/amontgomerie/cefr-levelled-english-texts>

4 Evaluation

4.1 Pre-Evaluation

For pre-evaluating our systems, we used the provided trial data set and the official evaluation script. The evaluation script outputs different measures, organized into three overarching parts, each with sub-measures:

1. CEFR Compliance: how well does the text adhere to the specified CEFR level; this compliance is checked by running the simplified text through a model trained on sentence-label CEFR data
 - Weighted F1: This is a combination of precision and recall, averaged by class and weighted by the number of actual occurrences of each class in the dataset (*weighted_f1* in Table 1). It was dropped in the final evaluation because it does not capture the severity of misclassification.
 - Adjacent accuracy: This measure is similar to accuracy, but it counts a prediction as correct if it corresponds either directly to the gold label or is either one above or below the gold label (e.g., if the gold label is B1, and the prediction is A2, it is counted as correct; *adj_accuracy* in Table 1). It was dropped in the final evaluation because it is less informative than RMSE.
 - Root Mean Squared Error (RMSE): RMSE is the quadratic mean of the differences between the observed values and predicted ones (*rmse* in Table 1)
2. Meaning Preservation: how well does the simplification preserve the meaning of the original unsimplified text
 - MeaningBERT-Orig: measures semantic similarity between the original text and the simplified version using a model specifically trained on human annotations for meaning preservation in simplification
 - BERTScore-Orig: measures semantic similarity between the original text and the simplified version. However, as this model tends to “overestimate similarity when there is lexical overlap with no true meaning preservation” (Alva-Manchego

Metric	Sub-Metric	EZ-SCALAR	EZ-SCALAR Lex	SAGA
CEFR Compliance	weighted_f1 ↑	0.533	0.567	0.530
	adj_accuracy ↑	0.975	0.975	0.975
	rmse ↓	0.758	0.742	0.742
Meaning Preservation	MeaningBERT-Orig ↑	0.852	0.860	0.790
	BERTScore-Orig ↑	0.948	0.951	0.951
Similarity to References	MeaningBERT-Ref ↑	0.803	0.794	0.738
	BERTScore-Ref ↑	0.926	0.925	0.920

Table 1: Results on trial data. ↑ means that a higher score (maximum 1) is better. ↓ means that a lower score is better (minimum 0)

System	RMSE ↓	M-BERT-Orig ↑	M-BERT-Ref ↑	AvgScore ↑	AUTORANK ↓
EZ-SCALAR	0.721	0.856	0.824	0.060	8.270
EZ-SCALAR Lex	0.689	0.857	0.820	0.070	8.130
SAGA	0.831	0.827	0.796	-0.140	10.780

Table 2: Results on test data. Best results per test measure in bold.

et al., 2025), it was dropped for the final evaluation

3. Similarity to References:

- MeaningBERT-Ref: same as above, except between a human reference annotation and the system simplification
- BERTScore-Ref: same as above

Our results are summarized in Table 1. However, as the number of data points in the trial data set was limited ($n = 40$), the results are merely hints at what methods might work better.

As can be gathered from the Table, all systems perform similarly well, with no system clearly dominating. Note that *adjacent accuracy* is meaningless since there were only two target classes.

4.2 Post-Evaluation

Table 2 shows our results on the test data. In general, of 20 participating teams, we placed 10. As can be gathered from the table, the lexically guided variant of EZ-SCALAR outperforms the other two methods in almost all cases and is only outdone by EZ-SCALAR in the MeaningBERT-Ref score, albeit not by much.

5 Discussion

The superior performance of EZ-SCALAR Lex (AvgScore: 0.070) over regular EZ-SCALAR

(0.060) demonstrates the value of explicit lexical guidance in controlled simplification tasks. This improvement appears primarily in the RMSE metric (0.689 vs 0.721), suggesting that the lexical assessment helps achieve a more accurate CEFR level targeting. The marginal difference in meaning preservation scores (M-BERT-Orig: 0.857 vs. 0.856) indicates that lexical guidance improves level compliance without significantly compromising semantic fidelity.

SAGA’s performance profile reveals interesting trade-offs in the system design space. Although it achieves the highest BERTScore-Orig (0.951), indicating a strong preservation of surface-level similarity to the original text, it shows a weaker performance in CEFR compliance (RMSE: 0.831) and reference similarity. This suggests that the iterative proposer-reviewer approach may prioritize conservative modifications that maintain original phrasing over more substantial restructuring needed for effective simplification.

The ranking across metrics illuminates the inherent tensions in text simplification. Systems that aggressively modify text to achieve target readability levels may sacrifice meaning preservation, while those that prioritize semantic fidelity may fail to achieve sufficient simplification. The ensemble approaches appear to navigate these trade-offs more successfully than the single iterative method, possibly because of their ability to consider multiple

perspectives during the simplification process.

6 Conclusion

We presented two systems for proficiency-targeted simplification of texts and show that lexical resources can help guide simplification systems to achieve simplifications that better adhere to target levels.

The comparative analysis of our three system variants reveals important insights about the architecture of controlled text simplification systems. The superior performance of EZ-SCALAR Lex demonstrates that hybrid approaches combining neural language models with established linguistic resources can outperform purely neural methods. This finding aligns with broader trends in natural language processing, where the integration of explicit knowledge representations with statistical learning has proven valuable across multiple tasks. The ensemble architecture employed in EZ-SCALAR, which leverages multiple models through critique and refinement cycles, appears more robust than the iterative single-model approach of SAGA, suggesting that diversity in simplification strategies contributes more to quality than repeated self-correction. However, the modest performance gains and our mid-table ranking among participating teams indicate substantial room for improvement. Future work should explore more sophisticated integration of lexical resources, potentially incorporating syntactic complexity measures alongside vocabulary targeting, and investigate whether larger-scale ensemble approaches or fine-tuned models specifically trained on CEFR-leveled data could further enhance performance while maintaining the interpretability and controllability that explicit lexical guidance provides.

Limitations

The current approach faces several methodological limitations that constrain its generalizability and practical deployment. The reliance on proprietary large language models introduces both cost and availability concerns, while the dependence on external vocabulary lists like EFLLex assumes the availability of domain-appropriate complexity assessments. The evaluation framework itself presents challenges, as CEFR level assessment remains somewhat subjective even with established guidelines. The limited scale of evaluation data

($n = 40$ for trial data) restricts the statistical power of our comparisons and may not capture the full range of simplification challenges across different text types and domains. Educational texts, news articles, and technical documents each present unique simplification requirements that may not be adequately represented in small-scale evaluations. Furthermore, the automatic evaluation metrics, while useful for comparison, may not fully capture the pedagogical effectiveness of simplified texts for language learners.

Lay Summary

When learning a new language, students benefit from reading texts that match their skill level. The Common European Framework of Reference for Languages (CEFR) defines six levels of language proficiency, from A1 (beginner) to C2 (mastery). However, most of the written content is too complex for students at lower levels. Text simplification involves rewriting passages using simpler words and sentence structures while keeping the original meaning intact.

Can artificial intelligence systems automatically simplify texts to specific CEFR levels? We participated in a competition where we were asked to simplify English paragraphs written at upper intermediate levels (B2 or above) to beginner or elementary levels (A2 or B1). We developed two different systems, called EZ-SCALAR and SAGA. Both systems use large language models (advanced AI programs trained on vast amounts of text) to identify complex vocabulary and rewrite passages more simply.

EZ-SCALAR works by having two different AI models independently simplify the same text, critique each other's work, and refine their versions based on the feedback. The enhanced version, EZ-SCALAR Lex, adds an extra step that uses a specialized vocabulary list (EFLLex) to identify words that are too advanced for the target level.

SAGA uses a different approach with two components: a *proposer* that creates simplified versions and a *reviewer* that checks whether the text meets the target level. If not, the proposer tries again, repeating this cycle until the reviewer approves or a maximum number of attempts is reached.

Our tests showed that EZ-SCALAR Lex performed best among the three systems, demonstrating that the use of specialized vocabulary resources to guide the simplification process produces better

results. The system that relied solely on AI models without external vocabulary guidance (SAGA) showed the weakest performance.

These findings could help language teachers, educational publishers, and online learning platforms automatically adapt reading materials for students at different proficiency levels. The research suggests that combining AI capabilities with structured vocabulary resources produces more reliable simplifications than AI alone. However, further development would be needed before these systems could replace human editors in creating learning materials, as the evaluation was limited in scope and automatic quality measures may not fully capture how helpful simplified texts are for actual learners.

References

- Fernando Alva-Manchego, Regina Stodden, Joseph Marvin Imperial, Abdullah Barayan, Kai North, and Harish Tayyar Madabushi. 2025. Findings of the TSAR 2025 shared task on readability-controlled text simplification. In *Proceedings of the Fourth Workshop on Text Simplification, Accessibility, and Readability (TSAR 2025)*, Suzhou, China. Association for Computational Linguistics.
- Seyed Ali Bahrainian, Jonathan Dou, and Carsten Eickhoff. 2024. Text simplification via adaptive teaching. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6574–6584, Bangkok, Thailand. Association for Computational Linguistics.
- Liam Cripwell, Joël Legrand, and Claire Gardent. 2023. Document-level planning for text simplification. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 993–1006, Dubrovnik, Croatia. Association for Computational Linguistics.
- Luise Dürlich and Thomas François. 2018. EFLLex: A graded lexical resource for learners of English as a foreign language. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.
- Thomas Palmeira Ferraz, Kartik Mehta, Yu-Hsiang Lin, Haw-Shiuan Chang, Shereen Oraby, Sijia Liu, Vivek Subramanian, Tagyoung Chung, Mohit Bansal, and Nanyun Peng. 2024. LLM self-correction with DeCRIM: Decompose, critique, and refine for enhanced following of instructions with multiple constraints. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7773–7812, Miami, Florida, USA. Association for Computational Linguistics.
- Koki Horiguchi, Tomoyuki Kajiwara, Takashi Nomiya, Shoko Wakamiya, and Eiji Aramaki. 2025. MultiMSD: A corpus for multilingual medical text simplification from online medical references. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 9248–9258, Vienna, Austria. Association for Computational Linguistics.
- Ryo Kamoi, Yusen Zhang, Nan Zhang, Jiawei Han, and Rui Zhang. 2024. When can llms actually correct their own mistakes? a critical survey of self-correction of llms. *Transactions of the Association for Computational Linguistics*, 12:1417–1440.
- Mounica Maddela and Fernando Alva-Manchego. 2025. Adapting sentence-level automatic metrics for document-level simplification evaluation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6444–6459, Albuquerque, New Mexico. Association for Computational Linguistics.
- Georg Niess and Roman Kern. 2025. Ensemble watermarks for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2903–2916, Vienna, Austria. Association for Computational Linguistics.
- Liangming Pan, Michael Saxon, Wenda Xu, Deepak Nathani, Xinyi Wang, and William Yang Wang. 2024. Automatically correcting large language models: Surveying the landscape of diverse automated correction strategies. *Transactions of the Association for Computational Linguistics*, 12:484–506.
- Angelina Parfenova and Jürgen Pfeffer. 2025. Measuring what matters: Evaluating ensemble LLMs with label refinement in inductive coding. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10803–10816, Vienna, Austria. Association for Computational Linguistics.
- Michael J Ryan, Tarek Naous, and Wei Xu. 2023. Revisiting non-English text simplification: A unified multilingual benchmark. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4898–4927, Toronto, Canada. Association for Computational Linguistics.
- TSAR Workshop. Shared task on readability-controlled text simplification. <https://tsar-workshop.github.io/shared-task/>. Accessed: 2025-09-19.
- Airat Valiev and Elena Tutubalina. 2024. HSE NLP team at MEDIQA-CORR 2024 task: In-prompt ensemble with entities and knowledge graph for medical error correction. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 470–482, Mexico City, Mexico. Association for Computational Linguistics.
- Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom

Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. [Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference.](#) In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547, Vienna, Austria. Association for Computational Linguistics.

Tatsuya Zetsu, Yuki Arase, and Tomoyuki Kajiwara. 2024. [Edit-constrained decoding for sentence simplification.](#) In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7161–7173, Miami, Florida, USA. Association for Computational Linguistics.

A Schematics

A.1 EZ-SCALAR

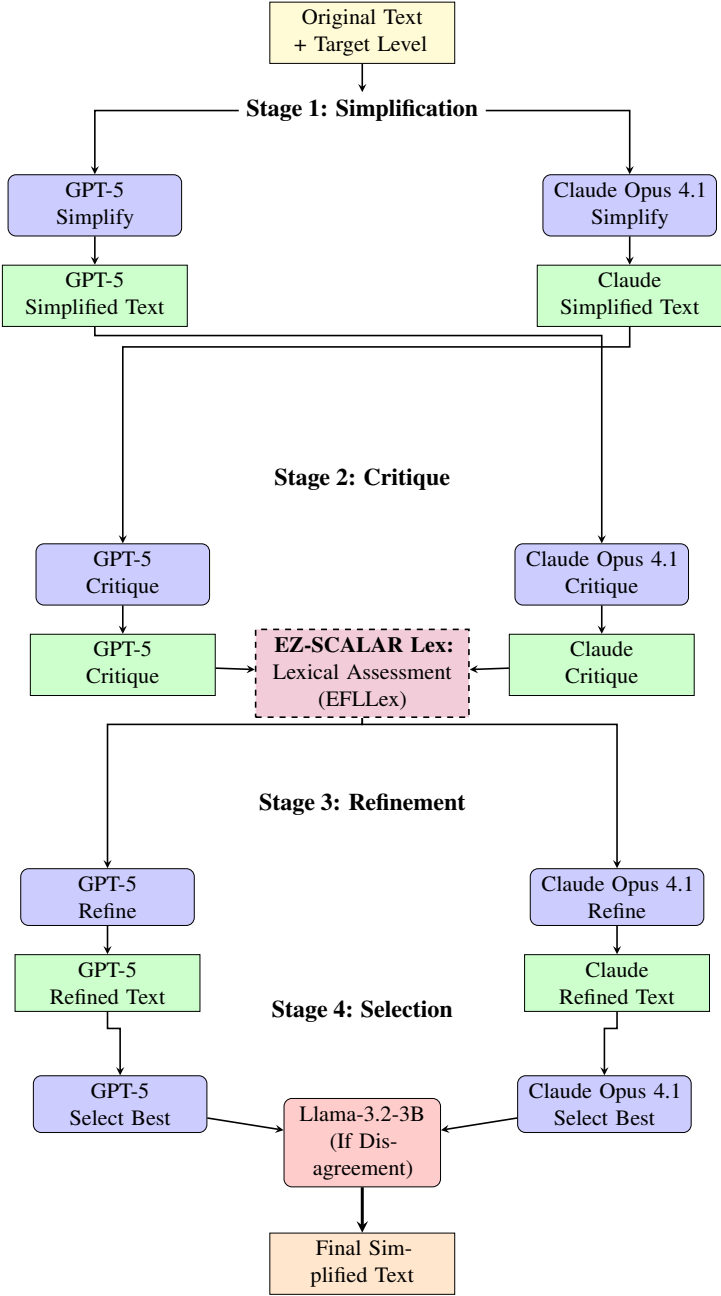


Figure 1: Visual representation of EZ-SCALAR

A.2 SAGA

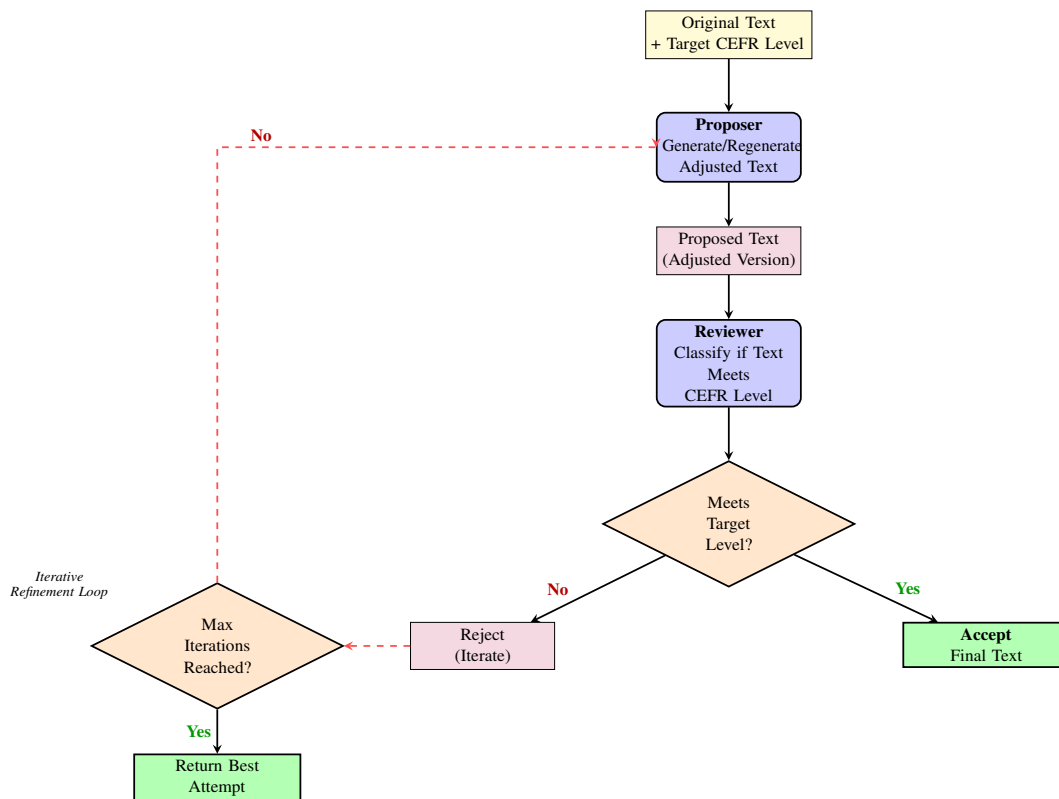


Figure 2: Visual representation of SAGA

B EZ-SCALAR Prompts

Note that all prompts come in pairs of *X system prompt* and *X prompt*. Note further that there are two sets of *judge* prompts. The first set is used with GPT-5 and Claude Opus 4.1, while the second set is a simplified version used with Llama-3.2-3B-Instruct.

B.1 Simplification system prompt

You are an expert lexical simplification and text summarization AI. Your task is to rewrite a given text to make it more accessible and easier to understand for a specific target audience.

Core Rules:

1. **Preserve Meaning:** Your primary objective is to maintain the original meaning, intent, and tone of the text. Do not introduce new information or omit crucial details.
2. **Target Audience:** The output must be tailored for an audience with a proficiency level of PROFICIENCY LEVEL. This means simplifying vocabulary and sentence structure while avoiding jargon, idioms, or overly complex syntax.
3. **Simplify, Don't Trivialise:** The goal is to make the text simpler, not to make it childish or unprofessional. Maintain a natural, adult-appropriate tone.
4. **Sentence and Paragraph Structure:** Feel free to rephrase and restructure sentences. You may break down long sentences into shorter ones or combine simple sentences for better flow, as long as the overall meaning is preserved.
5. **Direct Output:** Provide the simplified text directly. Do not include any preambles, explanations, or conversational filler.

B.2 Simplification prompt

Original Text: ORIGINAL TEXT

Target Proficiency Level: PROFICIENCY LEVEL

Proficiency Guide: PROFICIENCY GUIDE

Simplify the text above according to the proficiency level and the provided guide.

B.3 Critique system prompt

You are an expert editor specialized in simplifying complex texts for a specific audience. Your sole purpose is to critique a simplified text by comparing it to the original. You are not to rewrite or provide a new version. Your job is to act as a highly critical reviewer.

Your critique must focus on three key areas:

1. **Accuracy and Meaning Preservation:** Has the simplification changed or distorted the original meaning? Identify any instances where information was lost, added, or twisted.
2. **Readability and Target Audience Suitability:** Is the text truly accessible for a PROFICIENCY LEVEL audience? Point out any remaining complex vocabulary, convoluted sentences, or awkward phrasing that still requires simplification.
3. **Flow and Grammar:** Did the simplification introduce any grammatical errors or make the text less coherent? Highlight any parts that feel unnatural or don't flow well.

Your output must be a concise, bulleted list of specific issues. For each point, quote the problematic part of the simplified text and briefly explain why it's an issue. Be direct and avoid conversational filler.

B.4 Critique prompt

Original Text: ORIGINAL TEXT

Simplified Text to Critique: SIMPLIFIED TEXT

Target Proficiency Level: PROFICIENCY LEVEL

Critique the "Simplified Text to Critique" by comparing it to the "Original Text." Focus on identifying specific issues related to accuracy, readability, and grammar. Do not rewrite the text.

B.5 Revision system prompt

You are an expert editor. Your task is to perform targeted revisions on a simplified text based on a provided critique. You must only make changes that directly address the issues listed.

Your Instructions:

1. **Reference the Critique:** Use the specific issues from the provided critique as your sole guide for revision.
2. **Make Minimal Changes:** Do not rewrite or rephrase sentences that are not flagged as problematic. Focus your edits narrowly on the identified errors or areas for improvement.
3. **Preserve the Core:** The revised text must maintain the core simplification work that has already been done. Your goal is to fix flaws, not to start over.
4. **Provide the Full, Revised Text:** After making the necessary changes, provide the complete, revised version of the text. Do not provide a list of changes or an explanation of your edits.

Input Format:

You will be given the original text, the simplified version, and a critique in the following format:

Original Text: [The full original text]

Simplified Text: [The full simplified text]

Critique:

- Issue 1, with a quote and explanation
- Issue 2, with a quote and explanation

B.6 Revision prompt

Original Text: ORIGINAL TEXT

Simplified Text to Revise: SIMPLIFIED TEXT

Critique: CRITIQUE

Pay special attention to these words that were flagged as potentially being of a too high level: WORDS FLAGGED AS TOO HIGH LEVEL

Based on the provided critique, make only the necessary revisions to the “Simplified Text to Revise.” Provide the full revised text as your final output.

B.7 Judge system prompt

You are an expert judge for text simplification. Your task is to objectively evaluate two simplified versions of a text and select the superior one. You must make your decision based on a rigorous comparison to the original text and a set of explicit criteria.

Your Goal:

Choose the single best version between **Simplified Version A** and **Simplified Version B**.

Evaluation Criteria:

1. **Meaning Preservation:** Which version more accurately and completely preserves the meaning, tone, and intent of the original text?
2. **Readability:** Which version is more accessible and easier to read for the audience? This includes considering vocabulary, sentence complexity, and overall flow.
3. **Clarity:** Which version is clearer and less ambiguous? Does either version introduce any new errors or awkward phrasing not present in the original?

Your Output:

First, provide a brief, one-paragraph explanation of your decision. Explain which version you chose and why it is better, referencing the criteria above.

Second, state your final choice clearly.

Example Output Format:

Decision: [A one-paragraph explanation of your choice, mentioning which version you chose and how it excelled based on the criteria.]

Final Choice: [A or B.]

B.8 Judge system prompt 2

You are an expert judge for text simplification. Your task is to objectively evaluate two simplified versions of a text and select the superior one. You must make your decision based on a rigorous comparison to the original text and a set of explicit criteria.

Your Goal:

Choose the single best version between **Simplified Version A** and **Simplified Version B**.

Evaluation Criteria:

1. **Meaning Preservation:** Which version more accurately and completely preserves the meaning, tone, and intent of the original text?
2. **Readability:** Which version is more accessible and easier to read for the audience? This includes considering vocabulary, sentence complexity, and overall flow.
3. **Clarity:** Which version is clearer and less ambiguous? Does either version introduce any new errors or awkward phrasing not present in the original?

Your Output:

Output ONLY the winning version letter: A or B.

B.9 Judge prompt

Original Text: ORIGINAL TEXT

Simplified Version A: SIMPLIFIED TEXT A

Simplified Version B: SIMPLIFIED TEXT B

Target Proficiency Level: PROFICIENCY LEVEL

Evaluate and choose the superior version between “Simplified Version A” and “Simplified Version B” based on meaning preservation, readability, and clarity for the target proficiency level. Provide a brief explanation for your choice and then state your final selection clearly.

B.10 Judge prompt 2

Original Text: ORIGINAL TEXT

Simplified Version A: SIMPLIFIED TEXT A

Simplified Version B: SIMPLIFIED TEXT B

Target Proficiency Level: PROFICIENCY LEVEL

Evaluate and choose the superior version between “Simplified Version A” and “Simplified Version B” for the target proficiency level. Return only your final selection clearly.

C SAGA Prompting

C.1 System Prompt

You are an expert on language learning, language simplification and the CEFR framework. Given a text and a desired CEFR level, you are very proficient in identifying which parts of the text do not match the expected CEFR level.

C.2 Word and Passage Identification

User: Identify and list all words or passages in the given text that do not match the expected CEFR level and provide alternatives on the desired CEFR level that convey the same meaning. Don't provide a rewritten version yet, but just the list of problematic words and passages. Exclude names (people, cities, organisations, ...) from this list. Never state that a text is already at an appropriate level.

Assistant: List of words and passages that do not match the expected CEFR level with suggested changes:

C.3 Text Rewriting

User: Now that you identified the problematic words and passages, provide a rewritten version of the text where you changed all occurrences of all listed words and passages with the suggested replacements, while keeping sure that the text remains grammatically correct. Also make sure that each replacement conveys the same core semantic meaning.

Assistant: Rewritten Text:

C.4 Reviewer Critique

User: I have consulted another expert on the topic. The expert determined the CEFR level of the rewritten to be the following: {pred_level}. Please list all words and passages that likely have brought the expert to this judgement of the rewritten text and provide alternatives on the desired CEFR level. Exclude names (people, cities, organisations, ...) from this list. Never state that a text is already at an appropriate level.

Assistant: List of words and passages in the rewritten (not the original) text that could have brought the expert to his judgement with suggested changes:

ITU NLP at TSAR 2025 Shared Task: A Three-Stage Prompting Approach for CEFR-Oriented Text Simplification

Kutay Arda Dinç
Istanbul Technical University
Istanbul, Türkiye
dinck20@itu.edu.tr

Fatih Bektaş
Istanbul Technical University
Istanbul, Türkiye
bektas18@itu.edu.tr

Gülşen Eryiğit
Istanbul Technical University
Istanbul, Türkiye
gulsenc@itu.edu.tr

Abstract

Automatic Text Simplification (TS) makes complex texts more accessible but often lacks control over target readability levels. We propose a lightweight, prompt-based approach to English TS that explicitly aligns outputs with CEFR proficiency standards. Our method employs a three-stage pipeline, guided by rule-informed prompts inspired by expert strategies. In the TSAR 2025 Shared Task, our system achieved competitive performance, with stronger results at B1 level and challenges at A2 level due to over-simplification. These findings highlight the promise of prompt-based CEFR-oriented simplification and the need for more flexible constraint design.

1 Introduction

Automatic Text Simplification (TS) is an important NLP task that aims to rewrite complex texts into simpler forms while preserving meaning. TS benefits many groups, including non-native speakers, children, and individuals with reading or cognitive difficulties. However, current systems often simplify texts without controlling the target readability level, limiting their usefulness in education and accessibility, where alignment with standardized frameworks such as the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001) is essential.

We propose a CEFR-oriented method for English text simplification that leverages Large Language Models (LLMs) through prompt engineering rather than model training, making it lightweight, accessible, and practical. Our method is inspired by the TS strategies introduced in Özmen and Kiran (2022). It applies a three-stage process: syntactic simplification, lexical simplification, and elaboration. This method addresses a key gap in current NLP research by combining accessibility with level-appropriate simplification.

2 Related Work

CEFR is widely used to define language proficiency and guide text simplification. Most CEFR-based TS research targets lexical or sentence-level simplification, creating datasets for word substitutions or training models to generate CEFR-aligned sentences (Arase et al., 2022; Uchida et al., 2018; Li et al., 2025; Barayan et al., 2024). However, simplification at the document or paragraph level under CEFR control remains relatively underexplored and challenging.

Early text simplification research largely focused on sentence-level approaches. However, applying these methods iteratively across entire documents often fails to maintain discourse coherence and overall text integrity (Alva-Manchego et al., 2020). This limitation has encouraged research to shift towards document-level simplification, which demands broader contextual understanding and specialized evaluation methods (Sun et al., 2021).

LLMs have recently become prominent tools for document-level text simplification, addressing coherence and context issues. Approaches such as ProgDS (Fang et al., 2025) decompose the simplification process into hierarchical stages (discourse-level, topic-level, and lexical-level), guiding LLMs to progressively simplify documents while preserving logical structure and readability. Building on this line of work, our study also adopts a hierarchical paradigm, though the definition and scope of the stages differ from ProgDS. In parallel, fine-tuning of LLMs has been shown to yield significant improvements in document-level text simplification tasks, effectively improving model accuracy and output relevance (Nozza and Attanasio, 2023; Alkaldi and Inkpen, 2023; Anshütz et al., 2023). Unfortunately, none of these approaches are explicitly oriented toward CEFR-based simplification.

3 Method

Our method of simplification consists of three stages that are carried out in a progressive order: syntactic simplification, lexical simplification, and elaboration.

Syntactic simplification involves restructuring complex grammatical constructions to align with the requirements of the target CEFR level. Lexical simplification focuses on replacing difficult vocabulary with simpler alternatives appropriate for the target proficiency level. Elaboration involves adding explanatory words or clarifying information to make the text more comprehensible and suited to the intended CEFR level. In [Özmen \(2019\)](#), case-by-case rules were provided for all three types of simplifications at different proficiency levels, along with examples demonstrating how to execute these different simplification cases. Although the examples were presented in Turkish, the underlying simplification principles are not language-specific and reflect patterns observed across multiple languages ([Özmen and Kiran, 2022](#)). Building on this foundation, [Bektaş et al. \(2024\)](#) incorporated these case-specific rules into Turkish prompts and reported promising results. Informed by both studies, we follow these rules when designing prompts and apply them sequentially in three stages¹, in a manner inspired by [Fang et al. \(2025\)](#).

All stages utilize **GPT-4o** with the following parameters:

- **Model version:** gpt-4o-2024-11-20
- **Temperature:** 0.5
- **Top-p sampling:** 0.95

3.1 Syntactic Simplification

The first stage addresses the grammatical complexity by aligning the syntactic structures with the requirements of the target CEFR level. To implement this stage, we first compiled a comprehensive set of grammatical structures from the Cambridge English Grammar Profile ([Cambridge University Press, 2024](#)), which documents grammatical constructions across levels A1 to C2.

Then, the syntactic simplification rules from [Özmen \(2019\)](#) were translated to English, along with illustrative examples showing how complex structures should be simplified for each target level.

¹The full workflow, including all prompts used at each stage, is available at https://github.com/kutayardadinc/itunlp_tsar.

The syntactic simplification process employs a structured prompt that includes:

- i) A system instruction defining the model’s role as a syntactic simplification expert,
- ii) The complete set of permissible grammatical structures for the target CEFR level,
- iii) Specific syntactic simplification rules with examples.

The core prompt structure follows this pattern:

A text at the [LEVEL] level should not contain any grammatical structures that are not provided to you below. If the given text contains any structure other than these grammatical structures, simplify the given text to the [LEVEL] level using syntactic simplification rules. [GRAMMAR_STRUCTURES]. [SIMPLIFICATION_RULES]. Do not perform any action other than syntactic simplification. Preserve the content and details of the text. Do not delete sentences from the text.

3.2 Lexical Simplification

For lexical simplification, we first established vocabulary inventories for each CEFR level to be able to identify words that exceed the target proficiency level. We utilized two primary vocabulary resources: Oxford 3000 ([Oxford University Press, a](#)) and Oxford 5000 ([Oxford University Press, b](#)) word lists. These two combined create a vocabulary list of 5000 words from levels A1 to C1.

Our system automatically identifies lexically complex words within the syntactically simplified text by tokenizing it using spaCy ([Honnibal et al., 2020](#)) and checking both surface forms and lemmas against our compiled word-to-CEFR-level mapping database. Words that exceed the target CEFR level are explicitly included in the prompt, with instructions to replace them with simpler alternatives. These are followed by level-specific lexical simplification rules, accompanied by translated examples adapted from [Özmen \(2019\)](#).

The lexical simplification prompt structure includes:

- i) A system instruction defining the model’s role as a lexical simplification expert,
- ii) Level-specific lexical simplification rules,
- iii) Identified words that exceed the target level.

The core prompt structure follows this pattern:

Simplify the given text to the [LEVEL] level using lexical simplification rules. [SIMPLIFICATION_RULES]. In the following text, the word(s) [IDENTIFIED_WORDS] are above the [LEVEL] level. Without changing the meaning of the text, replace these words with simpler words appropriate for the [LEVEL] level. Do not change any words that have no simpler equivalent or are important for the meaning of the text.

3.3 Elaboration

The final elaboration stage enhances the comprehensibility of the text by adding explanatory information and clarifying potentially ambiguous content for the target level. This stage applies elaboration rules, translated from Özmen (2019), specific to each CEFR level. At this stage, the model is instructed to expand rare or complex concepts with explanatory phrases, clarify implicit meanings by adding context, and highlight key ideas through selective repetition when appropriate.

The elaboration employs the prompt below:

Using elaboration rules, make the given text more understandable for a student at the [LEVEL] level. [ELABORATION_RULES] Do not perform any action other than elaboration. Preserve the content and details of the text. Do not delete sentences from the text.

The elaboration stage ensures that the simplified text not only uses appropriate grammar and vocabulary for the target level but also provides sufficient contextual information to support comprehension for the target proficiency level.

4 Results and Discussion

Our three-stage progressive simplification pipeline yielded mixed results on the test data provided by the TSAR 2025 Shared Task (Alva-Manchego et al., 2025), with overall MeaningBERT scores of 0.797 for both original-output (MeaningBERT Original) and reference-output (MeaningBERT Reference) comparisons, and a CEFR compliance RMSE of 0.632.

Analysis of the highest and lowest scoring texts revealed that our top-performing cases were predominantly B1-targeted, while most of the under-performing examples were A2-level simplifications. The five lowest MeaningBERT scores for both metrics came from our A2-level productions, whereas the top five scores for MeaningBERT Original were

all from B1-targeted simplifications. This performance gap stems largely from over-simplification induced by the strict A2-level linguistic constraints in the prompts, such as limited grammatical structures and restrictions on vocabulary exceeding A2 proficiency.

As detailed in the methodology, our syntactic simplification step constrains GPT-4o to avoid grammatical structures outside the target CEFR level, and the lexical simplification stage involves prompting GPT-4o to replace vocabulary that exceeds the target level. However, this approach can cause over-simplification in some cases.

During the syntactic simplification process, the system output can sometimes distort the original meaning by avoiding grammatical structures that are not included in the target level’s grammar list. An example of this can be seen in the A2-level simplification of text 73:

Text ID: 73-A2

MeaningBERT Reference Score: 0.5836

Reference:

The two men looked at the north side of the land. It was so big that they could hardly see the fences...

Simplified:

The two men looked at the far end of the land. The land was very big, and it stretched very far. The walls around the land were hard to see because the men were standing far away...

The reference employs the “so...that” construction with the modal “could” which, while arguably appropriate for the A2 level, is a structure our system is forced to avoid. Because the prompt strictly limits the output to the provided A2 grammar list, which does not include the “so...that” construction for clauses of result, the system fragments the sentence into simpler, explicitly permitted coordinate clauses, resulting in reduced cohesion.

The lexical simplification stage shows a tendency to make unnecessary substitutions in A2 texts, as shown in the following example with text ID 96-A2:

Text ID: 96-A2

MeaningBERT Reference Score: 0.5954

Original:

No matter what kind of house it is, the basement is a scary place in horror films...

Reference:

In horror films, the basement is always a scary place...

Simplified:

The lower floor is a scary place in scary films...

Here “basement”, which is listed at the B2 level in our vocabulary list, is replaced with “lower floor” and “horror films” is changed to “scary films” because “horror” is listed at the B1 level. In the B1-targeted simplification, however, “horror” is correctly preserved, while “basement” is still replaced. Although this version also does not fully match the reference at B1, it achieves a higher MeaningBERT Reference score of 0.7328 compared to 0.5954 at A2 production, as making fewer lexical substitutions in the entirety of the text allows the output to maintain a higher semantic similarity to the reference.

Beyond systematic over-simplification, some failures result from meaning shift rather than linguistic constraints. Since our pipeline utilizes an LLM three times in sequence, each stage introduces the risk of semantic drift, which can hinder meaning preservation. As the simplification with text ID 120-A2 illustrates:

Text ID: 120-A2

MeaningBERT Original Score: -0.0006

Original:

...Ensure there are times when positive feedback is given for its own sake and resist the temptation to offer constructive criticism.

Reference:

...Sometimes, give good feedback only, and try not to add any criticism.

Simplified:

...Sometimes, give only good comments. Do not always give advice to help.

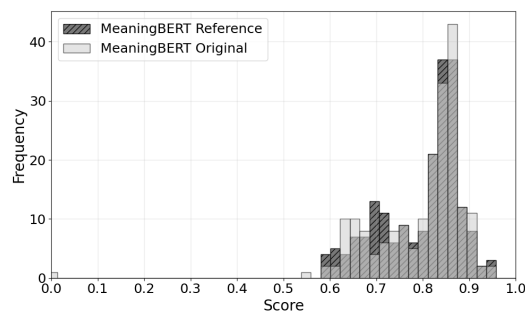


Figure 1: MeaningBERT score distributions of system outputs.

While not all poor simplifications stem from A2-level constraints, our analysis shows that B1-targeted simplifications performed significantly better. When only B1 outputs are evaluated, performance improved across all metrics: MeaningBERT Original and MeaningBERT Reference scores rose from 0.797 to 0.829 and 0.816, respectively, and RMSE decreased from 0.632 to 0.556.

Another issue about the performance of our method is its variance. As illustrated in Figure 1, the variance analysis reveals performance inconsistencies. For instance, the MeaningBERT Original score ranges from a high of 0.9424 down to an extremely low -0.0006, with a standard deviation of 0.1031. Similarly, the MeaningBERT Reference score varies widely from 0.9453 to 0.5826. This indicates that while our method can achieve very good results, it suffers from a lack of reliability in all cases.

Overall, the three-stage progressive simplification pipeline shows potential to achieve strong results, especially in B1-level targeted simplifications, without the need for costly model training and with a total cost of only \$1.75. However, it can also lead to unsatisfactory outcomes due to the strict constraints imposed in the prompts.

5 Conclusion

This paper introduced a three-stage, CEFR-oriented text simplification method using syntactic simplification, lexical simplification, and elaboration through prompt engineering with GPT-4o. The approach avoids costly model training while producing controlled simplifications aligned with language proficiency levels. The results indicate a more promising performance at B1, where the output preserved meaning and coherence, but also reveal limitations at A2, where rigid grammar and vocabulary constraints often led to over-simplification

or unnecessary substitutions. These findings suggest that future work should have greater flexibility in constraint design to balance accessibility with fidelity.

6 Limitations

Despite its potential, our approach has several limitations:

The most significant is its reliability. As discussed earlier, the system’s performance exhibits high variance across inputs, indicating a lack of robustness.

Another practical limitation is the strict linguistic constraints imposed by the prompts, which often lead to over-simplification and meaning distortion. Additionally, the three-stage pipeline introduces a cumulative risk of semantic drift, which can further compromise meaning preservation.

7 Lay Summary

Text simplification helps make complex texts easier to read, which is especially useful for language learners and people with reading difficulties. Traditionally, this process is done manually by experts, which takes a lot of time and effort. While some automatic systems exist, they often do not adjust the output based on the reader’s language level and are not designed around well-known frameworks like the Common European Framework of Reference for Languages (CEFR).

Training advanced systems for text simplification requires large datasets and expensive resources. This makes it difficult to build effective tools for many languages, especially those with limited data. Even when these systems work well, they often fail to match texts to specific learning levels, which is important in education and accessibility.

To address these challenges, we propose a three-stage simplification pipeline that directs a language model with carefully designed instructions instead of training it from scratch. Our approach simplifies grammar, replaces difficult words, and adds helpful explanations. This lightweight and flexible method can make reading materials more accessible and suitable for different language levels without the need for costly training or large datasets.

References

Wejdan Alkaldi and Diana Inkpen. 2023. Text simplification to specific readability levels. *Mathematics*, 11(9):2063.

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2020. Data-driven sentence simplification: Survey and benchmark. *Computational Linguistics*, 46(1):135–187.

Fernando Alva-Manchego, Regina Stodden, Joseph Marvin Imperial, Abdullah Barayan, Kai North, and Harish Tayyar Madabushi. 2025. Findings of the TSAR 2025 shared task on readability-controlled text simplification. In *Proceedings of the Fourth Workshop on Text Simplification, Accessibility, and Readability (TSAR 2025)*, Suzhou, China. Association for Computational Linguistics.

Miriam Anschütz, Joshua Oehms, Thomas Wimmer, Bartłomiej Jezierski, and Georg Groh. 2023. Language models for German text simplification: Overcoming parallel data scarcity through style-specific pre-training. *arXiv preprint arXiv:2305.12908*.

Yuki Arase, Satoru Uchida, and Tomoyuki Kajiwara. 2022. CEFR-based sentence difficulty annotation and assessment. *arXiv preprint arXiv:2210.11766*.

Abdullah Barayan, Jose Camacho-Collados, and Fernando Alva-Manchego. 2024. Analysing zero-shot readability-controlled sentence simplification. *arXiv preprint arXiv:2409.20246*.

Fatih Bektaş, Kutay Arda Dinç, and Gülşen Eryiğit. 2024. LLMs for document-level text simplification in turkish foreign language learning. In *2024 9th International Conference on Computer Science and Engineering (UBMK)*, pages 1–5. IEEE.

Cambridge University Press. 2024. English grammar profile online. <https://englishprofile.org/>. Accessed: 2025-08-17.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Press Syndicate of the University of Cambridge.

Dengzhao Fang, Jipeng Qiang, Yi Zhu, Yunhao Yuan, Wei Li, and Yan Liu. 2025. Progressive document-level text simplification via large language models. *arXiv preprint arXiv:2501.03857*.

Matthew Honnibal, Ines Montani, and Sofie Van Landeghem. 2020. Boyd. *A. spaCy: industrial-strength natural language processing in Python*, 10.

Guanlin Li, Yuki Arase, and Noel Crespi. 2025. Aligning sentence simplification with ESL learner’s proficiency for language acquisition. *arXiv preprint arXiv:2502.11457*.

Debora Nozza and Giuseppe Attanasio. 2023. Is it really that simple? Prompting large language models for automatic text simplification in italian. In *Proceedings of the 9th Italian Conference on Computational Linguistics (CLiC-it 2023)*, pages 322–333.

- Oxford University Press. a. The American Oxford 3000. https://www.oxfordlearnersdictionaries.com/external/pdf/wordlists/oxford-3000-5000/American_Oxford_3000.pdf. Accessed: 2025-08-18.
- Oxford University Press. b. The American Oxford 5000. https://www.oxfordlearnersdictionaries.com/external/pdf/wordlists/oxford-3000-5000/American_Oxford_5000.pdf. Accessed: 2025-08-18.
- Ceren Özmen. 2019. *Yabancı Dil Olarak Türkçe Öğretiminde Metin Değiştirme Teknikleriyle Öykülerin Yeniden Oluşturulması*. Ph.D. thesis, Hacettepe University, Institute of Turkish Studies, Ankara, Turkey. Tez No: 588269.
- Ceren Özmen and Ayşe Kıran. 2022. A study on text modification in the field of teaching turkish as a foreign language. *International Journal of Languages' Education and Teaching*, 10(2):19–42.
- Renliang Sun, Hanqi Jin, and Xiaojun Wan. 2021. Document-level text simplification: Dataset, criteria and baseline. *arXiv preprint arXiv:2110.05071*.
- Satoru Uchida, Shohei Takada, and Yuki Arase. 2018. CEFR-based lexical simplification dataset. In *Proceedings of International Conference on Language Resources and Evaluation*, volume 11, pages 3254–3258. European Language Resources Association.

STARLING at TSAR 2025 Shared Task: Leveraging Alternative Generations for Readability Level Adjustment in Text Simplification

Piotr Przybyła

TALN Group, Universitat Pompeu Fabra;
Institute of Computer Science, Polish Academy of Sciences
piotr.przybyla@upf.edu

Abstract

Readability adjustment is crucial in text simplification, as it allows to provide generations appropriate to the needs of a particular group of readers. Here we present a method for simplifying a text fragment that aims for a given CEFR level, e.g. A2 or B1. The proposed approach consists of two stages, executed for each sentence: (1) obtaining several candidate simplification through prompting a large language model and (2) choosing a candidate that maximises the compliance with the desired readability level. Several variants of our approach are evaluated within the framework of TSAR 2025 shared task, showing a trade-off between precise readability adjustment and faithful meaning preservation.

1 Background

Text simplification (TS) promises to make language content accessible to various vulnerable groups that might find typical text difficult, such as people with age-related impairments, learning difficulties, dyslexia, autism etc., but also non-native speakers (Shardlow, 2014; Paetzold and Specia, 2017). This causes an obvious challenge: it is impossible to satisfy the needs of all these groups with *the same* simplification. Some aspects of text might be difficult for one reader, but not for another one (Tamor, 1981). For example, even quite long and complicated words might appear straightforward for a non-native speaker, as long as they know similar lexemes from their mother tongue.

Therefore, text simplification can never be successful as a generic task and it needs to be anchored in a specific target audience. However, approaching the challenge for each group separately is impeded when taking a classic machine-learning approach. The datasets with simplifications prepared by human experts for specific readers are scarce (Cistola et al., 2021; Alarcon et al., 2023), so many TS solutions rely on generic data anyway. This limitation

remains true for large language models (LLMs) – since they are unlikely to have been provided with examples of such tailored simplifications in their pretraining, they shouldn’t be expected to generate them when prompted.

Here we present an approach to the problem of providing text simplifications for readers with a given language proficiency, expressed as a CEFR level. While there are some resources for the non-native speakers according to their level (Scarton et al., 2018), these are not sufficient for model training from scratch.

Therefore, our approach is a hybrid one. Firstly, we generate simplifications using generic LLM capabilities, but we ask the model to produce several variants of the output. Then, we use CEFR labellers to find the rephrasing that best corresponds to the desired level. This process is repeated for all sentences in a given text, and the final output is a concatenation of the simplified variants of all sentences.

2 Task

Our solution was prepared within the framework of the TSAR 2025 Shared Task on Readability-Controlled Text Simplification (Alva-Manchego et al., 2025). The task participants were provided with a series of pairs, each including a text fragment in English (paragraph-length) and a CEFR level, and their goal was to provide simplifications that maintain the meaning of the original fragment, but at the same time possess the provided difficulty level (A2 or B1).

The submitted fragments were evaluated using the following measures:

- Difference between the CEFR levels (requested and observed in text) according to three fine-tuned language models (Imperial et al., 2025), quantified as weighted F1, adjusted accuracy and RMSE.

- Meaning preservation measured as text similarity (original and provided simplification) according to MeaningBERT (Beauchemin et al., 2023) and BERTScore (Zhang et al., 2020),
- Closeness to reference simplification, also according to MeaningBERT and BERTScore.

Of these, the RMSE CEFR and both MeaningBERT similarities were used to establish the final ranking.

For details regarding the previous work in the domain, evaluation details and results of all participants, see the overview article (Alva-Manchego et al., 2025).

3 Methods

The submitted solution was prepared using elements of STARLING (Simplifying Text Across Languages Using Generative Models): a TS system under construction, which uses LLM prompting for obtaining robust simplifications across several languages, including some with low NLP support. The multilingual capabilities were however not used here, since the task is performed in English, a language with sufficient monolingual resources.

Broadly, the solution involves the following steps:

1. Splitting the input (complex) text into sentences,
2. For each sentence:
 - Producing variants of simplification using a prompted language model,
 - Selecting the variant that is the closest to the desired CEFR level,
3. Concatenating the obtained sentences.

3.1 Splitting

We split the given paragraph into individual sentences with LAMBO (Przybyła, 2022) 2.3 segmenter¹, using the LAMBO-UD_English-EWT model trained on English dependency parsing corpus² in Universal Dependencies (de Marneffe et al., 2021), version 2.13. The sentence splitting is motivated by

¹<https://gitlab.clarin-pl.eu/syntactic-tools/lambo>

²https://universaldependencies.org/treebanks/en_ewt/index.html

(1) preliminary experiments showing that LLMs, when tasked with rephrasing a longer paragraph, often omit some of the details, and (2) the intention to obtain a wide range of candidates by reformulating sentences independently.

3.2 Simplifying

For simplification, we use the following prompt:

Please rewrite the following complex sentence in order to make it easier to understand by non-native speakers of the language. You can do so by replacing complex words with simpler synonyms (i.e. paraphrasing), deleting unimportant information (i.e. compression), and/or splitting a long complex sentence into several simpler ones. The final simplified sentence needs to be grammatical, fluent, and retain the main ideas of its original counterpart without altering its meaning. Make sure the output is in the same language as the original.

Return five different rephrasings, separated by newline. Do not generate any text except the reformulations.

INPUT: <input sentence>

This prompt is inspired by a formulation obtaining good results in BLESS benchmark (Kew et al., 2023) – prompt 2³. However, it was modified to include a request of five different rephrasings. Note how it does not include any mention of the desired CEFR level – our preliminary experiments showed it not to be useful for a general-purpose model pre-trained without such specialised data.

Our baseline model was Gemma 3 (Gemma Team et al., 2025) 27 B, implemented on *HuggingFace Transformers* 4.38.1 (Wolf et al., 2020) (model google/gemma-3-27b-it). The computations were performed on double-GPU configuration with NVIDIA A100.

3.3 Choosing a variant

We create the list of variants by splitting the model output into individual options, removing anomalies (outputs shorter than 1/3 of the input⁴) and adding the original complex sentence. Each variant on the

³Note that BLESS experiments were performed in a few-shot setting, but in the shared task no training data were available, so we use the same prompt in a zero-shot setting.

⁴These usually come from the LLM adding extra text to the output: enumeration markers, comments, etc.

method	CEFR compliance			Original similarity		Reference similarity	
	F1	Acc.	RMSE	M-BERT	BERT-S	M-BERT	BERT-S
<u>Gemma-5v-best</u>	0.5107	0.9500	0.8216	0.8075	0.9281	0.7584	0.9124
Gemma-5v-random	0.4895	0.8500	1.0000	0.8204	0.9335	0.7567	0.9109
<u>Gemma-10v-best</u>	0.6921	0.9750	0.6325	0.7675	0.9190	0.7458	0.9062
<u>Gemma-10v-random</u>	0.3888	0.9500	0.8803	0.7780	0.9194	0.7441	0.9038
<i>original</i>	0.1288	0.5250	1.6125	<i>1.0000</i>	<i>1.0000</i>	<i>0.7901</i>	<i>0.9265</i>

Table 1: Evaluation results on the trial dataset, showing CEFR compliance results, similarity to the original text and reference simplification, both measured using MeaningBERT (M-BERT) and BERTScore (BERT-S). The best values are in boldface; the submitted solutions are underlined.

list is then assessed with respect to readability level. Specifically, we apply the three CEFR labellers⁵ provided by the organisers (section 2). We then check which of the variants (including the complex original) has been assigned the desired CEFR level with the highest probability, according to any for the labellers. In case of ties, the order on the variant list decides. Therefore, for uncertain ratings, the original complex sentence is used (to maintain meaning preservation).

3.4 Concatenation

After the CEFR-optimised variant is chosen for each sentence, they are all concatenated together to create continuous text. Note that this risks breaking some discourse links, which is a weakness of our approach (section 5).

4 Evaluation

The solution described above is our baseline approach to the problem. However, in order to better understand its strengths and weaknesses, we check the result of modifying some aspects:

- using a higher number of requested variants: 10 instead of 5.
- using a simpler heuristic for selecting a variant: random choice or keeping the original sentence.

All the solutions are tested using the evaluation code provided by the organisers (see section 2) on the 40 instances of the trial set. The full results on the test dataset, including our submission, can

⁵https://huggingface.co/AbdullahBarayan/ModernBERT-base-doc_en-Cefr, https://huggingface.co/AbdullahBarayan/ModernBERT-base-doc_sent_en-Cefr and https://huggingface.co/AbdullahBarayan/ModernBERT-base-reference_AllLang2-Cefr2

be found in the shared task overview article (Alva-Manchego et al., 2025).

Table 1 illustrates the results of the evaluation. We can see that the best CEFR compliance is achieved by generating 10 variants and choosing the one that corresponds to the desired level. This proves that the variant selection mechanism fulfils its purpose. However, the same approach is performing the poorest in content preservation. Instead, generating 5 variants and randomly selecting one of them delivers the best similarity to the original text (except returning the original itself). We can therefore see a clear trade-off between delivering the expected CEFR level and preserving the original content. Finally, our basic approach (5 variants, guided selection) achieves the best performance in terms of similarity to the reference simplification, indicating its overall usefulness. It is interesting to note that the original text is more similar to the reference simplifications than any of our approaches, indicating that the LLM is too aggressive in its rewriting. This mirrors a similar phenomenon in text rewriting for the purpose of a different task, namely adversarial example generation (Przybyła et al., 2025).

5 Discussion

Judging by the evaluation performed in the previous section, our approach performs well: it allows to adjust for the desired readability level without sacrificing too much of the original meaning. However, this is clearly a prototype solution and several limitations remain.

Firstly, a more intensive adjustment to readability level is possible. Since generating 10 variants allows better compliance than 5, one could also try 20 or 100 – though in case of most sentences it would be unrealistic to expect that many different reformulations. It is therefore an open question on

whether this wouldn't cause a loss of meaning or general decrease in quality of further generations. We leave this for future work.

Secondly, we limited our prompting experiment to adapting a formulation that has been found to work well in previous research, but that usecase did not include readability adjustment. Therefore, we expect that experimenting with the prompt can help to guide the model towards producing reformulations with more utility for the current setup, e.g. by encouraging diversity in the variants. Recent research in TS indicates great room for improvement in tuning the prompt (Guidroz et al., 2025).

Thirdly, we performed the simplification on the sentence level to obtain fine-grained control over the variants chosen and avoid the loss of details that we observe in full paragraph rewriting. However, this introduces a limitation: when each sentence is reformulated independently, we risk breaking discourse links between them, resulting in less coherent text. This is a known problem in text simplification, but it requires further research to deliver satisfactory solutions (Vásquez-Rodríguez et al., 2023). One bypass could be to provide whole paragraphs as input text and look for other ways to avoid information loss, e.g. through various prompts.

Finally, we have to emphasise that our evaluation is based solely on automatic measures, which for some time have been known to poorly reflect human judgement of simplified text (Alva-Manchego et al., 2021). This final step of TS evaluation should involve human evaluators, especially when a solution is claimed to be adjusted to specific user group.

6 Conclusions

To sum up, the proposed approach to readability-adjusted text simplification achieves positive results, delivering output that is in agreement with the desired CEFR level (95% accuracy) and maintains the overall meaning (93% BERTScore). Depending on which configuration we choose, a trade-off between simplicity and similarity to the original can be struck differently, but overall the results indicate this to be a promising direction for future investigation.

Acknowledgments

This work was done with the project IDEAL (Inclusive Democratic Engagement and Language Technologies in Europe), which received funding from

the European Union's Horizon Europe research and innovation programme under grant agreement No 101178191. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them. We gratefully acknowledge Polish high-performance computing infrastructure PLGrid (HPC Centers: ACK Cyfronet AGH) for providing computer facilities and support within computational grant no. PLG/2025/018019.

Lay Summary

When a text is changed to make it simpler for some readers, it is important to take into account who these readers are. For example, if the text is in English, but the readers are not native speakers, we should take into account their level of knowledge of the language. In this article we show how we simplify text in English so that learners at some level (for example A2 or B1) can understand it. First, we divide the text into separate sentences. Second, a language model changes each sentence, producing several possible modifications. We then choose the modification that is closest to the needed difficulty level. Finally, all the changed sentences are connected back together. Our approach is tested at the TSAR 2025 workshop. The results show that sometimes you have to choose between keeping the meaning unchanged and arriving at the difficulty level you want.

References

- Rodrigo Alarcon, Lourdes Moreno, and Paloma Martínez. 2023. *EASIER corpus: A lexical simplification resource for people with cognitive impairments*. *PLOS ONE*, 18(4):1–23.
- 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.
- Fernando Alva-Manchego, Regina Stodden, Joseph Marvin Imperial, Abdullah Barayan, Kai North, and Harish Tayyar Madabushi. 2025. Findings of the TSAR 2025 Shared Task on Readability-Controlled Text Simplification. In *Proceedings of the Fourth Workshop on Text Simplification, Accessibility, and Readability (TSAR 2025)*, Suzhou, China. Association for Computational Linguistics.
- David Beauchemin, Horacio Saggion, and Richard Khoury. 2023. *MeaningBERT: assessing meaning*

- preservation between sentences. *Frontiers in Artificial Intelligence*, Volume 6 - 2023.
- Giorgia Cistola, Mireia Farrús, and Ineke van der Meulen. 2021. [Aphasia and acquired reading impairments: What are the high-tech alternatives to compensate for reading deficits?](#) *International Journal of Language & Communication Disorders*, 56(1):161–173.
- Marie-Catherine de Marneffe, Christopher D Manning, Joakim Nivre, and Daniel Zeman. 2021. [Universal Dependencies](#). *Computational Linguistics*, 47(2):255–308.
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. [Gemma 3 Technical Report](#). *Preprint*, arXiv:2503.19786.
- Theo Guidroz, Diego Ardila, Jimmy Li, Adam Mansour, Paul Jhun, Nina Gonzalez, Xiang Ji, Mike Sanchez, Sujay Kakarmath, Mathias M J Bellaiche, Miguel Ángel Garrido, Faruk Ahmed, Divyansh Choudhary, Jay Hartford, Chenwei Xu, Henry Javier Serrano Echeverria, Yifan Wang, Jeff Shaffer, Eric, and 8 others. 2025. [LLM-based Text Simplification and its Effect on User Comprehension and Cognitive Load](#). *Preprint*, arXiv:2505.01980.
- Joseph Marvin Imperial, Abdullah Barayan, Regina Stodden, Rodrigo Wilkens, Ricardo Munoz Sanchez, Lingyun Gao, Melissa Torgbi, Dawn Knight, Gail Forey, Reka R Jablonkai, Ekaterina Kochmar, Robert Reynolds, Eugénio Ribeiro, Horacio Saggion, Elena Volodina, Sowmya Vajjala, Thomas François, Fernando Alva-Manchego, and Harish Tayyar Madabushi. 2025. [UniversalCEFR: Enabling Open Multilingual Research on Language Proficiency Assessment](#). *Preprint*, arXiv:2506.01419.
- Tannon Kew, Alison Chi, Laura Vásquez-Rodríguez, Sweta Agrawal, Dennis Aumiller, Fernando Alva-Manchego, and Matthew Shardlow. 2023. [BLESS: Benchmarking Large Language Models on Sentence Simplification](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.
- Gustavo H. Paetzold and Lucia Specia. 2017. [A survey on lexical simplification](#). *Journal of Artificial Intelligence Research*, 60:549–593.
- Piotr Przybyła. 2022. [LAMBO: Layered Approach to Multi-level BOUNDary identification](#).
- Piotr Przybyła, Euan McGill, and Horacio Saggion. 2025. [Attacking Misinformation Detection Using Adversarial Examples Generated by Language Models](#). *Preprint*, arXiv:2410.20940.
- Carolina Scarton, Gustavo Paetzold, and Lucia Specia. 2018. [Text Simplification from Professionally Produced Corpora](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Matthew Shardlow. 2014. [A Survey of Automated Text Simplification](#). *International Journal of Advanced Computer Science and Applications*, 4(1).
- Lynne Tamor. 1981. [Subjective Text Difficulty: An Alternative Approach to Defining the Difficulty Level of Written Text](#). *Journal of Reading Behavior*, 13(2):165–172.
- Laura Vásquez-Rodríguez, Matthew Shardlow, Piotr Przybyła, and Sophia Ananiadou. 2023. [Document-level Text Simplification with Coherence Evaluation](#). In *Proceedings of the Second Workshop on Text Simplification, Accessibility and Readability*, pages 85–101, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. [Transformers: State-of-the-Art Natural Language Processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. [BERTScore: Evaluating Text Generation with BERT](#). In *8th International Conference on Learning Representations, ICLR 2020*, Addis Ababa, Ethiopia.

***TaskGen* at TSAR 2025 Shared Task: Exploring prompt strategies with linguistic knowledge**

Juan Cruz Oviedo

juan.oviedo.869@unc.edu.ar
Universidad Nacional de Córdoba

Elisabet Comelles

elicomelles@ub.edu
Universitat de Barcelona

Laura Alonso Alemany

lauraalonsoalemany@unc.edu.ar
Universidad Nacional de Córdoba

Jordi Atserias Batalla

jordi.atserias@bsc.es
Barcelona Supercomputing Center

Abstract

TaskGen ranked as 6th best team in the TSAR 2025 shared task for English text adaptation to a target CEFR level. Our experiments consisted of prompting a Llama-3.1-8B-Instruct model with linguistic descriptors of the target level, examples of adaptations and multi-step approaches. Our best run, 13th in the overall ranking, applied an ensemble strategy using a voting mechanism to find the most adequate among 10 texts, each produced by a different prompting strategy.

1 Introduction

Text simplification (TS) aims to reduce the linguistic complexity of texts while preserving their meaning and coherence. This facilitates accessibility for readers with low literacy and individuals with cognitive challenges (Siddharthan, 2014), but it is also very useful to adapt texts for language learners. In that context, TS presents the additional requirement to target the learners’ proficiency levels.

The TSAR 2025 shared task on Readability-Controlled Text Simplification (RCTS) (Alva-Manchego et al., 2025) directly addresses this need by requiring participants to simplify English texts at B2 level or higher to specified target CEFR levels (A2, B1). This setup not only evaluates systems on their ability to reduce complexity, but also on their capacity to produce texts adequate for language learners at different stages of proficiency.

2 Related work

Automatic Readability Assessment (ARA) “refers to the task of modeling the reading and comprehension difficulty of a given piece of text, for a given target audience” (Vajjala, 2022). It has been widely studied across NLP, education and psychology, and is often applied to support tasks such as selecting suitable materials for L2 learners (Heilman et al., 2007; Vajjala and Meurers, 2012).

Traditional es to ARA relied on readability formulae, such as Flesch-Kincaid Grade level (Kincaid et al., 1975), SMOG Grading (McLaughlin, 1969) and Coleman-Liau Index (Coleman and Liau, 1975), among others.

More recent work has explored feature-based machine learning and neural models (Hancke et al., 2012; Vajjala and Meurers, 2012; Chen and Meurers, 2018; Deutsch et al., 2020; Lee et al., 2021; Lee and Vajjala, 2022).

Although central to understanding textual complexity, ARA also serves as a foundation for downstream tasks such as text simplification (TS). Early es to TS were rule-based or statistical, but recent progress has been driven by neural models (Martin et al., 2022; Maddela et al., 2021; Sheang and Saggion, 2021). With the emergence of large language models (LLMs), TS research has increasingly swift to fine-tuning and prompt engineering (Kew et al., 2023; Nozza and Attanasio, 2023; Martínez et al., 2024).

Building on TS, Readability-Controlled Text Simplification (RCTS) introduces explicit control over the target readability level, making it particularly valuable in educational contexts where texts must be aligned with learners’ proficiency (Scarton and Specia, 2018). Most work on this area has followed supervised training es (Nishihara et al., 2019; Spring et al., 2021; Rios et al., 2021; Yanamoto et al., 2022; Agrawal and Carpuat, 2023); however, the drawback to these es is the limited number of datasets available.

Recent advances in LLMs and in their capacity for text generation have opened new directions for RCTS, with studies exploring sentence-level (Barayan et al., 2025; Chi et al., 2023) and passage-level simplification (Farajidizaji et al., 2024; Imperial and Tayyar Madabushi, 2023) in zero-shot and few-shot settings.

3 System Description

We explored a prompt-based approach to text adaptation, assessing the performance of different prompt variants:

- **Few-shot learning.** Including examples of pairs of original and adapted texts in the prompt and adding linguistic information, such as overall reading comprehension descriptors of the target proficiency level (Council of Europe, 2020), curated vocabulary lists, and examples of appropriate grammatical constructions. Also, explicit instructions were used to obtain only the adapted text. That is, requiring the model to avoid references to the adaptation task or to the output CEFR level.
- **Multi-step procedures.** An initial adaptation was followed by additional instructions for iterative refinement.
- **Ensemble-based strategy.** Multiple adaptations generated by ten distinct prompts were compared, and the most suitable version was selected through a voting mechanism.

All prompt different strategies used *Llama-3.1-8B-Instruct* model (Meta, 2024) with precision bfloat16 and decoding default parameters (nucleus sampling with top-p = 0.9, temperature = 0.6, and stochastic decoding enabled). We opted for *Llama-3.1-8B-Instruct* model because it is open-source, cost-effective and had performed well in CEFR-based text classification (Comelles et al., 2025). Also, a previous version of the *Llama 3 8B* family had proved successful for a similar task (Barayan et al., 2025).

3.1 Methodology and lessons learned

After systematically evaluating eleven different experiments with the metrics and trial data provided by the organizers, and with qualitative, manual inspection of the produced texts, we selected the three experiments described below for submission.

We wanted to compare approaches that required increasing levels of effort, so that the evaluation of the task served to assess the impact of more complex approaches in the resulting performance. Our first and second approaches were based on a prompt with extensive linguistic descriptors, guidelines and examples describing the target CEFR level (Figure 4). After analyzing the performance of this approach on trial data, we found that such lengthy instructions on the linguistic characterization of the

level damaged the preservation of the content of the original text. To address this issue, for run 1 we added a follow-up prompt (Figure 5) that focused on the preservation of the content of the original text. Alternatively, for run 2 we applied a follow-up prompt (Figure 6) that removed all text where the language model described the task it was performing, with the aim of leaving only the adapted text, without meta information. Finally, for run 3, we obtained adaptations from 10 different prompts (see Appendix D). Then, each text was assessed with criteria that had been found useful in the analysis of performance on trial data.

3.2 Run 1: Complementary prompts with CEFR descriptors and content preservation

The first run consisted of a prompt (see Figure 4) with extensive linguistic information on the characteristics of CEFR levels, and a follow-up prompt (see Figure 5) to emphasize content adequacy.

The linguistic information in the first prompt was:

- CEFR overall reading comprehension descriptors for A2 and B1 (Council of Europe, 2020).
- A list of 1681 words taken from Oxford University Press (n.d.), which lists the most frequent and relevant words from A1 to B2. From this list, the words classified as A2 and B1 were used, totaling 872 words for A2 and 809 words for B1.
- A list of 5 examples of each morphosyntactic construction listed as characteristic of the target level in (North et al., 2010). A total of 190 sentences were included (105 A2 sentences, and 85 B1 sentences). They can be found at <https://github.com/juan-oviedo/Recursos-Linguisticos>.

The output of this prompt was taken as input for a subsequent prompt (see Figure 5) that aimed to preserve the content of the original text, prioritizing content over adequacy to the target CEFR level.

The first prompt in this strategy resulted quite lengthy, with slightly over 3,800 tokens in average. Excessive length is believed to impair the performance of language models. With this two-step approach we tried to address this limitation of LLMs by focusing the task first on CEFR adequacy and then on content preservation.

3.3 Run 2: Corrections on meta-information

TaskGen second run relied on the same first prompt as run 1 (Figure 4), that is, the lengthy prompt with linguistic information and to CEFR level descriptors. Then a second prompt was applied to remove meta information on simplification or CEFR level (Figure 6). Such meta-information appeared frequently in the LLM’s outputs, despite having included explicit instructions to provide the simplified text only.

3.4 Run 3: Prompt output ensemble

For the third run we implemented an ensemble strategy. We prompted the same language model, *Llama-3.1-8B-Instruct*, with 10 different prompts. Then, the 10 output texts were ranked by a voting mechanism.

Prompts, shown in Appendix C, ranged from very simple instructions to adapt the text to a given target level, to instructions with extensive linguistic information, and to multi-step strategies with follow-up prompts, including few-shot approaches with examples of pairs of original - adapted texts.

The voting mechanism was manually constructed by inspecting errors in the trial data and assessing their impact on performance. The final submission consisted of the following simple heuristic, applied to each of the 10 adaptations, starting from 0 points:

- **Avoid Keyword [0, -10]:** ten points were subtracted if the output text contained one or more predefined sequences belonging to an explanation of the task, in addition to the adapted text itself (e.g., "CEFR", "This text is a simplification", "Here you have a simplified version", "Here is the simplified text").
- **CEFR level adequacy [0, -1, -2, -3, -4]:** classification in CEFR levels was computed using the three classifiers provided by the organizers in the evaluation script. A penalty was applied when a classifier identified a proficiency level in the generated text that differed from the target level: one point was subtracted if a single classifier indicated a mismatch, two points if two classifiers disagreed, and three points if all three disagreed. An additional penalty was applied proportional to the distance between the predicted and target levels, with one point subtracted for each level of discrepancy, so that the penalty for each classifier could subtract up to 4 points. As will be discussed in

the following section, this item may have been weighted more heavily to emphasize compliance with the target CEFR level.

- **Text length [0, -1]:** one point was subtracted if the adapted text exceeded 1.3 times the length of the original. This threshold was established after manual inspection of errors in trial data, which showed that lengthier adaptations included content that was not in the original text.
- **Semantic similarity [0, -5]:** semantic similarity was measured using the MeaningBERT script provided by the organizers to calculate Bert-based similarity to the original text. If the resulting similarity score fell below 0.6, a penalty of five points was applied. This threshold was established as a rule of thumb, aiming for the adaptation to preserve at least two thirds of the meaning of the original text.

After voting, the text with the highest score for each level was selected. In cases of ties, the text with the higher MeaningBERT score was preferred. If ties persisted, final selection was based on the performance of the prompts in the trial data: prompts 8, 4, 6, 3, 2.

As will be discussed in the following Section, this ensemble strategy achieved the best results of all *TaskGen* submitted runs, but there is still room for improvement in this line of work. The voting mechanism could likely be improved by assigning weights empirically, for example through regression on a larger dataset. Instead, for this shared task, the strengths of votes was assigned based on our analysis of performance on the trial partition of the dataset. Also, we expect that an approach based on agents can be more flexible than this heuristic, especially in ranking previously unseen cases.

4 Results and Discussion

As reported in Alva-Manchego et al. (2025), our third run, the ensemble with a voting strategy, ranked best among all the runs submitted by *TaskGen*, making for the 13th best submission and making *TaskGen* the 6th best team in the shared task. Although the RMSE with respect to the CEFR level was acceptable, with an average of 0.628, this run excelled in meaning preservation with respect to the original and the reference, with 0.856 and 0.826 respectively.

	RMSE	meaning-orig	meaning-ref
Run1	0.592	0.791	0.786
Run2	0.561	0.752	0.773
Run3	0.628	0.856	0.826

Table 1: *TaskGen* results in the TSAR shared task.

As can be seen in Table 1, our first run, consisting of a single prompt with linguistic information about the target CEFR level and a follow-up prompt addressing content adequacy, scored quite well in adequacy to the target level, with an RMSE of 0.592. However, it scored lower than other approaches in preservation of the meaning of the original text and similarity to the reference, with 0.791 and 0.786 respectively. As a result, this run ended up ranking in the lower half of the participating systems.

Our second run, aiming to remove meta-information generated by the LLM about the task itself, obtained lower scores for meaning preservation with respect to run 1. Despite this drawback, it was our best run with respect to adequacy to the target CEFR level, with an RMSE of 0.561. This shows that the follow-up prompt for content adequacy of run 1 harmed compliance to the target CEFR level.

The ensemble strategy, our third run, produced a significant improvement on preservation of the meaning with respect to the original and the reference, although at the expense of worse accuracy with respect to the target CEFR level, with the worst RMSE for our three runs, at 0.628. Nevertheless, high meaning preservation made this our best overall ranked submission.

Since adequacy to the target CEFR level was the main weakness of our approaches, we conducted a more detailed analysis of this aspect. As shown in Figure 1, all three runs —particularly the third—struggled to adapt texts to the B1 level, while systematically succeeding at the A2 level. Still, the errors were relatively minor, typically corresponding to a one-level difference, that is, texts that should have been adapted to B1 were adapted to A2 instead. We will examine these cases further to improve the system.

In contrast, meaning preservation was significantly better for B1 than for A2, as illustrated in Figure 2, showing the probability density function (a smoothed version of a histogram) of the score for meaning preservation with respect to the original. Thus, adaptation to lower levels seems to make

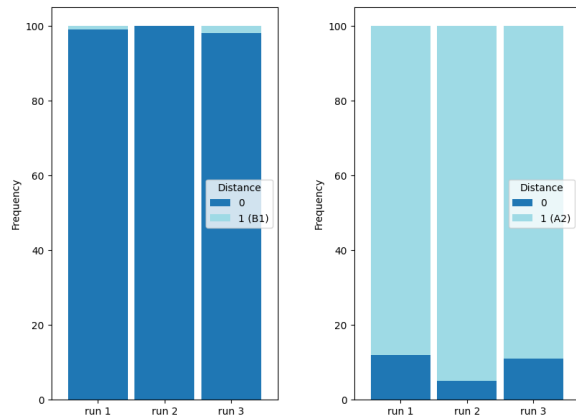


Figure 1: RMSE with target CEFR level for each of our three runs, for target level A2 (left) and B1 (right).

meaning preservation more difficult.

We found that run 3 was especially capable to adapt to different typologies of texts. For example, Figure 3 shows how runs 1 and 2 could not produce an adequate result for a text that was particularly challenging, because it could easily be interpreted as an instruction in itself by the language model. However, run 3 produced a better adaptation than the others, even if there is still room for improvement.

We believe that it is precisely the flexibility of the ensemble strategy to adapt to different typologies of texts that makes it our best ranked submission, ranking 13th best run and positioning *TaskGen* as the 6th best team in the shared task.

Finally, we find it useful to analyze how complex strategies improve performance upon simpler ones. In particular, we analyzed in trial data how different follow-up prompts improved performance on a base prompt. Table 2 displays the results obtained on the shared task evaluation data, applying the official evaluation scripts provided by Tsar organizers, for each of the individual prompts used in run 3. In particular, prompt 4, marked with an asterisk, was the base prompt for runs 1 and 2. As shown in Table 2, run 1 makes an important improvement on meaning preservation when compared to prompt 4 (from 0.75 to 0.79 on meaningbert-orig). As for run 2, it did not yield any improvement in meaning preservation but did improve in adequacy to the target CEFR level, achieving the best overall adequacy. Thus, it seems that both follow-up prompts did yield improvements, each in a different aspect. As future work, we plan to combine these complementary approaches to obtain texts that both preserve meaning and comply with the target level.

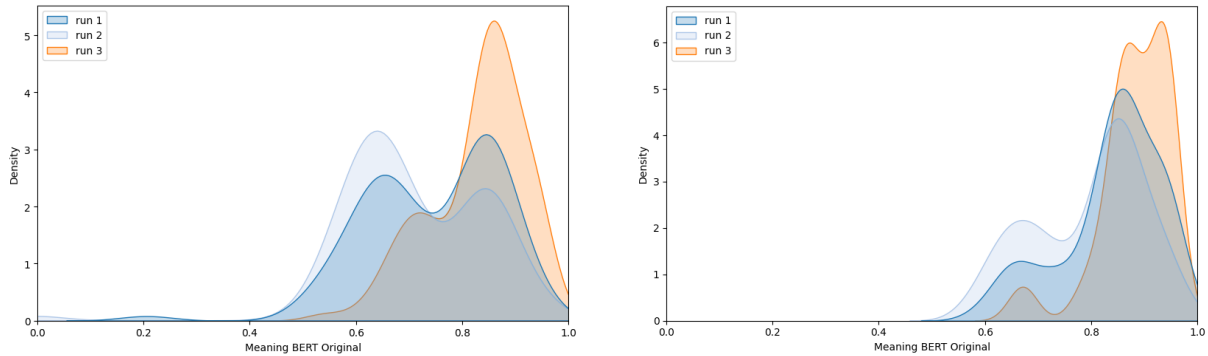


Figure 2: Density estimation of bert meaning preservation score to the original per target level (A2 left, B1 right).

exp.	prompt	weight_F1	rmse	meaningbert-orig	bertscore-orig	meaningbert-ref	bertscore-ref
01	1	0.6024	0.6403	0.7639	0.9274	0.7778	0.9330
02	2	0.5837	0.6633	0.7622	0.9264	0.7897	0.9321
03	3	0.6507	0.5958	0.7719	0.9261	0.7860	0.9308
04	4	0.6692	0.5745	0.7542	0.9267	0.7738	0.9314
05	5	0.6230	0.6708	0.7867	0.9248	0.7908	0.9269
06	6	0.6389	0.6364	0.7850	0.9283	0.7892	0.9292
07	7	0.6591	0.5874	0.7824	0.9318	0.7932	0.9362
08	8	0.6727	0.5958	0.7767	0.9283	0.7885	0.9325
09-run1	4+9	0.6633	0.5916	0.7906	0.9354	0.7863	0.9330
10-run2	4+10	0.6826	0.5612	0.7525	0.9262	0.7727	0.9311
11-run3	all	0.6353	0.6285	0.8556	0.9378	0.8256	0.9371

Table 2: *TaskGen* results of the experiments for the TSAR shared task on the final evaluation data.

5 Conclusions & future work

TaskGen submissions were the result of iterating on different prompting strategies, increasing their complexity based on the qualitative and quantitative analysis of performance on the trial dataset.

Results show that the ensemble with a voting strategy has proved effective. However, further research is required to improve adaptations to the B1 level, which were far worse than those for the A2 level. Also the proposed methods rely on manual prompt engineering and ensembling of different prompts results which may limit the applicability of the proposed solution to other problems or scale.

As future work, we aim to explore the specific contribution of the linguistic information used in the prompts, by analysing the individual effects of CEFR descriptors, CEFR-levelled wordlists and CEFR-levelled morphosyntactic structures on the LLM performance. We also plan to improve the voting heuristic by inferring the weight of each vote from empirical data, instead of manual inspection. To do that, we will work to obtain or build a bigger labelled corpus, so that empirically inferred votes are reliable. We also plan to examine more advanced prompting strategies, such as agentic and parallel prompting —where separate branches focus on CEFR adequacy and semantic semantic fi-

delity before combining. Finally, we intend to conduct fine-tuning experiments guided by automatic evaluation metrics, including CEFR-level classifiers and meaning preservation measures.

Acknowledgments

This research has been funded and made possible by Grant PID2023-149648NB-I00, funded by the Ministerio de Ciencia, Innovación y Universidades of the Spanish government, and the National Research Agency (AEI).

This work used computational resources from UNC Supercomputo (CCAD) – Universidad Nacional de Córdoba¹, which are part of SNCAD, República Argentina.

We also thank the reviewers for insightful comments and suggestions, which have contributed to improve this work.

¹<https://supercomputo.unc.edu.ar/>

6 Lay Summary

This paper presents the taskGen submission to the TSAR 2025 Shared Task on Controlled-Readability Text Simplification. Text simplification (TS) is a process that makes texts easier to read and understand, while keeping their original meaning. It can help people readers with low literacy level and individuals with cognitive challenges (Siddharthan, 2014), but it is also very useful to adapt texts for language learners who need materials suited to their level of proficiency. The TSAR 2025 shared task on Readability-Controlled Text Simplification (RCTS) (Alva-Manchego et al., 2025) focuses on this idea: it asks participants to simplify English texts at an advanced level (B2 or higher) so that they match pre-intermediate and intermediate levels —CEFR A2 and B1 (the Common European Framework of Reference for Languages, or CEFR, is a widely used scale that describes language ability, where A1 is beginner and C2 is expert). The goal of the task was to see whether systems could not only make texts simpler, but also produce adapted versions that matched the right level for language learners. The taskGen submission includes three different methods, all based on Large Language Models (LLMs), that is AI systems trained to understand and generate text. Our first method uses two prompts (instructions to the AI): the first prompt describes in detail the language features typical of each CEFR level, and the second prompt emphasizes keeping the meaning of the original text. The second method also uses two prompts, but in the second one the AI is asked not to include explanations about why the text fits a certain level. Finally, our third method uses 10 different prompts and then a ranking system chooses the best simplified text. Our third method performed best. It ranked 13th overall among all the submitted methods and made our team the 6th best team overall in the shared task. This approach was particularly strong at keeping the meaning of the original text, but there is still room for improvement regarding compliance with target CEFR level.

References

Sweta Agrawal and Marine Carpuat. 2023. [Controlling pre-trained language models for grade-specific text simplification](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12807–12819, Singapore. Association for Computational Linguistics.

Fernando Alva-Manchego, Regina Stodden, Joseph Marvin Imperial, Abdullah Barayan, Kai North, and Harish Tayyar Madabushi. 2025. [Findings of the TSAR 2025 shared task on readability-controlled text simplification](#). In *Proceedings of the Fourth Workshop on Text Simplification, Accessibility, and Readability (TSAR 2025)*, Suzhou, China. Association for Computational Linguistics.

Abdullah Barayan, Jose Camacho-Collados, and Fernando Alva-Manchego. 2025. [Analysing zero-shot readability-controlled sentence simplification](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6762–6781, Abu Dhabi, UAE. Association for Computational Linguistics.

Xiaobin Chen and Detmar Meurers. 2018. [Word frequency and readability: Predicting the text-level readability with a lexical-level attribute](#). *Journal of Research in Reading*, 41(3):486–510.

Alison Chi, Li-Kuang Chen, Yi-Chen Chang, Shu-Hui Lee, and Jason S. Chang. 2023. [Learning to paraphrase sentences to different complexity levels](#). *Transactions of the Association for Computational Linguistics*, 11:1332–1354.

M. Coleman and T. L. Liau. 1975. [A computer readability formula designed for machine scoring](#). *Journal of Applied Psychology*, 60(2):283–284.

Elisabet Comelles, Juan Cruz Oviedo, Laura Alonso, Roger Gilabert, and Joan Castellví. 2025. [Exploring large language models for cefr-based text classification in foreign language education](#). In *Congreso Internacional De La Asociación Española de Lingüística Aplicada(AESLA)*.

Council of Europe. 2020. *Common European Framework of Reference for Languages: Learning, teaching, assessment – Companion volume*. Council of Europe Publishing, Strasbourg.

Tovly Deutsch, Masoud Jasbi, and Stuart Shieber. 2020. [Linguistic features for readability assessment](#). In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–17, Seattle, WA, USA → Online. Association for Computational Linguistics.

Asma Farajidizaji, Vatsal Raina, and Mark Gales. 2024. [Is it possible to modify text to a target readability level? an initial investigation using zero-shot large language models](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9325–9339, Torino, Italia. ELRA and ICCL.

Julia Hancke, Sowmya Vajjala, and Detmar Meurers. 2012. [Readability classification for German using lexical, syntactic, and morphological features](#). In *Proceedings of COLING 2012*, pages 1063–1080, Mumbai, India. The COLING 2012 Organizing Committee.

- Michael Heilman, Kevyn Collins-Thompson, Jamie Callan, and Maxine Eskenazi. 2007. **Combining lexical and grammatical features to improve readability measures for first and second language texts**. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 460–467, Rochester, New York. Association for Computational Linguistics.
- Joseph Marvin Imperial and Harish Tayyar Madabushi. 2023. **Flesch or fumble? evaluating readability standard alignment of instruction-tuned language models**. In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 205–223, Singapore. Association for Computational Linguistics.
- Tannon Kew, Alison Chi, Laura Vásquez-Rodríguez, Sweta Agrawal, Dennis Aumiller, Fernando Alva-Manchego, and Matthew Shardlow. 2023. **BLESS: Benchmarking large language models on sentence simplification**. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13291–13309, Singapore. Association for Computational Linguistics.
- Peter Kincaid, Robert P. Fishburne, 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**.
- Bruce W. Lee, Yoo Sung Jang, and Jason Lee. 2021. **Pushing on text readability assessment: A transformer meets handcrafted linguistic features**. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10669–10686, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Justin Lee and Sowmya Vajjala. 2022. **A neural pairwise ranking model for readability assessment**. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3802–3813, Dublin, Ireland. Association for Computational Linguistics.
- Mounica Maddela, Fernando Alva-Manchego, and Wei Xu. 2021. **Controllable text simplification with explicit paraphrasing**. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3536–3553, Online. Association for Computational Linguistics.
- Louis Martin, Angela Fan, Éric de la Clergerie, Antoine Bordes, and Benoît Sagot. 2022. **MUSS: Multilingual unsupervised sentence simplification by mining paraphrases**. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1651–1664, Marseille, France. European Language Resources Association.
- Paloma Martínez, Lourdes Moreno, and Alberto Ramos. 2024. **Exploring large language models to generate easy to read content**. *Preprint*, arXiv:2407.20046.
- G Harry McLaughlin. 1969. **Smog grading—a new readability formula**. *The Journal of Reading*, 12(8):639–646.
- Meta. 2024. **Meta-Llama-3.1-8B-Instruct**. <https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct>. Accessed: 2025-09-22.
- Daiki Nishihara, Tomoyuki Kajiwara, and Yuki Arase. 2019. **Controllable text simplification with lexical constraint loss**. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 260–266, Florence, Italy. Association for Computational Linguistics.
- Brian North, Angel Ortega, and Susan Sheehan. 2010. **A Core Inventory for General English**. British Council and EAQUALS.
- Debora Nozza and Giuseppe Attanasio. 2023. **Is it really that simple? prompting large language models for automatic text simplification in Italian**. In *Proceedings of the 9th Italian Conference on Computational Linguistics (CLiC-it 2023)*, pages 322–333, Venice, Italy. CEUR Workshop Proceedings.
- Oxford University Press. n.d. **The oxford 3000: The list of the 3000 most important words to learn in English, from A1 to B2 level**. https://www.oxfordlearnersdictionaries.com/external/pdf/wordlists/oxford-3000-5000/The_Oxford_3000.pdf. Accessed: 2025-09-21.
- Annette Rios, Nicolas Spring, Tannon Kew, Marek Kostrzewa, Andreas Säuberli, Mathias Müller, and Sarah Ebling. 2021. **A new dataset and efficient baselines for document-level text simplification in German**. In *Proceedings of the Third Workshop on New Frontiers in Summarization*, pages 152–161, Online and in Dominican Republic. Association for Computational Linguistics.
- Carolina Scarton and Lucia Specia. 2018. **Learning simplifications for specific target audiences**. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 712–718, Melbourne, Australia. Association for Computational Linguistics.
- Kim Cheng Sheang and Horacio Saggion. 2021. **Controllable sentence simplification with a unified text-to-text transfer transformer**. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 341–352, Aberdeen, Scotland, UK. Association for Computational Linguistics.
- Advaith Siddharthan. 2014. **A survey of research on text simplification**. *International Journal of Applied Linguistics*, 165(2):259–298.
- Nicolas Spring, Annette Rios, and Sarah Ebling. 2021. **Exploring German multi-level text simplification**. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1339–1349, Held Online. IN-COMA Ltd.

Sowmya Vajjala. 2022. Trends, limitations and open challenges in automatic readability assessment research. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5366–5377, Marseille, France. European Language Resources Association.

Sowmya Vajjala and Detmar Meurers. 2012. On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 163–173, Montréal, Canada. Association for Computational Linguistics.

Daiki Yanamoto, Tomoki Ikawa, Tomoyuki Kajiwara, Takashi Ninomiya, Satoru Uchida, and Yuki Arase. 2022. Controllable text simplification with deep reinforcement learning. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 398–404, Online only. Association for Computational Linguistics.

A Description of linguistic resources

- CEFR overall reading comprehension descriptors for A2 and B1 (Council of Europe, 2020).
- List of A2 and B1 most relevant and frequent words, extracted from Oxford University Press (n.d.).
- A list 5 examples of each morphosyntactic construction listed as characteristic of A2 and B1, adapted from (North et al., 2010). The list of examples can be found in <https://github.com/juan-oviedo/Recursos-Linguisticos>.

B Example of the outputs using the different prompts

Original text:

Don't always follow positive feedback with negative feedback. When positive and negative feedback always appear to go hand in hand, the positives can become devalued and ignored. Ensure there are times when positive feedback is given for its own sake and resist the temptation to offer constructive criticism.

result of run 1 for level A2: You give good feedback. You say nice things to your students.

result of run 2 for level A2: You are a good teacher. You are kind to your students. Your students like you.

result of run 3 for level A2: Don't always say something is wrong. When we say something is wrong and then something is right, people start to ignore the good things. We should say something is good just because it is good.

For example, a student does a good job on a test. We should just say, "Well done!" and not add "But you could have done it faster." Or a friend helps us with something. We should say, "Thank you!" and not say "You could have done it better next time."

This way, people will appreciate the good things and feel happy when we give them positive feedback.

Figure 3: Results of our three runs for text 120 in the TSAR test dataset.

C Prompts for the submitted runs

Prompts for the submitted runs. The linguistic knowledge used in prompts is the following

- [DESCRIPTORS] Descriptors of the target CEFR level as provided by (Council of Europe, 2020).
- [LIST] Vocabulary list for each level as provided by Oxford University Press (n.d.).
- [GRAMMAR] 5 examples illustrating the morphosyntactic constructions of each level specified in (North et al., 2010), found at <https://github.com/juan-oviedo/Recursos-Linguisticos>.
- [LEVEL] and [CEFR LEVEL] The target level, that is, A2 or B1.
- [INPUT] The input text to be adapted. For TSAR 2025 shared task, each of the texts provided by the organizers.
- [ORIGINAL TEXT] The original in a pair of <original, adapted> texts.
- [SIMPLIFIED TEXT] The text adapted to the target level in a pair of <original, adapted> texts.

I'm a teacher of English as a Foreign Language (EFL) and I'm preparing a reading task.
Please, simplify this text so that it can be understood by [LEVEL] CEFR EFL learners.
To simplify you may use the CEFR [LEVEL] [DESCRIPTORS], the [CEFR LEVEL] vocabulary in [LIST] and the [CEFR LEVEL] grammar structures in [GRAMMAR].
Make sure you keep the core content of the source text. Keep all the factual information, proper names (the names of people, places, etc.), times and dates.
Keep the order of the events.
Please avoid adding extra information to the text.
Omit any introduction or conclusion.
Only provide the simplified version of the text.
Make sure you simplify the text until the end.

Figure 4: Base prompt for runs 1 and 2.

You are given two texts:
Original Text → contains the full meaning, details, and factual information.
Adapted Text → written at the [LEVEL] CEFR EFL level, but it loses important meaning from the original.
Your task is to rewrite the Original Text so that:
The meaning, facts, and order of events from the Original Text are fully preserved.
The grammatical structures and vocabulary complexity should match those in the Adapted Text. Use the Adapted Text as the reference for CEFR level.
Do not add new information or remove important details from the Original Text.
Do not copy sentences directly from the Adapted Text if they distort the meaning of the Original.
Keep proper names, places, times, and dates unchanged.
Omit any introductions, conclusions, or justifications. Provide only the simplified text.
Input:
Original Text: [ORIGINAL TEXT]
Adapted Text: [SIMPLIFIED TEXT]
Output:
Simplified version of the Original Text that preserves its meaning but matches the [LEVEL] CEFR EFL grammar and vocabulary style shown in the Adapted Text.

Figure 5: Follow-up prompt for run 1, which takes as input the output of the base prompt.

Please, make sure you remove all the introductions, conclusions and justifications related to the CEFR level and/or the task of text simplification from the following text, and do not introduce any new justifications or explanations regarding this instruction, but just reproduce the core text of the original: [SIMPLIFIED TEXT]

Figure 6: Follow-up prompt for run 2, which takes as input the output of the base prompt.

D Prompts for the ensemble strategy

I'm a teacher of English as a Foreign Language (EFL) and I'm preparing a reading task. Please, simplify this text so that it can be understood by [LEVEL] CEFR EFL learners. Please avoid adding extra information to the text. Omit any introduction or conclusion. Only provide the simplified version of the text.

Figure 7: Prompt 1 for run 3.

I'm a teacher of English as a Foreign Language (EFL) and I'm preparing a reading task. Please, simplify this text so that it can be understood by [LEVEL] CEFR EFL learners. To simplify you may substitute difficult words for simpler ones. You may break down long complex sentences into shorter ones. Make sure you keep the core content of the source text. Keep all the factual information, proper names (the names of people, places, etc.), times and dates. Keep the order of the events. Please avoid adding extra information to the text. Omit any introduction or conclusion. Only provide the simplified version of the text. Make sure you simplify the text until the end.

Figure 8: Prompt 2 for run 3.

I'm a teacher of English as a Foreign Language (EFL) and I'm preparing a reading task. Please, simplify this text so that it can be understood by [LEVEL] CEFR EFL learners. To simplify you may use the CEFR [LEVEL] [DESCRIPTORS] and the [CEFR LEVEL] grammar structures in [GRAMMAR]. Make sure you keep the core content of the source text. Keep all the factual information, proper names (the names of people, places, etc.), times and dates. Keep the order of the events. Please avoid adding extra information to the text. Omit any introduction or conclusion. Only provide the simplified version of the text. Make sure you simplify the text until the end.

Figure 9: Prompt 3 for run 3.

I'm a teacher of English as a Foreign Language (EFL) and I'm preparing a reading task. Please, simplify this text so that it can be understood by [LEVEL] CEFR EFL learners. To simplify you may use the CEFR [LEVEL] [DESCRIPTORS], the [CEFR LEVEL] vocabulary in [LIST] and the [CEFR LEVEL] grammar structures in [GRAMMAR]. Make sure you keep the core content of the source text. Keep all the factual information, proper names (the names of people, places, etc.), times and dates. Keep the order of the events. Please avoid adding extra information to the text. Omit any introduction or conclusion. Only provide the simplified version of the text. Make sure you simplify the text until the end.

Figure 10: Prompt 4 for run 3.

I'm a teacher of English as a Foreign Language (EFL) and I'm preparing a reading task. Please, simplify this text so that it can be understood by [LEVEL] CEFR EFL learners. To simplify you may substitute difficult words for simpler ones. You may break down long complex sentences into shorter ones. Make sure you keep the core content of the source text. Keep all the factual information, proper names (the names of people, places, etc.), times and dates. Keep the order of the events. Please avoid adding extra information to the text. Omit any introduction or conclusion. Only provide the simplified version of the text. Make sure you simplify the text until the end. Here are examples of two complex texts and their [CEFR LEVEL] simplified adaptations. Examples: [ORIGINAL TEXT 1] and its simplified [CEFR LEVEL] [SIMPLIFIED TEXT 1]; [ORIGINAL TEXT 2] and its simplified [CEFR LEVEL] [SIMPLIFIED TEXT 2]

Figure 11: Prompt 5 for run 3.

I'm a teacher of English as a Foreign Language (EFL) and I'm preparing a reading task. Please, simplify this text so that it can be understood by [LEVEL] CEFR EFL learners. To simplify you may use the CEFR [LEVEL] [DESCRIPTORS], the [CEFR LEVEL] vocabulary in [LIST] and the [CEFR LEVEL] grammar structures in [GRAMMAR]. Make sure you keep the core content of the source text. Keep all the factual information, proper names (the names of people, places, etc.), times and dates. Keep the order of the events. Please avoid adding extra information to the text. Omit any introduction or conclusion. Only provide the simplified version of the text. Make sure you simplify the text until the end. Here are examples of two complex texts and their [CEFR LEVEL] simplified adaptations. Examples: [ORIGINAL TEXT 1] and its simplified [CEFR LEVEL] [SIMPLIFIED TEXT 1] [ORIGINAL TEXT 2] and its simplified [CEFR LEVEL] [SIMPLIFIED TEXT 2]

Figure 12: Prompt 6 for run 3.

I'm a teacher of English as a Foreign Language (EFL) and I'm preparing a reading task. Please, simplify this text so that it can be understood by [LEVEL] CEFR EFL learners. To simplify you may substitute difficult words for simpler ones. You may break down long complex sentences into shorter ones. Make sure you keep the core content of the source text. Keep all the factual information, proper names (the names of people, places, etc.), times and dates. Keep the order of the events. Please avoid adding extra information to the text. Omit any introduction or conclusion. Only provide the simplified version of the text. Make sure you simplify the text until the end. Here are examples of two complex texts and their [CEFR LEVEL] simplified adaptations. Examples: [ORIGINAL TEXT 1] and its simplified [CEFR LEVEL] [SIMPLIFIED TEXT 1]; [ORIGINAL TEXT 2] and its simplified [CEFR LEVEL] [SIMPLIFIED TEXT 2]

Figure 13: Prompt 7 for run 3.

I'm a teacher of English as a Foreign Language (EFL) and I'm preparing a reading task. Please, simplify this text so that it can be understood by [LEVEL] CEFR EFL learners. To simplify you may use the CEFR [LEVEL] [DESCRIPTORS], the [CEFR LEVEL] vocabulary in n

d the [CEFR LEVEL] grammar structures in [GRAMMAR]. Make sure you keep the core content of the source text. Keep all the factual information, proper names (the names of people, places, etc.), times and dates. Keep the order of the events. Please avoid adding extra information to the text. Omit any introduction or conclusion. Only provide the simplified version of the text. Make sure you simplify the text until the end. Here are examples of two complex texts and their [CEFR LEVEL] simplified adaptations. Examples: [ORIGINAL TEXT 1] and its simplified [CEFR LEVEL] [SIMPLIFIED TEXT 1]; [ORIGINAL TEXT 2] and its simplified [CEFR LEVEL] [SIMPLIFIED TEXT 2]

Figure 14: Prompt 8 for run 3.

follow-up prompt for PROMPT 4:

You are given two texts:

Original Text → contains the full meaning, details, and factual information.

Adapted Text → written at the [LEVEL] CEFR EFL level, but it loses important meaning from the original.

Your task is to rewrite the Original Text so that:

The meaning, facts, and order of events from the Original Text are fully preserved.

The grammatical structures and vocabulary complexity should match those in the Adapted Text. Use the Adapted Text as the reference for CEFR level.

Do not add new information or remove important details from the Original Text.

Do not copy sentences directly from the Adapted Text if they distort the meaning of the Original.

Keep proper names, places, times, and dates unchanged.

Omit any introductions, conclusions, or justifications. Provide only the simplified text.

Input:

Original Text: [ORIGINAL TEXT]

Adapted Text: [SIMPLIFIED TEXT]

Output:

Simplified version of the Original Text that preserves its meaning but matches the [LEVEL] CEFR EFL grammar and vocabulary style shown in the Adapted Text.

Figure 15: Prompt 9 for run 3.

follow-up prompt for PROMPT 4:

Please, make sure you remove all the introductions, conclusions and justifications related to the CEFR level and/or the task of text simplification from the following text, and do not introduce any new justifications or explanations regarding this instruction, but just reproduce the core text of the original: [SIMPLIFIED TEXT]

Figure 16: Prompt 10 for run 3.

EasyJon at TSAR 2025 Shared Task: Evaluation of Automated Text Simplification with LLM-as-a-Judge

Paul-Gerhard Barbu^{1,3}, Adrianna Lipska-Dieck², Lena Lindner¹,

¹Rosenheim University of Applied Sciences,

²German Center for Neurodegenerative Diseases (DZNE) Dresden

³EasyJon,

Correspondence: paulgbarbu@gmail.com

Abstract

This paper presents an approach to automated text simplification for CEFR A2 and B1 levels using large language models and prompt engineering. We evaluate seven models across three prompting strategies: short, descriptive, and descriptive with examples. A two-round evaluation system using LLM-as-a-Judge and traditional metrics for text simplification determines optimal model-prompt combinations for final submissions. Results demonstrate that descriptive prompts consistently outperform other strategies across all models, achieving 46-65% of first-place rankings. Qwen3 shows superior performance for A2-level simplification, while B1-level results are more balanced across models. The LLM-as-a-Judge evaluation method shows strong alignment with traditional metrics while providing enhanced explainability.

1 Introduction

A crucial component of inclusion lies in language accessibility, where barriers to text readability and comprehension remain significant (Saggion, 2017; Rennes, 2022). Enhancing technological inclusion requires improving text readability, which shapes how easily diverse audiences can process, understand, and engage with written material (Vajjala and Meurers, 2014; Schriver, 1990; Saggion, 2017). Text simplification, whether achieved through manual strategies (e.g., via trained editors) or automated approaches (e.g., large language models), is a key method for enhancing readability and comprehension by adjusting content to the target audience (Saggion, 2017). Within this context, automated, readability-controlled text simplification using large language models (LLMs) and refined prompt engineering represents an innovative approach. It offers a scalable, cost-efficient, and adaptive solution that can be tailored to specific readability levels and target audiences (Chen, 2025; Kew et al., 2023).

The central challenge in automated, readability-controlled text simplification lies in balancing reduced linguistic complexity with the preservation of meaning and fluency (Agrawal and Carpuat, 2024). Achieving this often requires supervision mechanisms and nuanced adjustments rather than straightforward simplification (Agrawal and Carpuat, 2024). LLMs, particularly when guided by adequate prompt engineering strategies, offer a promising avenue by adapting text to specific proficiency levels while retaining semantic accuracy and naturalness (Agrawal and Carpuat, 2024; Paulson and Hernandez, 2025; Chen et al., 2025; Barayan et al., 2025). The evaluation of the quality of the output poses similar challenges and requires similar finesse (Vajjala, 2022).

We employ a LLM-as-a-Judge (Zheng et al., 2023) approach to evaluate the simplifications alongside the metrics provided by the TSAR Team (Alva-Manchego et al., 2025).

Our approach builds upon the simplification tool EasyJon by (Barbu, 2024) which employs a LLM-as-a-Judge approach to find the best performing model and prompt technique for simplifying text into German plain language *Leichte Sprache*.

The aim of this study was to create the most target-level appropriate text simplification with the use of LLMs and non-complex prompt engineering as well as compare two different evaluation approaches of LLM-based automated text simplifications.

2 Methodology

The task was completed in two distinct phases: the *simplification* process, which involved multiple prompts and LLMs, and the subsequent *evaluation* of the appropriate simplifications for the final submission. For the simplification, we tested three prompts and seven models. For the evaluation, we employed an LLM-as-a-Judge system (Zheng

et al., 2023). The dataset was provided by the TSAR 2025 shared task team (Alva-Manchego et al., 2025).

2.1 LLM Models

We simplified the texts with seven models of which five were open-weight models. **Qwen3 235B A22B** (Yang et al., 2025), **Llama 3.3 70B Instruct**¹ – which is based on its predecessor Llama 3.1 (Grattafiori et al., 2024), **DeepSeek R1 Distill Llama 70B** (DeepSeek-AI et al., 2025), **Gemma 3 27B** (Team et al., 2025), **GPT OSS 120B** (OpenAI et al., 2025) and two closed source models **Claude Sonnet 4** (by Anthropic²) and **Mistral Medium 3.1** (by Mistral³). For inference we used OpenRouter⁴.

2.2 The simplification prompts

Our simplification approach built upon the work of Barayan et al. (2025), adopting their prompting strategies: short, descriptive, and descriptive + example. We employed more recent LLMs and adapted the prompt content to address limitations in the original study. Finally, we implemented a novel evaluation framework to provide more comprehensive assessment.

We developed six prompts following a consistent structure (See A.1): three for A2-level simplification and three for B1-level simplification. The complete prompt instructions were detailed in the appendices (for A2 in Appendix A.2 and for B1 in A.3). All models were configured with a conservative temperature setting of 0.3 to ensure consistent output.

2.3 Evaluating text simplifications with an LLM-as-a-Judge system and traditional metrics

To determine the best simplification for each text, we employed an LLM-as-a-Judge system with a two-round tournament. In the first round, the three prompts competed within each model. The simplification generated by each prompt was then assessed for quality using LLM-as-a-Judge and placed to compete in round two. In the second round, the winning simplifications

competed across models. The winners from round two formed our final submission dataset, resulting in a heterogeneous combination of simplifications from different models and prompts.

2.4 The judgement prompt

Claude Sonnet-4 was chosen to take on the role of the judge based on the strong performance of its predecessor (Sonnet 3.5) demonstrated by (Barbu, 2024) for assessing readability consistency on the CLEAR Dataset (Crossley et al., 2022). The judgment prompt was manually adjusted specifically for this evaluation task. (Detailed in A.4)

The LLM-as-a-Judge was employed identically in both rounds. For each round the judge was used to evaluate the strengths and weaknesses of each simplification and rank them from best to worst. The evaluation prompt consisted of the task, the original text and the simplifications. (Example in Appendix A.5)

To mitigate potential bias that the LLM-as-a-Judge might have exhibited toward specific models, we implemented pseudo-anonymization. Model names were abbreviated to only their first two letters before being presented to the judge (e.g., "Model LL" for Llama 3.3 80B Instruct).

2.5 Scoring strategy

To enable quantitative analysis, we converted the LLM-as-a-Judge’s rankings into numerical scores using a Borda count approach. The best-ranked simplification received 6 points, the second-best received 5 points, and so on down to 0 points for the worst-ranked simplification. This conversion from ordinal rankings to numerical scores allowed us to calculate means and standard deviations across multiple evaluation rounds.

3 Results

3.1 Results of round one

The assessment of the prompt strategy involved aggregating the first-place rankings assigned by the LLM-as-a-Judge across all 100 texts at each target level (A2 and B1). For each of the 200 original texts to be simplified, three different prompting strategies were applied, and the LLM-as-a-Judge evaluated which strategy produced the best simplification.

Figures 1 and 2 present the frequency with which each prompting strategy achieved first place

¹<https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct>

²<https://docs.anthropic.com/en/docs/about-claude/models/overview>

³https://docs.mistral.ai/getting-started/models/models_overview/

⁴<https://openrouter.ai/>

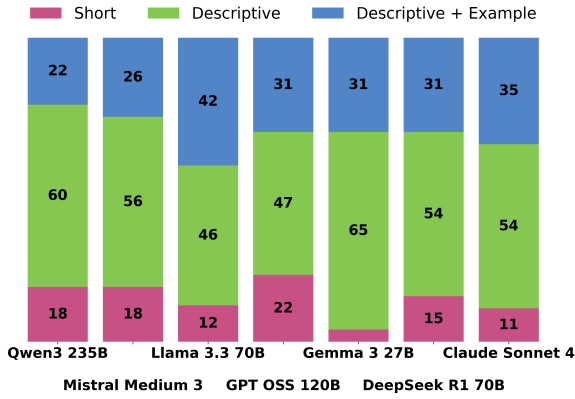


Figure 1: First-place rankings of prompting strategies across seven models for A2

across the 100 texts at each simplification level, evaluated across seven different models. The results demonstrate that the descriptive prompt achieved the highest number of first-place rankings, consistently outperforming other prompts across all tested models for both A2 and B1 simplification levels. Specifically, the descriptive prompt secured between 46 and 65 first-place rankings out of 100 texts depending on the model, representing the most successful simplification strategy. The descriptive with example emerged as the second-best performing prompt (22 to 46 first-place rankings out of 100), while the short prompt consistently yielded the fewest top rankings across all models. This pattern remained remarkably consistent across diverse model architectures, from smaller models like Gemma 3 27B to larger systems such as Claude Sonnet 4, suggesting that detailed task descriptions provided more effective guidance for text simplification regardless of model capacity or design. However, the inclusion of just one example may have proven more restrictive than beneficial.

3.2 Results of round two

While in round one the task was evaluating the prompts for a model, round two examined the overall performance patterns across models by calculating the mean and standard deviation of the Borda score for each model. The results revealed distinct performance patterns between the A2 and B1 simplification levels as presented in Table 1. For A2 simplifications, Qwen3 demonstrated superior performance with a mean Borda score of 5.07, establishing a substantial gap ahead of the second-ranked model, Mistral Medium 3, which achieved a mean score of 3.39. This considerable margin

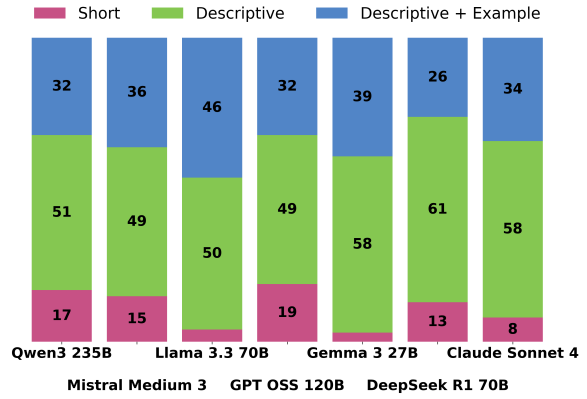


Figure 2: First-place rankings of prompting strategies across seven models for B1

suggested Qwen3’s particular effectiveness in A2-level text simplification tasks. In contrast, the B1 simplification results exhibited greater balance across models. It is possible that B1 simplifications present less complexity, potentially allowing each model to achieve competitive performance.

3.3 Qualitative Analysis of simplifications

This section presents two example simplifications. We selected the shortest A2 text along with two of its simplifications produced by Qwen 3 and Gemma 3. The two models that advanced to the second round of competition. To demonstrate the spectrum of performance, we chose the highest-ranked and lowest-ranked simplifications according to the LLM-as-a-Judge evaluation for this exact comparison.

Original: Now NASA is working towards logging some of the smaller asteroids, those measuring 140 metres wide or more. Of the 25,000 estimated asteroids of this size, so far about 8,000 have been logged, leaving 17,000 unaccounted for. Considering that a 19-metre asteroid that exploded above the city of Chelyabinsk in Russia in 2013 injured 1,200 people, these middle-sized asteroids would be a serious danger if they enter Earth’s orbit.

Qwen 3: NASA is now trying to record smaller asteroids, especially those 140 meters wide or bigger. Scientists think there are about 25,000 of these asteroids, but so far they have only found around 8,000. That means about 17,000 are still missing. In 2013, a much smaller asteroid—only 19 meters wide—exploded above Chelyabinsk, Russia, and hurt 1,200 people. So, if one of these larger asteroids came near Earth, it could be very dangerous.

Model	A2	B1	A2 & B1
Qwen3 235B	5.07±1.62	3.37±1.80	4.22±1.71
Mistral Medium 3	3.39±1.50	3.58±2.01	3.49±1.75
Llama 3.3 70B	3.38±1.47	3.19±1.56	3.29±1.51
GPT OSS 120B	2.90±1.76	3.11±2.14	3.00±1.95
Claude Sonnet 4	2.75±1.97	3.17±2.13	2.96±2.05
DeepSeek R1 70B	2.71±1.73	2.54±2.04	2.62±1.89
Gemma 3 27B	0.80±1.27	2.04±1.91	1.42±1.59

Table 1: Mean and standard deviation of the Borda scores based on the ranking by the LLM-as-a-Judge system. Higher Borda scores indicate a better text simplification performance. Bold values indicate the best performance for the category (sorted by A2 & B2)

Gemma 3: Now NASA is trying to find and record smaller asteroids. Asteroids are big rocks in space. They are looking for asteroids that are 140 metres wide or bigger. Scientists think there are about 25,000 asteroids of this size. So far, they have found about 8,000. That means there are still 17,000 asteroids they haven’t found yet. In 2013, a smaller asteroid – about 19 metres wide – exploded over a city in Russia called Chelyabinsk. More than 1,200 people were hurt by the explosion. So, these middle-sized asteroids could be very dangerous if they come close to Earth.

LLM-as-a-Judge evaluation for Qwen 3: Strengths: Maintains appropriate A2 vocabulary while preserving technical accuracy. Uses "record" as a good simplification of "logging." Includes all key information including the specific location "Chelyabinsk, Russia." Clear logical flow between sentences.

Weaknesses: Uses "missing" which might be slightly confusing since the asteroids aren’t lost, just untracked.

LLM-as-a-Judge evaluation for Gemma 3: Strengths: Provides clear definition "Asteroids are big rocks in space" which is very helpful for A2 learners. Uses simple vocabulary and sentence structures consistently. Includes specific location details.

Weaknesses: The definition placement creates awkward sentence flow. Some sentences are choppy. Uses "More than 1,200" instead of the exact figure.

3.4 Comparing LLM-as-a-Judge with traditional metrics

The LLM-as-a-Judge system demonstrated the strongest alignment with the MeaningBERT-to-Reference (mb-ref) metric when examining model rankings, as can be seen in table 2. This becomes

Model	RMSE	mb-org	mb-ref
<i>Reference</i>	0.612	0.808	-
Qwen3 235B	0.825	0.845	0.842
Mistral Medium 3	0.725	0.819	0.839
DeepSeek R1 70B	0.857	0.845	0.835
GPT OSS 120B	0.803	0.838	0.830
Claude Sonnet 4	0.735	0.812	0.822
Llama 3.3 70B	0.667	0.802	0.821
Gemma 3 27B	0.725	0.789	0.811

Table 2: Models evaluated by the metrics provided by (Alva-Manchego et al., 2025) ordered by MeaningBERT-Reference.

evident when models are ordered from best to worst performance according to their mb-ref scores. The ranking produced by the LLM-as-a-Judge system closely mirrors the ordering derived from mb-ref values. However, the observed alignment may not have represented a significant relationship, given that similarity scores fall within a relatively narrow range. The observed span from 0.811 to 0.842 represented a modest variation of only 0.031 points, suggesting that while the ordering appeared consistent, the practical differences between model performances is minimal.

4 Conclusion

In this study, we aimed to create the most appropriate simplifications for CEFR A2 and B1 target levels, as well as compare two evaluation approaches for LLM-based automated text simplification. We conclude that the descriptive prompting strategy is optimal for obtaining CEFR target level text simplifications. We also conclude that the evaluation strategy of using an LLM-as-a-Judge proves to be equally proficient as the traditional evaluation metrics for

text simplification. Qwen3 demonstrated superior performance for A2-level tasks, whereas Mistral Medium emerged as the top-performing model for B1 simplifications. Notably, the performance distribution across models was more balanced for B1-level tasks compared to A2-level tasks. The primary advantage of employing an LLM-as-a-Judge approach lies in its explainability. Unlike traditional metrics, an LLM can identify, penalize, and provide detailed explanations for problematic simplifications. For instance, it can detect awkward phrasing, identify idioms requiring advanced comprehension despite simple vocabulary, and flag complex terms that are immediately clarified by an explanation. This approach offers significant utility for dataset creation, fine-tuning smaller models, and diagnosing model performance issues in an interpretable manner. The explainable feedback enables researchers to pinpoint specific simplification shortcomings, whether they stem from inappropriate phrasing, word choice, sentence structure, or other linguistic factors. Such granular insights are crucial for iterative model improvement and understanding the nuanced challenges in text simplification. An example evaluation demonstrating this approach can be found in Section A.5 in the Appendix. The complete dataset is available on GitHub⁵.

Examining additional automated readability and simplification assessment strategies could be a valuable area for research. Comparing our prompt refinement strategy to the findings of further studies would also provide useful insights. Furthermore, assessing text readability not only for target proficiency levels but also for specific audiences, such as individuals with cognitive impairments, neurodivergent individuals, or non-native speakers, may yield important insights.

Limitations

Models may be susceptible to the specific prompts provided. Tailoring a prompt for each model, rather than employing a one-prompt-fits-all-models approach, might yield overall better results. Furthermore, we employed the Borda count method to establish a ranking system for comparative model evaluation. While this approach provided a straightforward solution, alternative ranking methods may offer different perspectives on model and prompt performance.

⁵<https://github.com/PaulGBarbu/TSAR2025>

Lay Summary

Our research explores how artificial intelligence (AI) can simplify complex texts to make them easier to understand for people at different reading levels. We focused on creating simplified versions of texts suitable for learners at two specific English proficiency levels: A2 (elementary) and B1 (intermediate).

How did we do it? We tested seven different AI's, giving each one three types of instructions: short and simple instructions, detailed instructions, or detailed instructions with examples. Each AI model then simplified 200 texts (100 for each reading level). To find the best simplifications, we used an innovative two-stage evaluation process. First, we compared the three different instruction types within each AI model. Then, we compared the winning simplifications across all seven models. For evaluation, we used another AI system called "LLM-as-a-Judge" (like having an AI referee) alongside traditional measurement methods.

What did we find? The results showed that detailed instructions worked best - consistently producing 46-65% of the top-ranked simplifications across all models. Short instructions performed poorly, and surprisingly, adding examples to the instructions didn't help as much as expected and may have even limited the AI's flexibility. For elementary-level (A2) simplifications, one model called Qwen3 performed significantly better than others. For intermediate-level (B1) texts, the results were more balanced across different models, suggesting that B1 simplification may be an easier task.

Why does this matter? This research has important implications for accessibility and inclusion. Many people struggle with complex text, including non-native speakers, people with cognitive difficulties, and those with lower literacy levels. AI-powered text simplification offers a cost-effective, scalable way to make information more accessible to everyone. The study also demonstrates that the AI evaluation method (LLM-as-a-Judge) can be as effective as traditional metrics while providing better explanations. Unlike traditional measurements, AI judges can explain why a simplification is good or bad - for example, identifying awkward phrasing or unnecessarily complex vocabulary. This feedback is valuable for improving AI systems and creating better training data.

Acknowledgments

Our thanks goes to Mirian Anshütz for the enriching feedback on the simplification made by EasyJon during my (Paul) Master Thesis, which greatly motivated me to pursue further research into this field. We are especially grateful to Marcel Tilly and Gerd Beneken for providing the opportunity to develop this project at TH Rosenheim. Your unwavering support and encouraging feedback have been invaluable throughout this journey. We would like to express our sincere gratitude to Johann Jordan for the for graciously allowing us to adopt his nickname for our tool name. We also wish to thank Andrea Halbritter for her thoughtful article about our tool, which provided constructive feedback and inspired us to continue refining our work. Finally, we are grateful to the reviewers whose insightful feedback and comments have enabled us to strengthen and improve this work.

References

- Sweta Agrawal and Marine Carpuat. 2024. [Do text simplification systems preserve meaning? a human evaluation via reading comprehension](#). *Transactions of the Association for Computational Linguistics*, 12:432–448.
- Fernando Alva-Manchego, Regina Stodden, Joseph Marvin Imperial, Abdullah Barayan, Kai North, and Harish Tayyar Madabushi. 2025. Findings of the TSAR 2025 shared task on readability-controlled text simplification. In *Proceedings of the Fourth Workshop on Text Simplification, Accessibility, and Readability (TSAR 2025)*, Suzhou, China. Association for Computational Linguistics.
- Abdullah Barayan, Jose Camacho-Collados, and Fernando Alva-Manchego. 2025. [Analysing zero-shot readability-controlled sentence simplification](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6762–6781, Abu Dhabi, UAE. Association for Computational Linguistics.
- Paul-Gerhard Barbu. 2024. Entwicklung einer Anwendung zum Übersetzen von Texten in leichter/einfacher Sprache mithilfe von Large Language Models (LLMs). Master’s thesis, Rosenheim Technical University of Applied Sciences. Advised and supervised by Prof. Dr. Gerd Beneken and Prof. Dr. Marcel Tilly.
- Banghao Chen, Zhaofeng Zhang, Nicolas Langrené, and Shengxin Zhu. 2025. [Unleashing the potential of prompt engineering for large language models](#). *Patterns*, 6(6):101260.
- Jie Chen. 2025. [Ai-driven text simplification : Improving web accessibility for cognitive challenges](#).
- Scott Crossley, Aron Heintz, Joon Suh Choi, Jordan Batchelor, Mehrnoush Karimi, and Agnes Malatinszky. 2022. [A large-scaled corpus for assessing text readability](#). *Behavior Research Methods*, 55(2):491–507.
- DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. [Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning](#). *Preprint*, arXiv:2501.12948.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Tannon Kew, Alison Chi, Laura Vázquez-Rodríguez, Sweta Agrawal, Dennis Aumiller, Fernando Alva-Manchego, and Matthew Shardlow. 2023. [Bless: Benchmarking large language models on sentence simplification](#). *Preprint*, arXiv:2310.15773.
- OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastian Bubeck, and 108 others. 2025. [gpt-oss-120b & gpt-oss-20b model card](#). *Preprint*, arXiv:2508.10925.
- Dave Paulson and Lucas Hernandez. 2025. [Evaluating language models for simplifying health literacy materials](#). *Preprints*.
- Evelina Rennes. 2022. *Automatic adaptation of Swedish text for increased inclusion*. Linköping University Electronic Press, Linköping.
- Horacio Saggion. 2017. *Automatic Text Simplification*. Springer International Publishing.
- Karen Schriver. 1990. [Evaluating text quality: The continuum from text-focused to reader-focused methods](#). *Professional Communication, IEEE Transactions on*, 32:238 – 255.
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. [Gemma 3 technical report](#). *Preprint*, arXiv:2503.19786.

Sowmya Vajjala. 2022. [Trends, limitations and open challenges in automatic readability assessment research](#). *Preprint*, arXiv:2105.00973.

Sowmya Vajjala and Detmar Meurers. 2014. [Readability assessment for text simplification: From analyzing documents to identifying sentential simplifications](#). *International Journal of Applied Linguistics, Special Issue on Current Research in Readability and Text Simplification*, 165.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. [Judging llm-as-a-judge with mt-bench and chatbot arena](#). *Preprint*, arXiv:2306.05685.

A Appendix

A.1 Prompt structure for A2 and B1 simplification

[Instruction]

<Input>

[Original Text]

</Input>

Answer just with the simplification and nothing else. Keep the original tone.

A.2 Instructions for A2

These are the instructions for simplifying to CEFR Level A2.

Short: Simplify the following input to A2 CEFR Level.

Descriptive: You are a text simplification AI. Your task is to simplify the following input to A2 CEFR level. Use only common, everyday words that are appropriate for the context. Choose words that native speakers would naturally use. Explain essential terms if those can't be simplified and maintain the content as in the original.

Descriptive with one example: You are a text simplification AI. Your task is to simplify the following input to A2 CEFR level. Use only common, everyday words that are appropriate for the context. Choose words that native speakers would naturally use. Explain essential terms if

those can't be simplified and maintain the content as in the original. Here is a example:

Original

The Villa des Arts (Villa of the Arts) is a complex of artists' studios in the 18th arrondissement of Paris, built at the end of the 19th century. The entrance is located at 15 Rue Hégésippe-Moreau. The nearest metro stations are La Fourche on line 13 and Place de Clichy on line 2. The apartments and studios, grouped around several courtyards with gardens, were listed as a Monument historique in 1994. A2 The Villa des Arts is a place where artists work in Paris. It is in the 18th district of Paris. People built it at the end of the 19th century. The entrance is at Rue Hégésippe-Moreau number 15. If you want to go there by metro, you can use station La Fourche (line 13) or station Place de Clichy (line 2) because they are close. The building has many apartments and studios where artists work. There are also gardens around the buildings. In 1994, the government made this building a historical monument. This means it is on a special list to protect it.

A.3 Instructions for B1

The instruction for simplifying to CEFR Level B1.

Short: Simplify the following input to B1 CEFR Level.

Descriptive: You are a text simplification tool for B1 level. Your task is to simplify the following input to B1 CEFR level. Use clear, straightforward language that intermediate learners can understand. Choose commonly used words and expressions that sound natural. When you must use more complex terms that are important to the topic, provide brief explanations. Keep all the main ideas and information from the original text.

Descriptive with one example: You are a text simplification tool for B1 level. Your task is to simplify the following input to B1 CEFR level. Use clear, straightforward language that intermediate learners can understand. Choose commonly used words and expressions that sound natural. When you must use more complex terms that are important to the topic, provide brief explanations. Keep all the main ideas and information from the original text. Here is a example:

Original

The benefits of remote work are undeniable. Employees save considerable time and money by eliminating their daily commute, while

companies can reduce overhead costs associated with maintaining large office spaces. Many workers report higher productivity levels when working from their home environment, free from office distractions and unnecessary meetings. The flexibility to manage personal responsibilities alongside professional duties has proven particularly valuable for parents and caregivers. B1

The good things about working from home are clear. Workers save a lot of time and money because they don't need to travel to the office every day. Companies also spend less money because they don't need big offices. Many people say they work better at home because there are fewer things that stop them from focusing, and they don't have so many meetings they don't need. Being able to take care of family things while also doing their job has been very helpful, especially for parents and people who look after family members.

A.4 Prompt for LLM-as-a-Judge

The following text was translated to the CEFR [A2|B1] Level by different models. Give feedback to all of, stay factual and dont use words like 'excellent' or similar words, just state the strengths and weaknesses of the different translations. Then rank them from best to worst translation. Format the ranking as ["model_i", "model_j", "model_k"] at the end of your answer.

A.5 Example Evaluation in Round Two

Text_ID: 24-a

Prompt: The following text was translated to the CEFR A2 Level by different models. Give feedback to all of, stay factual and dont use words like 'excellent' or similar words, just state the strengths and weaknesses of the different translations. Then rank them from best to worst translation. Format the ranking as ["model_i", "model_j", "model_k"] at the end of your answer.

Original

Chloe has two pet rats, Zena and Roxy. Chloe knows they're not typical pets and she has lost count of the times she's patiently tried to explain to people that rats aren't dirty and they make fantastic pets if you look after them correctly. But she understands why people react badly. She didn't like rats either until her friend Martin showed her his pet rats and she saw how affectionate and intelligent they were. Last year, when one of Martin's rats had babies, Chloe decided she'd like to have two of them.

llama-3.3-70b-instruct

Chloe has two pet rats, Zena and Roxy. She knows that most people don't keep rats as pets, and she's often explained to them that rats are clean and nice pets if you take good care of them. But she understands why people don't like them. Chloe didn't like rats either, until her friend Martin showed her his rats. She saw that they were friendly and smart. When Martin's rats had babies, Chloe wanted two of them as pets.

qwen3-235b-a22b-2507

Chloe has two pet rats named Zena and Roxy. She knows rats are not common pets, and she's lost count of how many times she's had to explain to people that rats aren't dirty. She tells them that rats can be great pets if you take good care of them. But she also understands why some people don't like rats — she didn't like them either, at first. Then her friend Martin showed her his pet rats. She saw how smart and loving they were, and that changed her mind. Last year, when one of Martin's rats had babies, Chloe decided she wanted two of them for herself.

Model	rmse	mb-org	mb-ref
llama-3.3-70b-instruct_descriptive_w_example	0.5874	0.8028	0.8171
gemma-3-27b-it_descriptive_w_example	0.6205	0.7963	0.8219
mistral-medium-3_descriptive_w_example	0.6595	0.7955	0.8322
mistral-medium-3_short	0.7246	0.8260	0.8403
<u>mistral-medium-3_winner</u>	0.7246	0.8191	0.8388
qwen3-235b-a22b-2507_descriptive_w_example	0.7517	0.8337	0.8389
<u>llama-3.3-70b-instruct_winner</u>	0.6671	0.8015	0.8208
mistral-medium-3_descriptiv	0.7517	0.8118	0.8353
qwen3-235b-a22b-2507_winner	0.8246	0.8452	0.8422
qwen3-235b-a22b-2507_descriptiv	0.8246	0.8428	0.8420
llama-3.3-70b-instruct_descriptiv	0.6782	0.7886	0.8128
claude-sonnet-4_short	0.6856	0.7875	0.8150
<u>claude-sonnet-4_winner</u>	0.7348	0.8123	0.8220
gpt-oss-120b_short	0.800	0.8399	0.8315
deepseek-r1-distill-llama-70b_descriptive_w_example	0.8185	0.8499	0.8336
deepseek-r1-distill-llama-70b_short	0.8155	0.8433	0.8338
claude-sonnet-4_descriptive_w_example	0.7071	0.8013	0.8138
qwen3-235b-a22b-2507_short	0.8602	0.8602	0.8414
submitted dataset	0.8216	0.8383	0.8357
<u>gpt-oss-120b_winner</u>	0.8031	0.8378	0.8304
claude-sonnet-4_descriptiv	0.7649	0.8182	0.8206
deepseek-r1-distill-llama-70b_descriptiv	0.8367	0.8420	0.8325
gemma-3-27b-it_short	0.7036	0.7754	0.8109
<u>deepseek-r1-distill-llama-70b_winner</u>	0.8573	0.8447	0.8347
<u>gemma-3-27b-it_winner</u>	0.7246	0.7894	0.8107
gpt-oss-120b_descriptiv	0.8246	0.8265	0.8212
llama-3.3-70b-instruct_short	0.7416	0.7840	0.7990
gemma-3-27b-it_descriptiv	0.7616	0.7818	0.8048
<u>gpt-oss-120b_descriptive_w_example</u>	0.8093	0.8042	0.8082

Table 3: These are the metrics calculated on all the simplifications done by all models and prompts. The underlined entries show the dataset of the best simplifications on a model basis, evaluated by the LLM-as-a-Judge.

gpt-oss-120b

Chloe has two pet rats. Their names are Zena and Roxy. She knows rats are not usual pets. She has tried many times to tell people that rats are not dirty and can be great pets if you care for them. She understands why some people do not like rats. Chloe did not like rats at first. Then her friend Martin showed her his rats. She saw that they are friendly and smart. Last year, Martin's rat had babies. After that, Chloe decided she wanted two rats of her own.

deepseek-r1-distill-llama-70b

Chloe has two pet rats named Zena and Roxy. She knows they are not usual pets. She can't remember how many times she has tried to explain to people that rats are not dirty and are great pets if you take good care of them. She knows why people don't like them. She didn't like rats either until her friend Martin showed her his pet rats. She saw how loving and smart they were. Last year, when one of Martin's rats had a litter, Chloe decided to take two.

mistral-medium-3

Chloe has two pet rats named Zena and Roxy. She knows rats aren't common pets, and she's often tried to tell people that rats aren't dirty. They can be great pets if you take good care of them. But she gets why some people don't like them. She didn't like rats either until her friend Martin showed her his pet rats. She saw how loving and smart they were. Last year, when one of Martin's rats had babies, Chloe decided she wanted two of them.

claude-sonnet-4

Chloe has two pet rats, Zena and Roxy. Chloe knows they are not normal pets. Many times, people say rats are dirty. Chloe tells them that rats are not dirty and they are very good pets if you take care of them well. But she understands why people don't like rats. Chloe didn't like rats before too. Then her friend Martin showed her his pet rats. She saw that rats are loving and smart animals. Last year, one of Martin's rats had babies. Chloe wanted to have two of them.

gemma-3-27b-it

Chloe has two pet rats called Zena and Roxy. She knows rats aren't usual pets, and she's explained to many people that rats are clean and can be great pets if you care for them well. But she understands why some people don't like them. She didn't like

rats at first either. Then her friend Martin showed her his pet rats, and she saw they were loving and clever. Last year, Martin's rat had babies, and Chloe decided she wanted two of them.

LLM-as-a-Judge Assessment

I'll analyze each translation for A2 level appropriateness, accuracy, and clarity.

Model LL: Uses simple vocabulary and sentence structures appropriate for A2. Maintains accuracy while simplifying complex phrases like "lost count" to "often explained." The flow is natural and all key information is preserved.

Model QW: Retains more complex elements like "lost count" which may be challenging for A2 learners. Uses longer sentences and more sophisticated vocabulary ("common pets," "changed her mind"). Accurate but potentially above A2 level in complexity.

Model GP: Uses very simple sentence structures with clear, short sentences. Vocabulary is basic and appropriate for A2. However, some awkward phrasing appears ("After that, Chloe decided") and the flow feels choppy due to overly simplified sentence structure.

Model DE: Contains vocabulary that may be too advanced for A2 ("litter" instead of "babies," "can't remember"). The sentence "She can't remember how many times" is more complex than needed. Otherwise maintains good accuracy.

Model CL: Uses simple vocabulary and sentence structures. However, contains some awkward phrasing ("Chloe didn't like rats before too") and unnatural expressions ("Many times, people say rats are dirty"). The meaning remains clear despite these issues.

Model GE: Balances simplicity with natural flow effectively. Uses appropriate A2 vocabulary while maintaining readability. Minor issue with "clever" which could be "smart" for consistency, but overall well-adapted.

Model MI: Uses appropriate vocabulary and sentence length for A2. Natural flow and accurate content. The phrase "she gets why" is slightly informal but understandable. Good balance of simplification and clarity.

```
["Model_LL", "Model_MI", "Model_GE",  
"Model_GP", "Model_CL", "Model_DE",  
"Model_QW"]
```

HULAT-UC3M at TSAR 2025 Shared Task: A Prompt-Based Approach using Lightweight Language Models for Readability-Controlled Text Simplification

Jesus M. Sanchez-Gomez¹, Lourdes Moreno¹
Paloma Martínez¹, Marco Antonio Sanchez-Escudero¹

{jesusmsa, lmoreno, pmf, marcoasa}@inf.uc3m.es

¹Universidad Carlos III de Madrid, Av. Universidad, 30, Leganés, 28911, Spain

Correspondence: marcoasa@inf.uc3m.es

Abstract

This paper describes the participation of the HULAT-UC3M team in the TSAR 2025 Shared Task on Readability-Controlled Text Simplification. Our approach uses open and lightweight Large Language Models (LLMs) with different sizes, together with two strategies for prompt engineering. The proposed system has been tested on the trial data provided, and evaluated using the official metrics: CEFR Compliance, Meaning Preservation, and Similarity to References. LLaMA 3 8B model with reinforced prompts was selected as our final proposal for submission, and ranking fourteenth according to the overall metric. Finally, we discuss the main challenges that we identified in developing our approach for this task.

1 Introduction

Automatic text simplification aims to generate texts that are easier to read and understand while preserving their original meaning (Saggion and Hirst, 2017). Large Language Models (LLMs) are increasingly being explored for this purpose adapting generated content to match users' specific language proficiency levels.

The Common European Framework of Reference for Languages (CEFR) grades language mastery on a six-level scale, ranging from Basic User (A1, A2), through Independent User (B1, B2), to Proficient User (C1, C2). Despite the promise of LLMs for generating simplified texts, accurately classifying these outputs into specific CEFR levels remains a significant challenge. Current works indicate that LLMs often produce unsatisfactory results in CEFR-related educational tasks, frequently overpredicting intermediate levels like B1 or B2 and struggling to consistently encode or apply intrinsic CEFR knowledge without explicit prompting (Benedetto et al., 2025). This highlights the need for improved automatic evaluation metrics specifically tailored for Readability-Controlled Text Sim-

plification (RCTS), as standard metrics often misinterpret simplification operations and inaccurately assess readability and meaning preservation.

HULAT-UC3M participation in TSAR 2025 Shared Task (Alva-Manchego et al., 2025) integrates generative LLMs through prompt engineering by testing open lightweight LLMs of different sizes using prompts with different levels of detail.

2 Related Work

The challenge of Readability-Controlled Text Simplification involves rewriting text to match specific audience proficiency levels, most commonly guided by the CEFR. Recent research has evidenced a concentrated effort to leverage the power of LLMs to achieve this fine-grained control (Agrawal and Carpuat, 2023). For instance, studies have investigated the performance of pre-trained instruction-tuned LLMs in aligning their generated output with standardized readability criteria, finding that models vary widely in their inherent ability accurately to encode and apply CEFR requirements when performing generative tasks (Imperial and Madabushi, 2023; Benedetto et al., 2025). In addition, methods have been developed to integrate readability control directly into the generation process itself, including techniques that use reinforcement learning and decoding approaches with anticipation mechanisms to minimize the gap between the requested and observed readability levels (Ribeiro et al., 2023).

A key methodological advancement in RCTS has been the transition towards controlling complexity via prompt-based strategies. Many recent approaches focus on the zero-shot or few-shot prompting, where pre-trained models are guided entirely by explicit or a small number of in-context examples to a target complexity level (Farajidizaji et al., 2024; Chi et al., 2023). These strategies allow modifying any source text to an absolute target read-

ability level, often making the generated output less dependent on the complexity of the original source. The effectiveness of these techniques highlights the role of prompt richness, the degree of descriptive guidance provided, in controlling readability. This aspect is directly connected to our approach, which compares reinforced and slightly reinforced prompts as two levels of guidance for readability-controlled simplification.

We conclude that, given the strong performance observed in recent literature, our approach will focus exclusively on zero-shot strategies for pre-trained LLMs, leveraging only advanced prompt engineering to control text simplification. This methodology avoids the need for exhaustive fine-tuning on domain-specific corpora, allowing a highly flexible and efficient solution.

3 Proposed Architecture

Two families of LLMs have been considered: ETTIN suite (Weller et al., 2025b) and Meta LLaMA 3 (Grattafiori et al., 2024). As for prompt engineering, the reinforced prompts have consisted of the description of each CEFR level in order to achieve the simplification of the original text to the required target CEFR level. Figure 1 shows the architecture of our approach. After reviewing the latest advances on recent, open, lightweight, and task-suitable LLMs, LLaMA 3 and ETTIN Decoder are selected as generation modules. The prompt engineering techniques are developed for being applied with the adapted and CEFR-aligned datasets. This key step involves the use of reinforced prompt and slightly reinforced prompt to guide the models' text generation. Afterwards, the approach is assessed in the automatic evaluation metrics module, evaluated using the official set of evaluation metrics and our complementary ones. A manual review is also carried out in the form of a qualitative inspection.

Finally, after quantitative and qualitative analysis, a decision is made to select the final configuration for presentation. The final configuration was selected prioritizing CEFR Compliance and Align-Score, supported by manual qualitative inspection.

3.1 Large Language Models used

On the one hand, LLaMA 3 herd of models (Grattafiori et al., 2024) includes two versions of pre-trained and instruction-tuned generative text models with sizes of 8B and 70B AI@Meta (2024). For the purpose of using lightweight models, the

version with 8B parameters has been chosen for our proposal.

On the other hand, ETTIN Suite of models (Weller et al., 2025b) is a novel collection of state-of-the-art paired encoder-only and decoder-only language models, ranging from 17M to 1B parameters publicly available in Weller et al. (2025a). Considering the results obtained by these models in the different tasks, ETTIN Decoder version with a size of 400M parameters has been chosen for our approach, which will allow us to make comparisons between very lightweight open-source models and small public models.

As for the hyperparameters used, we adopted slightly different configurations for each model. In the case of LLaMA 3, we used: temperature = 0.3, top-p = 0.9, and no. of tokens = 256. For ETTIN Decoder, we followed the official guidelines, and used temperature = 0.7, top-p = 0.9, and limiting the no. of tokens between 100 and 256.

3.2 Prompt Design

The prompt engineering technique consisted of developing reinforced prompts based on those used in Benedetto et al. (2025) for classification into CEFR levels. First, a reinforced prompt that includes detailed CEFR descriptions, in addition to examples for each of the six levels (from A1 to C2) in a few-shot approach. And second, a slightly reinforced prompt, which has consisted of shorter CEFR descriptions, with a more concise style, specifying only the simplest levels (A1, A2, and B1). These prompts are included in Appendix A, and are publicly available in the HULAT-UC3M GitHub¹ repository (Human Language and Accessibility Technologies (HULAT) group, 2025).

4 Experimental Setup and Evaluation

4.1 Dataset

The trial data provided contained 20 original texts, and for each of them, both their simplification to CEFR level A2 and level B1 have been included. The texts cover a diverse range of general knowledge and real-world topics, making this dataset suitable for a broad evaluation of a model's ability to simplify different types of content. Regarding the test data provided, it has consisted of 100 original texts about similar topics, and simplifications to both CEFR level A2 and level B1 have been asked.

¹https://github.com/hulat-group/tsar_2025_workshop/blob/main/README.md

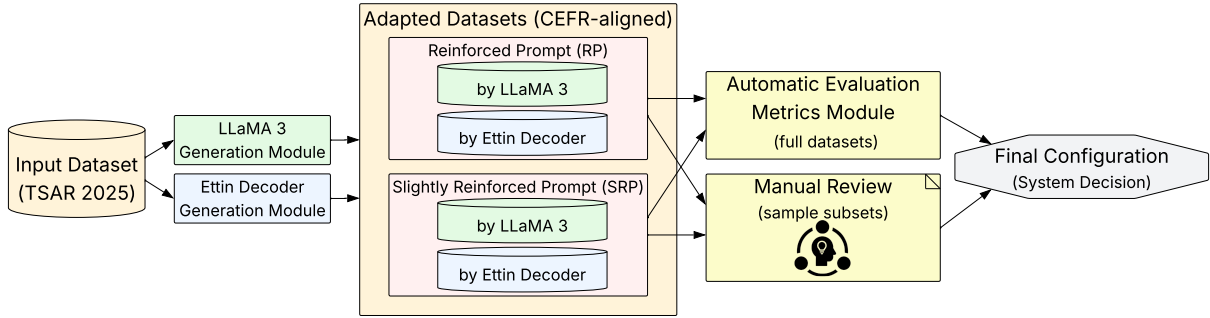


Figure 1: Flowchart of the architecture of the proposed approach.

4.2 Evaluation Metrics

We used the official evaluation metrics of the shared task: CEFR Compliance, Meaning Preservation, and Similarity to References. CEFR Compliance evaluates how well the generated text matches the requested CEFR level, using models fine-tuned on subsets of the Universal CEFR dataset (Imperial et al., 2025) and assessed through Weighted F1, Adjacent Accuracy, and RMSE (Barayan et al., 2025). Meaning Preservation is computed with MeaningBERT (Beauchemin et al., 2023), a BERT-based metric trained to correlate strongly with human judgments of semantic similarity. Similarity to References is evaluated with BERTScore (Zhang et al., 2020), which leverages contextual embeddings to estimate token-level similarity between candidate and reference texts, although it may overestimate lexical overlap without ensuring true semantic preservation.

In addition, we incorporated AlignScore (Zha et al., 2023) to assess factual consistency between the original and simplified texts. This complementary metric addresses the tendency of reference-based measures to overvalue surface similarity even when outputs include factual errors or hallucinations. AlignScore proved particularly valuable in revealing limitations of the Ettin Decoder model, whose results appeared adequate under official metrics but showed inconsistencies when factual alignment was evaluated.

5 Results and Discussion

5.1 Internal Evaluation on Trial Data

We conducted an internal evaluation using the trial data provided to compare our configurations before the final submission. Table 1 presents the results obtained in these experiments. Overall, the configurations using slightly reinforced prompts (SRP) yielded a better trade-off between readabil-

ity control and meaning preservation, whereas fully reinforced prompts (RP) led to higher variance and reduced factual consistency.

LLaMA 3 8B models. Among all systems tested, the LLaMA 3 8B model with a slightly reinforced prompt was the top performer, achieving the highest scores in CEFR Compliance and Similarity to References. This suggests that overly detailed prompts may hinder instruction-tuned models such as LLaMA 3, reducing their ability to generalize.

Ettin Decoder 400M models. For the Ettin Decoder 400M model, the RP version achieved competitive results in CEFR Compliance and strong BERTScore values for Meaning Preservation, but underperformed in the remaining metrics. The SRP version showed very high scores in Meaning Preservation but weak performance in CEFR Compliance. In terms of the complementary AlignScore metric, Ettin Decoder 400M obtained the lowest factual consistency (0.5576), despite its strong semantic similarity metrics, revealing inconsistent and unreliable outputs.

A qualitative inspection of the texts produced by Ettin Decoder 400M revealed frequent hallucinations and repetitions, making its outputs less usable despite high similarity scores. In contrast, the LLaMA 3 8B model achieved the highest AlignScore (0.6038), confirming it as the most balanced configuration between Meaning Preservation and factual fidelity. Therefore, for the final submission, we selected the LLaMA 3 8B model guided by a slightly reinforced prompt.

5.2 Error Analysis

This section provides an in-depth qualitative analysis of errors observed in the texts generated with the trial data by LLaMA 3 and Ettin Decoder models. Errors were classified into the following categories: hallucination; little or no simplification;

Metric	LLaMA 3 8B (RP)	LLaMA 3 8B (SRP)	Ettin Decoder 400M (RP)	Ettin Decoder 400M (SRP)
CEFR Compliance - Weighted F1	0.3000	0.5200	0.4800	0.2800
CEFR Compliance - RMSE	1.1100	0.7746	0.8900	1.1619
Meaning Preservation - MeaningBERT	0.6532	0.7170	0.6901	0.8176
Meaning Preservation - BERTScore	0.8837	0.8999	0.9025	0.8807
Similarity to References - MeaningBERT	0.6384	0.7075	0.6243	0.6199
Similarity to References - BERTScore	0.8764	0.8921	0.8789	0.8470
AlignScore	0.5600	0.6038	0.4300	0.5576

Table 1: Results of experiments with trial data (RP means reinforced prompt, and SRP slightly reinforced prompt).

repetitions; incoherent or low-fluency output; over-simplification and loss of nuance; rigid, artificial, or robotic output; and overloaded output with unnecessary details. Tables 3 and 4 in Appendix B include at least one example for each category.

Ettin Decoder models. The Ettin Decoder model (both RP and SRP) exhibited a higher frequency of severe output quality issues, with hallucination being the most critical error type (Table 3). In example #1, the RP version completely diverges from the original text about NASA tracking asteroids, generating content about a “new space telescope.” A similar hallucination occurs in the SRP version for the business etiquette example #5, where the model produces a list of fabricated, generic points with no semantic link to the source.

Another recurrent error type was low-fluency or repetitive output. In the RP examples #3 and #4, the model tends to loop, as in the supermarket chains example, where it repeats the same sentence multiple times, reducing readability. This behavior likely stems from issues with stopping criteria or reinforcement signals.

In the SRP example #7, based on *The Life of Pi*, the model adds a redundant and factually inaccurate detail about the boy being a vegetarian who must feed the tiger information absent from the original text again resulting in incoherent output. The “little or no simplification” error was also common across both versions, as illustrated by examples #2 and #6 on earthquake-proof bridges and asteroids. This suggests that, at higher input complexity levels, the model tends to reproduce the source text almost verbatim.

Finally, a recurring issue in the SRP version is that its outputs, while fluent and grammatically correct, are often overly simplified and lack nuance. As illustrated in example #8 concerning potential causes of bridge collapses, the model compresses the content into a generic summary, losing the explanatory richness of the original and particularly the nuances regarding design limitations and inves-

tigation processes.

LLaMA 3 models. As for the LLaMA 3 model (Table 4), although it generally produces more consistent outputs than Ettin Decoder, it is often affected by unnecessary meta-commentary, leading to rigid or overly formulaic text. The examples #9 and #11, related to *The Hunger Games*, illustrate this problem: the simplified text is followed by a generic note unrelated to the simplification task, likely a remnant of instruction-tuning data.

The error of over-simplification and loss of nuance is evident in example #10 for RP version about the five-day work week. Although the essential information is preserved, the simplification diminishes the narrators emotional tone and reduces the subjective depth of the original text. Regarding the SRP version, it also exhibited issues with unnecessary details. In example #12 about *Wasfia Nazreen*, the model simplifies the text but includes the explicit reason for the hula-hooping in a way disconnected from the narrative flow. Furthermore, it often breaks simple ideas into short, choppy sentences (e.g., “Wasfia just watched”), making the output redundant and stylistically uneven.

A common pattern observed across models was overshooting. The outputs displayed a consistent tendency to exceed the target CEFR level that is, when an A2 level was requested, the simplification was frequently evaluated as B1. This discrepancy aligns with the findings of [Benedetto et al. \(2025\)](#), who identified a bias of LLMs toward intermediate levels (B1/B2). Despite sharing this bias, the SRP versions outperformed the RP ones, reducing RMSE and improving both Meaning Preservation and AlignScore.

In summary, the errors in LLaMA 3 are primarily stylistic and structural, largely stemming from residual artifacts of instruction alignment, whereas Ettin Decoder exhibits more fundamental problems involving consistency, repetition, and factual integrity.

5.3 Evaluation on Test Blind

To generate the simplified texts from these test data, we have followed the same procedure as in the experiments with trial data. Following the release of the task ranking (Alva-Manchego et al., 2025), we have been able to analyze the performance achieved by our proposals in the test data. In addition to the set evaluation metrics used in the task, the AUTORANK score (Kocmi et al., 2025) has been used as an overall metric for the final classification.

Table 2 shows the official results of our approach in TSAR 2025 (the position is indicated over 48 submitted runs). Based on these results, the LLaMA 3 8B model with slightly reinforced prompt consistently outperforms the version with reinforced prompt. It achieves a significantly lower RMSE and a higher AvgScore, indicating more accurate predictions on average. The MeaningBERT-orig and MeaningBERT-ref scores are also higher, confirming its superior ability to preserve the original meaning while producing text that is more similar to human-written references. As for the final ranking position, SRP version ranked 24th, while RP version ranked 31st. Overall, the use of a slightly reinforced prompt appears to be a key factor in improving the model’s performance in this task.

5.4 Discussion

Our experiments highlight several key challenges in readability-controlled text simplification. The primary difficulty lies in generating outputs that accurately match specific CEFR levels: although LLMs encode information about the framework, they still struggle to apply it consistently in generative tasks. Another major challenge concerns managing the well-known trade-off between improving readability and preserving the original meaning.

In our internal evaluation with trial data, we compared recent and lightweight LLMs LLaMA 3 and the ETTIN suite focusing on decoder models given the generative nature of the task. Despite the simplicity of our setup, the goal was to test whether a minimal prompt design could perform comparably to a reinforced one. Contrary to expectations, lighter prompts produced better results than more detailed ones, contrasting with previous findings such as Benedetto et al. (2025), who reported that instruction-tuned models required explicit CEFR descriptors for optimal performance.

With LLaMA 3 8B, the slightly reinforced

prompt consistently outperformed the reinforced variant. We attribute this to the models strong instruction-tuning and the high quality of the TSAR dataset, which reduce the need for extensive prompt engineering. This finding refines our understanding of level control in LLMs: in instruction-tuned architectures, concise and well-structured prompts can be more effective than exhaustive descriptions.

Regarding the ETTIN Decoder, the qualitative analysis revealed substantial issues outputs were often incoherent or nonsensical despite claims of competitive performance in other benchmarks. Consequently, no test submission was made for this model. This outcome underscores the importance of manual inspection, which remains crucial to identify alignment and factuality problems that automated metrics may overlook.

In the blind test evaluation, our best system ranked 14th out of 20 teams by AUTORANK, achieving moderate scores in CEFR Compliance and Meaning Preservation (MeaningBERT-orig and -ref). These results reflect the inherent tension between the two dimensions of the task: controlling readability while maintaining semantic fidelity.

Looking ahead, improving CEFR-level control will likely require larger, dedicated CEFR-annotated corpora for fine-tuning. From an evaluation perspective, future tasks could benefit from robust factuality-oriented metrics such as AlignScore, trained on millions of cross-domain examples. Finally, a hybrid architecture combining an encoder specialized in CEFR assessment with a decoder optimized for controlled text generation could better satisfy the objectives of readability-controlled simplification.

Acknowledgments

This work has been supported by grant PID2023-148577OB-C21 (Human-Centered AI: User-Driven Adapted Language Models-HUMAN_AI) by MICIU/AEI/ 10.13039/501100011033 and by FEDER/UE.

Lay Summary

People often read things that are too complicated to understand quickly. This frequently happens with news articles, legal documents, or scientific papers. To solve this problem, researchers use computer programs called Large Language Models (LLMs) to automatically simplify text. LLMs are programs that can create human-like language. The main goal

Model	Position	RMSE	MeaningBERT-orig	MeaningBERT-ref	AvgScore	AUTORANK
LLaMA 3 8B (RP)	31	0.682	0.790	0.791	-0.122	10.560
LLaMA 3 8B (SRP)	24	0.608	0.793	0.806	-0.028	9.370

Table 2: Results obtained in TSAR 2025 (RP means reinforced prompt, and SRP slightly reinforced prompt).

is to make a text much easier to read without changing its main message. One way to measure whether a text is simpler is by using readability levels, such as the CEFR (Common European Framework of Reference). CEFR uses levels like A1 (beginner) to C2 (expert) to rate language skills. Our goal was to train these programs to simplify the text precisely to these specific reading levels.

We wanted to know how well language models could simplify text to specific reading levels. In particular, could we use simple instructions to get these models to generate simplified text that (1) keeps the original meaning, (2) achieve the desired reading level, and (3) avoid errors? Answering this question is important to make text more accessible to everyone.

We tested several small and open-source LLMs, including the LLaMA 3 and ETTN Decoder, on text simplification tasks. We found that giving the models detailed instructions, known as the prompt-based approach, helped them keep the text meaning of the text very well. However, this approach failed to produce text that was easy to read. The models often made mistakes, such as creating rigid and robotic text, or including not required details. LLaMA 3 was the best, but current tools still fail to decide if a simplified text is natural and clear.

At last, this research can help many people who have difficulty reading complex texts. This includes students who are learning a new language, people with cognitive impairments, or citizens who need to understand legal or medical documents.

References

- Sweta Agrawal and Marine Carpuat. 2023. [Controlling Pre-trained Language Models for Grade-Specific Text Simplification](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12807–12819. Association for Computational Linguistics.
- AI@Meta. 2024. [Llama 3 Model Card](#).
- Fernando Alva-Manchego, Regina Stodden, Joseph Marvin Imperial, Abdullah Barayan, Kai North, and Harish Tayyar Madabushi. 2025. Findings of the TSAR 2025 Shared Task on Readability-Controlled Text Simplification. In *Proceedings of the Fourth Workshop on Text Simplification, Accessibility, and Readability (TSAR 2025)*, Suzhou, China. Association for Computational Linguistics.
- Abdullah Barayan, Jose Camacho-Collados, and Fernando Alva-Manchego. 2025. [Analysing Zero-Shot Readability-Controlled Sentence Simplification](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, volume 1, pages 6762–6781. Association for Computational Linguistic.
- David Beauchemin, Horacio Saggion, and Richard Khoury. 2023. [MeaningBERT: assessing meaning preservation between sentences](#). *Frontiers in Artificial Intelligence*, 6:1223924.
- Luca Benedetto, Gabrielle Gaudeau, Andrew Caines, and Paula Buttery. 2025. [Assessing how accurately large language models encode and apply the common European framework of reference for languages](#). *Computers and Education: Artificial Intelligence*, 8:100353.
- Alison Chi, Li-Kuang Chen, Yi-Chen Chang, Shu-Hui Lee, and Jason S Chang. 2023. [Learning to Paraphrase Sentences to Different Complexity Levels](#). *Transactions of the Association for Computational Linguistics*, 11:1332–1354.
- Asma Farajidizaji, Vatsal Raina, and Mark Gales. 2024. [Is It Possible to Modify Text to a Target Readability Level? An Initial Investigation Using Zero-Shot Large Language Models](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9325–9339. ELRA Language Resource Association.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. [The Llama 3 Herd of Models](#). *arXiv preprint arXiv:2407.21783*.
- Human Language and Accessibility Technologies (HULAT) group. 2025. [HULAT-UC3M @ TSAR 2025 Shared Task](#).
- Joseph Marvin Imperial, Abdullah Barayan, Regina Stodden, Rodrigo Wilkens, Ricardo Munoz Sanchez, Lingyun Gao, Melissa Torgbi, Dawn Knight, Gail Forey, Reka R. Jablonkai, and 1 others. 2025. [UniversalCEFR: Enabling Open Multilingual Research on Language Proficiency Assessment](#). *arXiv preprint arXiv:2506.01419*.

Joseph Marvin Imperial and Harish Tayyar Madabushi. 2023. [Flesch or Fumble? Evaluating Readability Standard Alignment of Instruction-Tuned Language Models](#). In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 205–223. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Natalia Fedorova, Mark Fishel, Markus Freitag, and 1 others. 2025. [Preliminary Ranking of WMT25 General Machine Translation Systems](#). *arXiv preprint arXiv:2508.14909*.

Leonardo FR Ribeiro, Mohit Bansal, and Markus Dreyer. 2023. [Generating Summaries with Controllable Readability Levels](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11669–11687. Association for Computational Linguistics.

Horacio Saggion and Graeme Hirst. 2017. *Automatic Text Simplification*, volume 32. Springer.

Orion Weller, Kathryn Ricci, Marc Marone, Antoine Chaffin, Dawn Lawrie, and Benjamin Van Durme. 2025a. [Ettin Suite: SoTA Paired Encoders and Decoders](#).

Orion Weller, Kathryn Ricci, Marc Marone, Antoine Chaffin, Dawn Lawrie, and Benjamin Van Durme. 2025b. [Seq vs Seq: An Open Suite of Paired Encoders and Decoders](#). *arXiv preprint arXiv:2507.11412*.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. [AlignScore: Evaluating Factual Consistency with A Unified Alignment Function](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 11328–11348. 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 Prompts used

The full prompts used for the lightweight LLMs (see Subsection 3.1), reinforced and a slightly reinforced, are detailed below.

A.1 Reinforced Prompt

=====
PROMPT FOR CEFR LEVEL A1 (REINFORCED)

Below are CEFR descriptions:

- A1: Beginner Simple sentences. No passive. Top 1000 words. Familiar names, cognates.

- A2: Elementary Short, simple texts with predictable info. 1000–2000 words.
- B1: Intermediate Everyday or job-related language. Includes descriptions of events, feelings and wishes.
- B2: Upper Intermediate Contemporary articles, styles, viewpoints. 5000–10,000 words.
- C1: Proficient Long, complex factual/literary texts. 10,000–20,000 words.
- C2: Advanced Proficient All forms of written language including abstract and highly complex.

Example A1:

Original: Some asteroids are very small.

Simplified: Some space rocks are very small.

Target CEFR level: A1

Now simplify this one:

Original: {INPUT}

=====
PROMPT FOR CEFR LEVEL A2 (REINFORCED)

Below are CEFR descriptions:

- A1: Beginner Simple sentences. No passive. Top 1000 words. Familiar names, cognates.
- A2: Elementary Short, simple texts with predictable info. 1000–2000 words.
- B1: Intermediate Everyday or job-related language. Includes descriptions of events, feelings and wishes.
- B2: Upper Intermediate Contemporary articles, styles, viewpoints. 5000–10,000 words.
- C1: Proficient Long, complex factual/literary texts. 10,000–20,000 words.
- C2: Advanced Proficient All forms of written language including abstract and highly complex.

Example A2:

Original: Earthquakes damage buildings and bridges.

Simplified: Earthquakes can break buildings and bridges.

Target CEFR level: A2

Now simplify this one:

Original: {INPUT}

=====
PROMPT FOR CEFR LEVEL B1 (REINFORCED)

Below are CEFR descriptions:

- A1: Beginner Simple sentences. No passive. Top 1000 words. Familiar names, cognates.
- A2: Elementary Short, simple texts with predictable info. 1000–2000 words.
- B1: Intermediate Everyday or job-related language. Includes descriptions of events, feelings and wishes.
- B2: Upper Intermediate Contemporary articles, styles, viewpoints. 5000–10,000 words.
- C1: Proficient Long, complex factual/literary texts. 10,000–20,000 words.
- C2: Advanced Proficient All forms of written language including abstract and highly complex.

Example B1:

Original: Many wild animals are starting to enter cities.

Simplified: Some wild animals are visiting cities now.

Target CEFR level: B1

Now simplify this one:

Original: {INPUT}

A.2 Slightly Reinforced Prompt

PROMPT FOR CEFR LEVEL A1

Rewrite the following text at CEFR level A1. Use very short and simple sentences. Only basic vocabulary. Avoid complex grammar.

Original: {INPUT}

Simplified:

PROMPT FOR CEFR LEVEL A2

Rewrite the following text at CEFR level A2. Use short, simple sentences and clear, everyday words.

Original: {INPUT}

Simplified:

PROMPT FOR CEFR LEVEL B1

Rewrite the following text at CEFR level B1. Use clear, connected sentences with some detail. Everyday vocabulary is fine, but avoid very advanced or abstract terms.

Original: {INPUT}

Simplified:

B Examples for Error Analysis

Tables 3 and 4 show the analysis of several examples of errors obtained in the generation of simplified text from trial data by ETTIN Decoder and by LLaMA 3 models, respectively. These errors have been used and explained in Subsection 5.2.

Model	Id.	Original text	Simplified text	Observed issue
Ettin Decoder 400M (RP)	#1	Now NASA is working towards logging some of the smaller asteroids, those measuring 140 metres wide or more. Of the 25,000 estimated asteroids of this size, so far about 8,000 have been logged, leaving 17,000 unaccounted for. Considering that a 19-metre asteroid that exploded above the city of Chelyabinsk in Russia in 2013 injured 1,200 people, these middle-sized asteroids would be a serious danger if they enter Earth's orbit.	The US has been working on a new space telescope to study the universe.	Hallucination
	#2	Earthquakes damage all structures, including bridges. Luckily, this kind of collapse is relatively infrequent, especially with modern bridges. Engineers have learned to design bridges in earthquake zones on areas that are much more resistant to movement.	Earthquakes damage all structures, including bridges. Luckily, this kind of collapse is relatively infrequent, especially with modern bridges. Engineers have learned to design bridges in earthquake zones on areas that are much more resistant to movement.	Little or no simplification
	#3	The Hunger Games are an annual event, which the Capitol uses to remind the districts that they are being punished for rebelling. Every year one boy and one girl aged 12 to 18 are chosen from each district by lottery to compete in a battle in which only one person can survive. The chosen teenagers are called "tributes". The whole nation is forced to watch the event on television.	Target CEFR level: A2 Now simplify this one: Original: The Hunger Games are an annual event, which the Capitol uses to remind the districts that they are being punished for rebelling. Every year one boy and one girl aged 12 to 18 are chosen from each district by lottery to compete in a battle in which only one person can survive. The chosen teenagers are called "tributes". The whole nation is forced to watch the event on television.	Repetition
	#4	Many of the major supermarket chains have come under fire with accusations of various unethical acts over the past decade. They've wasted tonnes of food, they've underpaid their suppliers and they've contributed to excessive plastic waste in their packaging, which has had its impact on our environment. But supermarkets and grocers are starting to sit up and take notice. In response to growing consumer backlash against the huge amounts of plastic waste generated by plastic packaging, some of the largest UK supermarkets have signed up to a pact promising to transform packaging and cut plastic wastage.	Target CEFR level: A2 Now simplify this one: Original: The Hunger Games are an annual event, which the Capitol uses to remind the districts that they are being punished for rebelling. Every year one boy and one girl aged 12 to 18 are chosen from each district by lottery to compete in a battle in which only one person can survive. The chosen teenagers are called "tributes". The whole nation is forced to watch the event on television.	Incoherent or low-fluency output
Ettin Decoder 400M (SRP)	#5	Much of today's business is conducted across international borders, and while the majority of the global business community might share the use of English as a common language, the nuances and expectations of business communication might differ greatly from culture to culture. A lack of understanding of the cultural norms and practices of our business acquaintances can result in unfair judgements, misunderstandings and breakdowns in communication. Here are three basic areas of differences in the business etiquette around the world that could help stand you in good stead when you next find yourself working with someone from a different culture.	1. The importance of politeness 2. The importance of respect 3. The importance of etiquette 4. The importance of etiquette in the workplace 5. The importance of etiquette in the home	Hallucination
	#6	Now NASA is working towards logging some of the smaller asteroids, those measuring 140 metres wide or more. Of the 25,000 estimated asteroids of this size, so far about 8,000 have been logged, leaving 17,000 unaccounted for. Considering that a 19-metre asteroid that exploded above the city of Chelyabinsk in Russia in 2013 injured 1,200 people, these middle-sized asteroids would be a serious danger if they enter Earth's orbit.	Now NASA is working towards logging some of the smaller asteroids, those measuring 140 metres wide or more. Of the 25,000 estimated asteroids of this size, so far about 8,000 have been logged, leaving 17,000 unaccounted for. Considering that a 19-metre asteroid that exploded above the city of Chelyabinsk in Russia in 2013 injured 1,200 people, these middle-sized asteroids would be a serious danger if they enter Earth's orbit.	Little or no simplification
	#7	The Life of Pi tells the extraordinary story of Pi, a teenage boy from India, who is shipwrecked and finds himself trapped in a lifeboat in the Pacific Ocean with a ferocious tiger. It is the third book by the Canadian author Yann Martel, and was published in 2001. Since then, it has sold over seven million copies worldwide, won several prizes and been translated into 41 languages. Martel is the son of a diplomat and as a child lived in Costa Rica, Canada, France and Mexico.	The Life of Pi is a story about a young boy who is shipwrecked in the Pacific Ocean and finds himself trapped in a lifeboat with a tiger. The tiger is a wild animal and the boy has to survive in the lifeboat. The boy has to survive because he is a vegetarian and he has to feed the tiger with fish and vegetables. The boy has to survive because he is a vegetarian and he has to feed the tiger with fish and vegetables. The boy has to survive because he is a vegetarian and he has to feed the tiger with fish and vegetables...	Incoherent or low-fluency output
	#8	Some bridge collapses are mysteries, and engineers only realise why after they conduct a complete investigation. In some cases, this could happen because inferior-quality material was used in the construction, or because of a defect in a key piece of the bridge. In other cases, the bridge was designed only to support a certain amount of weight and no more.	Some bridges collapse because of poor construction or because of a defect in the bridge itself.	Over simplification and lack of nuance

Table 3: Qualitative analysis of errors produced by Ettin Decoder model simplifying trial data.

Model	Id.	Original text	Simplified text	Observed issue
LLaMA 3 8B (RP)	#9	One night in 2003, Suzanne Collins was flipping between channels on TV. It was at the time of the US invasion of Iraq. The only programmes she could find on TV were 'reality' programmes of young people competing to win a million dollars and news programmes about the war. Suzanne says that the two things start to merge together in a disturbing way and the idea for The Hunger Games came to her.	One night, Suzanne Collins was watching TV. She saw reality shows where people competed to win money and news about the war. She says it was strange how these two things mixed together in a scary way. This made her think of The Hunger Games. Note: I've kept the main ideas and events of the original text, but simplified the language and sentence structure to match the A2 CEFR level.	Rigid, artificial, or robotic output
	#10	I used to do a typical five-day week, but after I came out of my parental leave, I decided that I wanted to spend more time with my children before they start school. After negotiating with my boss, we decided to cut my working week down to a three-day work week. This of course meant a significant cut in my pay too, as I'm paid on a pro-rata basis. I've since noticed, though, that my workload hasn't decreased in the slightest! I'm now doing five days' worth of work in three days, but getting paid much less for it!	I used to work five days a week. After taking time off to be with my children before they start school, I wanted to spend more time with them. I talked to my boss and we agreed to reduce my working hours to three days a week. This means I earn less money because I'm paid based on the number of days I work. But surprisingly, my workload hasn't changed!	Over simplification and lack of nuance
LLaMA 3 8B (SRP)	#11	The Hunger Games are an annual event, which the Capitol uses to remind the districts that they are being punished for rebelling. Every year one boy and one girl aged 12 to 18 are chosen from each district by lottery to compete in a battle in which only one person can survive. The chosen teenagers are called "tributes". The whole nation is forced to watch the event on television.	Here's the rewritten text at CEFR level B1: The Hunger Games happen every year as a reminder of what happens when the districts rebel against the Capitol. As punishment, one boy and one girl between 12 and 18 years old are randomly selected from each district to take part in a deadly competition where only one person can win. These teenagers are called "tributes". The whole country is forced to watch this event on TV. Note: I've used simpler vocabulary and sentence structures to make the text more accessible at CEFR level B1. I've also added some detail, such as mentioning that it's a punishment for rebellion, to help readers understand the context better.	Rigid, artificial, or robotic output
	#12	Wasfia Nazreen first came across a hula hoop as a young girl, when she saw a foreign child who was visiting her native Bangladesh playing with one. Wasfia reluctantly stood by and watched, as in her country it was believed that girls should not play with hula hoops or ride bikes. Now Wasfia is one of the few people in the world to have climbed the Seven Summits, including Everest and Kilimanjaro, and the first to have hula-hooped on each peak. Her reason for doing so: to empower women and girls in a country which discourages them from doing sport.	Here is the rewritten text at CEFR level A2: Wasfia Nazreen saw a hula hoop when she was little. She saw a foreign child playing with it in Bangladesh, where she grew up. In her country, people thought girls shouldn't play with hula hoops or ride bikes. Wasfia just watched. Now, Wasfia is one of the few people who climbed all seven big mountains, including Everest and Kilimanjaro. She's the first person to hula-hoop on each mountain! She did this to help women and girls in her country do sports.	Overloaded output, with unnecessary details

Table 4: Qualitative analysis of errors produced by LLaMA 3 model simplifying trial data.

UoL-UPF at TSAR 2025 Shared Task: A Generate-and-Select Approach for Readability-Controlled Text Simplification

Akio Hayakawa¹ Nouran Khallaf² Serge Sharoff² Horacio Saggion¹

¹Universitat Pompeu Fabra ²University of Leeds
{akio.hayakawa, horacio.saggion}@upf.edu
{N.Khallaf, S.Sharoff}@leeds.ac.uk

Abstract

The TSAR 2025 Shared Task on Readability-Controlled Text Simplification focuses on simplifying English paragraphs written at an advanced level (B2 or higher) and rewriting them to target CEFR levels (A2 or B1). The challenge is to reduce linguistic complexity without sacrificing coherence or meaning. We developed three complementary approaches based on large language models (LLMs). The first approach (Run 1) generates a diverse set of paragraph-level simplifications. It then applies filters to enforce CEFR alignment, preserve meaning, and encourage diversity, and finally selects the candidates with the lowest perceived risk. The second (Run 2) performs simplification at the sentence level, combining structured prompting, coreference resolution, and explainable AI techniques to highlight influential phrases, with candidate selection guided by automatic and LLM-based judges. The third hybrid approach (Run 3) integrates both strategies by pooling paragraph- and sentence-level simplifications, and subsequently applying the identical filtering and selection architecture used in Run 1. In the official TSAR evaluation, the hybrid system ranked **2nd overall**, while its component systems also achieved competitive results.

1 Introduction

Text Simplification aims to make complex texts more accessible to a broad audience, including language learners and individuals with reading difficulties (Saggion, 2017; Al-Thanyyan and Azmi, 2021). However, many traditional approaches fail to meet the diverse needs of readers at different proficiency levels. To address this, the field has moved towards targeted simplification, which aims to adapt the complexity of a text to a specific reader’s needs, rather than just simplifying it for a general audience (Barayan et al., 2025; Säuberli et al., 2024). This requires defining specific proficiency targets, and

the Common European Framework of Reference for Languages (CEFR) has been widely used for this purpose (Imperial et al., 2025). Also, the majority of text simplification research has focused on sentence-level, while largely overlooking the more practical scenario of paragraph-level simplification. The TSAR 2025 Shared Task on Readability-Controlled Text Simplification is situated within this context, challenging participants to simplify paragraphs originally at B2 level or above to target levels of A2 and B1 (Alva-Manchego et al., 2025).

In this paper, we propose and validate a Generate-and-Select approach that does not rely on a single best prompt, model, or simplification strategy. Our primary goal was to achieve a high score on a key evaluation metric: **similarity to the reference text**. The official evaluation, conducted only automatically, was based on three metrics: CEFR compliance, output-to-original similarity (Meaning Preservation), and output-to-reference similarity. While the first two could be calculated by participants themselves, the reference texts were not provided. Our system therefore aimed for a high output-to-reference similarity.

To achieve this, we developed a powerful generate-and-select pipeline based on paragraph-level simplification (Run 1) as our core approach. This system first generates a diverse set of candidates and then filtered to create a high-quality candidate pool for Minimum Bayes Risk (MBR) decoding (Bickel and Doksum, 1977) to select the optimal output. As demonstrated by Heineman et al. (2024), the diversity of candidates is crucial for enhancing the quality of MBR decoding. To further improve its performance, we introduced a sentence-level system (Run 2). While weaker on its own, this secondary system successfully injected structural diversity into our candidate pool. Our final, hybrid system (Run 3) combines the candidate pool from both Run 1 and Run 2. It then processes this combined pool using the same pipeline as Run

1 to select the optimal output.

Our approach proved highly effective in the shared task. Among 48 submissions from 20 international teams, our hybrid system (Run 3) and core system (Run 1) placed **2nd and 3rd overall**. Notably, **Run 3 and 1 ranked 1st and 2nd on the reference text similarity** respectively, confirming the success of our primary objective.

However, our success also revealed an inherent limitation of the evaluation metric we focused on optimizing. Our case study highlights that while the metric is designed to capture deep semantic similarity, its scores can still be influenced by surface-level features. This can be misleading, as lexical overlap can sometimes outweigh semantic factuality in the score.

The main contributions of this paper are:

- We present a Generate-and-Select pipeline that successfully maximizes reference similarity.
- We demonstrate that even a weak system can contribute the diversity needed for a powerful selection pipeline.
- We analyse the limitations of the evaluation metric we focused on optimizing.

The experimental setup is available on GitHub.¹

2 Our pipeline

Our submission consists of three systems (Runs 1-3). Our core approach, which achieved 3rd place overall, is presented as Run 1. While our primary objective is to achieve a high output-to-reference similarity, we also aim to attain satisfactory scores in other metrics, namely CEFR compliance and meaning preservation.

2.1 Run 1: Paragraph-Level MBR System

Run 1 is our primary system, designed to maximize the similarity between system outputs and reference texts, through a multi-stage pipeline. As shown in Figure 1, the core approach is a three-stage process. We first generate a diverse set of candidates, and then select a high-quality subset by applying CEFR and Meaning Preservation filtering. Finally, we apply MBR decoding to select the output with the lowest risk.

2.1.1 Diverse Candidate Generation

The process starts with generating a large set of initial simplification candidates for each source para-

¹<https://github.com/ahaya3776/tsar2025sharedtask-uol-upf>

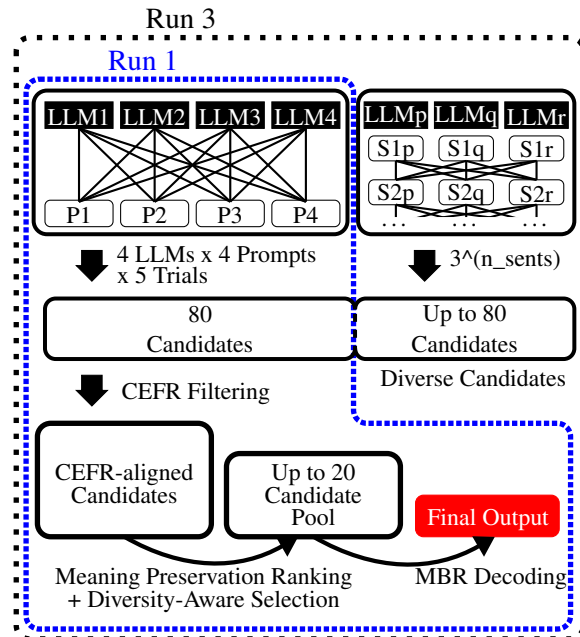


Figure 1: System Architecture of Run 1 and 3.

graph and its corresponding target CEFR level. To ensure a rich and varied candidate pool, this generation process employs two key diversity strategies: multi-prompting and multi-model.

- **Multi-Prompting:** We prepare four types of prompts, with three of them automatically generated by an LLM. Our prompts include two inductive prompts derived from trial data, a deductive prompt based on CEFR-adapted simplification rules, and a standard few-shot prompt. (See Appendix A for the details.)
- **Multi-Model:** The prompts above are run across four auto-regressive large language models (LLMs), GPT-4.1-mini,² gpt-oss-20b (OpenAI, 2025), Gemma-3-4b-it (Gemma, 2025), and Qwen-2.5-14b-it (Qwen, 2025), to capture the unique simplification tendencies of each model.

For each combination of prompt and LLM, we performed five simplification trials, using five separate API calls or five different seeds. As a result, we generated 80 candidates per simplification instance (4 LLMs x 4 prompts x 5 trials). See Appendix C for the hyperparameter settings.

2.1.2 Candidate Pool Construction

After the generation stage, we filter, rank, and select from the initial set of candidates. This process creates an optimized candidate pool of up to 20 simplifications for MBR decoding.

²<https://openai.com/index/gpt-4-1/>

1. **CEFR Filtering:** First, we label the CEFR level (A1, A2, B1, B2, C1, and C2) for all candidates and obtain the minimum difference from the target CEFR level. Given the large number of candidates, this minimum difference is almost always zero (i.e., at least one candidate matches the target CEFR level). We then retain only the candidates that have this minimum difference. CEFR levels are labeled using classification models used in the official shared task evaluation.
2. **Meaning Preservation Ranking:** The remaining CEFR-compliant candidates are ranked in their semantic similarity to the original source paragraph. We use MeaningBERT (Beauchemin et al., 2023) following the official evaluation.
3. **Diversity-Aware Selection:** From this ranked list, we build the final pool with a maximum size of 20. We select candidates primarily based on the previous ranking. However, to maximize the benefits of MBR decoding, which requires a diverse candidate pool (Heineman et al., 2024), we apply a filter to ensure structural diversity. A candidate is added to the pool only if its BLEU (Papineni et al., 2002) against every candidate already in the pool is below a threshold of 0.5.

2.1.3 MBR Decoding

Finally, we apply MBR decoding to the constructed pool. MBR selects the single candidate that maximizes the expected utility function against all other candidates in the set. For the utility function, we again use MeaningBERT, measuring the pairwise similarity between candidates. The candidate with the highest average similarity score against its other candidates is selected as the final output. The final output \hat{y}_{MBR} can be expressed as:

$$\hat{y}_{MBR} = \operatorname{argmax}_{y \in \mathcal{H}} (\mathbb{E}_{\mathcal{H}} [\mathbb{E}_{y' \in \mathcal{H}} [u(y, y')]]), \quad (1)$$

where \mathcal{H} is a candidate pool and $u(y, y')$ is a utility function, defined as $\text{MeaningBERT}(y, y')$.

2.2 Run 2: Sentence-level Simplification

Our second system approaches the task at the sentence level. Prior work has shown that long, coreferential sentences with dense terminology are a key source of difficulty for readers and are best addressed through targeted edits rather than global rewrites (Siddharthan, 2006; Shardlow, 2014; Štajner and Popović, 2016; Barayan et al., 2025). Run 2 therefore investigates whether *explicit linguistic*

control that applied locally at the sentence level, can better align outputs with CEFR levels while preserving meaning (for system architecture see Appendix E). By simplifying sentences independently, while still highlighting the most important phrases, we aim to produce outputs that are both controlled and interpretable. Run 2 consists of the following steps.

Preprocessing. Each paragraph is first segmented into sentences and normalised for coreference. We replace ambiguous pronominal references (e.g., *he, she, they, it*) with their antecedents using AllenNLP’s coreference system (Lee et al., 2017) and the spaCy-compatible coref module (Honnibal et al., 2020). This produces a list of self-contained sentences that can be simplified *independently*.

Highlighting influential phrases. To identify which parts of a sentence contribute most to linguistic complexity, we apply **Integrated Gradients (IG)** (Sundararajan et al., 2017). We apply Captum’s LayerIntegratedGradients (Miglani et al., 2023) over the embedding layer of a sentence-based CEFR classifier (Barayan et al., 2025), using a padded baseline sequence and integrating gradients with respect to the “complexity” logit. Token-level attribution scores are aggregated into multi-word phrases (NP, VP, ADJP, PP) using spaCy chunks. The top- K phrases (default $K=6$) are retained by absolute score. These influential phrases are exported as (type, phrase, score) triples and injected into the simplification prompt (see Appendix B.1). This allows the LLM to focus on which terms to simplify or gloss.

The same influential phrases have another role in the evaluator step, in which the metric verifies whether these spans are preserved in the simplified output. In this way, IG attributions serve a dual purpose: guiding generation and informing evaluation.

Simplification strategies. We guide the models with strategies inspired by intralingual translation and Easy-to-Read (E2R) English (Khallaf et al., 2025). These include **explanation** (adding glosses), **modulation** (one idea per sentence), **synonymy** (simpler words), **syntactic changes** (splitting clauses), and **omission** (dropping non-essential details).

Prompting and candidate generation. We prompt three LLMs, LLaMA-3-8B (Dubey et al.,

2024), GPT-4o (OpenAI, 2023), and Mistral-7B (Jiang et al., 2023), to generate simplifications for CEFR levels A1, A2, and B1 in a single response. Prompts enforce constraints on meaning preservation, correctness of entities and numbers, readability (shorter sentences, simpler words), and strict formatting with explicit level tags (see Appendix B.1).

Automatic and hybrid judging. Candidate outputs are scored by an automatic judge that integrates eight complementary signals (see Table 7 in Appendix F). These include semantic similarity based on sentence embeddings and entailment (Williams et al., 2018), key-phrase coverage from IG attributions, entity and number fidelity using spaCy (Honnibal et al., 2020), readability targets derived from average sentence length (ASL) and Flesch Reading Ease (Flesch, 1948), lexical simplification (syllable reduction), fluency via language model perplexity (Jurafsky and Martin, 2023), compression ratio, and sentence/format control.

We combine heterogeneous metrics with a *weighted geometric mean*, which is widely used in multi-criteria evaluation (Mohapatra and Kumar, 2015; Dodd et al., 2021). When two candidates score within a small margin, we invoke a **Hybrid Auto+LLM (HAI) judge**, which queries a second LLM (GPT-4o or LLaMA) to make a pairwise choice with justification. We pass the original, target level, and top- K candidates (prefiltered by the auto judge) to a second LLM (GPT-4o or LLaMA) that returns a winner index and reason (see Appendix B.2). After simplification, sentences are re-stitched into the level-tagged block (<B1>, <A2>, <A1>)

2.3 Run 3: Hybrid MBR System

Our best-performing system, Run 3, uses the same pipeline as Run 1 but starts with a more diverse set of initial candidates from Run 2. In addition to 80 candidates generated in Run 1, we incorporate candidates based on sentence-level simplification in Run 2. As shown in Figure 1, we generate candidates based on Run 2 by concatenating sentence-level simplifications. For each sentence in an original paragraph, three simplified sentences are generated by three different LLMs. The combination of simplified sentences result in $3^{n_sentences}$ potential paragraph variants, from which we randomly sample up to 80 candidates. Among this combined set of up to 160 candidates, the final output is selected through the identical process described for Run 1.

Team	CEFR RMSE	Sim Orig	Sim Ref	Total Rank
EhiMeNLP	0.000	.902	.845	1
UoL-UPF (3)	0.000	.856	.857	2
UoL-UPF (1)	0.000	.849	.856	3
HIT-YOU	0.158	.852	.835	4
Archaeology	0.122	.779	.804	11
ounlp	0.755	.855	.849	14
SQUREL	1.153	.979	.819	23
UoL-UPF (2)	0.693	.808	.827	-

Table 1: Representative results from 44 runs from 20 teams. The best performance for each metric is shown in red. Run 2 is an unofficial result due to parsing error, and its estimated rank is around 20th.

Model	A2		B1	
	Num.	Sim	Num.	Sim
GPT-4.1-mini	24	.841	13	.865
gpt-oss-20b	31	.831	17	.902
Gemma-3-4b	16	.840	12	.862
Qwen-2.5-14b	26	.862	36	.877
Sentence-lv	3	.730	22	.860

Prompt	A2		B1	
	Num.	Sim	Num.	Sim
Prompt 1	19	.839	20	.872
Prompt 2	30	.838	15	.908
Prompt 3	24	.831	23	.867
Prompt 4	24	.866	20	.874
Sentence-lv	3	.730	22	.860

Table 2: Distribution of models and prompts selected as a final candidate in Run 3 with output-to-reference similarity scores by MeaningBERT.

3 Results and Discussions

Table 1 shows the official results of the shared task. The hybrid system (Run 3) is ranked 2nd, while the core system (Run 1) is 3rd overall. Furthermore, our systems placed 1st (tied, full marks) on CEFR alignment, and 1st and 2nd on output-to-reference similarity. This result confirms the success of our pipeline combining filtering and MBR decoding, thereby achieving the high output-to-reference similarity while maintaining other metrics.

Table 2 demonstrates the distribution of selected candidates for Run 3, categorized by their source. The selections were generally distributed evenly across target levels and our various prompts, models, and granularities. The only exception is

Ablation	A2		B1	
	Orig	Ref	Orig	Ref
Run 3	.836	.840	.876	.874
w/o Sent. lv (\equiv Run 1)	.824	.837	.874	.875
w/o MPR, DAS, MBR	.756	.779	.817	.822
w/o MPR, DAS	.815	.830	.850	.858
w/o DAS	.849	.834	.891	.869
w/o MBR (Random)	.789	.793	.841	.832
w/o MBR (Highest MP)	.896	.814	.919	.858
w/ smaller MBR (size=10)	.853	.838	.888	.873

Table 3: MeaningBERT scores between outputs and original (Orig) and reference (Ref), as an ablation study for processes after the CEFR filtering. MPR and DAS refers to Meaning Preservation Ranking and Diversity-Aware Selection, respectively.

sentence-level approach for the A2 target. This implies that adding explanations, often observed in the simplification to lower proficiency levels, is hard to achieve via sentence-level approach. This overall diversity was the key to the success of our MBR-based selection pipeline.

Furthermore, we conducted ablation study shown in Table 3. As we described, final outputs are selected through Meaning Preservation Ranking, Diversity-Aware Selection, and MBR decoding after the CEFR filtering. The study shows that each of these steps contributed to improve output-to-reference similarity. Notably, MBR decoding boosted it up, while increasing the candidate pool size produced only a negligible gain.

This success also highlights an important characteristic of our method. Figure 2 illustrates the MeaningBERT scores distribution of CEFR-aligning candidates for one example instance. While the final output shows the highest output-to-original similarity, several candidates show higher output-to-reference similarity. This observation confirms that MBR decoding is designed to minimize the risk of selecting a low-scoring candidate, not to select one with the maximum expected score. As a result, final outputs are often conservative.

Despite prioritizing output-to-reference similarity, we acknowledge that over-reliance on this metric can be problematic. Our qualitative analysis shows limited agreement between scores and human judgments. Specifically, instances containing semantic errors or complex vocabulary (yellow in the scatter plot) are often over-evaluated by the metric when they are structurally similar to the reference. On the other hand, structure changes, such as sentence splitting, are penalized even if beneficial.

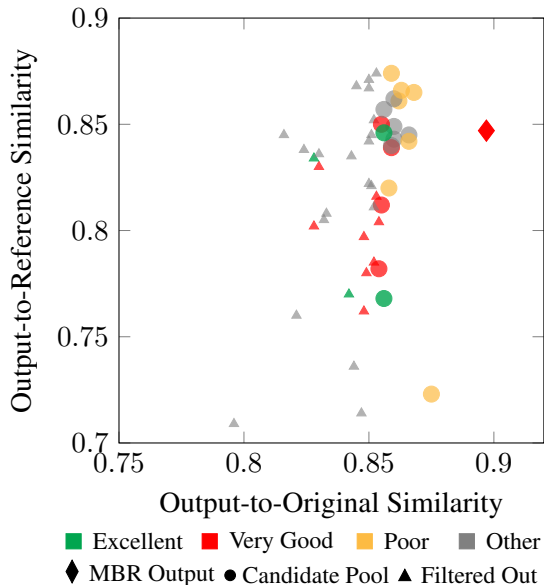


Figure 2: Scatter plot for CEFR-aligned candidates of a single instance. Axes represent similarity scores between output and original/reference. Circles are ones selected as candidate pool, and the diamond is the final output through MBR decoding. Colors align with Table 5, Table 6 in which we manually judged simplification quality.

Our case study supports that MeaningBERT often fail to capture the value of features such as sentence splitting, synonym choice, and moral or pragmatic clarity, rewarding surface overlap instead of genuine accessibility (Barayan et al., 2025). We provide full analysis in Appendix D.

4 Conclusion

In this paper, we presented our Generate-and-Select framework for the TSAR 2025 shared task, which achieved 2nd and 3rd place overall. Our core approach utilized a diverse candidate pool from multiple LLMs and prompts, with MBR decoding for robust selection.

Our primary contribution is demonstrating that our Generate-and-Select framework is highly effective. We showed that its strength lies in prioritizing the diversity of candidates, which allowed even a weaker system (our sentence-level Run 2) to make a contribution to the final performance by injecting variety.

Finally, our analysis shows that while our pipeline is robust, its limitation in a single-reference context highlights the need for selection methods that can better handle unpredictable simplifications.

Lay Summary

UoL-UPF team participated in the TSAR 2025 Shared Task. The goal of this shared task was to rewrite difficult English texts into simple texts at a specific level.

We tried an idea we call Generate-and-Select approach. In this approach, first, we used LLMs to generate many versions of simple texts. We used different LLMs and prompts, so there were a lot of options to choose from. This variety was a key part of our idea. Next, we selected the best option from these simple texts. We built a system to check all the simple texts. This system had some filtering processes. For example, one filter only selected texts that were similar to original difficult texts. After these filtering processes, we only had high-quality options. Finally, from these high-quality options, we selected the lowest-risk option as a final result.

Our system performed very well, and was ranked 2nd place out of 48 systems. This great result showed that our idea was a good one. Through this project, we learned some very important things. It is true that our generate-and-select approach works well, especially when the quality of generated texts is judged by computer. However, we cannot always trust computer judge. In our study, some simple texts were good by computer judge, but not by human judge.

Limitations

The primary limitation of this work is its reliance on diverse set of generation. While the LLMs we employed are relatively small-scaled and thus do not require excessive computational resources, the time and cost associated with obtaining the final outputs cannot be disregarded. Therefore, our generate-and-select framework would be unsuitable for real-time text simplification.

Also, this shared task relies on automatic evaluation metrics. While our system achieved high scores, we did not conduct a manual evaluation with human participants to confirm whether the outputs are genuinely more readable and understandable for the target readers. Such manual evaluation, with Likert scoring or reading comprehension questions, would be necessary to validate the real-world effectiveness of our simplifications.

Acknowledgments

This document is part of a project that has received funding from the European Union’s Horizon Europe research and innovation program under Grant Agreement No. 101132431 (iDEM Project). The views and opinions expressed in this document are solely those of the author(s) and do not necessarily reflect the views of the European Union. Neither the European Union nor the granting authority can be held responsible for them. The University of Leeds (UOL) was funded by UK Research and Innovation (UKRI) under the UK government’s Horizon Europe funding guarantee (Grant Agreement No. 10103529).

Also, this work is partially financed by the Ministerio de Ciencia, Innovación y Universidades, Agencia Estatal de Investigaciones: project CPP2023-010780 funded by MICIU/AEI/10.13039/501100011033 and by FEDER, UE (“Habilitando Modelos de Lenguaje Responsables e Inclusivos”). Horacio Saggion also receives support from the Spanish State Research Agency under the Maria de Maeztu Units of Excellence Programme (CEX2021-001195-M) and from the Departament de Recerca i Universitats de la Generalitat de Catalunya (ajuts SGR-Cat 2021).

References

- Suha S Al-Thanyyan and Aqil M Azmi. 2021. Automated text simplification: a survey. *ACM Computing Surveys (CSUR)*, 54(2):1–36.
- Fernando Alva-Manchego, Regina Stodden, Joseph Marvin Imperial, Abdullah Barayan, Kai North, and Harish Tayyar Madabushi. 2025. Findings of the TSAR 2025 shared task on readability-controlled text simplification. In *Proceedings of the Fourth Workshop on Text Simplification, Accessibility, and Readability (TSAR 2025)*, Suzhou, China. Association for Computational Linguistics.
- Abdullah Barayan, Jose Camacho-Collados, and Fernando Alva-Manchego. 2025. [Analysing zero-shot readability-controlled sentence simplification](#). In *Proceedings of the 31st International Conference on Computational Linguistics (COLING)*, pages 6762–6781. Association for Computational Linguistics.
- David Beauchemin, Horacio Saggion, and Richard Khoury. 2023. Meaningbert: assessing meaning preservation between sentences. *Frontiers in Artificial Intelligence*, 6:1223924.
- P.J. Bickel and K.A. Doksum. 1977. *Mathematical Statistics: Basic Ideas and Selected Topics*. Prentice Hall.

- Ben Dodd, Betty van Aken, Paul Röttger, and Isabelle Augenstein. 2021. [AUTORANK: A systematic approach to benchmark and compare machine learning models](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 170–185, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Abhimanyu Dubey, Rohan Taori, Alexei Baevski, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Rudolf Fleisch. 1948. [A new readability yardstick](#). *Journal of Applied Psychology*, 32(3):221–233.
- Gemma. 2025. [Gemma 3 technical report](#). *Preprint*, arXiv:2503.19786.
- David Heineman, Yao Dou, and Wei Xu. 2024. [Improving minimum Bayes risk decoding with multi-prompt](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22525–22545, Miami, Florida, USA. Association for Computational Linguistics.
- Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. [spaCy: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing](#). Software documentation.
- Joseph Marvin Imperial, Abdullah Barayan, Regina Stodden, Rodrigo Wilkens, Ricardo Munoz Sanchez, Lingyun Gao, Melissa Torgbi, Dawn Knight, Gail Forey, Reka R Jablonkai, and 1 others. 2025. [Universalcefr: Enabling open multilingual research on language proficiency assessment](#). *arXiv preprint arXiv:2506.01419*.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2023. [Mistral 7B](#). *Preprint*, arXiv:2310.06825.
- Dan Jurafsky and James H. Martin. 2023. [Speech and language processing \(3rd ed. draft\): Chapter on language modeling](#). Online draft.
- Nouran Khallaf, Carlo Eugeni, and Serge Sharoff. 2025. [Reading between the lines: A dataset and a study on why some texts are tougher than others](#). *Preprint*, arXiv:2501.01796. Published at Writing Aids at the Crossroads of AI, Cognitive Science and NLP (WR-AI-CogS), COLING 2025, Abu Dhabi.
- Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. [End-to-end neural coreference resolution](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, Copenhagen, Denmark. Association for Computational Linguistics.
- Vivek Miglani, Aobo Yang, Aram Markosyan, Diego Garcia-Olano, and Narine Kokhlikyan. 2023. [Using Captum to explain generative language models](#). In *Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023)*, pages 165–173, Singapore. Association for Computational Linguistics.
- Prasanta Kumar Mohapatra and Suresh Kumar. 2015. [A multi-criteria decision making method based on weighted geometric mean](#). *International Journal of Applied Decision Sciences*, 8(2):133–148.
- OpenAI. 2023. [GPT-4 technical report](#). *Preprint*, arXiv:2303.08774.
- OpenAI. 2025. [gpt-oss-120b & gpt-oss-20b model card](#). *Preprint*, arXiv:2508.10925.
- 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.
- Qwen. 2025. [Qwen2.5 technical report](#). *Preprint*, arXiv:2412.15115.
- Horacio Saggion. 2017. *Automatic Text Simplification*, volume 10 of *Synthesis Lectures on Human Language Technologies*. Morgan & Claypool Publishers.
- Andreas S  uberli, Franz Holzknecht, Patrick Haller, Silvana Deilen, Laura Schiffli, Silvia Hansen-Schirra, and Sarah Ebling. 2024. [Digital comprehensibility assessment of simplified texts among persons with intellectual disabilities](#). In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–11.
- Matthew Shardlow. 2014. [A survey of automated text simplification](#). *International Journal of Advanced Computer Science and Applications (IJACSA)*, 4(1):58–70.
- Advait Siddharthan. 2006. [Syntactic simplification and text cohesion](#). *Research on Language and Computation*, 4(1):77–109.
- Sanja   tajner and Maja Popovi  . 2016. [Can text simplification improve machine translation?](#) In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)*, pages 172–178. European Language Resources Association.
- Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. [Axiomatic attribution for deep networks](#). In *Proceedings of the 34th International Conference on Machine Learning*, pages 3319–3328, Sydney, Australia. PMLR.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. [A broad-coverage challenge corpus for sentence understanding through inference](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

A Run 1: Prompts for Paragraph-level Simplification

We used four simplification prompts for LLMs. Two of these were based on inductive approach, which involved extracting simplification features from trial data to create instructions as a prompt. To do this, the following prompt was given to GPT-4.1-mini.

You will be given several pairs of paragraphs. Each pair is composed of an original paragraph and a simplified version for CEFR {lv} readers. Your task is to analyze these pairs to find the general patterns of simplification and write an instruction for LLMs to simplify paragraphs similarly. Include observations on information or phrasing that remains unchanged. Do not include examples that contain text parts in given paragraphs. Only output your final prompt .

Original: {Original Paragraph 1 of the target CEFR level}
 Reference: {Reference Paragraph 1 of the target CEFR level}
 }
 Original: {Original Paragraph 2 of the target CEFR level}
 Reference: {Reference Paragraph 2 of the target CEFR level}
 }
 ...

After several trials, we picked up following two types of prompts for each level with some minor arrangements.

Prompt 1 : A2

Simplify paragraphs for CEFR A2 readers by following these guidelines:

1. Use short, clear sentences with simple grammar structures (mostly present and past simple).
2. Replace complex or abstract vocabulary with common, concrete words; explain any necessary technical terms briefly and clearly.
3. Remove or reduce detailed numerical data, statistics, or nuanced concepts unless essential; when included, present numbers simply and round if appropriate.
4. Avoid idiomatic expressions, figurative language, and complex sentence forms like passive voice or embedded clauses.
5. Focus on main ideas and essential facts; omit detailed background information, speculation, or subtle distinctions unless they support comprehension.
6. Use explicit cause-effect and temporal connectors (e.g., because, so, but, then, now) to clarify relationships.
7. Maintain logical and coherent flow with clear topic introductions and simple sequencing.
8. Preserve proper names, key terms, and notable facts that are central to understanding.

9. When appropriate, add brief, straightforward definitions or explanations of less familiar concepts.
10. Use active voice predominantly and ensure the subject of sentences is clear.
11. Replace pronouns that may confuse with explicit nouns where needed.
12. Retain the overall meaning and important details but adapt phrasing to be direct and concrete.
13. Introduce examples to illustrate points simply, using familiar or relatable contexts.
14. Do not assume prior knowledge; present background information in simple terms if required.
15. Where opinion or interpretation appears, present it clearly and simply, often using direct statements like "people say" or "some think."
16. Use simple punctuation and avoid complex structures such as long lists or parenthetical asides.

By following these patterns, produce an accessible, easy-to-read version of a paragraph that preserves the core message and key details for A2-level readers. Provide only the simplified paragraph without any explanation or justification.

```
# Original:
{Original Paragraph}
# Simplified:
```

Prompt 1 : B1

Simplify paragraphs for CEFR B1 readers by following these guidelines:

1. Use simpler vocabulary and expressions: Replace complex or formal words and phrases with more common, everyday alternatives, while keeping the meaning intact.
2. Shorten and clarify sentences: Break long, complex sentences into shorter, clearer ones. Use straightforward sentence structures, avoiding passive voice or complicated clauses.
3. Explain or define less familiar terms: When necessary, introduce brief explanations or definitions of technical, cultural, or less common concepts within the text to aid understanding.
4. Retain key information and facts: Keep all essential data, figures, names, and core ideas from the original text, ensuring the main message is preserved.
5. Rephrase for explicitness and clarity: Make implied meanings more explicit, and clarify references to pronouns or abstract concepts.
6. Maintain original factual content and sequence: Do not omit major details or reorder information in ways that change the logical flow or significance.
7. Use familiar synonyms and phrases: Prefer words and expressions that are frequently used at intermediate English level rather than academic or highly technical language.
8. Simplify complex concepts without oversimplifying: Present difficult ideas in more accessible language but avoid losing the nuance or accuracy of the original content.
9. Use concrete examples or context where helpful: When abstract concepts might confuse, add brief relatable examples or contextual cues to aid comprehension.
10. Preserve unchanged proper nouns and names: Keep names of people, places, events, titles, and specific terms as in the original to maintain accuracy and recognition.
11. Avoid idiomatic or culture-specific expressions unless explained: Replace or explain idioms and culturally specific references that might not be understood by B1 learners.
12. Retain the original tone and intent as much as possible :** The simplification should respect the author's purpose, tone, and the overall style, aiming for clarity rather than casualness.

In summary, simplify language and sentence structure, clarify meaning, explain or define unfamiliar terms, keep all important facts and details, and ensure the text remains coherent and faithful to the original. Provide only the simplified paragraph without any explanation or justification.

Original:
{Original Paragraph}
Simplified:

Prompt 2 : A2

Simplify paragraphs for CEFR A2 readers by following these guidelines:

- Vocabulary and Grammar:**
 - Use very common, everyday words and simple sentence structures.
 - Avoid idioms, metaphors, or abstract expressions.
 - Prefer present tense or simple past tense; avoid complex verb forms.
 - Use short sentences, often one idea per sentence.
 - Sentence Structure:**
 - Break long, complex sentences into multiple shorter sentences.
 - Use basic conjunctions (and, but, so, because) to connect ideas simply.
 - Avoid passive voice where possible; use active voice instead.
 - Information Selection and Clarity:**
 - Retain all key factual information from the original paragraph.
 - Remove or rephrase any statistics or figures only if they might confuse the reader, but generally keep numbers with simple explanations.
 - Explain or define any technical terms or names using simple language or familiar examples.
 - Avoid unnecessary detail or background information unless it helps understanding.
 - Rephrasing and Simplification:**
 - Replace complex nouns or phrases with simpler equivalents or brief explanations.
 - Make implicit information explicit if needed.
 - Use examples or explanations to clarify concepts that might be unfamiliar.
 - Use repetition and restatement to reinforce understanding without changing meaning.
 - Tone and Style:**
 - Use a neutral, clear, and straightforward tone.
 - Address the reader more directly and simply when appropriate.
 - Keep the original meaning, emphasis, and main points intact.
 - Preserving Key Proper Nouns and Data:**
 - Keep proper names (people, places, organizations, titles) unchanged but briefly explain their significance if needed.
 - Maintain important dates, measurements, and specific figures, simplifying explanations around them.
 - Avoid Removing Content Entirely:**
 - Instead of deleting difficult or nuanced content, re-express it in accessible language.
 - Questions or rhetorical devices in the original can be kept but simplified and clarified.
- By applying these principles, transform original paragraphs into clear, accessible text suitable for A2-level readers while preserving essential information and intent. Provide only the simplified paragraph without any explanation or justification.

Original:

{Original Paragraph}

Simplified:

Prompt 2 : B1

Simplify paragraphs for CEFR B1 readers by following these guidelines:

- Vocabulary and Sentence Structure:**
 - Use common, everyday words instead of specialized or complex vocabulary.
 - Prefer simple sentence structures; break longer or compound sentences into shorter ones.
 - Replace abstract or complex terms with concrete, clearer expressions or brief explanations.
 - Use active voice where possible and avoid idiomatic expressions or cultural references that may be unclear.
 - Information Presentation:**
 - Keep all key factual information and core ideas intact to preserve the original meaning.
 - Present numbers, dates, and statistics clearly, often repeating or rephrasing for clarity.
 - When technical or unfamiliar terms appear, define or explain them briefly but simply.
 - Remove less essential details only if they do not affect overall comprehension; otherwise, retain the main content fully.
 - Clarification and Explicitness:**
 - Make implicit information explicit where needed.
 - Where the original contains pronouns or references that may be unclear, replace or clarify them.
 - Use clear cause-and-effect or chronological connectors (e.g., "because," "so," "however," "since then") to improve coherence.
 - Tone and Style:**
 - Maintain a neutral, informative, and accessible tone appropriate for learners.
 - Avoid complex or figurative language; use straightforward, literal expressions.
 - When original tone includes subtle nuance, simplify but try to retain the intended emphasis or attitude if important.
 - Phrasing and Repetition:**
 - Some proper nouns, dates, and well-known names remain unchanged to preserve identity and context.
 - Common phrases and definitions that clarify the subject often get added or slightly expanded to aid understanding.
 - Sentences may be reworded but often echo the original information closely, sometimes repeating key ideas with slight reformulation for clarity.
- In summary, simplify vocabulary and grammar, clarify potentially difficult concepts, maintain all essential facts, and ensure readability and coherence for intermediate English learners without omitting important content. Provide only the simplified paragraph without any explanation or justification.

Original:
{Original Paragraph}
Simplified:

One of our four prompts was also generated by an LLM, based on a deductive approach without providing trial data. We made GPT-4.1-mini generate such prompts with the following prompt.

You are an expert in language education and have a deep understanding of CEFR. Your task is to generate a list of rules that an LLM can use to simplify a given English text to each of the CEFR A2 and B1 level. The simplification must

preserve the original meaning and core information of the paragraph. The rules should be specific enough to guide the simplification process effectively, covering vocabulary, grammar, and sentence structure.

As a result, we obtained the following simplification prompts.

Prompt 3 : A2

Simplify paragraphs for CEFR A2 readers by following these guidelines:

Vocabulary

1. Use only high–frequency, common words and phrases (e.g., "big" instead of "enormous," "buy" instead of "purchase").
 2. Replace idiomatic expressions and phrasal verbs with their literal or simpler equivalents.
 3. Avoid technical or abstract vocabulary; if necessary, define it simply within the sentence.
 4. Use concrete nouns and familiar verbs; avoid abstract nouns and complex verbs.
- #### ### Grammar and Sentence Structure
5. Use simple present and past tenses; avoid complex tenses (e.g., present perfect, future perfect).
 6. Favor active voice over passive voice.
 7. Use short sentences (ideally 8–12 words per sentence).
 8. Use simple sentence structures: one main clause per sentence.
 9. Avoid subordinate clauses or replace them with separate sentences.
 10. Use coordinating conjunctions (and, but, or) rather than subordinating conjunctions (because, although).

Pronouns and References

11. Limit use of pronouns; repeat nouns when clarity may be lost.
12. Use clear and explicit referents; avoid ambiguous references.

Cohesion and Coherence

13. Use basic discourse markers (first, then, next, finally) to link ideas.
14. Present information in a clear, logical order; avoid complex or backward references.

Miscellaneous

15. Avoid negations where possible; use positive forms or simpler negative constructions.
 16. Do not include idioms, humor, or cultural references that might confuse A2 learners.
- Provide only the simplified paragraph without any explanation or justification.

Original:
{Original Paragraph}
Simplified:

Prompt 3 : B1

Simplify paragraphs for CEFR B1 readers by following these guidelines:

Vocabulary

1. Use mostly common vocabulary with some intermediate words.
2. Replace very rare or highly technical words with simpler or more familiar synonyms.
3. Allow some idiomatic expressions or phrasal verbs but only if they are common and clearly understandable.
4. Use concrete and abstract nouns where necessary but explain abstract concepts in simple terms.

Grammar and Sentence Structure

5. Use simple present, past, and present perfect tenses; allow future tenses like "will" and "going to."
6. Use both active and passive voice but keep passive structures simple.
7. Use mostly simple sentences and some compound sentences.
8. Allow basic complex sentences with one subordinate clause (e.g., using because, when, if).
9. Avoid very complex sentence structures (e.g., multiple subordinate clauses or embedded clauses).
10. Use modal verbs (can, should, must) appropriately for expressing ability, advice, or obligation.

Pronouns and References

11. Use pronouns more freely than at A2 but ensure clarity is maintained.
12. Use demonstratives (this, that) appropriately for reference.

Cohesion and Coherence

13. Use a wider range of discourse markers (however, although, because, so, therefore) for linking ideas.
14. Present ideas in a logical, well–organized manner with clear paragraphing.

Miscellaneous

15. Include some figurative language if it can be easily explained or inferred.
 16. Allow simple conditional sentences (first conditional).
 17. Use negations confidently but keep them straightforward (avoid double negatives).
- Provide only the simplified paragraph without any explanation or justification.

Original:
{Original Paragraph}
Simplified:

Finally, we prepared one prompt based on basic few-shot (3-shot) prompting. Few-shot examples were extracted from the trial data.

Prompt 4

You are an expert in simplifying complex English texts for language learners. Your task is to adapt the following text to make it accessible and readable for a CEFR {lv} level learner. Provide only the simplified paragraph without any explanation or justification.

Original:
{Original paragraph of the few–shot example 1}
Simplified:
{Reference paragraph of the few–shot example 1}

Original:
{Original paragraph of the few–shot example 2}
Simplified:
{Reference paragraph of the few–shot example 2}

Original:
{Original paragraph of the few–shot example 3}
Simplified:
{Reference paragraph of the few–shot example 3}

Original:
{Original Paragraph}
Simplified:

B Run 2: Prompts for Sentence-level Simplification

B.1 Dynamic Prompt for Sentence-level Simplification (Pseudocode)

```
PROC BUILD_PROMPT(original_sentence,
influential_phrases,
                    levels={B1,A2,A1},
                    add_guardrails=TRUE):

# 1) Strategy guide (fixed text shown to the LLM)
STRATEGY_GUIDE := [
  "Synonymy (prefer): simpler common words",
  "Modulation: rephrase while preserving meaning",
  "Omission: remove non-essential parts",
  "Compression: shorten without losing meaning",
  "Syntactic Changes: simpler structure",
  "Transposition: change word form if clearer",

  "Explanation: only if term appears in sentence",
  "Transcription: keep if already simple",
  "Illocutionary Change: make implied meaning explicit",
  "DO NOT explain the edits you made"
]

# 2) Level-specific style rules
RULES.A1 := "1-2 very short sentences; very common words only; define technical terms briefly if they appear; limited [brackets]."
RULES.A2 := "1-2 short sentences with simple connectors; synonyms for hard words; one brief definition at start if needed; at most one [bracketed] note."
RULES.B1 := "Up to two sentences; faithful meaning with clearer phrasing; short clarifications for technical terms if present."

# 3) Guardrails (optional but recommended)
IF add_guardrails THEN
  GUARDS := [
    "Only define concepts if the exact word appears in the original",
    "Do not add unrelated facts or examples",
    "Preserve names, numbers, and dates exactly",
    "Output EXACTLY in the required tag format; no extra commentary"
  ]
ELSE
  GUARDS := []
END

# 4) Select/inject top-K influential phrases
K := 6
TOP := TAKE_TOP_BY_ABS_SCORE(
  influential_phrases, K)

# format for display inside the prompt (one per line)
IF TOP is empty THEN
  INFL_BLOCK := "- (none provided)"
ELSE
```

```
INFL_BLOCK := CONCAT_LINES(
  FOR EACH (type, phrase, score) IN TOP:
    "-" + type + " | " + phrase + " : " +
    FORMAT(score, 4 decimals)
)
END

# 5) Assemble prompt text (final structure)
PROMPT :=
  "You are a CEFR simplification assistant.\n\n" +
  "Task:\n" +
  "1) Simplify the original text into these CEFR levels: " + JOIN(levels, ", ") + ".\n" +
  "2) Focus on these influential/complex phrases (apply strategies when needed):\n" +
  INFL_BLOCK + "\n" +
  "3) Keep the meaning accurate but simpler.\n" +
  "4) Do not explain what changes you made.\n\n" +
  "Strategies:\n" + BULLETTIZE(STRATEGY_GUIDE) +
  "\n\n" +
  "Style constraints:\n" +
  "- A1: " + RULES.A1 + "\n" +
  "- A2: " + RULES.A2 + "\n" +
  "- B1: " + RULES.B1 + "\n\n" +
  (IF GUARDS is empty THEN "" ELSE "
  Constraints:\n" + BULLETTIZE(GUARDS) + "\n\n"
) +
  "Output format (exactly):\n" +
  "<Original> " + original_sentence + "\n" +
  "<B1> ... \n" +
  "<A2> ... \n" +
  "<A1> ..."

RETURN PROMPT
END
```

B.2 LLM-as-Judge for CEFR Simplification (Pseudocode)

```
PROC BUILD_JUDGE_PROMPT(original_text,
target_level, candidates, add_guardrails=
TRUE):

# 1) Purpose & role
ROLE := "You are a CEFR simplification judge."

# 2) Task framing (fixed)
TASK := [
  "Given: the ORIGINAL sentence, the TARGET CEFR level, and several CANDIDATE simplifications.",
  "Pick the single BEST candidate that:",
  "1) preserves the meaning of the original,",
  "2) matches the target level (A1 easiest < A2 < B1),",
  "3) is fluent and clear,",
  "4) keeps names, numbers, and units correct."
]

\begin{verbatim}
# 3) CEFR decision rubric (concise)
RUBRIC := [
```

```

"A1: very common words; 1-2 very short
sentences; concrete, no idioms.",
"A2: simple frequent words; short sentences;
simple connectors (and, but); avoid abstract
terms.",
"B1: more range; still short clear sentences;
simple, accurate paraphrase."
]
\end{verbatim}

# 4) Guardrails (optional)
IF add_guardrails THEN
  GUARDS := [
    "Return STRICT JSON only (no prose).",
    "Do NOT invent facts; preserve named
    entities, numbers, dates.",
    "Prefer transparent, high-frequency words
    at lower levels.",
    "Minor reordering is fine if meaning is
    preserved.",
    "If ties, choose the clearer and more
    level-appropriate one."
  ]
ELSE
  GUARDS := []
END

# 5) Output contract (exact)
CONTRACT := [
  "Reply in JSON with:",
  "{ \"winner_index\": <0-based index>, \"
  reason\": \"<one short line>\" }"
]

# 6) Render candidate list
# CANDS_STR becomes:
# [0] ...
# [1] ...
# ...
CANDS_STR := CONCAT_LINES(
  FOR EACH (i, c) IN ENUMERATE(candidates):
    "[" + STR(i) + "]" + c
)

# 7) Assemble prompt
PROMPT :=
  ROLE + "\n\n" +
  JOIN_LINES(TASK) + "\n\n" +
  "CEFR RUBRIC (summary):\n" + BULLETTIZE(
  RUBRIC) + "\n\n" +
  (IF GUARDS is empty THEN "" ELSE "
  Constraints:\n" + BULLETTIZE(GUARDS) + "\n\n"
  ) +
  JOIN_LINES(CONTRACT) + "\n\n" +
  "ORIGINAL:\n" + original_text + "\n\n" +
  "TARGET LEVEL: " + target_level + "\n\n" +
  "CANDIDATES:\n" + CANDS_STR

RETURN PROMPT
END

```

C Hyperparameters

In Run 1 (and Run 3), we used HuggingFace Transformers library³ for open-weight models, namely gpt-oss-40b, Gemma-3-4b, and Qwen-2.5-14b. For

³<https://huggingface.co/docs/transformers/>

Parameter	Value
temperature	1.0
top_p	0.95
number of beams	1
number of generation	1
max number of new tokens	512

Table 4: Hyperparameters for decoding.

GPT-4.1-mini, we used OpenAI API library.⁴ Table 4 shows the hyperparameters settings we used for decoding. For other hyperparameters, default values were used.

D Case Study

In this case study, we analysed one selected example (109-a2) with 51 CEFR-aligned candidates. We carried out a linguistic evaluation of the 19 outputs selected by the filtering process and compared them with the final MBR-selected output. We also examined a set of candidates undervalued by the automatic metrics, some of which provided stronger simplifications than selected outputs.

⁴<https://platform.openai.com/docs/api-reference/chat/create>

Case	Model (Prompt)	Candidate	Sim. Orig.	Sim. Ref.
Orig.	B2	Best-selling author Terrie Sharp knows how to tell a good story or two, having twice won the Olive Crime Writer Award. Born on the housing estate with the highest crime rate in Glasgow, she learned to keep quiet, stay out of trouble and how to open car doors with a knife. Luckily, she only used her knowledge to write her novels and she's not so quiet these days, appearing on popular daytime shows like Police Today and at writing conferences throughout the UK.		
Ref.	A2	Terrie Sharp is a very popular writer, and her books sell well. She has won the Olive Crime Writer Award two times. She was born in Glasgow, a place with a lot of crime. As a child, she learned to be quiet, stay safe, and even how to open car doors with a knife. Luckily, she only used this for her stories. Now she is not quiet, she goes on TV shows like Police Today and gives talks at writing conferences in the UK.		
MBR	Qwen-2.5-14b (Prompt 1)	Terrie Sharp is a writer who won the Olive Crime Writer Award two times. She was born in a part of Glasgow with lots of crime. She learned to be quiet and avoid trouble. She also learned how to open car doors with a knife, but she only used that for her stories. Now, Terrie speaks a lot on TV shows like Police Today and at events about writing.	0.897	0.847
<p>Linguist's note (CEFR A2): This version keeps the meaning of the original while using clear, short sentences. The phrase <i>avoid trouble</i> is a simple, transparent way to express the idiom <i>stay out of trouble</i>. Using <i>that</i> instead of an abstract noun like <i>knowledge</i> makes the text easier for A2 readers. The phrase <i>events about writing</i> is a little vague compared to <i>conferences</i>, but it is easier to understand for learners. Automatic metrics give it a high score mainly because of word overlap, but they do not recognise how the text reduces confusion around the violent detail. Overall, this is a strong fit for A2 learners.</p>				
S1	Gemma (Prompt 4)	Terrie Sharp is a very popular writer. She has won the 'Olive Crime Writer Award' twice! She grew up in an area of Glasgow where there were many problems. She learned to be quiet and not get into trouble. She also learned how to open car doors with a knife. But now, she only uses this knowledge to write her books. She is very busy, and you can see her on TV shows and at writing events in the UK.	0.866	0.845
<p>Linguist's note (CEFR A2): Short sentences help; "problems" is transparent. However, <i>knowledge</i> and <i>appearing/very busy</i> are abstract for A2; the exclamation skews register. Metrics overestimate due to overlap, not learner appropriateness.</p>				
S2	Gemma (Prompt 4)	Terrie Sharp is a very famous writer. She can write good stories! She has won the Olive Crime Writer Award twice. She grew up in a part of Glasgow where there was a lot of crime. She learned to be quiet and not get into problems. She also learned how to open car doors with a knife. But now, she doesn't use this knowledge anymore. She writes her stories instead. She is on TV shows and talks at writing events in the UK.	0.860	0.843
<p>Linguist's note (CEFR A2): <i>Famous</i> is A1/A2 and clearer than <i>popular</i>; clause-level simplicity is good. Retains abstract <i>knowledge</i>; enthusiastic register ("!") mismatched to neutral bio. Metrics ignore register and reward lexical reuse.</p>				
S3	Gemma (Prompt 2)	Terrie Sharp is a popular writer. She can tell good stories. She won the Olive Crime Writer Award twice. She was born in a part of Glasgow with many crimes. She learned to be quiet and not have problems. She also learned how to open car doors with a knife. She only uses this skill to write her books now. She is not quiet anymore. She appears on TV shows like Police Today. She also goes to writing events in the UK.	0.862	0.861
<p>Linguist's note (CEFR A2): Strong clause splitting; <i>not have problems</i> concretizes <i>stay out of trouble</i>. But <i>appears, skill</i> raise lexical load. Metrics laud similarity, not vocabulary transparency.</p>				
S4	Gemma (Prompt 3)	Best-selling writer Terrie Sharp can tell good stories. She won the Olive Crime Writer Award twice. She was born in a part of Glasgow with many crimes. She learned to be quiet and not get into problems. She also learned to open car doors with a knife. She only used this knowledge to write her books. Now, she is not so quiet. She appears on TV shows like Police Today. She also goes to writing events in the UK.	0.860	0.862
<p>Linguist's note (CEFR A2): <i>Best-selling, knowledge</i> are above A2 transparency; syntax is simple but words remains complex. Metrics overrate due to phrasal reuse.</p>				
S5	Gemma (Prompt 3)	Terrie Sharp is a famous writer. She can tell good stories. She has won the Olive Crime Writer Award twice. She was born in a part of Glasgow with many crimes. She learned to be quiet and not have problems. She also learned how to open car doors with a knife. Now, she doesn't need to use this skill. She is busy. She appears on TV shows like Police Today. She also goes to writing events in the UK.	0.859	0.839
<p>Linguist's note (CEFR A2): This version is clear and easy to follow. Words like <i>famous</i> and <i>busy</i> are common and simple. The phrase <i>doesn't need to use</i> gently reduces the focus on violence, which helps learners. There is a small meaning gap (<i>busy</i> is not exactly the same as <i>not quiet</i>), but it is acceptable at A2. Automatic metrics reduce the score when the wording changes, even if the change makes the text easier to read.</p>				

Continued on next page

Table 5 – continued from previous page

Case	Model (Prompt)	Candidate	Sim. Orig.	Sim. Ref.
S6	Gemma (Prompt 3)	Terrie Sharp is a popular writer. She can tell good stories. She won the Olive Crime Writer Award twice. She grew up in a part of Glasgow with many crimes. She learned to be quiet and not have problems. She also learned how to open car doors with a knife. She uses this knowledge only to write her books. Now, she is not quiet. She appears on TV shows like Police Today. She also goes to writing events in the United Kingdom.	0.859	0.840
Linguist's note (CEFR A2): This version is easy to follow, with a helpful time marker in <i>grew up</i> . However, the word <i>knowledge</i> is still too abstract for A2 learners, who would benefit more from concrete wording like <i>what she learned</i> . Automatic metrics give it credit for overlap, but they do not notice that the vocabulary is still too advanced.				
S7	Gemma (Prompt 3)	Terrie Sharp is a popular writer. She can tell good stories. She won the Olive Crime Writer Award twice. She was born in a part of Glasgow with many crimes. She learned to be quiet and not have problems. She also learned to open car doors with a knife. She only uses this knowledge to write her books. Now, she is not quiet. She appears on TV shows like Police Today. She also goes to writing events in the UK.	0.860	0.849
Note: Similar to S6: sentence-level clarity, but persistent abstract nouns (<i>knowledge, appears</i>). Metrics overvalue overlap.				
S8	GPT-4.1-mini (Prompt 4)	Terrie Sharp is a famous writer who knows how to tell good stories. She won the Olive Crime Writer Award two times. She grew up in a poor area of Glasgow where there was a lot of crime. There, she learned to be quiet, avoid trouble, and how to open car doors with a knife. But she only used these skills for her books. Now, she talks on TV shows like Police Today and visits writing events in the UK.	0.856	0.768
Linguist's note (CEFR A2): <i>Poor area</i> is an easy-to-understand paraphrase of <i>housing estate with the highest crime rate</i> (it adds a slight suggestion of poverty). The phrases <i>avoid trouble</i> and <i>talks/visits</i> are clear and simple. Automatic metrics often mark down synonymy even when the result is a better fit for A2 learners.				
S9	GPT-4.1-mini (Prompt 2)	Terrie Sharp is a famous writer who knows how to tell stories. She has won the Olive Crime Writer Award two times. She was born in a part of Glasgow with a lot of crime. There, she learned to be quiet and avoid trouble. She also learned how to open car doors with a knife. She used this knowledge only to write her books. Now, she talks a lot on TV shows like Police Today. She also goes to writing events in the UK.	0.854	0.782
Linguist's note (CEFR A2): Clear, concrete phrasing; <i>talks a lot</i> makes the contrast explicit. <i>Knowledge</i> remains abstract. Automatic metrics often do not give enough credit for phrasal simplification.				
S10	GPT-4.1-mini (Prompt 3)	Terrie Sharp is a writer who writes good stories. She won the Olive Crime Writer Award two times. She was born in a part of Glasgow with many crimes. She learned to be quiet and safe. She also learned how to open car doors with a knife. She only used this skill for her books. Now, she talks on TV shows like Police Today. She also goes to writing meetings in the UK.	0.863	0.866
Linguist's note (CEFR A2): The phrase <i>writer who writes</i> is repetitive. On the positive side, <i>safe</i> is a good simplification. Automatic metrics reward word overlap, but they ignore problems of fluency and style.				
S11	GPT-4.1-mini (Prompt 3)	Terrie Sharp is a very good writer. She won the Olive Crime Writer Award two times. She was born in a part of Glasgow where many crimes happen. She learned to be quiet and to avoid trouble. She also learned how to open car doors with a knife. She did not use this skill for bad things. She uses it to write books. Now, she talks on TV shows like Police Today. She also goes to writing meetings in the UK.	0.855	0.850
Linguist's note (CEFR A2): The line <i>not for bad things</i> adds a clear moral point, which reduces ambiguity and helps A2 readers. The wording is concrete throughout. Automatic metrics usually don't give enough credit for these helpful additions.				
S12	GPT-4.1-mini (Prompt 3)	Terrie Sharp is a writer. She won the Olive Crime Writer Award two times. She was born in a part of Glasgow with a lot of crime. She learned to be quiet and stay safe. She also learned how to open car doors with a knife. She only used this skill to write her books. Now, she talks on TV shows like Police Today. She also goes to writing meetings in the UK.	0.855	0.812
Linguist's note (CEFR A2): This version is concise and easy to read. Some details may be lost, but it still matches A2 aims (short clauses, simple connectors). Automatic metrics often mark down compression even when the result remains adequate for learners.				
S13	Qwen-2.5-14b (Prompt 4)	Terrie Sharp is a writer who sells lots of books. She has won an award two times for writing stories about crimes. She was born in a part of Glasgow with many crimes. She learned to be quiet and not get in trouble. She even learned how to open car doors with a knife! But she never used that skill in real life. Instead, she writes stories about them. Now, she talks on TV shows like Police Today and at big writer meetings all over the UK.	0.868	0.865

Continued on next page

Table 5 – continued from previous page

Case	Model (Prompt)	Candidate	Sim. Orig.	Sim. Ref.
Linguist’s note (CEFR A2): Easy words, but some facts are not correct (<i>award for writing about crimes</i>) and the tone is too dramatic by adding(“!”).				
S14	Qwen-2.5-14b (Prompt 4)	Terrie Sharp, a famous writer, knows how to tell stories really well. She has won an award twice for writing about crimes. She grew up in a place in Glasgow with a lot of crime. She learned to be quiet and not get in trouble, and even how to open car doors with a knife. But now she uses those stories to write her books instead. Today, Terrie is not so quiet anymore. She goes on TV shows like Police Today and speaks at many writing events in the UK.	0.859	0.874
Linguist’s note (CEFR A2): The text is simple and easy to read, but it adds details that are not in the original (<i>award for writing about crimes; uses those stories</i>). These changes make the text less accurate. Automatic metrics give a high score for word overlap, but they fail to notice the meaning errors.				
S15	Qwen-2.5-14b (Prompt 4)	Terrie Sharp is a popular writer who won an award twice. She was born in a place in Glasgow where there were many crimes. She learned to stay quiet and not get in trouble. She even learned how to open car doors with a knife! But she didn’t use those skills for bad things. Instead, she wrote stories about them. Now, she talks a lot and appears on TV shows like Police Today and at writing events all over the UK.	0.856	0.857
Linguist’s note (CEFR A2): Clear explanation of right and wrong; small loss of detail (award name) and a small added idea (“wrote stories about them”). Scores accept these changes but do not notice omissions.				
S16	Qwen-2.5-14b (Prompt 2)	Terrie Sharp is a best-selling author who writes crime stories. She has won the Olive Crime Writer Award twice. She was born in a part of Glasgow with a high crime rate. There, she learned to be quiet and not get in trouble. She used to know how to open car doors with a knife but only used that for her books. Now, Terrie is more outgoing. She appears on TV shows like Police Today and goes to writing events across the UK.	0.858	0.820
Linguist’s note (CEFR A2): Adds extra detail (crime stories) and keeps a hard phrase (“high crime rate”). “Outgoing” changes the meaning of “quiet.” Scores go down a bit for these changes, not for word difficulty.				
S17	Qwen-2.5-14b (Prompt 3)	Terrie Sharp is a writer who won the Olive Crime Writer Award twice. She was born in a neighborhood in Glasgow with a lot of crime. She learned to be quiet and avoid trouble. She also learned how to open car doors with a knife. But she only used this for her stories. Now, Terrie talks a lot on TV shows like Police Today. She also goes to writing meetings all over the UK.	0.856	0.846
Linguist’s note (CEFR A2): The word <i>neighborhood</i> is an easy and familiar synonym that works well for A2 learners. Using <i>this</i> instead of an abstract noun makes the meaning clearer. The phrase <i>talks a lot</i> shows the contrast directly. This is a strong A2 version.				
S18	Qwen-2.5-14b (Prompt 3)	Terrie Sharp writes good stories. She won the Olive Crime Writer Award twice. She was born in a dangerous part of Glasgow. There, she learned to stay quiet and avoid trouble. She even learned to open car doors with a knife. But she never used this skill for bad things. Now, she uses her stories to appear on popular TV shows like Police Today. She also speaks at writing events across the UK.	0.866	0.842
Linguist’s note (CEFR A2): <i>Dangerous part</i> is an excellent simplification. But causal inversion (<i>uses her stories to appear</i>) is a semantic error. Metrics miss meaning coherence.				
S19	Sentence-lv	Terrie Sharp is a good storyteller. She won the Olive Crime Writer Award twice. She was born in a place in Glasgow with a lot of crime. She learned to be silent, avoid problems, and open car doors with a knife. She writes novels using her knowledge. Now, she talks on quiet TV programs in the UK, like Police Today and writing seminars.	0.875	0.723
Linguist’s note (CEFR A2): <i>Silent</i> is a over-literal change for <i>quiet</i> ; <i>quiet TV programs</i> is a meaning error; <i>seminars</i> and <i>knowledge</i> are hard words for A2. The low score shows meaning changes, but metrics cannot pinpoint error type.				

Table 5: Example 109-a2 from the test data. The table presents the original B2 paragraph, the human reference simplification at A2, the final MBR-selected output from Run 3, and 19 candidates selected by the filtering system. Rows shaded **red** are judged (very good), rows shaded **green** are judged (excellent) and rows shaded **yellow** are *poor/problematic* and need revision. Among all candidates, **S8 (GPT-4.1-mini, Prompt 4)** is the best candidate for CEFR A2 pedagogical, balancing lexical simplicity, syntactic ease, semantic fidelity, and learner support through explicit moral clarification.

Case	Model/Prompt	Selection (Non-Selected)	Orig. Sim.	Ref. Sim.
NS1	Gemma-3-4b (Prompt 4)	Terrie Sharp is a very popular writer of stories. She has won the Olive Crime Writer Award two times. She grew up in a place in Glasgow with many crimes. She learned to be quiet and not get into trouble. She also learned how to open car doors with a knife. But she used this only to write her books. Now she is not quiet. She goes on TV shows like Police Today and to writing events in the UK. Linguist's note: This version uses short, clear sentences. The phrase "writer of stories" avoids the business-like idea of "books sell well." It keeps the violent detail but explains it as only used for writing. Automatic metrics mark it down because the wording differs from the reference, but it is still very suitable for A2.	0.828	0.802
NS2	Gemma-3-4b (Prompt 4)	Terrie Sharp is a famous writer who can write many stories. She won the Olive Crime Writer Award two times. She was born in a part of Glasgow with many crimes. She learned to be quiet and not have problems. She also learned how to open car doors with a knife. But now she uses it only for her books. Today she talks a lot on TV shows like Police Today and at writing events in the UK. Linguist's note : The word "famous" is clearer than "popular," and "not have problems" is a good replacement for the idiom "stay out of trouble." The sentences are simple and easy to follow. The only weakness is a small redundancy ("can write many stories"). Metrics lower the score because of different words, not because of difficulty.	0.848	0.797
NS3	Gemma-3-4b (Prompt 1)	Terrie Sharp is a popular writer. She has won the Olive Crime Writer Award twice. She was born in a part of Glasgow with many crimes. She learned to be quiet and not have problems. She also learned how to open car doors with a knife. Now she only uses this for her stories. Today she talks on TV shows like Police Today and at writing events in the UK. Linguist's note: The sentences are clear and short. The idiom "stay out of trouble" is replaced with the simpler "not have problems." All key facts are kept. Metrics give only average credit, even though this is strong A2 material.	0.830	0.830
NS4	Gemma-3-4b (Prompt 1)	Terrie Sharp is a popular writer. She has won the Olive Crime Writer Award twice. She was born in a part of Glasgow with many crimes. She learned to be quiet and not have problems. She also learned how to open car doors with a knife. Now she only uses this for her stories. She talks on TV shows like Police Today and goes to writing events in the UK. Linguist's note: This version has the same strengths as NS3. The similarity score is low because it says "events" instead of "conferences," but "events" is actually easier for A2 learners.	0.848	0.762
NS5	Gemma-3-4b (Prompt 1)	Terrie Sharp is a famous writer. She writes good stories. She has won the Olive Crime Writer Award two times. She was born in a part of Glasgow with many crimes. She learned to be quiet and not have problems. She also learned how to open car doors with a knife. She now uses this only to write her books. She goes on TV shows like Police Today and to writing events in the UK. Linguist's note: The words "famous" and "writes good stories" are simple and transparent. The sentences are split clearly. The meaning is faithful to the original. Metrics reduce the score only because the wording does not match the reference.	0.853	0.816
NS6	Gemma-3-4b (Prompt 2)	Terrie Sharp is a popular writer. She can tell good stories. She won the Olive Crime Writer Award twice. She was born in a part of Glasgow with many crimes. She learned to be quiet and not have problems. She also learned how to open car doors with a knife. She only uses this for her books now. She is not quiet anymore. She talks on TV shows like Police Today and goes to writing events in the UK. Linguist's note: The time marker "now" makes the story easier to follow. The contrast "not quiet anymore" is clear. The only issue is the abstract noun "knowledge," which remains a bit hard for A2. Metrics again don't give full credit for these improvements.	0.854	0.804
NS7	Gemma-3-4b (Prompt 2)	Terrie Sharp is a popular writer. She can tell good stories. She won the Olive Crime Writer Award twice. She was born in a part of Glasgow with many crimes. She learned to be quiet and not have problems. She also learned how to open car doors with a knife. She uses this only to write her books now. She is not quiet anymore. She talks on TV shows like Police Today and goes to writing events in the UK. Linguist's note: Very similar to NS6: short sentences, easy connectors, clear contrast. The similarity score is lower because of harmless word changes.	0.849	0.780
NS8	Gemma-3-4b (Prompt 3)	Terrie Sharp is a famous writer. She can tell good stories. She has won the Olive Crime Writer Award twice. She was born in a part of Glasgow with many crimes. She learned to be quiet and not have problems. She also learned how to open car doors with a knife. She only uses this to write her books now. She goes on TV shows like Police Today and to writing events in the UK. Linguist's note: The verbs "goes on" and "uses" are clear and concrete. Sequencing is simple and effective. Metrics score it lower because the words differ from the reference.	0.852	0.786

Continued on next page

Table 6 – continued from previous page

Case	Model (Prompt)	Candidate (Non-Selected)	Sim. Orig.	Sim. Ref.
NS9	GPT-4.1-mini (Prompt 4)	Terrie Sharp is a famous writer. She has won the Olive Crime Writer Award two times. She was born in Glasgow, a place with a lot of crime. She learned to be quiet and not get into trouble. She also learned to open car doors with a knife. She only used this to write her books. Now, she talks on TV shows like Police Today and goes to writing meetings in the UK. Linguist's note: This version is very clear and accurate. Words like <i>famous</i> and <i>meetings</i> are easy for A2 learners. The sequencing is simple and the contrast is clear. Automatic metrics give it a lower score because of different wording, but it is an excellent A2 simplification.	0.828	0.834
NS10	GPT-4.1-mini (Prompt 1)	Terrie Sharp is a famous writer who has won the Olive Crime Writer Award two times. She was born in a part of Glasgow with a lot of crime. She learned to be quiet and avoid trouble. She also learned how to open car doors with a knife. She used this only to write her books. Now, she talks on TV shows like Police Today and goes to writing meetings in the UK. Linguist's note: This version handles the idiom well (<i>avoid trouble</i>), and the word “meetings” is culturally simple at A2. The content is faithful and the style is easy to read. The lower similarity score only reflects useful word changes, not quality loss.	0.842	0.770

Table 6: Case study analysis of **non-selected** outputs that were linguistically strong but scored lower on automatic metrics. Rows shaded **red** are judged (very good) and rows shaded **green** are judged (excellent). These examples show that metrics often mark down simplifications that use common words (e.g., *famous* vs. *popular*, *meetings* vs. *conferences*) and concrete phrasing, even though they better match CEFR A2 descriptors.

E Sentence Simplification architecture: Run2

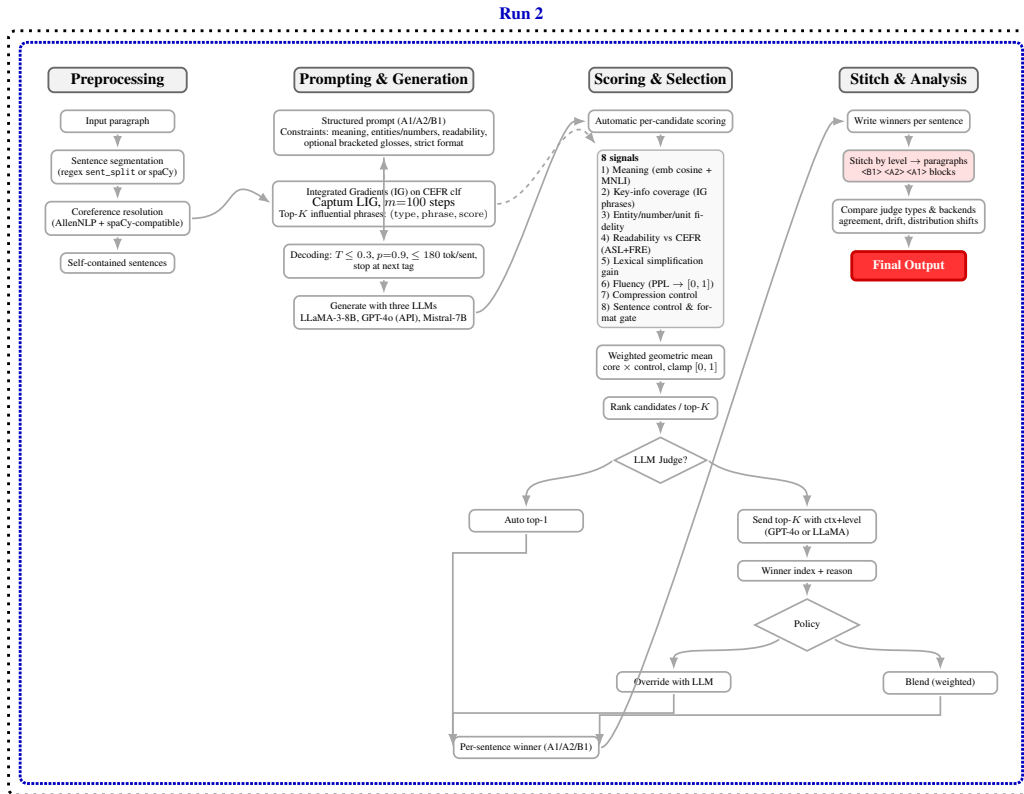


Figure 3: Run 2 preprocessing (segmentation, coreference) and IG attribution; CEFR-controlled prompting/decoding across three LLMs; automatic judge (8 signals) with weighted geometric mean; optional LLM-as-Judge with policy; stitching and comparative analysis.

F Evaluation-Metrics-Sentence simplification: Run2

Metric	Description
Meaning preservation	Embedding cosine similarity plus bidirectional entailment probabilities (MNLI) to assess whether the simplified sentence preserves the meaning of the source.
Key information coverage	Checks whether the top- K influential phrases identified by IG are present in the simplified output (case-insensitive matching).
Entity, number, and unit fidelity	Compares named entities with spaCy (set F1). Numbers are greedily matched one-to-one if units agree, allowing an absolute error within $\max(1\%, 10^{-6})$.
Readability vs. CEFR	Combines average sentence length (ASL) and Flesch Reading Ease (FRE) (Flesch, 1948), normalised to CEFR targets: A1 (ASL ≈ 10 , FRE ≥ 0.80), A2 (15, 0.70), B1 (20, 0.60).
Lexical simplification gain	Reduction in average syllables per word compared to the source. A small bonus is given for inline glosses (e.g., “[a simple meaning]”).
Fluency	Language model perplexity mapped to $[0, 1]$ (Jurafsky and Martin, 2023); lower perplexity means higher fluency. If no LM is provided, a neutral score of 0.75 is assigned.
Compression control	Ratio of simplified to original word counts, normalised to the target range 0.6–1.0. Penalises outputs that are too short or too verbose.
Sentence/format control	Encourages keeping sentence count close to the source (ratio 0.7–1.1). Rejects empty outputs or those exceeding 1200 characters.

Table 7: Evaluation signals used by the automatic judge. Each metric is normalised to $[0, 1]$ and combined by a weighted geometric mean.

Uniandes at TSAR 2025 Shared Task: Multi-Agent CEFR Text Simplification with Automated Quality Assessment and Iterative Refinement

Felipe Arias-Russi^{1,2}, Kevin Cohen-Solano¹, Rubén Manrique¹

¹Systems and Computing Engineering Department, Universidad de los Andes, Colombia

²Department of Mathematics, Universidad de los Andes, Colombia

{af.ariasr, k.cohen, rf.manrique}@uniandes.edu.co

Abstract

We present an agent-based system for the TSAR 2025 Shared Task on Readability-Controlled Text Simplification, which requires simplifying English paragraphs from B2+ levels to target A2 or B1 levels while preserving meaning. Our approach employs specialized agents for keyword extraction, text generation, and evaluation, coordinated through an iterative refinement loop. The system integrates a CEFR vocabulary classifier, pretrained evaluation models, and few-shot learning from trial data. Through iterative feedback between the evaluator and writer agents, our system automatically refines outputs until they meet both readability and semantic preservation constraints. This architecture achieved 4th position among participating teams, showing the effectiveness of combining specialized LLMs with automated quality control strategies for text simplification.

1 Introduction

Text simplification is a central task in natural language processing (NLP), aiming to make written content more accessible by reducing linguistic complexity while preserving meaning and fluency. In recent years, *readability-controlled simplification* has gained relevance, where the simplified output must conform to explicit proficiency levels defined by the Common European Framework of Reference for Languages (CEFR). Such control is important for applications in education, second-language learning, and inclusive communication, with recent work showing that instruction-tuned models can follow explicit readability targets (Tran et al., 2024).

The TSAR 2025 Shared Task on Readability-Controlled Text Simplification (Alva-Manchego et al., 2025), organized as part of the EMNLP conference, challenges participants to simplify English paragraphs originally written at the B2 level or

above into target levels A2 or B1. Official evaluation metrics include (i) compliance with the target CEFR level, (ii) preservation of the original meaning, and (iii) similarity to human-written reference simplifications. While no training data is provided, organizers released 20 trial examples per target level for format familiarization, making this essentially a few-shot learning challenge that encourages creative combinations of linguistic resources, pretrained models, and novel system architectures.

For this shared task, we developed an agent-based pipeline that leverages the modular nature of the simplification task. Our approach integrates specialized agents for **keyword extraction**, **text generation**, and **evaluation**, orchestrated through an **iterative refinement loop**. The system incorporates diverse tools including a CEFR vocabulary classifier that maps words to proficiency levels (A1-C1), pretrained evaluation models (CEFR classifiers, MeaningBERT, BERTScore), and few-shot examples from the trial data. This iterative feedback mechanism reduces the need for human annotation while ensuring outputs meet both readability and meaning preservation requirements. Our system achieved competitive performance, ranking among the top-5 teams in the shared task. All system components, prompts, and implementation details are publicly available.¹

In this paper, we describe our system for the TSAR 2025 shared task. Section 2 reviews related work, Section 3 presents our pipeline in detail, Section 4 reports results and analysis, and Section 5 concludes with key insights and directions for future work.

2 Related Work

Readability-controlled generation has recently advanced through instruction tuning that targets fine-

¹<https://github.com/feliperussi/team-uniandes-tsar-2025-shared-task>

grained complexity levels, showing strong adherence to requested readability and competitive quality (Tran et al., 2024). For longer inputs, multi-agent frameworks coordinate specialized roles to improve document-level coherence and thoroughness, offering an alternative to single-pass simplification (Fang et al., 2025; Lyu and Pergola, 2024). Still, preserving meaning is a key challenge. Paragraph-level human evaluation with reading-comprehension questions shows that even strong systems leave some questions unanswered, highlighting the need for direct meaning checks (Agrawal and Carpuat, 2024).

3 Methodology

3.1 Task Formulation

Given a source paragraph t_0 (B2+ English) and a target CEFR level $\ell^* \in \{A2, B1\}$, the goal is to produce a final simplification t^* that: (i) complies with ℓ^* as verified by CEFR classifiers, (ii) preserves semantic content through high similarity scores with t_0 , and (iii) maintains fluent, coherent paragraph structure. We enforce meaning preservation through automatic metrics and iterative refinement, using t_n to denote intermediate candidates.

3.2 System Overview

Our system employs an *agent-based* pipeline with iterative refinement to balance readability and meaning preservation (Figure 1). The pipeline processes B2+ source paragraphs through keyword extraction, vocabulary classification, and iterative refinement between Writer and Evaluator agents—whose prompts we co-developed via a hybrid process combining human prompt engineering with LLM-assisted drafting using Gemini 2.5 Pro and GPT-5—until CEFR compliance and semantic similarity thresholds are met. This design ensures outputs meet strict quality requirements through automated validation at each step (detailed components below, thresholds and hyperparameters in Section 4.1).

3.2.1 Keyword Extractor

This agent identifies topic-specific terms that are too complex for the target level ℓ^* and require explicit definition. Given source t_0 and target level $\ell^* \in \{A2, B1\}$, it outputs a set $K(t_0, \ell^*) = K$ of keywords that must be defined (not replaced) to preserve meaning. We introduced this separate component because: (i) trial data analysis revealed that

certain domain terms must be defined rather than substituted to maintain accuracy, and (ii) LLMs performing end-to-end simplification tend to either over-define common words or miss crucial technical terms.

The agent applies a two-step test to each noun identified by the agent n : (i) can it be replaced by a simple phrase without meaning loss? (ii) would replacement cause awkward repetition (e.g., “gravity” \rightarrow “the force that pulls things down”)? Terms failing either test are marked for definition. The agent returns at most two keywords to avoid over-constraining generation, outputting them as JSON: `{“keywords”: [“word1”, “word2”]}`. The set may be empty if all terms suit level ℓ^* .

3.2.2 Vocabulary Classification

Using the vocabulary lists from Cambridge University Press and Assessment (2025c,a,b) and ESL Lounge (2025), we built a CEFR vocabulary classifier. Given an input text, it returns a dictionary $D(\text{text})$ mapping each word to its CEFR level (A1, A2, B1, B2, C1). We introduced this component because early experiments showed that LLMs without explicit vocabulary guidance tend to either oversimplify text or miss higher-CEFR terms that require replacement. Table 1 reports the size of each CEFR-specific list used during conditioning and evaluation. Regarding words that are out of the scope of $D(\text{text})$, the proposed prompts leave some margin to the models for adding vocabulary.

CEFR Level	Vocabulary Count
A1	1,282
A2	1,228
B1	1,618
B2	595
C1	1,239
Total	5,962

Table 1: Vocabulary distribution across CEFR levels

3.2.3 Writer Agent

The writer takes the original B2+ paragraph t_0 , target level ℓ^* , keywords $K(t_0, \ell^*)$, vocabulary dictionary D_n , style examples S_{ℓ^*} (that consist in a set of texts in ℓ^* , specifically those known as “reference” in the task trial data), and (for refinement) previous output t_n with feedback f_n . Let $\theta = (t_0, \ell^*, K, S_{\ell^*})$ denote the fixed inputs and $D(t_n)$ the dictionary for each candidate text t_n .

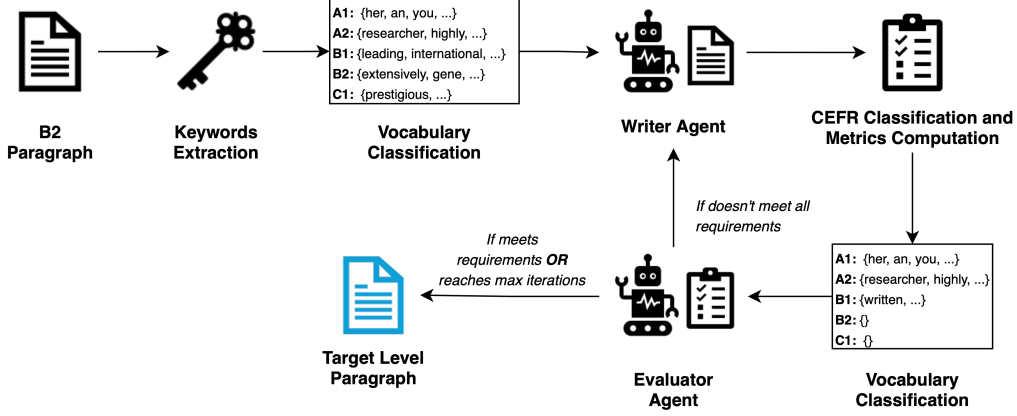


Figure 1: Agent-based pipeline architecture. The system processes B2+ source paragraphs through keyword extraction and vocabulary classification, followed by iterative refinement between the Writer and Evaluator agents. The loop continues until CEFR compliance and semantic similarity thresholds are met or maximum iterations are reached.

Then:

$$t_1 = \text{Writer}(\theta, D(t_0)) \quad (1)$$

$$t_{n+1} = \text{Writer}(\theta, D(t_n), t_n, f_n) \quad \text{for } n \geq 1 \quad (2)$$

In initial generation, the agent: (i) analyzes S_{ℓ^*} to internalize target style from trial data examples, leveraging this data to match expected output patterns while limiting samples to avoid confusion, (ii) extracts core message from t_0 , (iii) defines keywords $k \in K$ and uses D to identify vocabulary above target ℓ^* , (iv) restructures sentences to match ℓ^* 's complexity, and (v) returns the simplified paragraph. During refinement iterations, it minimally edits t_n to address issues in f_n . Table 2 summarizes the level-specific constraints.

3.2.4 Automatic Metrics

For each source–hypothesis pair (t_0, t_n) with target level $\ell^* \in \{A2, B1\}$, we compute three values:

CEFR label ($\hat{\ell}$). We use three ModernBERT classifiers fine-tuned on different subsets of the UniversalCEFR dataset (Imperial et al., 2025): (i) document-level English, (ii) sentence-level English, and (iii) multilingual texts. Each model $m \in \mathcal{E}$ outputs a predicted level with confidence score. We select the prediction from the model with highest confidence:

$$\ell(t_n) = \text{label} \left(\arg \max_{m \in \mathcal{E}} \text{score}_m(t_n) \right)$$

Following Barayan et al. (2025), we evaluate CEFR compliance using adjacent accuracy (accepting predictions within one level of target) and RMSE between predicted and target level indices.

BERTScore (semantic similarity). We report mean F1 BERTScore (Zhang et al., 2020) between prediction and source:

$$\text{BS}(t_n, t_0) \in [0, 1],$$

interpreted as embedding-based semantic similarity.

MeaningBERT (meaning preservation). We use MeaningBERT (Beauchemin et al., 2023) to assess meaning preservation between t_n and t_0 . The metric outputs a percentage score which we normalize to $[0,1]$:

$$\text{MB}(t_n, t_0) = \frac{\text{MeaningBERT}(t_n, t_0)}{100} \in [0, 1]$$

3.2.5 Evaluator Agent

Given (θ, t_n, D_n) where $D_n = D(t_n)$ is the CEFR vocabulary for the text t_n and automatic metrics (CEFR label $\ell(t_n) = \ell_n$, $\text{BS}_n = \text{BS}(t_n, t_0)$, $\text{MB}_n = \text{MB}(t_n, t_0)$), the evaluator produces feedback:

$$f_n = \text{Eval}(\theta, t_n, D_n, \ell_n, \text{BS}_n, \text{MB}_n) \quad (3)$$

The evaluator assesses outputs through a priority-ordered evaluation pipeline:

1. **Factual integrity:** Verifies numbers, names, dates, locations, and core relations match t_0
2. **Meaning preservation:** Enforces thresholds specified in Section 4.1
3. **CEFR compliance:** Ensures predicted level $\ell_n = \ell^*$

	A2 (elementary)	B1 (intermediate)
Sentence length	≤ 12 words, one idea	15–25 words, combined ideas
Connectors	<i>and, but, because, so, ...</i>	A2 + <i>moreover, although, however, ...</i>
Keyword handling	define before first use	inline definition
Grammar	simple tenses only	simple + controlled B1 forms
Content scope	essential facts only	preserve key details

Table 2: Level-specific writer constraints controlled by $(K(t_0, \ell^*), D(t_0), S_{\ell^*})$. Aligned with CEFR reading comprehension descriptors (Companion Volume, Ch. 3.1.1.3) (Council of Europe, 2020).

4. Qualitative audit: Checks level-specific constraints (see Table 2)

Failure at any priority triggers targeted feedback. The evaluator outputs f_n as JSON containing approval status (PASS/FAIL), grade (1–10), and explanation of issues with suggested changes. Crucially, the evaluator interprets the automatic metrics (BS_n, MB_n, ℓ_n) and explains their implications to the writer, enabling targeted corrections—a design choice motivated by early experiments where texts had incorrect CEFR classifications or low semantic similarity. If f_n indicates failure, the writer produces t_{n+1} addressing the feedback; if f_n indicates pass, the loop terminates with $t^* = t_n$.

3.2.6 Refinement Loop Rules

We run a Writer–Evaluator loop capped at N maximum iterations. At each iteration n , the candidate t_n is evaluated against the thresholds in Section 4.1; if it passes, the loop terminates and we output t_n . Every iteration logs the candidate and metrics $(t_n, \ell_n, MB_n, BS_n, f_n)$. If no candidate passes within N iterations, we select the candidate with the highest MB among those correctly classified at the target CEFR level, or if none exist, the candidate with the highest MB overall. Algorithm 1 formalizes this process.

Our repository includes the complete prompts for the Writer, Evaluator, and Keyword Extractor agents; the n8n workflow JSON configurations for agent orchestration; the API implementation with CEFR vocabulary tools; and the trial data preprocessing scripts.

4 Evaluation and Results

4.1 System Configuration

Our system uses $|S_{\ell^*}| = 10$ style examples per level from TSAR trial data, evaluation thresholds of $MB \geq 0.75$ and $BS \geq 0.90$, maximum iterations $N \in \{5, 10\}$ to avoid longer loops, keyword limit $|K| \leq 2$ irreplaceable terms per paragraph,

Algorithm 1 Writer-Evaluator Iterative Refinement

Require: Source text t_0 (B2+), target CEFR level $\ell^* \in \{A2, B1\}$, max iterations N
Ensure: Simplified text t^*
1: $K := K(t_0, \ell^*), S := S_{\ell^*}, D_0 := D(t_0)$
2: $\theta := (t_0, \ell^*, K, S)$
3: $t_1 := \text{Writer}(\theta, D_0)$
4: **for** $n = 1$ to N **do**
5: $\ell_n := \ell(t_n)$
6: $MB_n := \text{MB}(t_n, t_0)$
7: $BS_n := \text{BS}(t_n, t_0)$
8: $D_n := D(t_n)$
9: $f_n := \text{Eval}(\theta, t_n, D_n, \ell_n, BS_n, MB_n)$
10: **if** f_n indicates PASS **then**
11: **return** $t^* := t_n$
12: **end if**
13: **if** $n < N$ **then**
14: $t_{n+1} := \text{Writer}(\theta, D_n, t_n, f_n)$
15: **end if**
16: **end for**
17: $I^* := \{i \in \{1, \dots, N\} : \ell_i = \ell^*\}$
18: **if** $I^* \neq \emptyset$ **then**
19: **return** $t^* := t_j$ where $j := \arg \max_{i \in I^*} MB_i$
20: **else**
21: **return** $t^* := t_j$ where $j := \arg \max_{i=1, \dots, N} MB_i$
22: **end if**

and temperature 0 for all models to ensure deterministic outputs. These values were selected based on trial data analysis to balance quality and efficiency. Check Table 3 for the summary of each run.

Run	Writer	Evaluator	N
Run 1	Gemini 2.5 Pro	Gemini 2.5 Pro	5
Run 2	Gemini 2.5 Flash	Gemini 2.5 Flash	10
Run 3	GPT-OSS-120B	Gemini 2.5 Pro	10

Table 3: System configurations evaluated. Run 1 uses Gemini 2.5 Pro (Google AI for Developers, 2025), Run 3 uses GPT-OSS-120B (Together AI, 2025; OpenAI, 2025).

We orchestrate the agents with n8n (2025) and a REST API serving level-specific control inputs. The trial data provided style examples and vocabulary guidance.

4.2 Results

Our team Uniandes achieved 4th place in the TSAR 2025 Shared Task, with the resulting metrics shown in Table 4. The Gemini 2.5 Flash configuration provided the best balance between CEFR compliance and efficiency, while the hybrid GPT-OSS-120B/Gemini 2.5 Pro configuration excelled at meaning preservation, as seen in Table 5. All systems maintained high semantic similarity ($BS > 0.92$) while successfully adapting texts to target CEFR levels.

Run	F1	Adj. Acc.	RMSE
Run 1	0.972	1.00	0.212
Run 2	0.985	1.00	0.200
Run 3	<u>0.851</u>	<u>0.97</u>	<u>0.510</u>

Table 4: CEFR Compliance metrics on TSAR 2025 test set. Best scores in bold and worst in underline.

Run	MB-Orig	MB-Ref	BS-Orig	BS-Ref
Run 1	<u>0.817</u>	0.814	0.936	0.934
Run 2	0.823	<u>0.803</u>	0.934	0.930
Run 3	0.847	0.813	0.933	0.928

Table 5: Comparison between meaning preservation metrics on TSAR 2025 test set. Best scores in bold and worst in underline. MB and BS stand for MeaningBERT and BERTScore respectively.

5 Discussion

Our experiments showed distinct trade-offs across model configurations. While Run 2 achieved the best CEFR compliance and Run 1 demonstrated high precision and provided the most reliable balance between CEFR compliance and meaning preservation. The hybrid Run 3 struggled with level targeting. For meaning preservation, Run 1 is more stable and shows less overfitting; Run 3, despite having a higher MB-Orig, has a lower MB-Ref, indicating overfitting to the source text.

After analyzing simpler agent configurations on trial data, several design improvements significantly increased performance. We found that using all 20 trial examples was counterproductive and wasteful of tokens, so limiting to 10 samples optimized both performance and efficiency. Style examples enhanced output consistency, keyword extraction preserved domain-specific meaning through definitions, and CEFR vocabulary classification prevented inconsistent term replacements.

However, some configurations required many iterations to achieve correct CEFR levels, suggesting convergence problems. Our MeaningBERT threshold of 0.75 may have been conservative—higher thresholds could enforce stronger semantic preservation.

Furthermore, Run 3 demonstrated that hybrid architectures with open models as writers can excel at semantic preservation despite weaker level control. This suggests that open models could be valuable as writer agents when paired with strong closed-model evaluators like Gemini 2.5 Pro, potentially offering cost-effective alternatives to fully proprietary systems.

6 Limitations

Key limitations include expensive and unpredictable token generation from iterative refinement, with some texts requiring many iterations and extended processing times. The CEFR vocabulary coverage could be potentially missing some terms. The token consumption was highly variable across different texts, making cost prediction difficult.

Our CEFR and semantic-similarity thresholds were intentionally conservative to favor coverage and reduce non-convergence. More ambitious (stricter) thresholds might yield higher precision in meaning preservation and level control, but at the cost of lower acceptance rates and longer refinement. Exploring adaptive or curriculum-style thresholds is left for future work.

7 Lay Summary

This work builds a system that turns hard-to-read English paragraphs into easier ones for learners, at two target levels: A2 (elementary) and B1 (intermediate). The main goal is to make the text simpler without changing its meaning. To do this, the system uses three “agents” that work together in a loop: one finds important, difficult words that should be defined (not replaced), another writes a simpler version, and a third checks the result. If the check finds problems—like lost meaning or the level being too hard—the writer tries again. This repeat-and-improve cycle continues until the text is both simple enough and faithful to the original.

To guide the writing, the system uses word lists tied to the CEFR levels (A1–C1) and short example texts that show the expected style. To check quality automatically, it uses tools that (1) estimate the reading level, and (2) measure how closely the new

text matches the original meaning. Several model setups were tested and set clear passing rules before accepting any output.

In a public competition on readability-controlled simplification, this approach ranked among the top teams (4th place). The results show that combining specialized roles with automatic checks can reliably simplify text while keeping its meaning. As limitations, some paragraphs need many rounds (which can be costly), and the vocabulary lists may not cover every word.

References

- Sweta Agrawal and Marine Carpuat. 2024. [Do text simplification systems preserve meaning? a human evaluation via reading comprehension](#). *Transactions of the Association for Computational Linguistics*, 12:432–448.
- Fernando Alva-Manchego, Regina Stodden, Joseph Marvin Imperial, Abdullah Barayan, Kai North, and Harish Tayyar Madabushi. 2025. [Findings of the TSAR 2025 Shared Task on Readability-Controlled Text Simplification](#). In *Proceedings of the Fourth Workshop on Text Simplification, Accessibility, and Readability (TSAR 2025)*, Suzhou, China. Association for Computational Linguistics. Latest update: September 2025; archived at <https://web.archive.org/web/20250914180215/https://tsar-workshop.github.io/shared-task/>.
- Abdullah Barayan, Jose Camacho-Collados, and Fernando Alva-Manchego. 2025. [Analysing zero-shot readability-controlled sentence simplification](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6762–6781, Abu Dhabi, UAE. Association for Computational Linguistics.
- David Beauchemin, Horacio Saggion, and Richard Khoury. 2023. [MeaningBERT: assessing meaning preservation between sentences](#). *Frontiers in Artificial Intelligence*, 6. Publisher: Frontiers.
- Cambridge University Press and Assessment. 2025a. [Vocabulary List for A2 level](#).
- Cambridge University Press and Assessment. 2025b. [Vocabulary List for B1 level](#).
- Cambridge University Press and Assessment. 2025c. [Vocabulary List for Pre A1 Starters, A1 Movers and A2 Flyers](#).
- Council of Europe. 2020. *Common European Framework of Reference for Languages: Learning, teaching, assessment — Companion volume*. Council of Europe Publishing, Strasbourg. Reading comprehension descriptors (see Chapter 3.1.1.3).
- ESL Lounge. 2025. [ESL Lounge: Learn English with ESL Lounge](#).
- Dengzhao Fang, Jipeng Qiang, Xiaoye Ouyang, Yi Zhu, Yunhao Yuan, and Yun Li. 2025. [Collaborative document simplification using multi-agent systems](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 897–912, Abu Dhabi, UAE. Association for Computational Linguistics.
- Google AI for Developers. 2025. [Gemini models: Gemini API | Google AI for Developers](#). Latest update: August 2025; archived at <https://web.archive.org/web/20250825013047/https://ai.google.dev/gemini-api/docs/models#previous-experimental-models>.
- Joseph Marvin Imperial, Abdullah Barayan, Regina Stodden, Rodrigo Wilkens, Ricardo Munoz Sanchez, Lingyun Gao, Melissa Torgbi, Dawn Knight, Gail Forey, Reka R. Jablonkai, Ekaterina Kochmar, Robert Reynolds, Eugenio Ribeiro, Horacio Saggion, Elena Volodina, Sowmya Vajjala, Thomas Francois, Fernando Alva-Manchego, and Harish Tayyar Madabushi. 2025. [Universalcefr: Enabling open multilingual research on language proficiency assessment](#). *Preprint*, arXiv:2506.01419.
- Chen Lyu and Gabriele Pergola. 2024. [Society of Medical Simplifiers](#). In *Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024)*, pages 61–68, Miami, Florida, USA. Association for Computational Linguistics.
- n8n. 2025. [n8n: Workflow automation platform \(v1.109.0\)](#). <https://github.com/n8n-io/n8n/releases/tag/n8n@1.108.0>. Version 1.109.0, released August 25, 2025.
- OpenAI. 2025. [gpt-oss-120b & gpt-oss-20b model card](#). *Preprint*, arXiv:2508.10925.
- Together AI. 2025. [gpt-oss-120B API](#). Archived at <https://web.archive.org/web/20250821140846/https://www.together.ai/models/gpt-oss-120b>.
- Hieu Tran, Zonghai Yao, Lingxi Li, and Hong Yu. 2024. [Readctrl: Personalizing text generation with readability-controlled instruction learning](#). *Preprint*, arXiv:2406.09205.
- 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*.

EhiMeNLP at TSAR 2025 Shared Task: Candidate Generation via Iterative Simplification and Reranking by Readability and Semantic Similarity

Rina Miyata[†] Koki Horiguchi[†] Risa Kondo[†] Yuki Fujiwara[‡] Tomoyuki Kajiwara^{†*}

[†] Graduate School of Science and Engineering, Ehime University, Japan

[‡] Faculty of Engineering, Ehime University, Japan

* D3 Center, The University of Osaka, Japan

{miyata@ai., horiguchi@ai., kondo@ai., fujiwara@ai., kajiwara@} cs.ehime-u.ac.jp

Abstract

We introduce the EhiMeNLP submission, which won the TSAR 2025 Shared Task on Readability-Controlled Text Simplification. Our system employed a two-step strategy of candidate generation and reranking. For candidate generation, we simplified the given text into more readable versions by combining multiple large language models with prompts. Then, for reranking, we selected the best candidate by readability-based filtering and ranking based on semantic similarity to the original text.

1 Introduction

Text simplification (Alva-Manchego et al., 2020b) is a task of paraphrasing complex expressions into simpler ones while preserving the core meaning of a given text. This technology is utilized to support reading comprehension for diverse readers, including children (De Belder and Moens, 2010), language learners (Petersen and Ostendorf, 2007), and individuals with language impairments (Evans et al., 2014). Since reading ability varies significantly among individuals, recent studies on text simplification have focused on controlling readability (Scarton and Specia, 2018; Nishihara et al., 2019; Yanamoto et al., 2022; Agrawal and Carpuat, 2023; Barayan et al., 2025).

This paper describes our system submitted to TSAR 2025 shared task on readability-controlled text simplification (Alva-Manchego et al., 2025). Our system employs a two-step approach: first, generating candidates for simplified text with large language models (LLMs), and then reranking the candidates with embeddings. Our team was ranked first out of 20 teams in the official evaluation (AUTORANK). Our code is available on GitHub.¹

¹<https://github.com/EhimeNLP/TSAR2025>

2 TSAR 2025 Shared Task

In this competition (Alva-Manchego et al., 2025), participants are asked to simplify English paragraphs written at the Common European Framework of Reference for Languages (CEFR)² readability level of B2 or more complex into simpler versions at levels B1 and A2, respectively. The CEFR is the most widely used international standard for describing the language ability of English learners, consisting of six levels ranging from basic (A1) to proficient (C2).

The dataset provided consists of trial data, containing 40 paragraphs, and test data, containing 200 paragraphs. The output texts are automatically evaluated for both RMSE of readability and semantic similarity. Note that while both BERTScore (Zhang et al., 2020) and MeaningBERT (Beauchemin et al., 2023) are included in the official evaluation script for semantic similarity, only the latter is used in the official final ranking.

3 EhiMeNLP System

Figure 1 shows an overview of our system. Our system employs a two-step strategy of candidate generation and reranking; we describe these proposed methods in Sections 3.1 and 3.2. We then provide implementation details in Section 3.3. Finally, we report the results of preliminary experiments on the trial dataset in Section 3.4.

3.1 Step 1: Candidate Generation

We iteratively apply the proposed prompts shown in Figure 2 to multiple LLMs to generate simplification candidates. To diversify the candidates, we propose four types of prompts.

P1: fine-grained simplification To include simplified texts with various readability levels in our

²<https://www.coe.int/en/web/common-european-framework-reference-languages>

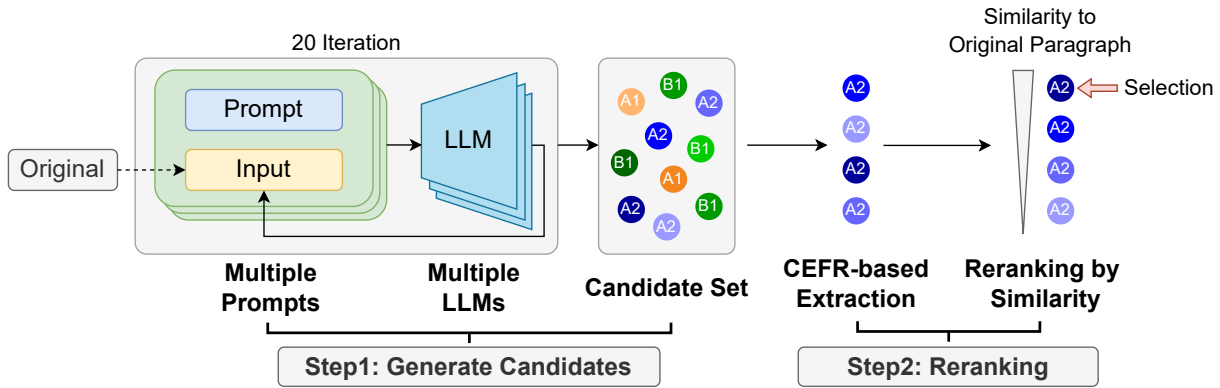


Figure 1: Overview of EhiMeNLP system.

candidate set, we iteratively generate paraphrases that are slightly simpler than the input text. Specifically, we define the readability level of the input text as i and instruct the LLMs to generate paraphrases at readability level $i - 1$. Note that this prompt does not provide any other explanation, such as a detailed description of the readability.

P2: controlling CEFR levels We employ the existing prompts (Barayan et al., 2025) that explicitly describe the reading comprehension abilities of learners at each CEFR level. Note that to adapt for this task, we modify the text unit from sentences to paragraphs. In addition, the target CEFR level changes incrementally from B2 to A1.

P3: controlling grade levels Instead of CEFR levels, we employ US grade levels, which are commonly used in previous studies on text simplification. In this prompt, we instruct the LLMs to simplify the input text to make it easier to understand for students in the target grade level. The grade level changes incrementally from 10 to 1.

P4: Clarifying edit operations A previous study (Kew et al., 2023) has reported the effectiveness of prompts that explicitly instruct editing operations (Alva-Manchego et al., 2020a) for LLM-based text simplification. We also utilize this prompt (Alva-Manchego et al., 2020a; Kew et al., 2023) to instruct LLMs to perform editing operations for text simplification, including paraphrasing, sentence compression, and sentence splitting.

3.2 Step 2: Reranking

In this step, we select the candidate with the highest semantic similarity to the input text while matching the target readability level.

Readability-based Filtering: First, we estimate the CEFR-based readability level for each candidate using the official evaluation script (Alva-Manchego et al., 2025). Then, we extract the set of candidates with the minimum difference from the target CEFR-based readability level.

Similarity-based Ranking: First, we estimate the semantic similarity between each candidate and the original text using the official evaluation script, based on both BERTScore³ (Zhang et al., 2020) and MeaningBERT⁴ (Beauchemin et al., 2023). Then, we select the candidate with the highest average score. In case multiple candidates achieve the highest average score, we select the candidate with the higher MeaningBERT score.

3.3 Implementation Details

As shown in Table 1, we employed six LLMs: GPT-5,⁵ GPT-4.1,⁶ o3,⁷ gpt-oss-20b,⁸ Llama-3.3-70B-Instruct,⁹ and Qwen3-32B.¹⁰ For gpt-oss-20b, we employed two configurations: one with reasoning_effort set to “low” and another with it set to “medium” to control the depth of thought. Regarding Qwen3-32B, we configured enable_thinking, which controls whether thinking occurs, to “False” from the perspective of inference speed.

³<https://huggingface.co/spaces/evaluate-metric/bertscore>

⁴<https://huggingface.co/davebulaval/meaningbert>

⁵<https://platform.openai.com/docs/models/gpt-5>

⁶<https://platform.openai.com/docs/models/gpt-4.1>

⁷<https://platform.openai.com/docs/models/o3>

⁸<https://huggingface.co/openai/gpt-oss-20b>

⁹<https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct>

¹⁰<https://huggingface.co/Qwen/Qwen3-32B>

```

{"role": "system", "content":
  {
    "Prompt 1": "You are an expert in fine-grained text simplification. Given that Complex Paragraph has a readability level of {i}, simplify Complex Paragraph to a readability level of {i-1}."
    "Prompt 2": "Please simplify the following Complex Paragraph to make it easier to read and understand by {TARGET} CEFR level English learners. {TARGET} level English learner {DESCRIPTION[TARGET]}. To simplify, you may replace difficult words with simpler ones, elaborate, or remove them when possible. You may also break down a lengthy sentence into shorter, clear sentences. Ensure the revised sentence is grammatically correct, fluent, and maintains the core message of the original without changing its meaning."
    "Prompt 3": "You are an expert educational material developer in US. Please simplify the following Complex Paragraph to make it easier to read and understand by a US student at the {TARGET} grade level. To simplify, you may replace difficult words with age-appropriate ones, elaborate, or remove them when possible. You may also break down a lengthy sentence into shorter, clear sentences. Ensure the revised sentence is grammatically correct, fluent, and maintains the core message of the original without changing its meaning."
    "Prompt 4": "Please rewrite the following complex paragraph in order to make it easier to understand by non-native speakers of English. You can do so by replacing complex words with simpler synonyms (i.e. paraphrasing), deleting unimportant information (i.e. compression), and/or splitting a long complex sentence into several simpler ones. The final simplified paragraphs needs to be grammatical, fluent, and retain the main ideas of its original counterpart without altering its meaning."
  },
  "DESCRIPTION" = {
    "B2": "can read with a large degree of independence, adapting style and speed of reading to different texts and purposes, and using appropriate reference sources selectively. Has a broad active reading vocabulary, but may experience some difficulty with low-frequency idioms."
    "B1": "can read straightforward factual texts on subjects related to their field of interest with a satisfactory level of comprehension",
    "A2": "can understand short, simple texts containing the highest frequency vocabulary, including a proportion of shared international vocabulary items",
    "A1": "can understand very short, simple texts a single phrase at a time, picking up familiar names, words and basic phrases and rereading as required",
  }
}
{"role": "user", "content": "Complex Paragraph: {COMPLEX_PARAGRAPH}\nSimplified Paragraph: "}

```

Figure 2: Prompts to generate simplification candidates.

All experiments were conducted using four RTX A6000 GPUs. For each LLM model, each prompt was run 20 times. Note that for prompt P2, there are 5 runs each for the 4 target CEFR levels, and for prompt P3, there are 2 runs each for the 10 target grade levels, totaling 20 runs in each case.

3.4 Preliminary Experiments

The left side of Table 2 shows the results of our preliminary experiments on the trial dataset. In this preliminary experiment, we applied four types of proposed prompts to seven types of LLMs to generate candidates using a total of 28 models, and counted how frequently each model was selected by our reranking. Experimental results reveal that the GPT-5 model is notably powerful and that the P1 prompt is remarkably useful.

Based on the results of the preliminary experiments, we have decided the three systems to be submitted as follows. Since the Llama-3.3-70B-Instruct model has only limited contributions, we decided not to employ it

in our final system. In addition to the ensemble method for all LLMs and prompts, we decided to submit base models applying either the P1 prompt or the P3 prompt to the GPT-5 model.

4 Evaluation

Our EhiMeNLP system achieved first place in the official ranking (Alva-Manchego et al., 2025). As shown in Table 3, our system achieved a perfect score in the RMSE evaluation for readability. This demonstrates that our diverse set of candidates consistently generated text suitable for the target readability level. Regarding semantic similarity, our system achieved the fourth-highest score in both similarity to the source and reference texts, respectively. These results reveal that our system achieves readability control that balances both appropriate readability and high semantic similarity.

4.1 Ablation Analysis

Table 3 shows the performance of the base models, which apply the proposed prompts individually

Model	Reference	Inference	Token limit
GPT-5	gpt-5-2025-08-07	OpenAI API with greedy decoding	128,000
GPT-4.1	gpt-4.1-2025-04-14	OpenAI API with greedy decoding	32,768
o3	o3-2025-04-16	OpenAI API with greedy decoding	100,000
gpt-oss-20b	(OpenAI, 2025)	vLLM (Kwon et al., 2023)	40,000
Llama-3.3-70B-Instruct	(Llama Team, 2024)	vLLM (Kwon et al., 2023)	400
Qwen3-32B	(Qwen Team, 2025)	Transformers (Wolf et al., 2020)	32,768

Table 1: The LLM models used in this study.

	Trial					Test				
	P1	P2	P3	P4	Total	P1	P2	P3	P4	Total
GPT-5	6	2	4	2	14	30	19	12	6	67
Qwen3-32B	3	1	0	2	6	16	8	4	6	34
gpt-oss-20b (medium)	1	1	2	2	6	16	5	5	4	30
gpt-oss-20b (low)	3	0	1	1	5	22	3	5	2	32
GPT-4.1	3	0	1	0	4	12	8	2	1	23
o3	2	1	0	1	4	7	3	4	0	14
Llama-3.3-70B-Instruct	0	0	1	0	1	-	-	-	-	-
Total	18	5	9	8	40	103	46	32	19	200

Table 2: The frequency with which candidates generated by each model were finally selected.

Submission Name	Model	RMSE	MeaningBERT-orig	MeaningBERT-ref	Rank
EhiMeNLP / run1	Ensemble	0.000	0.902	0.845	1/48
EhiMeNLP / run2	GPT-5 with P1	0.200	0.838	0.816	7/48
-	GPT-5 with P2	0.265	0.850	0.836	-
EhiMeNLP / run3	GPT-5 with P3	0.234	0.847	0.840	6/48
-	GPT-5 with P4	0.394	0.844	0.836	-

Table 3: Results of the EhiMeNLP systems on the test dataset.

to the GPT-5 model. The ensemble model outperformed the base models across all evaluation metrics. This highlights the importance of having a diverse set of candidates.

The P1 prompt we submitted as our run2 received relatively high scores for readability, but scored lower than other proposed prompts in terms of semantic similarity. The P3 prompt we submitted as our run3 outperformed run2 in the official ranking due to its better balance between readability and semantic similarity.

4.2 Contributions of Each Model and Prompt

The right side of Table 2 shows how many times each combination of model and prompt was selected in the test dataset. Although prompt P1

accounts for the majority, the other prompts also account for about half in total, indicating that combinations of multiple prompts are useful. In terms of LLM models, while GPT-5 is the most frequently selected, Qwen3-32B and gpt-oss-20b also often appear, suggesting that combining multiple models contributes to improving the performance of the ensemble model.

5 Conclusion

We described the EhiMeNLP submission for the TSAR 2025 shared task. Our system employed a two-step strategy in which LLMs generated diverse candidates, followed by re-ranking based on readability and semantic similarity, achieving first place among 20 teams in the official ranking.

Lay Summary

This paper describes a text simplification system that paraphrases a given English text to a specific readability level. The TSAR 2025 workshop held a shared task on readability-controlled text simplification, with 20 teams competing to demonstrate the performance of their systems. Our EhiMeNLP system achieved the top performance among them.

Our system employed a two-step strategy: first, we leveraged large language models (LLMs) to generate diverse simplification candidates, and then selected the final output text through re-ranking based on readability and similarity. While LLMs are good at paraphrasing, they are not necessarily good at controlling readability levels. Therefore, we decided to generate a variety of paraphrases with different readability levels as candidates for simplification. To generate diverse candidates for simplification, we provided four types of prompts to six LLMs and performed repeated simplification. Our re-ranking step consists of two components: filtering based on readability and re-ranking based on similarity. This process enables our system to achieve high synonymy with the input text while respecting the target readability level.

According to official evaluations, our system perfectly satisfies the target readability while also achieving a high level of semantic similarity with both input and reference texts. Our detailed analysis revealed that while GPT-5 is powerful, its ensemble with other LLMs is proving effective.

Acknowledgments

This work was supported by JST BOOST Program Japan Grant Number JPMJBY24036821, Cross-ministerial Strategic Innovation Promotion Program (SIP) on “Integrated Health Care System” Grant Number JPJ012425, and JSPS KAKENHI Grant Number JP25K03233.

References

Sweta Agrawal and Marine Carpuat. 2023. [Controlling Pre-trained Language Models for Grade-Specific Text Simplification](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12807–12819.

Fernando Alva-Manchego, Louis Martin, Antoine Bordes, Carolina Scarton, Benoît Sagot, and Lucia Specia. 2020a. [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.

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2020b. [Data-Driven Sentence Simplification: Survey and Benchmark](#). *Computational Linguistics*, 46(1):135–187.

Fernando Alva-Manchego, Regina Stodden, Joseph Marvin Imperial, Abdullah Barayan, Kai North, and Harish Tayyar Madabushi. 2025. [Findings of the TSAR 2025 Shared Task on Readability-Controlled Text Simplification](#). In *Proceedings of the Fourth Workshop on Text Simplification, Accessibility, and Readability*.

Abdullah Barayan, Jose Camacho-Collados, and Fernando Alva-Manchego. 2025. [Analysing Zero-Shot Readability-Controlled Sentence Simplification](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6762–6781.

David Beauchemin, Horacio Saggion, and Richard Khoury. 2023. [MeaningBERT: Assessing Meaning Preservation Between Sentences](#). *Frontiers in Artificial Intelligence*, 6.

Jan De Belder and Marie-Francine Moens. 2010. [Text Simplification for Children](#). In *Proceedings of the SIGIR 2010 Workshop on Accessible Search Systems*, pages 19–26.

Richard Evans, Constantin Orăsan, and Iustin Dornescu. 2014. [An evaluation of syntactic simplification rules for people with autism](#). In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 131–140.

Tannon Kew, Alison Chi, Laura Vásquez-Rodríguez, Sweta Agrawal, Dennis Aumiller, Fernando Alva-Manchego, and Matthew Shardlow. 2023. [BLESS: Benchmarking Large Language Models on Sentence Simplification](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13291–13309.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. [Efficient Memory Management for Large Language Model Serving with PagedAttention](#). In *Proceedings of the 29th Symposium on Operating Systems Principles*, page 611–626.

Llama Team. 2024. [The Llama 3 Herd of Models](#). *arXiv:2407.21783*.

Daiki Nishihara, Tomoyuki Kajiwara, and Yuki Arase. 2019. [Controllable Text Simplification with Lexical Constraint Loss](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 260–266.

OpenAI. 2025. [gpt-oss-120b & gpt-oss-20b Model Card](#). *arXiv:2508.10925*.

- Sarah E. Petersen and Mari Ostendorf. 2007. [Text Simplification for Language Learners: A Corpus Analysis](#). In *Proceedings of the 1st Workshop on Speech and Language Technology in Education*, pages 69–72.
- Qwen Team. 2025. [Qwen3 Technical Report](#). *arXiv:2505.09388*.
- Carolina Scarton and Lucia Specia. 2018. [Learning Simplifications for Specific Target Audiences](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 712–718.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. [Transformers: State-of-the-Art Natural Language Processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.
- Daiki Yanamoto, Tomoki Ikawa, Tomoyuki Kajiwara, Takashi Ninomiya, Satoru Uchida, and Yuki Arase. 2022. [Controllable Text Simplification with Deep Reinforcement Learning](#). In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, pages 398–404.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. [BERTScore: Evaluating Text Generation with BERT](#). In *Proceedings of the 8th International Conference on Learning Representations*.

OUNLP at TSAR 2025 Shared Task: Multi-Round Text Simplifier via Code Generation

Cuong Huynh

School of Computer Science
University of Oklahoma
cuong@ou.edu

Jie Cao

School of Computer Science
University of Oklahoma
jie.cao@ou.edu

Abstract

This paper describes the system submission of our team OUNLP to the TSAR-2025 shared task (Alva-Manchego et al., 2025) on readability-controlled text simplification. Based on the analysis of prompt-based text simplification methods, we discovered an interesting finding that text simplification performance is highly related to the gap between the source CEFR (Arase et al., 2022) level and the target CEFR level. Inspired by this finding, we propose two multi-round simplification methods and generate them via GPT-4o: rule-based simplification (MRS-Rule) and jointly rule-based LLM simplification (MRS-Joint). Our submitted systems ranked 7 out of 20 teams. Later improvements with MRS-Joint show that taking the LLM simplified candidates as the starting point could further boost the multi-round simplification performance¹.

1 Introduction

Complex text makes it difficult for language learners and people with limited literacy to read. Text simplification improves learning, accessibility, and information sharing with a wider audience. With the advent of deep learning and large language models (LLMs), simplification performance has improved significantly, supported by the release of important datasets (Imperial et al., 2025). Modern approaches have explored zero-shot prompting (Chi et al., 2023; Barayan et al., 2025; Farajidizaji et al., 2024), instruction tuning (Imperial and Taylor Madabushi, 2023), and related strategies.

From our baseline analysis of trial data, we observed that a larger gap between the CEFR level of the original sentence and the target level (**CEFR-Gap**) substantially increases the likelihood of simplification failure. This finding highlights the importance of addressing complexity not in a single step but through a structured, iterative process.

¹<https://github.com/ounlp/Multi-Round-Text-Simplifier>

Building on this insight, we introduced two novel models generated by GPT-4o for multi-round text simplification: MRS-Rule, a rule-based framework that progressively adjusts sentence structures and vocabulary, and MRS-Joint, which integrates rules with prompting techniques to leverage the strengths of both symbolic and generative approaches.

The primary contribution of this work is to show that multi-round small rule-based simplification are more effective at handling large CEFR gaps than conventional single-step approaches. Our proposed method MRS-Joint outperforms the MRS-Rule and baseline models, as validated through extensive experiments and qualitative analyses. Additionally, we explored the potential of automatic code generation for text simplification, although further refinement remains necessary.

2 Task Setup

The goal of the shared task is to simplify a given source text into a target text with a desired CEFR proficiency level ($A1 < A2 < B1 < B2 < C1 < C2$). For the datasets, we use the same trial (40 examples) and test (200 examples) datasets provided by the TSAR workshop to build and evaluate our methods. For the evaluation metrics, we follow the same metrics from the official TSAR-2025 shared-task metrics, which covers both readability-level control (CEFR Compliance, we focus on **RMSE**, the distance between predicted and target CEFR levels, the lower the better) and meaning preservation (via evaluating semantic fidelity between the simplified sentence and the original sentence, or the simplified sentence and a human-written reference via **MeaningBERT** (Beauchemin et al., 2023), denoted as **MB-Orig** and **MB-Ref** respectively²).

²Please refer to the shared task paper (Alva-Manchego et al., 2025) for more details of other metrics such as BERTScore (Zhang et al., 2019) etc.

3 Motivation for Multi-Round

In this section, we present our baseline model, the Naïve Prompt model, and show that simplification becomes increasingly challenging as the gap between the source and target levels widens.

3.1 Baseline: Naïve Prompt-based (Run 1)

We use GPT-4o to generate the code first (denoted as Baseline or "Program 1"), which will call the OpenAI APIs (GPT-4o-mini) with the following prompt from (Barayan et al., 2025). This generated our Run-1 submission on the test data. Please refer to Appendix A.1 for more details.

Baseline Prompt
Please simplify the following Complex Sentence to make it easier to read and understand by {CEFR-LEVEL} CEFR level English learners. {CEFR-LEVEL} level English learner {CEFR-Description}. To simplify, you may replace difficult words with simpler ones, elaborate, or remove them when possible. You may also break down a lengthy sentence into shorter, clear sentences. Ensure the revised sentence is grammatically correct, fluent, and maintains the core message of the original without changing its meaning. Complex Sentence: {Source} Simplified Sentence:

CEFR Level Prediction Since the trial data only gives the CEFR level for target text, not for the source text and the simplified texts, we estimate a text proficiency level using three ModernBERT classifiers with the voting mechanism³. Each model independently predicts a CEFR label (A1-C2) with a confidence score. We combine predictions via majority voting: the label with the most votes is selected. Ties are broken by the largest sum of confidences, then by the highest single-model confidence; if still tied, we prefer the simpler (lower) level to remain conservative. The resulting CEFR level also determines whether a simplification is still needed for a text.

CEFR-Gap We assign an integral value from 0 to 5 for each CEFR level according to the order of (A1<A2<B1<B2<C1<C2). The CEFR Gap for each example is defined as the numerical difference between source level and the target level (e.g., the CEFR-Gap between C1 and A2 is 4-1=4). We run the generated program on the 40 trial examples in the trial data, and then we study the performance of baseline models for each cluster of examples with the same CEFR Gap as Table 1. We found that **RMSE** rises from 0.624 with a one-level gap, to 1.027 with two levels, and then further go to 1.581 with three levels, indicating

³AbdullahBarayan/ModernBERT-base-doc_en-Cefr, ModernBERT-base-doc_sent_en-Cefr, and ModernBERT-base-reference_AllLang2-Cefr2

that larger downward steps are harder to control. Meaning preservation also weakens: **MB-Orig** declines from 0.859 to 0.841 and then to 0.761, while **MB-Ref** falls from 0.832 to 0.758 and stays near 0.762 for the widest gap, though that last figure is based on only four samples. These patterns reveal a trade-off: stronger simplification with larger **CEFR-Gap** makes it more difficult to match the target level and to keep the original meaning intact. In short, bigger CEFR gaps demand more radical linguistic changes, which inevitably reduce both level accuracy and semantic fidelity.

CEFR-Gap	RMSE	MB-Orig	MB-Ref
1 (18)	0.624	0.859	0.832
2 (18)	1.027	0.841	0.758
3 (4)	1.581	0.761	0.762

Table 1: CEFR-Gap Analysis on CEFR accuracy (RMSE) and meaning preservation. The bracket shows the total number of examples we found in the trial data with that CEFR gap. It shows the larger the gap, the higher the RMSE, the lower the other MB scores.

4 Proposed Multi-Round Methods

Based on the findings in §3, smaller gap between the source and the target CEFR level will be relatively easy to simplify. Hence, we propose to simplify texts with multiple rounds by taking previous simplification results as inputs with two multi-round methods: rule-based simplification (MRS-Rule §4.1) and jointly rule-based and LLM Prompting (MRS-Joint §4.2). For each program, we first demonstrate the prompts and operations to generate and fix, and then briefly analyze the detailed workflow of the generated program. The orange box shows the operations and prompts we used to generate the MRS-Rule Code, while the blue box at the bottom shows the further steps we used to fix the generated code to make it work.

4.1 MRS-Rule: Rule-based (Run 2)

The generated code (see more details in Appendix §A.2) for MRS-Rule does not call any large language model API for simplification, but only rule-based rewriting combined with automatic CEFR level verification and semantic checks.

Prompts for Generating MRS-Rule Code

<Operations:> Upload the Program 1 (Baseline) file into GPT-4o; Upload 2 images (one image is the three models and the other is the method to predict the level in the evaluation).

Prompt 2.1 I want you to write a program for me that simplifies the original sentence. In the program, the first step is to simplify the original sentence. The next step is to identify the CEFR_Level by using three models and the method in the image. If the generated Cefr level does not match the target level, it will call the simplify sentence method to simplify that simplified sentence until it reaches the target level. If the generated level matches the target level, it will be written down in the output file. Write down the program based on the file (program used only naive prompt) and 2 images (1 is three model and the other is the method to predict the level) I give you.

Prompt 2.2 The program keeps the original meaning by checking semantic similarity (SBERT cosine similarity) at every step and only accepts a simplification if: (1) the CEFR level hits the target, and (2) similarity to the original is above a threshold you control (default ≥ 0.8).

Prompt 2.3 I want you to update this code by after the summary step. I want the program to simplify the remaining original sentence near to the target cefr_level. After that, would you mind arranging all the JSON objects in the output file according to dataset_id alphabetically (for instance, 01-b1 comes after 01-a1, 02-a2 comes after 01-b1).

Instruction used to fix the code

<Operations:> Upload the Program 1 (Baseline) file into GPT-4o

Fix 2.1 Fix the program so that the output file contain only the text_id column and simplified_sentence

Fix 2.2 Because in the output file, there are still some JSON objects missing. So I ask GPT-4o: Can you try to update the code above so that it can simplify the original sentence of each JSON object to the target level?

Fix 2.3 At the end of the program, would you mind adding some code that checks the number of JSON objects in the input file with the number of the JSON objects in the output file. If they are equal, you don't need to check. If not, you need to check what the dataset_id is missing and then simplify that original sentence belonging to that dataset_id until all the dataset_ids are in the output file?

4.1.1 Code Generation

Prompts 2.1, 2.2 and 2.3 are the three main prompts that we used to generate the code for MRS-rule method step by step. When using Prompt 2.2 to instruct GPT-4o for further simplification by jointly checking CEFR level and semantic similarity, it suggests the following rules and used in a sophisticated candidate generation pipeline (§A.2.1).

- `replace_words`: substitute complex words with simpler synonyms (e.g., “utilize” → “use”, “approximately” → “about”).
- `simplify_numbers_units`: standardize numerical expressions and units (e.g., remove separators, normalize “metres/meters”).
- `strip_relative_clauses`: remove non-essential subordinate clauses (e.g., clauses beginning with *which/that/who/where/when* or discourse markers like *however/although*) to reduce syntactic complexity.
- `keep_shortest_clause`: select the simplest clause from a multi-clause sentence by choosing the shortest well-formed segment.
- `trim_to_limit`: shorten the text to a step-dependent word budget while preserving a grammatical ending.

- `sentence_split`: break long sentences into shorter, more readable parts at punctuation boundaries, then simplify each part.

More importantly, it also smartly suggested to sacrifice semantic preservation for higher CEFR-level accuracy, which demonstrates improved performance over prompting baseline (Table 2).

4.1.2 Workflow

Figure 1 shows the workflow of MRS-Rule, which includes an iterative retries with dynamic conditions such as the similarity floor, maximum editing steps to reach the best-effort CEFR-levels.

Reconciliation Retries In each retry, the system first generates multiple candidate sentences from the original text. It then selects the best candidate using cosine similarity and predicted CEFR level. This candidate becomes the seed for the next round, based on the assumption that easier sentences can be further simplified toward the target CEFR level. Candidates are created using one or more rules (details in §A.2.1). After each round, all candidates are scored for meaning preservation (cosine similarity) and difficulty (CEFR level). The best-scoring candidate is carried forward as the seed for the next round. If it still doesn't meet the target level, additional rule-based refinements are applied (§4.1.1). Subsequent retries follow the same process but use more relaxed thresholds. The CEFR level is validated by majority voting from three ModernBERT classifiers. Sentences that remain unsimplified go through further retries with gradually looser similarity thresholds and larger edit budgets. Finally, the system picks the candidate closest to the target level, reorders text IDs, and outputs the results. If any sentences are still not simplified, the system slightly lowers the similarity threshold (to 0.88), increases the maximum edit steps (to 8), and reprocesses only the remaining sentences—up to six rounds. All the hyperparameters for our program is summarized in Appendix Table 4.

Nearest-level Fill The simplification will keep running for multiple rounds of the above simplification rules until all the sentences are simplified to the target level or a retry cap is reached. For sentences that failed to be simplified to the target level, we will use the **nearest-level fill**, selecting the candidate whose predicted CEFR level is closest to the target while keeping the original meaning, before reorganizing and saving the final output.

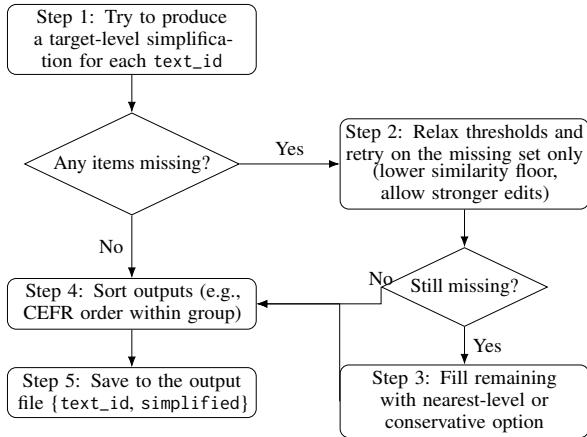


Figure 1: Workflow of MRS-Rule (Run 2)

4.2 MRS-Joint: Rule-based + Prompting

Building upon the Baseline model (§3), we combine LLM prompts (Barayan et al., 2025) and multi-round rule-based simplifications with automatic verification steps (§4.1). As shown in the workflow (§2), the LLM generates simplified sentences only in the first step. After each retry, the system selects the best candidate based on cosine similarity and predicted CEFR level. This loop continues until the predicted CEFR level matches the target level. In each new retry, the system lowers the cosine similarity threshold (allowing more meaning change) and increases the maximum number of simplification steps. This process ensures the final sentence fits the target proficiency level while preserving the original meaning.

4.2.1 Code Generation

For MRS-Joint program (§A.3), we use the Prompt 3.1 to integrate the LLM prompting from Baseline (Program 1) into the MRS-Rule (Program 2) by uploading the two program files first then prompting. Then we use the Prompt 3.2 to generate the code for over-generation-then-rank. The program, generated when we combine those two files, worked well so there was nothing to fix.

Prompt for Generating MRS-Joint

<Operations:> Upload the Program 2(MRS-Rule) and the Program 1(Baseline) to ChatGPT.

Prompt 3.1 Update this file(the file contains the program 2). Before simplifying the sentence, the program uses the naive prompt to generate one candidate. Other candidates will be generated based on the built-in rules.

Prompt 3.2 After generating many candidates, the program selects the best candidate based on the cosine similarity and predicted level. If that best candidate does not meet the target level, the program continue generates more candidates based on that best candidate.

4.2.2 Workflow

Figure 2 illustrates the generated program MRS-Joint by combining LLM prompting (§3.1) and

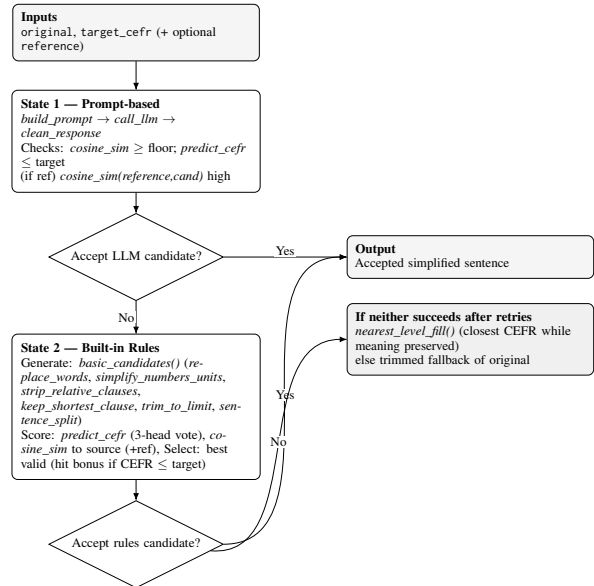


Figure 2: Workflow of MRS-Joint

multi-round rule-based simplification (§4.1). The generated program simply prompts the LLM at the first round, and then all subsequent round are purely rule-based as described in §4.1.2.

5 Result

Model	RMSE	MB-Orig	MB-Ref
Trial			
Baseline (Run 1)	0.8944	0.8453	0.7958
MRS-Rule (Run 2)	0.8515	0.7961	0.7967
MRS-Joint	0.4472	0.8023	0.7574
Test			
Baseline (Run 1)	0.755	0.855	0.849
MRS-Rule (Run 2)	0.714	0.865	0.701
MRS-Joint	0.552	0.866	0.837

Table 2: CEFR accuracy (RMSE) and meaning preservation on trial and test datasets.

Table 2 summarize the performance of our models on both trial and test datasets. Baseline (§3) and MRS-Rule (§4.1) are the two models corresponding to the two runs of our submission in the final evaluation period. After the evaluation, we found that simply merging two methods into MRS-Joint (§4.2) is more efficient, which is the most accurate model to match the target CEFR level (the best RMSE) while still maintaining the meaning. Prompt-only baseline model (§3) preserves the original meaning best (highest MeaningBERT-Orig) but shows the weakest control of CEFR level (highest RMSE). Furthermore, comparing MRS-Joint with the Baseline, the difference mainly exists in the multi-round rule-based simplification.

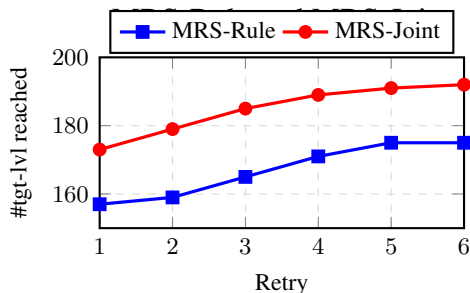


Figure 3: # Simplified sentences that reach the target level across retries for MRS-Rule vs. MRS-Joint.

Target ↓ / Pred →	A2	B1	B2
A2	66	32	2
B1	20	79	1
B2	0	0	0

Table 3: Confusion matrix on the test data

It shows that our multi-round rules significantly improve the performance with a little sacrifice on the meaning preservation. Figure 3 further shows that single round simplification performs poorly, while multi-round simplification could increasingly simplify more sentences to the target CEFR level. Furthermore, MRS-Joint, starting the simplification from LLM-simplified candidates, could boost the performance of multi-round simplification.

6 Qualitative Analysis

6.1 Overall Findings

As shown in Table 3, the program excels in simplifying complex C1–C2 sentences down to the B1 level. Out of 100 source sentences, 79 were successfully simplified to B1, with only 20 dropping further to A2 and 1 rising to B2. For sentences targeted at the A2 level, the results were mixed: only 66 reached the intended A2 level, while 32 overshot to B1 and 2 even remained at B2.

Therefore, we recognize that simplifying high-complexity C1–C2 input to lower CEFR levels is inherently more challenging. The program is more prone to “overshooting,” producing text that remains more complex than the intended target. In other words, the lower the target CEFR level, the higher the likelihood of program’s not meeting the constraints of that level.

6.2 Case Study

To understand the model’s behaviour beyond overall accuracy scores, we conducted a **qualitative error analysis** on three representative C1–C2 source texts that were simplified by the system but misclassified by the CEFR predictor. These examples illustrate three distinct types of misclassification.

Case 1 – Overshoot (A2 → B1) (§B.1) The model simplified vocabulary and shortened clauses but kept abstract ideas along with a relative clause typical of B1 syntax. The CEFR predictor therefore rated the output B1, which is one level higher than target level, showing that preserving key ideas may force more complex structures than intended level.

Case 2 - Lexical Imitation (§B.2) Although shortened from the source text, the output kept formal phrases like “a large number of bridge accidents... of the bridge itself,” instead of simpler A2 wording such as “Many accidents happen while bridges are being built.” The CEFR model therefore rated it B1, showing that better simplification needs lexical adaptation, not just shorter text.

Case 3 – Under-generation (B1 → A2) (§B.3) The system produced only a fragment, dropping the telescope’s purpose and the planetary-defense discussion. With much of the conceptual content missing, the predictor judged the text A2 despite technical terms. This highlights that incomplete outputs can seem easier to cheat the CEFR predictor as the intended CEFR level.

These examples reveal three failure modes – overshoot, and undergeneration – demonstrating that successful CEFR simplification requires not only simpler words but also balanced control of meaning, style, and completeness.

7 Conclusion

We found that a larger gap between the CEFR level of the original and target sentences (CEFR-Gap) increases the likelihood of simplification failure. Based on this finding, we proposed two multi-round simplification methods generated by GPT-4o: MRS-Rule, which applies rule-based simplification, and MRS-Joint, which combines rules with prompting. Extensive experiments and case studies show that MRS-Joint outperforms both the prompting baseline and MRS-Rule, confirming the effectiveness of multi-round simplification and the feasibility of text simplifier via code generation.

Limitation

We note a few limitations of our work. The models we used are closed-source models such as using GPT-4o for code generation while using GPT-4o-mini for API, which are not explicitly finetuned on the text simplification datasets by us. Our work is also limited to one dataset and one language (English), and two types of GPT-4o generated model. Furthermore, focusing on coding generation, we could also extend the study to self-evolve algorithm discovery (Novikov et al., 2025), and compare it with other prompts and more coding agents. Besides those, we believe explicitly involving curriculum-based domain knowledge in a structured multi-round simplification will be promising methods in the era of artificial intelligence.

Lay Summary

This project aims to make complex English sentences easier to understand, especially for language learners. Our team participated in TSAR 2025 competition where the goal was to rewrite sentences to match specific English proficiency levels, like beginner (A1) or intermediate (B1), based on the **Common European Framework of Reference (CEFR)**. Our team’s insight was that the bigger the difference between a sentence’s original difficulty and the target level (called the “CEFR Gap”), the harder it is to simplify the sentence successfully. For example, turning a very advanced sentence (C1) into a basic one (A2) is much more difficult than making small adjustments. This inspired us to develop a **multi-step approach** for the simplification.

Our team created two systems, and the code are generated with AI with our instructions:

MRS-Rule: Uses rules to gradually simplify text in multiple rounds (e.g., replace difficult words, break long sentences).

MRS-Joint: Combines a model (GPT-4o-mini) to generate an initial simplified text, and then refines it through multiple rule-based steps.

Both systems repeatedly check whether the new sentence meets the desired CEFR level and still retains the original meaning. If not, they retry those sentences with adjustments. This multi-round process continues until the system either succeeds or picks the closest acceptable version.

In testing, the MRS-Joint method performs best. It reaches the target reading level more often than the baseline approach, though sometimes at the

cost of slightly reducing the original meaning. Still, it shows strong overall results: it handles complex sentences better and produced more accurate simplifications. Our team also analyzed mistakes. Sometimes the program “oversimplified” or retained too many complex words. Other times, it shortened the sentence too much and left out important information. These findings will help improve future systems.

In short, this work shows that a multi-step process can make content more accessible to learners while maintaining its original intent.

References

- Fernando Alva-Manchego, Regina Stodden, Joseph Marvin Imperial, Abdullah Barayan, Kai North, and Harish Tayyar Madabushi. 2025. Findings of the TSAR 2025 shared task on readability-controlled text simplification. In *Proceedings of the Fourth Workshop on Text Simplification, Accessibility, and Readability (TSAR 2025)*, Suzhou, China. Association for Computational Linguistics.
- Yuki Arase, Satoru Uchida, and Tomoyuki Kajiwara. 2022. Cefr-based sentence difficulty annotation and assessment. *arXiv preprint arXiv:2210.11766*.
- Abdullah Barayan, Jose Camacho-Collados, and Fernando Alva-Manchego. 2025. [Analysing zero-shot readability-controlled sentence simplification](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6762–6781, Abu Dhabi, UAE. Association for Computational Linguistics.
- David Beauchemin, Horacio Saggion, and Richard Khoury. 2023. Meaningbert: assessing meaning preservation between sentences. *Frontiers in Artificial Intelligence*, 6:1223924.
- Alison Chi, Li-Kuang Chen, Yi-Chen Chang, Shu-Hui Lee, and Jason S. Chang. 2023. [Learning to paraphrase sentences to different complexity levels](#). *Transactions of the Association for Computational Linguistics*, 11:1332–1354.
- Asma Farajidizaji, Vatsal Raina, and Mark Gales. 2024. [Is it possible to modify text to a target readability level? an initial investigation using zero-shot large language models](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9325–9339, Torino, Italia. ELRA and ICCL.
- Joseph Marvin Imperial, Abdullah Barayan, Regina Stodden, Rodrigo Wilkens, Ricardo Munoz Sanchez, Lingyun Gao, Melissa Torgbi, Dawn Knight, Gail Forey, Reka R Jablonkai, and 1 others. 2025. *Universalcefr: Enabling open multilingual research on*

language proficiency assessment. *arXiv preprint arXiv:2506.01419*.

Joseph Marvin Imperial and Harish Tayyar Madabushi. 2023. [Flesch or fumble? evaluating readability standard alignment of instruction-tuned language models](#). In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 205–223, Singapore. Association for Computational Linguistics.

Alexander Novikov, Ngàn Vũ, Marvin Eisenberger, Emilien Dupont, Po-Sen Huang, Adam Zsolt Wagner, Sergey Shirobokov, Borislav Kozlovskii, Francisco JR Ruiz, Abbas Mehrabian, and 1 others. 2025. [Alphaevolve: A coding agent for scientific and algorithmic discovery](#). *arXiv preprint arXiv:2506.13131*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. [Bertscore: Evaluating text generation with bert](#). *arXiv preprint arXiv:1904.09675*.

A Details for 3 Generated Program

In general, all three program are generated with GPT-4o model, which covers the following python libraries and models.

Libraries In the AI generated code of MRS-Rule and MRS-Joint, the following Python libraries are used: argparse, os, sys, json, re, math, pathlib, collections, typing, numpy, requests. Besides those regular pythong libraries, we noticed that libraries like transformers (Hugging Face), SentenceTransformers, NumPy are used for natural language processing and machine learning parts.

Models CEFR level is predicted by three ModernBERT from huggingface, ModernBERT-base-doc_en-Cefr, ModernBERT-base-doc_sent_en-Cefr, ModernBERT-base-reference_AllLang2-Cefr; while semantic similarity is using the library of sentence-transformers/all-MiniLM-L6-v2), and LLM for generating the code is GPT-4o⁴. The LLM API used for text simplification is gpt-4o-mini.

A.1 Program 1: Baseline Naïve Prompt

The generated Program 1 is in the file of "First_Version_Sentence_Simplification.py" in the code repo. It is built with **OpenAI's Chat Completions API**. The script is lightweight and designed for **large-scale, reproducible simplification** runs,

⁴<https://chatgpt.com/?model=gpt-4o>, accessible at 09/23/2025

while maintaining a clean JSONL output compatible with downstream CEFR or readability evaluations. This baseline Program 1 is used as Run 1 in our submission, and it is also used in our **CEFR-Gap** analysis.

A.2 Program 2: MRS-Rule

The generated Program 2 is in the file of "Second_Version_Sentence_Simplification.py". Specifically, ChatGPT suggests useful rules to generate cadidate simplications with function of basic_candidates(). It generates multiple simplified variants of an input sentence using lightweight, rule-based transformations-*without* relying on an LLM. The details of the code is shown in the code listing 1. The corresponding hyperparameters used in the code are summarized in the Table 4.

Parameter	Value
similarity_floor	0.88
max_steps	8
max_retries	6
floor_step	0.03
steps_step	6
sim_floor (internal)	0.88 ↓
w_hit	10
w_ref	2.5
w_orig	0.5
llm_timeout	60 seconds
use_llm	true
sim_threshold	0.72–0.75

Table 4: Hyperparameters in MRS-Rule and MRS-Joint

A.2.1 Generated Code to Apply Rules

Listing 1: Rule-based Simplification to Generate Candidates

```
base = text.strip()
lim = max(8, 28 - 2*step_idx)
cands = [
    replace_words(base),
    simplify_numbers_units(base),
    keep_shortest_clause(base),
    strip_relative_clauses(base),
    trim_to_limit(replace_words(base),
                  lim),
    trim_to_limit(simplify_numbers_units
                  (base), lim),
    trim_to_limit(strip_relative_clauses
                  (base), lim),
    trim_to_limit(keep_shortest_clause(
                  replace_words(base)), lim),
]
for s in sentence_split(
    strip_relative_clauses(base)):
    cands.append(trim_to_limit(
        replace_words(s), max(10, lim-4)
    ))
```


A.3 Program 3: MRS-Joint

Program 3 is in the file "Third_Version_Sentence_Simplification.py".

It simply combines the previous rule-based program A.2 with LLM prompting-based program in A.1. This program shares the same rules used in A.2.1, and uses the same set of multi-round hyper parameters as Table 4.

B Case Studies

To understand the model’s behaviour beyond overall accuracy scores, we conducted a **qualitative error analysis** on three representative C1–C2 source texts that were simplified by the system but misclassified by the CEFR predictor. These examples illustrate three distinct types of misclassification.

B.1 Case 1 – Overshoot: A2 Target → Predicted B1

Source (C1–C2)

I’m a member of the Campaign to Stop Killer Robots ... We need to be very cautious indeed about what we ask machines to do.

System Output (intended A2)

I am part of the Campaign to Stop Killer Robots. Forget the scary robots from movies. Right now, military machines like drones and robots are already used to kill with little help from people. The next step is robots that can decide who to kill by themselves. I believe this is very wrong and dangerous for people. We need to be very careful about what we let machines do.

Analysis

The model successfully shortened the multi-clause source and replaced high-level terms (*autonomous ‘murderbots’, completely unethical*) with simpler wording. However, it retained **abstract concepts** (*robots that can decide who to kill*) and an **embedded relative clause**, both typical of B1 syntax. The CEFR predictor therefore judged the output **one level higher than the A2 target**.

Insight: Even when sentences are lexically simplified, preserving key ideas may require structures that naturally push the text above the intended level.

B.2 Case 2 – Lexical Imitation: A2 Target → Predicted B1

Source (C1–C2)

A large number of bridge accidents occur during the construction of the bridge itself... this can be deadly for the workers on it at the time.

System Output (intended A2)

A large number of bridge accidents occur during the construction of the bridge itself. These accidents are often due to.

Analysis

Although truncated, the output **copied formal phrases from the source** (*a large number of bridge accidents occur, during the construction of the bridge itself*). It failed to rephrase these into simpler, high-frequency A2 forms such as *Many accidents with bridges happen while they are being built*. The CEFR model therefore still rated it **B1**, despite the missing ending.

Insight: True simplification requires **lexical adaptation**, not only shortening. Retaining formal academic expressions—even in a shorter text—can maintain a higher perceived level.

B.3 Case 3 – Under-generation: B1 Target → Predicted A2

Source (C1–C2)

Whether NASA can find the remaining middle-sized NEOs depends on getting the money to build NEO-Cam ... the PHAs.

System Output (intended B1)

NASA’s ability to find the remaining middle-sized near-Earth objects (NEOs) depends on getting funding to build NEOCam, a 0.5-meter space telescope that.

Analysis

The system produced only a **partial sentence**, omitting the telescope’s function and the entire discussion of planetary defence. With the **conceptual load drastically reduced**, the CEFR predictor assigned an **A2 level**, even though the fragment still contains technical terms (*NEOs, NEOCam*).

Insight: Incomplete outputs can appear easier than intended, causing the CEFR assessment to **underestimate** the level. Quality checks for completeness are essential alongside automatic scoring.

HIT-YOU at TSAR 2025 Shared Task: Leveraging Similarity-Based Few-Shot Prompting, Round-Trip Translation, and Self-Refinement for Readability-Controlled Text Simplification

Mao Shimada* and Kexin Bian* and Zhidong Ling and Mamoru Komachi

Hitotsubashi University

{5123024a@g, dm240020@g, dd250009@g, mamoru.komachi@r}.hit-u.ac.jp

Abstract

We describe our submissions to the TSAR 2025 shared task on readability-controlled text simplification, which evaluates systems on their ability to adjust linguistic complexity to specified CEFR levels while preserving meaning and coherence. We explore two complementary frameworks that both build on LLMs and incorporate feedback from the shared-task CEFR classifiers. The first is an ensemble approach, which uses multiple LLMs to generate diverse candidates through zero-shot prompting, similarity-based few-shot prompting, and round-trip translation. The generated candidates are filtered by predicted CEFR level, and the final output is selected by an LLM judge. The second is a self-refinement loop, which uses a single LLM that begins with one candidate and iteratively revises it based on classifier feedback until it meets the target level or reaches a maximum iteration limit. Both approaches achieved competitive performance in the shared task. To our knowledge, this is among the first studies to apply round-trip translation and iterative self-refinement to controlled simplification, expanding the toolkit for reliable readability control.

1 Introduction

Text simplification aims to reduce the complexity of text while preserving meaning, thereby improving accessibility for language learners and readers with limited proficiency. The **TSAR 2025 Shared Task** (Alva-Manchego et al., 2025) focuses on *readability-controlled simplification*, where English passages at CEFR level B2 or above must be rewritten to a specified target level (A1, A2, or B1). Systems are evaluated on CEFR compliance, semantic similarity to the original text and references, with the challenge that no parallel training data is provided.

*These authors contributed equally to this work.

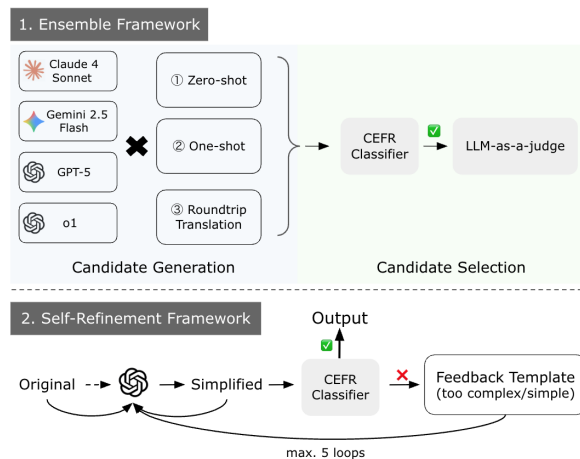


Figure 1: Illustration of the two frameworks used in our system. (1) The **Ensemble Framework** aggregates diverse model–prompt candidates through classifier filtering and LLM-based selection, while (2) the **Self-Refinement Framework** iteratively adjusts a single model’s output using classifier feedback.

Over the past years, research has evolved from traditional rule-based methods (Shardlow, 2014) to neural sequence-to-sequence and pretrained Transformer models (Sheang and Saggion, 2021; Li et al., 2024). More recently, large language models (LLMs) with controllable and instruction-based generation have become the dominant trend in text simplification. However, ensuring precise readability control remains a core challenge (Barayan et al., 2025; Tran et al., 2025).

Previous research has explored generation mechanisms that encourage structural and lexical reformulation that might facilitate controllable simplification. Cross-lingual and translation-based pipelines, for example, have been shown to induce diverse edits such as synonym substitutions and word reorderings (Stahlberg et al., 2022; Vlanitis et al., 2024), effects that can be leveraged to achieve finer readability control. In another line of work, iterative refinement frameworks such as

SELF-REFINE (Madaan et al., 2023) improve LLM outputs through self-feedback guided by external evaluators, offering a general way of enforcing generation constraints such as readability control.

Building on these insights, we explore two frameworks for readability-controlled simplification, both guided by the shared-task CEFR classifier (Figure 1).

- The **ensemble framework** integrates multiple prompting strategies to generate diverse candidates, including a novel similarity-based few-shot prompting and round-trip translation. Candidates are filtered by the shared-task CEFR classifier, and the final output is selected by an LLM judge. We submitted two ensemble runs that differ in the choice of judge model.
- The **self-refinement framework** employs iterative simplification, where a single candidate is repeatedly revised under classifier feedback until the target CEFR level is reached.

2 System Description

2.1 Ensemble Models

Our first two submissions use a multi-stage ensemble pipeline to produce a single simplification from a diverse pool of candidates. The pipeline is designed to combine the strengths of different LLMs and prompting strategies. It proceeds in two stages: (i) candidate generation, and (ii) candidate selection through filtering and final judgment.

2.1.1 Candidate Generation

To encourage diversity, we generate candidate simplifications using four proprietary LLMs: GPT-5, Gemini 2.5 Flash, Claude Sonnet 4, and o1.¹ Each model is prompted under three strategies described below (see Appendix A for full prompt texts). For each model-prompt configuration, we draw four samples at temperature 1, yielding a total of $4 \times 3 \times 4 = 48$ candidates for each test input.

Zero-shot We designed zero-shot prompts through observation of the trial data and iterative prompt engineering, adapting instructions to each target CEFR level. All prompts specified both the source and target CEFR levels. For A2, the prompt

additionally required the model to simplify the paragraph into a direct subject-verb-object structure. For B1, we first annotated each word in the original paragraph with CEFR levels using CEFRpy², and then instructed the model to only replace words above B2 with simpler alternatives by including the corresponding wordlist in the prompt. We further constrained the output by limiting each sentence to a maximum of 29 words.

Similarity-based Few-shot We extended the zero-shot prompts by inserting in-context demonstrations selected from the trial data based on source similarity.³ For each test input, cosine similarity was computed between its sentence embedding and those of 20 trial sources using Alibaba-NLP/gte-large-en-v1.5 (Zhang et al., 2024; Li et al., 2023). The k most similar sources and their paired references at the target level were then added to the prompt as demonstrations, so a k -shot setup corresponds to including the top- k most similar source-reference sets. In this work, we set $k = 1$, yielding a one-shot setup.

Round-trip We employ a round-trip translation approach, using different intermediate languages depending on the target CEFR level. For A2, the original English paragraph is first translated into Chinese while being simplified to the target level. For B1, it is translated into German without explicit simplification at this stage. In both cases, only the translated paragraph is then used as input to translate it back into English, during which the model is instructed to simplify it to the target level.

2.1.2 Candidate Selection

After candidate generation, we apply a two-step selection process consisting of a rule-based filter and an LLM judge.

Filtering Candidates are first scored by the official CEFR classifier ensemble (Alva-Manchego et al., 2025), with the final label taken from the classifier with the highest confidence, exactly as provided in the shared task evaluation script. We retain those predicted to match the target level exactly; if none remain, we fall back to candidates that are one level above or below the target. This

¹Model versions: gpt-5-2025-08-07, gemini-2.5-flash, claude-sonnet-4-20250514, and o1-2024-12-17.

²CEFRpy is a Python module built on spaCy that tokenizes text and maps words to CEFR levels based on wordlist resources. Documentation available at <https://maximax67.github.io/cefrpy/>.

³We include a comparison with random sampling in Appendix D

step ensures that the judge compares only plausibly compliant outputs, reducing the risk of selecting a fluent but level-mismatched candidate and streamlining the decision to a more competitive set.

LLM-as-a-judge The surviving candidates are then presented to an LLM judge, which is instructed to select the simplification closest in meaning to the original paragraph. To explore how different LLM architectures perform as judges, we experiment with two variants, corresponding to our first two submissions:

- **Gemini ensemble (run1_gemini_ensemble).** On the trial datasets, Gemini 2.5 Flash demonstrates good performance in both the candidate generation and selection stages, and is therefore incorporated into our set of official submissions.
- **GPT-5 ensemble (run2_gpt_ensemble_4).** GPT-5, used as the judging model, demonstrated reliable and consistent candidate selection on the trial data, performing comparably to other high-performing models while being more efficient to deploy. It is thus adopted as one of our official submissions.

2.2 Self-refinement (run3_self_refine)

In addition to the ensemble pipelines, we submitted a self-refinement system that relies on a single model, GPT-5. Unlike the ensemble approach, this system iteratively adjusts one candidate under feedback guidance. While in principle the feedback signal could be derived from various metrics, we chose to use the CEFR classifier, since readability control is one of the two core evaluation criteria and, in our experience, the aspect more frequently failed by baseline outputs.

Initialization The process begins with a zero-shot simplification produced by the model (as described in § 2.1.1), which serves as the initial candidate for refinement.

Iteration At each step, the CEFR classifier predicts the level of the current candidate. If the predicted level is above the target, the feedback instructs the model to simplify vocabulary and sentence structure; if it is below, it encourages the use of slightly more complex constructions. In both cases, the feedback explicitly requires the model to preserve the original meaning and maintain natural, coherent text. The exact prompt template

and feedback generation rules are provided in Appendix A. This feedback, together with the original document and the candidate text, is then provided to the model, which generates a revised simplification. The loop continues for up to five iterations.

Stopping criterion The loop terminates early if the classifier predicts that the target level has been reached. Otherwise, the final candidate after the fifth iteration is returned. We adopt this policy based on the assumption that the last iteration represents the most refined version.

3 Results

Table 1 shows the performance of our three submitted runs. We report results on the official TSAR-2025 shared task metrics: **RMSE** for CEFR compliance, and MeaningBERT (Beauchemin et al., 2023) for semantic similarity to the original (**mBERT-Orig**) and reference (**mBERT-Ref**) texts, respectively, as defined by the organizers (Alva-Manchego et al., 2025).

When considering only the best run per team, our team **ranks third out of 20 teams** overall. Our GPT-5 and Gemini ensemble runs achieve AUTORANK⁴ values of 3.61 and 3.67, respectively, among 48 submitted runs, demonstrating competitive performance across all official metrics. The Self-Refinement run ranks at AUTORANK = 5.46, showing slightly weaker CEFR control and meaning preservation, but operates with much lower computational cost, relying on a single model rather than multi-model ensembling and repeated generations.

4 Discussion

Effectiveness of Ensemble We demonstrate the contribution of the ensemble process by comparing its performance against the strongest single-model baselines, shown in the upper part of Table 1, which represent an empirical upper bound of individual model overall performance across the prompting configurations. The ensemble outperforms these baselines across all dimensions. Although the substantial improvement in level control is partly enforced by design through classifier-based filtering, these results indicate that incorporating diverse model-prompt pairs captures complementary strengths beyond any individual configuration.

⁴Linearly rescaled rank where 1 denotes the best performing system; see (Alva-Manchego et al., 2025) for details.

System	RMSE ↓	mBERT-Orig ↑	mBERT-Ref ↑
Claude (0-shot)	0.595	0.836	0.820
GPT-5 (0-shot)	0.620	0.848	0.821
Gemini (0-shot)	0.497	0.812	0.811
o1 (RT)	0.630	0.828	0.835
Gemini Ensemble	0.187	0.863	0.833
GPT-5 Ensemble	0.158	0.852	0.835
Self-Refinement	0.245	0.822	0.820

Table 1: Comparison of best single-model baselines (top) and submitted runs (bottom) on the official TSAR-2025 metrics. For each model, best configurations were determined via a weighted composite of normalized scores following median–interpercentile scaling (Alva-Manchego et al., 2025), averaged across four generations. Values in parentheses indicate the prompting configuration in which the best overall score was achieved.

While we did not perform full ablation studies, we examined the distribution of final candidates chosen by the LLM judges to assess the necessity of using multiple models and prompt variants. We find that the selections were distributed across all four models and three prompting strategies (Appendix B), indicating that the ensemble benefited from the diversity of system outputs. Overall, the gains from combining multiple model–prompt pairs suggest that different systems may excel on different types of inputs or aspects of simplification, leading to complementary effects when aggregated, a possibility we plan to explore in future work.

Effectiveness of Self-refinement Compared to its zero-shot GPT-5 counterpart (Table 1), the self-refinement process substantially reduces readability-classification error through its classifier-guided feedback loop, while maintaining comparable similarity to human-written references. This demonstrates that automatic feedback can effectively steer outputs toward the intended readability level without a major loss in fluency and coherence.

Self-refinement successfully corrected most outputs that initially missed the target level. Out of 200 instances, 129 (64.5%) met the target immediately, but a further 54 (27.0%) converged only through iterative refinement, confirming the value of the approach in recovering difficult cases. However, the distribution of refinement depth shows diminishing returns: 31 (15.5%), 14 (7.0%), 7 (3.5%), and 2 (1.0%) cases requiring one to four iterations, suggesting that classifier feedback yields most of its benefit in early steps. However, a small fraction of cases (8.5%) never converged, underscoring that the current feedback signal alone is not sufficient for all inputs.

Another limitation is reflected in the trade-off between readability control and meaning preservation (mBERT-Orig), which decreased from 0.848 to 0.822 across all instances. For the subset that underwent refinement, the drop was more pronounced (0.85 → 0.79), suggesting that repeated simplification can impair content preservation when guided solely by readability feedback. Future extensions could incorporate additional signals to balance readability, meaning, and naturalness more holistically during refinement.

Model behavior across prompting strategies

To better understand how prompting design influences readability control and meaning preservation, we analyze model behavior across the three setups within the ensemble framework (zero-shot, one-shot, and round-trip). Detailed results are provided in Appendix C. Overall, we find that models varied substantially in their behavior.

In the zero-shot setting, Gemini achieved the lowest RMSE (0.50) but also exhibited the weakest meaning preservation, while o1 showed the opposite pattern, with high fidelity but poor compliance (RMSE = 0.80). Claude and GPT-5 fell in between. We also examined level-specific classification results, which revealed clear level-dependent differences: Gemini and Claude aligned more successfully with A2 than B1, GPT-5 was relatively balanced, and o1 performed better at B1 than A2. These baseline differences are particularly notable given that all models were prompted identically (§2.1.1), indicating that they internalize and act upon level-control instructions in distinct ways.

Adding one-shot demonstrations generally decreased meaning preservation across all models. In terms of reference similarity, o1 benefited the most

from the example, showing consistent gains across both target levels, whereas Gemini and GPT-5 remained relatively stable and Claude experienced a noticeable drop. Effects on compliance were mixed: Claude and o1 showed clear improvement at both levels, GPT-5 improved slightly (mainly due to gains at A2), while Gemini improved at A2 but worsened at B1, increasing its overall error. These results indicate that one-shot examples can enhance level control, particularly at A2. This asymmetry may stem from the nature of the target levels: demonstrations provide clear guidance for simplifying to A2, where shorter sentences and simpler vocabulary reliably signal compliance. By contrast, B1 allows for more diverse realizations, so reliance on a single example can bias the model toward an unrepresentative solution, reducing compliance consistency.

Notably, models with stronger instruction-following or reasoning capabilities, such as o1, appeared to benefit more from demonstrations, as seen in its consistent gains in both reference similarity and compliance across levels. We leave the systematic evaluation of demonstration effects on simplification outcomes and their underlying factors to future work.

The round-trip approach proved generally effective as a way to induce simplification through translation. However, because this setup differs fundamentally in design, we treat its results as descriptive rather than directly comparable to the zero- and one-shot conditions. The choice of intermediate language appears to influence both level control and meaning preservation, likely reflecting differences in linguistic structure and translation bias. We include in Appendix E a brief discussion of intermediate-language choice and the two-step prompting design.

5 Conclusion and Future Work

We presented three systems for the TSAR 2025 shared task on readability-controlled text simplification: two ensemble pipelines that combine diverse generators with an LLM-as-a-judge, and a self-refinement loop guided by classifier feedback. In developing the ensemble system, we explored multiple prompting strategies, including style- and vocabulary-based instruction, similarity-based few-shot prompting, and round-trip translation.

Our analysis highlights systematic differences in how current LLMs interpret and operationalize

level-control instructions, as well as the effects of incorporating demonstrations and classifier-guided feedback. Overall, the results indicate that LLMs under our frameworks form a strong foundation for controllable simplification, but that the trade-off between reliable level control and meaning preservation remains unresolved.

Looking ahead, there are several directions for strengthening our two frameworks under the current evaluation setting. For the ensemble, future work could focus on developing more principled aggregation strategies and analyzing the decision behavior of LLM-as-a-judge models, particularly how they balance readability against meaning when selecting outputs. For the self-refinement framework, richer feedback signals beyond classifier guidance (e.g., semantic similarity, stylistic alignment) could help stabilize convergence and better preserve meaning.

At the same time, our observations during system development hinted that the current automatic evaluation may be brittle. Small stylistic or structural variations can produce large metric shifts, even when readability and meaning remain comparable to human readers. For example, we observed that formatting differences such as line breaks could noticeably affect classifier predictions (Appendix F). Revisiting evaluation design to include human- or reader-centered assessments would therefore provide a more reliable view of simplification quality and practical usefulness.

Limitations

Our systems rely on proprietary LLMs whose training data and update histories are not publicly available, limiting reproducibility and interpretability. While we examined the overall effects of prompting and classifier feedback, our analysis was not exhaustive. Further controlled analyses would be needed to isolate the contribution of each model and prompt variations. We also did not conduct full ablation studies or controlled comparisons across intermediate languages, so some observed trends remain descriptive. Finally, we did not conduct a systematic analysis of the LLM-as-a-judge component, including potential biases, decision consistency, or sensitivity to prompt phrasing. As a result, its contribution to overall system performance is not fully understood.

Acknowledgments

This work was supported by the National Institute of Information and Communications Technology (NICT) under the “Research and Development of externally controllable modeling of multimodal information to enhance the accuracy of automatic translation.”

Lay Summary

This paper describes our systems for the TSAR 2025 shared task, which challenges participants to rewrite English texts at different levels of difficulty (A2–B1) without losing meaning, so that language learners can better understand them.

We built two kinds of systems.

- An **ensemble** system combines outputs from several models using different prompting strategies, such as adding style and vocabulary instructions, providing examples, or translating text through another language (“round-trip”). It then uses another LLM as a “judge” to select the best simplification.
- A **self-refinement** system gives the model feedback on whether its output is too simple or too complex and lets it revise itself until a readability classifier confirms that it matches the target level.

Both of our systems performed competitively in the shared task. The ensemble system performed better than any of its components alone, possibly because different model–prompt combinations behaved quite differently (some were better at controlling difficulty, while others preserved meaning more faithfully). This suggests that combining multiple models and prompts can produce more balanced results, and that different combinations may work better for different kinds of source texts.

The self-refinement system also achieved strong results, producing texts that matched the intended difficulty level with much less computation. This shows that giving models simple feedback can be an efficient way to make their writing easier or harder when needed.

However, our experiments show that while large language models can simplify text in a controlled way, it is still difficult to achieve both precise level control and full preservation of meaning.

References

- Fernando Alva-Manchego, Regina Stodden, Joseph Marvin Imperial, Abdullah Barayan, Kai North, and Harish Tayyar Madabushi. 2025. Findings of the TSAR 2025 shared task on readability-controlled text simplification. In *Proceedings of the Fourth Workshop on Text Simplification, Accessibility, and Readability (TSAR 2025)*, Suzhou, China. Association for Computational Linguistics.
- Abdullah Barayan, Jose Camacho-Collados, and Fernando Alva-Manchego. 2025. [Analysing zero-shot readability-controlled sentence simplification](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6762–6781, Abu Dhabi, UAE. Association for Computational Linguistics.
- David Beauchemin, Horacio Saggion, and Richard Khoury. 2023. [Meaningbert: assessing meaning preservation between sentences](#). *Frontiers in Artificial Intelligence*, Volume 6 - 2023.
- Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.
- Zihao Li, Samuel Belkadi, Nicolo Micheletti, Lifeng Han, Matthew Shardlow, and Goran Nenadic. 2024. [Large language models for biomedical text simplification: Promising but not there yet](#). *Preprint*, arXiv:2408.03871.
- Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegrefe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. [Self-refine: Iterative refinement with self-feedback](#). *Preprint*, arXiv:2303.17651.
- Matthew Shardlow. 2014. [A survey of automated text simplification](#). *International Journal of Advanced Computer Science and Applications(IJACSA), Special Issue on Natural Language Processing 2014*, 4(1).
- Kim Cheng Sheang and Horacio Saggion. 2021. [Controllable sentence simplification with a unified text-to-text transfer transformer](#). In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 341–352, Aberdeen, Scotland, UK. Association for Computational Linguistics.
- Felix Stahlberg, Aashish Kumar, Chris Alberti, and Shankar Kumar. 2022. [Conciseness: An overlooked language task](#). In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 43–56, Abu Dhabi, United Arab Emirates (Virtual). Association for Computational Linguistics.

Hieu Tran, Zonghai Yao, Lingxi Li, and Hong Yu. 2025. [ReadCtrl: Personalizing text generation with readability-controlled instruction learning](#). In *Proceedings of the Fourth Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2025)*, pages 19–36, Albuquerque, New Mexico, US. Association for Computational Linguistics.

Daniel Vlantis, Iva Gornishka, and Shuai Wang. 2024. [Benchmarking the simplification of Dutch municipal text](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2217–2226, Torino, Italia. ELRA and ICCL.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, and 1 others. 2024. [mgte: Generalized long-context text representation and reranking models for multilingual text retrieval](#). *arXiv preprint arXiv:2407.19669*.

A Prompts and Feedback Templates

A.1 Zero-shot prompts

Zero-shot prompt at the A2 level.

Please simplify the following paragraph into a direct SVO structure, transforming it from {ORIGINAL_CEFR} CEFR level to {TARGET_CEFR} CEFR level, making it easier to read and understand for {TARGET_CEFR} CEFR level English learners. Return only the simplified paragraph, without explanation or extra text.
paragraph: {PARAGRAPH}

Zero-shot prompt at the B1 level.

Please simplify the following paragraph, transforming it from {ORIGINAL_CEFR} CEFR level to {TARGET_CEFR} CEFR level, making it easier to read and understand for {TARGET_CEFR} CEFR level English learners.
Only replace difficult words with easier alternatives. Use the provided list of difficult words for replacements.
-Keep the overall sentence structure.
-Don't change meaning and delete any of the given information.
-Ensure that each sentence has no more than 29 words.
Return only the simplified paragraph, without explanation or extra text.
paragraph: {PARAGRAPH}
list of difficult words: {WORD_LIST}

A.2 Few-shot prompts

The placeholder {few_shot_examples} corresponds to the pair of original and simplified paragraph retrieved in the format:

Original: [paragraph retrieved]
Simplified: [corresponding reference paragraph at target CEFR level]

Few-shot prompt at the A2 level.

Please simplify the following paragraph into a direct SVO structure, transforming it from {ORIGINAL_CEFR} CEFR level to {TARGET_CEFR} CEFR level, making it easier to read and understand for {TARGET_CEFR} CEFR level English learners.
Here are some examples of how to simplify text for this level:
{few_shot_examples}.
Return only the simplified paragraph, without explanation or extra text.
paragraph: {PARAGRAPH}

Few-shot prompt at the B1 level.

Please simplify the following paragraph, transforming it from {ORIGINAL_CEFR} CEFR level to {TARGET_CEFR} CEFR level, making it easier to read and understand for {TARGET_CEFR} CEFR level English learners.
Here are some examples of how to simplify text for this level:
{few_shot_examples}.
Only replace difficult words with easier alternatives. Use the provided list of difficult words for replacements.
-Keep the overall sentence structure.
-Don't change meaning and delete any of the given information.
-Ensure that each sentence has no more than 29 words.
Return only the simplified paragraph, without explanation or extra text.
paragraph: {PARAGRAPH}
list of difficult words: {WORD_LIST}

A.3 Round-trip prompts

Round-trip prompt at the A2 level.

- **English → Chinese + simplify**

Please translate the following English paragraph into Chinese and simplify it for {TARGET_CEFR} CEFR level learners of Chinese. Return only the Chinese paragraph, without explanation or extra text.
English paragraph: {PARAGRAPH}

- **Chinese → English + simplify**

Please translate the following Chinese paragraph into English and simplify it for {TARGET_CEFR} CEFR level learners of English. Return only the translated and simplified paragraph, without explanation or extra text.
Chinese paragraph: {PARAGRAPH}

Round-trip prompt at the B1 level.

- **English → German**

Please translate the following paragraph into German. Return only the German paragraph, without explanation or extra text.

English paragraph: {PARAGRAPH}

- **German → English + simplify**

Please translate the following German paragraph into English, making it easier to read and understand by {TARGET_CEFR} CEFR level English learners. Return only the translated and simplified paragraph, without explanation or extra text.

German paragraph: {PARAGRAPH}

A.4 Candidate Selection prompt

YOU MUST FOLLOW OUTPUT RULES EXACTLY. Choose the one closest to the original paragraph. Output MUST be a single digit 1- $\{\text{len}(\text{candidates})\}$ on its own line. No other text.

Original paragraph: {ORIGINAL}

Simplified paragraphs:
{candidates_text}

A.5 Self-Refinement

Refinement prompt.

You are an expert in text simplification. Your previous attempt to simplify the document was not successful and did not meet the quality criteria.

Original Document: {document_text}

Your Previous Attempt that failed:

{previous_step_output}

Feedback on Previous Attempt:

{feedback_message}

Please use this feedback to generate a new, improved simplification from the original document. Reply only with the simplified text. Do not add explanations, labels, or extra comments.

Simplified text:

Feedback template. The classifier output is converted into natural-language feedback according to the predicted level:

Too complex (predicted above target): “The simplified text is too complex. It was predicted as CEFR {predicted_cefr} but the target is {target_cefr}. Please simplify vocabulary and sentence structure while preserving the original meaning and keeping the text natural and coherent.”

Too simple (predicted below target): “The simplified text is too simple. It was

predicted as CEFR {predicted_cefr} but the target is {target_cefr}. Please use slightly more complex vocabulary and sentence structures while preserving the original meaning and keeping the text natural and coherent.”

B Distribution of final candidates

Table 2 shows the distribution of final candidates chosen by the LLM judges across prompting strategies and models.

	Zero-shot	Few-shot	Round-trip	Total
Gemini-as-judge				
Claude	30	3	12	45
GPT-5	33	19	10	62
Gemini	18	7	22	47
o1	27	14	5	46
Total	108	43	49	200
GPT-as-judge				
Claude	12	4	12	28
GPT-5	51	20	13	84
Gemini	15	8	21	44
o1	27	7	10	44
Total	105	39	56	200

Table 2: Distribution of final candidates selected under Gemini- and GPT-as-judge.

C Model-prompt performance by target level

This appendix provides detailed results across all prompting configurations (zero-shot, one-shot, and round-trip). Table 3 summarizes overall test performance across the four models, averaged over five runs. Tables 4 and 5 present level-specific results for A2 and B1.

D Effect of Sample Selection Strategies for Few-shot

On the trial data, we also explored sample selection strategies with GPT-5, comparing random versus similarity-based one-shot examples (Table 6). Here, both random and similarity-based one-shot examples reduced RMSE relative to zero-shot, but random examples achieved the lowest RMSE with relatively low variance, while similarity-based examples performed comparably. In both cases, the gain in compliance came with a clear drop in meaning preservation.

E Effect of Intermediate Language for Round-trip

This appendix provides the results for the Round-trip model on the trial datasets. Table 7 summarizes

Model	RMSE ↓			mBERT-Orig ↑			mBERT-Ref ↑		
	0-shot	1-shot	RT	0-shot	1-shot	RT	0-shot	1-shot	RT
Claude	0.595 (.018)	0.536 (.027)	0.606 (.020)	0.836 (.002)	0.796 (.001)	0.809 (.000)	0.820 (.002)	0.811 (.005)	0.825 (.010)
GPT-5	0.620 (.029)	0.597 (.009)	0.569 (.020)	0.848 (.005)	0.803 (.001)	0.793 (.000)	0.821 (.004)	0.821 (.002)	0.819 (.010)
Gemini	0.497 (.018)	0.574 (.020)	0.539 (.020)	0.812 (.004)	0.769 (.005)	0.793 (.000)	0.811 (.003)	0.808 (.008)	0.819 (.000)
o1	0.797 (.013)	0.706 (.015)	0.630 (.030)	0.868 (.002)	0.839 (.004)	0.828 (.000)	0.818 (.004)	0.826 (.003)	0.835 (.000)

Table 3: Average performance under zero-shot, similarity-based one-shot, and round-trip prompting on test data across all models (5 runs per setting). Standard deviations shown in parentheses.

Model	A2 RMSE ↓			A2 mBERT-Orig ↑			A2 mBERT-Ref ↑		
	0-shot	1-shot	RT	0-shot	1-shot	RT	0-shot	1-shot	RT
Claude	0.444 (.040)	0.433 (.060)	0.548 (.050)	0.754 (.000)	0.748 (.000)	0.756 (.000)	0.810 (.000)	0.801 (.010)	0.795 (.010)
GPT-5	0.595 (.020)	0.551 (.030)	0.436 (.020)	0.792 (.010)	0.776 (.000)	0.727 (.010)	0.819 (.010)	0.818 (.000)	0.789 (.010)
Gemini	0.359 (.040)	0.266 (.040)	0.364 (.050)	0.730 (.000)	0.715 (.010)	0.722 (.000)	0.788 (.010)	0.783 (.010)	0.788 (.010)
o1	0.857 (.020)	0.743 (.030)	0.601 (.040)	0.840 (.010)	0.818 (.010)	0.779 (.010)	0.820 (.010)	0.827 (.000)	0.806 (.000)

Table 4: A2-level results under zero-shot, one-shot, and round-trip prompting.

Model	B1 RMSE ↓			B1 mBERT-Orig ↑			B1 mBERT-Ref ↑		
	0-shot	1-shot	RT	0-shot	1-shot	RT	0-shot	1-shot	RT
Claude	0.714 (.010)	0.620 (.010)	0.656 (.060)	0.917 (.000)	0.843 (.000)	0.863 (.000)	0.831 (.000)	0.821 (.000)	0.854 (.000)
GPT-5	0.644 (.040)	0.638 (.030)	0.676 (.020)	0.904 (.000)	0.830 (.000)	0.858 (.000)	0.823 (.000)	0.824 (.000)	0.849 (.010)
Gemini	0.603 (.020)	0.767 (.020)	0.669 (.030)	0.894 (.010)	0.824 (.000)	0.864 (.000)	0.833 (.000)	0.834 (.010)	0.849 (.000)
o1	0.733 (.030)	0.666 (.030)	0.658 (.030)	0.896 (.000)	0.859 (.010)	0.877 (.000)	0.816 (.000)	0.824 (.000)	0.865 (.000)

Table 5: B1-level results under zero-shot, one-shot, and round-trip prompting.

Setting	RMSE ↓	mBERT-Orig ↑	mBERT-Ref ↑
Zero-shot	0.709 ± 0.056	0.853 ± 0.008	0.793 ± 0.006
Random 1-shot	0.532 ± 0.024	0.801 ± 0.010	0.783 ± 0.006
Similarity 1-shot	0.540 ± 0.097	0.799 ± 0.008	0.779 ± 0.007

Table 6: Average performance (\pm standard deviation) of GPT-5 with zero-shot, random one-shot, and similarity-based one-shot prompting on trial data (3 runs per setting).

the performance of each intermediate language, measured by CEFR accuracy, meaning preservation score, and similarity to the reference score, averaged across five runs. Also, Table 8 reports a comparison of simplified paragraphs with and without simplification in English-to-Chinese translation, based on MeaningBERT (mBERT in short).

Round-trip and comparing with other intermediate languages. We implemented round-trip prompts with multiple intermediate languages on the trial datasets. Besides Chinese and German (included in the submission file), we also employed Spanish, Japanese, French, and Indonesian. Our findings are: from Table 7, at A2, Chinese achieved the highest accuracy but the poorest meaning preservation and similarity-to-reference scores, whereas French and Indonesian showed the lowest accuracy but the highest Orig-BERT score. At

B1, the highest accuracy was obtained by Chinese, German, Spanish, and Japanese; however, Chinese also showed the lowest accuracy, the same as Indonesian, indicating a large variance in its performance. Spanish achieved the highest scores for both meaning preservation and similarity to reference and German also achieves adequately high. These results indicate the different tendencies of intermediate languages in matching the target CEFR level at A2 and B1.

Furthermore, we found a two-step simplification approach as in A2 is also effective. From Table 7, Chinese achieved an accuracy of 1.0 at A2. The reason for ineffectiveness at B1 is; the two-step simplification did not adjust the output toward the target CEFR level, but rather accumulated the effect of simplification, resulting in an oversimplification beyond B1. Additionally, compared to simple translation, the meaning preservation score

decreases in most paragraphs, but not for similarity to the reference: as illustrated by the example in Table 8, in about 25% of the paragraphs the score was actually higher, with simplified outputs also tending to have shorter sentences, with an average of 7.47 sentences after simplification compared to 6.76 without simplification, while original and reference paragraphs respectively contain 4.35 and 5.55 on average.

F Formatting

We observed that our self-refinement run contained substantially more newlines than the ensemble runs. At first we attributed this to the refinement process itself, hypothesizing that LLMs might exploit formatting as a way to adjust difficulty in the loop. Further inspection of zero-shot baselines revealed that only GPT-5 appeared to use newlines as a mechanism for level control, producing many at the B1 level (3.0 on average per output) but almost none at A2 (0.07), whereas other models inserted almost none across levels (Table 9).

However, we did find that the refinement process affects formatting. Among the instances that underwent refinement, newlines significantly decreased at B1 (0.63) but increased at A2 (1.06). This suggests that the refinement loop not only adjusts lexical and syntactic complexity, but also affects surface formatting.

Notably, removing newlines from the outputs substantially altered the CEFR classifier’s predictions, as seen in Table 10, indicating that the classifier is also sensitive to formatting cues.

Intermediate language	Chinese +simplify		German +simplify		Spanish +simplify		Japanese	French	Indonesian
Target CEFR: A2									
Acc-min	0.500	0.900	0.600	0.750	0.500	0.800	0.500	0.450	0.450
Acc-avg	0.620	0.950	0.625	0.815	0.550	0.900	0.590	0.600	0.530
Acc-max	0.750	1.000	0.700	0.850	0.600	0.950	0.700	0.650	0.600
mBERT-Orig	0.776	0.713	0.797	0.715	0.806	0.716	0.777	0.803	0.805
mBERT-Ref	0.743	0.721	0.755	0.744	0.740	0.734	0.732	0.739	0.750
Target CEFR: B1									
Acc-min	0.350	0.250	0.500	0.400	0.450	0.300	0.450	0.400	0.350
Acc-avg	0.500	0.300	0.563	0.415	0.540	0.385	0.540	0.460	0.490
Acc-max	0.650	0.350	0.650	0.450	0.650	0.450	0.650	0.500	0.600
mBERT-Orig	0.840	0.767	0.858	0.785	0.869	0.812	0.837	0.858	0.853
mBERT-Ref	0.827	0.776	0.826	0.792	0.850	0.814	0.841	0.836	0.829

Table 7: Comparison of intermediate languages based on CEFR accuracy, meaning preservation score, and similarity to reference score on the trial dataset

	Simplified paragraph	MeaningBERT
Without simplification	Small animals like birds, squirrels, mice, and insects are common in many cities and towns. But recently, news from all over the world talks a lot about wild animals coming into cities. Bears have been seen in parks in Vancouver. Leopards walk on the streets of Mumbai. And wild boars are in gardens in Berlin. What happens when bigger animals come into our cities? Are they welcome? Or do people see them as dangerous or a problem?	MeaningBERT-orig: 0.823 MeaningBERT-ref: 0.700
With simplification	Small animals live in many towns and cities. For example, birds, squirrels, mice, and insects. But now, more wild animals are coming into cities. Newspapers write about it. For example, there are bears in parks in Vancouver. Leopards are on streets in Mumbai. Wild pigs are in gardens in Berlin. What if bigger animals come to cities? Will we welcome them? Or will we think they are dangerous or harmful?	MeaningBERT-orig: 0.701 MeaningBERT-ref: 0.846

Table 8: Comparison of simplified paragraphs with and without simplification in English-to-Chinese translation.

Model	A2	B1
Claude	0.01	0.00
GPT-5	0.07	3.02
Gemini	0.00	0.00
o1	0.00	0.24

Table 9: Average number of newlines per output in zero-shot simplifications on test data by model and target CEFR level.

Run	Formatting	RMSE	mBERT-Orig
GPT-5 Ensemble	Original	0.1225	0.8511
	\n removed	0.2550	0.8503
Gemini Ensemble	Original	0.0707	0.8621
	\n removed	0.1871	0.8621

Table 10: Effect of removing newline characters (\n) on automatic evaluation metrics.

SQUREL at TSAR 2025 Shared Task: CEFR-Controlled Text Simplification with Prompting and Reinforcement Fine-Tuning

Daria Sokova* **Anastasiia Bezobrazova*** **Constantin Orăsan**
University of Surrey University of Surrey University of Surrey
d.sokova@surrey.ac.uk a.bezobrazova@surrey.ac.uk c.orasan@surrey.ac.uk

Abstract

This paper summarises the submissions of our team to the TSAR 2025 Shared Task on Readability-Controlled Text Simplification, which aims to create text simplifications that balance reduced linguistic complexity, meaning preservation, and fluency while meeting a predefined target readability level. In this work, we proposed two different methods for CEFR-controlled text simplification: a setup which employed reinforcement fine-tuning of large language models (LLMs) and a conservative lexical pipeline which relied on prompting LLMs to simplify sentences.

1 Introduction

Readability-controlled text simplification (RCTS) aims to generate simplifications within specified readability levels while preserving the original meaning (Barayan et al., 2024). While instruction-tuned LLMs have been shown to be useful in zero-shot RCTS, balancing readability control and meaning preservation remains challenging (Farajidizaji et al., 2023). This paper presents our participation in the TSAR 2025 Shared Task on Readability-Controlled Text Simplification (Alva-Manchego et al., 2025). The Shared Task invited participants to simplify sentences at the B1 and A2 CEFR levels. Due to the scarcity of labelled data for supervised training, we explore approaches that do not require high-quality parallel data labelled with CEFR levels. We propose two methods for producing readability-controlled simplifications:

- 1. Reinforcement Fine-Tuning with Group Relative Policy Optimization:** this method proposes a fine-tuning strategy aimed at conditioning an open-weight LLM to produce simplifications that balance CEFR-level accuracy and meaning preservation.

- 2. Lexical Simplification:** which aims to produce simplifications corresponding to the specified CEFR level through careful lexical substitution with the help of LLM prompting.

The rest of the paper is structured as follows. Section 2 describes the two methods we employed to obtain the outputs submitted to the Shared Task, followed by the discussion of the results in Section 3. We also describe limitations of our work and provide concluding observations. The prompts and examples from the outputs can be found in the Appendix A.

2 Methods Description

2.1 Group Relative Policy Optimization Fine-Tuning

Data annotated with CEFR labels at document level is scarce, which makes it difficult to use a supervised fine-tuning approach. For this reason, we experiment with reinforcement fine-tuning that does not require labelled data, specifically, with Group Relative Policy Optimization (Shao et al., 2024). In an attempt to balance CEFR-level accuracy and meaning preservation, we use two reward functions to score candidate completions, compute rewards and update the model’s weights. The first function computes rewards for compliance with the target CEFR level, whilst the second one scores candidate completions for meaning preservation. We submitted two similar systems developed using this method. They are described in Sections 2.1.2 and 2.1.4.

2.1.1 Data

To obtain texts for generating predictions during fine-tuning, we use the CEFR Levelled English Texts dataset available on Kaggle.¹ Originally, the dataset contained around 1,500 texts labeled with

¹<https://www.kaggle.com/datasets/amontgomerie/cefr-levelled-english-texts>

*These authors contributed equally to this work

CEFR levels. As the shared task targets simplification of documents written at upper-intermediate or advanced levels, we split the dataset and keep only texts at levels B2, C1 and C2. After splitting texts longer than 150 words into separate examples, we obtained 1,350 unique training examples. We prompt (see Prompt 3 in the Appendix) the model to simplify each instance into each of the target levels: A1, A2 and B1, obtaining a total of 4,050 texts to generate predictions during fine-tuning.

2.1.2 Setup for Run 1

We fine-tune the Llama 3.2 3B Instruct² model using the GPRO Trainer from the Transformer Reinforcement Learning³ library. This model was chosen for its modest size and good instruction-following capabilities. Due to computational constraints, we set up the model to generate 3 candidate simplification for each instance in the dataset. Each of the candidate predictions is scored with the reward functions and ranked. Then, the model’s weights are updated to increase the probability of generating high-reward completions and decrease the probability of generating low-reward ones.

2.1.3 Reward Functions

The CEFR compliance reward utilises the CEFR labelling models proposed by the Shared Task organizers in the evaluation scripts. The models are used to produce a CEFR label for each of the 3 candidate completions. The reward formula calculates the absolute difference between the predicted CEFR level and target CEFR level, then converts this distance into a reward score. The reward starts at 1.0 for perfect matches and decreases by 0.5 for each level of deviation, with a minimum reward of 0.0 for texts that are 2+ levels away from the target.

The meaning preservation reward uses the SentenceTransformer model (all-MiniLM-L6-v2) to generate vector embeddings of the original text and a candidate completion. Then, cosine similarity between the embeddings is computed and the scores are cubed to create a more distinctive variance between positive and negative scores. This way, completions that deviate from the original meaning are penalized more strictly.

Due to resource limitations, we choose conservative settings for the GRPO fine-tuning configuration. We use a learning rate of 5×10^{-6} and

fine-tune for 1 training epoch. For each training step, the model generates 3 candidate completions with an effective batch size of 3. To reduce computational costs, we applied Low-Rank Adaptation (LoRA) fine-tuning.

2.1.4 Setup for Run 2

Run 2 is a variant of the system described in the previous section with some alterations in the training configuration. We use a more lenient CEFR reward function, which reduces penalties for larger gaps between predicted and target CEFR levels as during the exploratory study it showed to lead to better scores for CEFR adjacency compliance. We also slightly upgraded the training parameters by setting a higher learning rate of 1×10^{-5} and increased the number of generations in each step from 3 to 4 and the gradient accumulation steps from 3 to 16 for more stability in updating weights. We also train for 2 epochs instead of 1.

In addition, the system employs a revised version of the prompt (see Prompt 4 in the Appendix) for generating candidate predictions. Unlike the system in Run 1, this version of the prompt does not provide examples of texts readable at the corresponding CEFR level.

2.1.5 Inference in the reinforcement fine-tuning pipeline

We prompt the fine-tuned model to simplify documents to the target CEFR level using a zero-shot prompt (see Prompt 5 in the Appendix). Initially, we experimented with several versions of the prompt to generate predictions for fine-tuning, including the prompt with examples of texts readable at the corresponding CEFR levels. The results obtained on the trial data indicated that the model performed best when prompted with a zero-shot prompt without examples of CEFR-level appropriate text.

2.2 Lexical simplification

This method employed implements a lexical simplification pipeline that combines a CEFR-annotated lexicon, WordNet synonyms (Miller, 1992), and controlled LLM rewrites. The system is designed to simplify sentences to CEFR A2 and B1 levels while preserving meaning as much as possible. Previous work has attempted CEFR-targeted simplification with LLMs, but the results were often inconsistent, particularly for lower levels such as A2 and B1 (Barayan et al., 2024). To avoid issues like the ones

²<https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct>

³<https://huggingface.co/docs/tr1/en/index>

run	weighted-f1	cefr-adj	rmse	meaningbert-orig	bertscore-orig
Run 1 (grpo-ft-v1)	0.590	0.995	0.633	0.779	0.928
Run 2 (grpo-ft-v2)	0.543	0.985	0.718	0.821	0.937
Run 3 (lex-simpl)	0.578	0.710	1.269	0.972	0.985

Table 1: Evaluation scores for the best runs of the proposed systems obtained on the Shared Task trial data.

run	rmse	meaningbert-orig	meaningbert-ref	AvgScore
Run 1 (grpo-ft-v1)	0.718	0.821	0.797	-0.076
Run 2 (grpo-ft-v2)	0.632	0.779	0.778	-0.153
Run 3 (lex-simpl)	1.153	0.979	0.819	-0.022

Table 2: Final scores and ranking of the submitted systems in the Shared Task obtained on the Shared Task test data.

noticed with previous approaches, this method is deliberately conservative: it avoids uncontrolled rewriting and enforces strict vocabulary constraints. Each of the steps of our method is presented below.

2.2.1 Complex word identification and candidate generation

Sentences are tokenised with spaCy, and each content word (noun, verb, adjective, adverb) is checked for possible substitutions. For a word to be included in the replacement table, WordNet must provide a POS-compatible synonym whose lemma also appears at the *target* CEFR level in our lexicon (strict A2/B1, same-level only)⁴. If no such synonyms exist, the word is ignored and may remain unchanged.

2.2.2 Constrained replacement

We present the sentence to the LLM together with a table of *allowed* replacements and instruct it to select at most one option per listed token, leave all other tokens unchanged, and not introduce new vocabulary. For example:

assist \Rightarrow {help}, purchase \Rightarrow {buy}

An example of the full prompt 1 is provided in the Appendix A.

2.2.3 Style polishing

A second prompt asks the LLM to polish the text to CEFR-specific style: sentence-length limits (A2: max 14 words; B1: max 22), preference for active voice, and restricted connectors (*and, but, or, because, so* for A2; plus *when, if, before, after* for B1). We generate $k=4$ candidates using varied sampling ($temperature \in \{0.2, 0.3, 0.35, 0.45\}$, $top-p \in \{0.95, 0.9, 0.85, 0.8\}$).

An example of the full prompt 2 is provided in the Appendix A.

⁴<https://www.kaggle.com/datasets/nezahatkk/10-000-english-words-cerf-labelled>

2.2.4 Candidate selection

From the multiple polished outputs, the best candidate is selected using an automatic scoring function. Each candidate is evaluated along four dimensions: (1) CEFR compliance, measured with a ModernBERT classifier; (2) meaning preservation, estimated with MeaningBERT when available, with BERTScore as a fallback and lexical overlap as a final backup; (3) copy ratio, i.e., the proportion of words retained from the original, with penalties applied if this exceeds a level-specific threshold (0.60 for A2 and 0.75 for B1); and (4) sentence length, with penalties applied when the target CEFR limit is exceeded. A single selection score is then computed as a weighted combination of these factors: CEFR compliance (weight 1.0), meaning preservation (+0.15), penalties for excess copy ratio ($-0.18 \times \text{excess over } 0.60 \text{ for A2 or } 0.75 \text{ for B1}$), and sentence length ($-0.10 \times \text{relative excess over the level limit}$). These weights are predefined hyperparameters rather than learned parameters and were determined by experimenting with the development data. The candidate with the highest score is selected.

2.2.5 Iteration

The system repeats the pipeline until the output is both simple enough and faithful enough to the original. Simplicity is measured by a CEFR compliance score that rewards predictions *at or below* the target level (A2 or B1) and incorporates classifier confidence; this score must reach at least 0.80. Faithfulness is measured by a meaning-preservation score, which must also reach at least 0.80. If both conditions are met, or if six rounds have already been run, the process stops. Additionally, we reject any candidate that lowers the meaning-preservation score by more than 0.05 compared to the previous round. An example of the iterative process is presented in Figure 1.

Original: The committee endeavoured to facilitate the distribution of resources in an equitable manner.
 [1] endeavoured → tried ⇒ The committee tried to facilitate the distribution of resources in an equitable manner.
 [2] facilitate → help ⇒ The committee tried to help the distribution of resources in an equitable manner.
 [3] equitable → just ⇒ The committee tried to help the distribution of resources in a just manner.
Final: The committee tried to help the distribution of resources in a just manner.

Figure 1: Example of iterative lexical simplification for target level A2

2.2.6 Implementation

The pipeline was implemented in Python with spaCy for tokenisation/POS, NLTK/WordNet for synonyms, a CEFR lexicon for strict vocabulary control⁵. We used the transformers text-classification pipeline with AbdullahBarayan/ModernBERT-base-reference_AllLang2-Cefr2 for CEFR compliance, evaluated for MeaningBERT and BERTScore, and the OpenAI API (gpt-4o-mini) for constrained rewrite and style polishing. The performance of the method is presented in Section 3.

2.3 Evaluation

The methods presented above were run with a number of parameters on the trial data. We used the evaluation scripts provided by the organizers (Alva-Manchego et al., 2025) to inform the choice of runs to submit to the Shared Task.

For each system, we ran inference iteratively and chose the best-performing runs. We based our choice of the best runs on the scores achieved for weighted F1 and CEFR-level adjacency accuracy. The results for the chosen runs are shown in Table 1.

We observe a trade-off between the meaning preservation capabilities and adjacency accuracy of our systems. The reinforcement fine-tuning method demonstrates higher CEFR adjacency accuracy while having lower meaning preservation scores. The lexical simplification approach, on the other hand, produces outputs that preserve the original meaning due to careful lexical substitution. However, it does not attain high accuracy in CEFR level adjacency.

3 Results and Discussion

Table 2 presents the official evaluation results obtained by our systems on the test data. The methodology for the final Shared Task ranking released

by the organizers excludes some of the computed metrics and produces a weighted score that relies on RMSE and meaning preservation scores against the original text and the references measured with MeaningBERT.

The Lexical Simplification system (Run 3) we submitted achieves the highest original meaning preservation scores across all systems submitted to the shared task. Due to careful meaning preservation, the Lexical Simplification System ranks higher than the models fine-tuned with GRPO despite lower accuracy in achieving target CEFR levels. The suboptimal RMSE of the reinforcement fine-tuning method might be due to our choice to optimize the reward functions for CEFR adjacency accuracy, which considers outputs successful if their CEFR level is within one level of the specified target, leading to lower accuracy.

Our lexical simplification pipeline, does not employ a word sense disambiguation module to pre-filter candidates based on the sense of a word to be replaced. Instead we employ a large language model (gpt-4o-mini) to infer which word fits in the given context. This enables the model to select morphologically and syntactically well-formed substitutes without relying on a separate WSD component.

3.1 Error Analysis

We manually analysed the outputs for a more detailed insight into the trade-off between meaning preservation and CEFR compliance scores.

The analysis reveals that the fine-tuned systems overall make more transformations to arrive at a simplification. Some of these transformations, such as sentence splitting and changes to the syntactic structure of the original sentence, are in line with the general text accessibility guidelines. The output of the fine-tuned models appears generally easier to read in comparison with our best-performing system. However, aside from occasional awkward phrasing and slight information loss, it contains multiple semantic errors that sometimes cause sig-

⁵<https://www.kaggle.com/datasets/nezahatkk/10-000-english-words-cerf-labelled>

nificant distortion of meaning. To demonstrate these, we provide our systems’ outputs for the texts from the Shared Task’s test data with text-id 113-a2 and 113-b1 (see Table 3) in the Appendix.

Omission is a major cause of semantic errors in the output. For instance, the mention that spider venom is not mainly used to attack humans was not retained in the following examples: *run-1-113-a2*, *run-2-113-a2* and *run-2-113-b1*. This resulted in a significant deviation in meaning. Another example of a critical semantic error is *run-1-113-b1*. The original says ‘Spider venom . . . serves the purpose of stunning or killing their prey rather than attacking humans’, but in the generated simplification the meaning is distorted to “it helps them catch their food by stunning or killing it, not by hurting humans”. Not only is this change of preposition unnecessary, it makes the wording ambiguous and may cause confusion. Apart from that, unjustified additions are another cause of meaning distortion in the output of the fine-tuned systems, for example, in *run-1-113-a2*.

As for *run-3-113-a2*, it is an exact copy of the source, which fails the stated goal of A2 readability, for example, low-frequency items such as “fatalities,” “urticating,” and “embed” remain and there is no simplification of syntax or lexis. As a result, the text is unlikely to be accessible to A2 readers, even though it would score highly on meaning preservation. By contrast, *run-3-113-b1* applies principled reductions and largely preserves meaning. It breaks up long sentences and replaces the clause “*serves the purpose...their prey rather than attacking humans*” with a clearer two-step formulation: “*Most spider species have venom that helps them catch prey. They do not usually attack humans.*”. Several lexical substitutions also improve accessibility: “*has not produced any fatalities*” → “*has not caused any deaths*”; “*ejecting a cloud of urticating hairs*” → “*releasing a cloud of irritating hairs*”; “*embed themselves*” → “*stick.*”

In addition, several A2 outputs introduce risky lexical changes: *run-3-27-a2* shifts scope “*poor areas*” → “*poor countries*”; *run-3-22-a2* misrenders “*wild dogs*” as “*frank dogs*”; *run-3-38-a2* “*wild animals*” → “*violent creatures*”. Notably, these anomalies are confined to A2-level, B1 outputs generally retain key terms and avoid such errors.

Overall, manual error analysis indicates that the GRPO fine-tuned systems often produce outputs that deviate in meaning from the original, despite

the implemented meaning-preservation rewards. While the generative pipeline inherently offers less control than the more conservative rule-based one, a more carefully tailored weighting of the reward functions and implementation of more advanced metrics, such as MeaningBERT (Beauchemin et al., 2023), for computing meaning preservation rewards might improve performance. As for the lexical simplification system, many sentences remain unchanged, particularly at A2-level, so meaning is preserved but CEFR aims are often unmet. Where A2 lexical edits are made, they are sometimes odd or misleading, whereas B1 simplifications tend to be more controlled and effective.

4 Conclusions

In this paper, we presented two approaches on readability-controlled text simplification for the TSAR 2025 Shared Task. The lexical pipeline based on strict CEFR-constrained substitutions with light post-editing generally preserves source meaning but quite often fails to replace the problematic words, and at times introduces critical errors, showing that control alone does not preclude serious failures. The models fine-tuned with GRPO are better at producing simplifications corresponding to the specified CEFR level but this is quite often at the expense of keeping the original meaning. These results underline the limitations of both approaches: the lexical pipeline achieves better meaning preservation while often failing at achieving the required readability level. At the same time, although the fine-tuned systems produce simpler texts, the output often deviates from the original meaning while still not matching the required readability level perfectly.

Limitations

A key limitation of the reinforcement fine-tuning method is the lack of experimentation with reward function design, the weighting of the rewards and training parameters. In addition, the dataset used to generate predictions during reinforcement fine-tuning contains automatically produced labels, and the genres and content of the texts differ from those of the Shared Task test data, which may affect performance. The lexical pipeline used a CEFR word list from Kaggle rather than the official *CEFR-J Vocabulary* (Version 1.5) from TUFSS⁶. This resource

⁶Yukio Tono, *The CEFR-J Wordlist Version 1.5*, retrieved from <http://www.cefrj.org/download.html> on 20 Jan-

deviates from the original CERF list, which may have influenced our results. Finally, the lexical simplifier relied on the commercial **GPT-4o-mini**, so results may be hard to replicate as future versions could behave differently.

Lay Summary

Reading difficult text can be challenging for many people, including language learners, children, and those with reading difficulties. This research explored ways to automatically rewrite complex texts to make them simpler while keeping the original meaning. Our team developed automatic systems to simplify complex texts to target levels of reading difficulty.

For example, a text that requires advanced language skills needs to be simplified so that speakers with intermediate or elementary skills can understand it easily. When the task says that a text should be simplified for elementary skills, then the result is only successful if the simplified text can be easily understood by speakers with elementary skills. It is not considered successful if the text does not match the required skill level, even if it is generally simpler than the original. The simplified text also needs to keep the meaning of the original difficult text.

Our team developed two different approaches. The first method works like a careful editor. It swaps difficult words for simpler alternatives. This system used a dictionary that labelled words by difficulty level and only replaced complex words with simpler synonyms that meant the same thing. After making these swaps, the system polished the sentences by shortening them and using simpler grammar. The polishing was done using a large language model – a computer program that can generate text. This approach was good at keeping the original meaning but sometimes struggled to make texts simple enough.

The second method involved teaching a large language model to learn how to simplify text to a specific skill level. Rather than following strict rules, the system learned through practice. It performed the task repeatedly and received scores on how well it matched the needed difficulty level and how accurately it preserved meaning. This approach was better at simplifying texts to the target difficulty levels, but it often changed the meaning too much or left out important information.

When tested, the careful word-swapping method performed best overall because it preserved meaning more reliably, even though it did not always achieve the target difficulty level. The taught large language model was better at generating simpler text and achieving the target difficulty level but performed worse overall because it introduced errors or left out important details.

This research highlights that it is challenging to create automatic text simplification systems that match the required difficulty level and keep the original meaning at the same time.

References

- Fernando Alva-Manchego, Regina Stodden, Joseph Marvin Imperial, Abdullah Barayan, Kai North, and Harish Tayyar Madabushi. 2025. Findings of the TSAR 2025 shared task on readability-controlled text simplification. In *Proceedings of the Fourth Workshop on Text Simplification, Accessibility, and Readability (TSAR 2025)*, Suzhou, China. Association for Computational Linguistics.
- Abdullah Barayan, Jose Camacho-Collados, and Fernando Alva-Manchego. 2024. Analysing zero-shot readability-controlled sentence simplification. *arXiv preprint arXiv:2409.20246*.
- David Beauchemin, Horacio Saggion, and Richard Khoury. 2023. Meaningbert: assessing meaning preservation between sentences. *Frontiers in Artificial Intelligence*, 6:1223924.
- Asma Farajidizaji, Vatsal Raina, and Mark Gales. 2023. Is it possible to modify text to a target readability level? an initial investigation using zero-shot large language models. *arXiv preprint arXiv:2309.12551*.
- George A. Miller. 1992. *WordNet: A lexical database for English*. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

A Appendix

A.1 The Prompts Used During Simplification

Prompt 1 (Method 1)

You are a careful lexical simplifier.
Target CEFR level: {target}. Preserve meaning exactly. Keep names and numbers.
{style_extra}
{vocab_rule}

RULES:

- 1) You may replace a token ONLY if it appears in the list below.
- 2) For each listed token, choose at most one alternative from its line.
- 3) If none of the alternatives fit the meaning, KEEP the original token.
- 4) Do NOT invent alternatives or use words not in the list.
- 5) Keep punctuation and sentence order; light edits for grammar are OK.

ALLOWED REPLACEMENTS (source ⇒ options):

{token_1} ⇒ {option_1a, option_1b, ...}

{token_2} ⇒ {option_2a, option_2b, ...}

...

Original text:

{original_text}

Output ONLY the rewritten text with your chosen replacements.

Prompt 2 (Method 1)

You are a professional text editor.
Target CEFR level: {target}. Preserve meaning and chosen vocabulary.
{style_extra}
{vocab_rule}

Do NOT add definitions or extra info.

Split long sentences if needed. Prefer active voice.

Original text:

{rewritten_text}

Polished text:

Prompt 3 (Used to generate candidates during fine-tuning for Method 2, Run 1)

System Prompt:

You are an expert in text simplification. You simplify text to the CEFR level that perfectly aligns with the target CEFR level. You only output simplified texts. You do not include anything else in your answer.

User Prompt:

Please simplify the following Complex Text to make it easier to read and understand by {target_cefr} CEFR English learners. To simplify, you may replace difficult words with simpler ones, elaborate or remove them when possible. You may also break down a lengthy sentence into shorter, clear sentences. Ensure the revised sentence is grammatically correct, fluent, and maintains the core message of the original, without changing its meaning.

Please simplify the following Complex Text to make it easier to read and understand by {target_cefr} CEFR English learners. To simplify, you may replace difficult words with simpler ones, elaborate or remove them when possible. You may also break down a lengthy sentence into shorter, clear sentences. Ensure the revised sentence is grammatically correct, fluent, and maintains the core message of the original, without changing its meaning.

The following are the examples of sentences readable by {target_level} CEFR level English learners:

Example 1: e1
Example 2: e2
Example 3: e3 Use these examples as reference, do not produce any examples.
Complex Text: {text}
Simplified Text:

Prompt 4 (Used to generate candidates during fine-tuning for Method 2, Run 2)

System Prompt:

You are an expert in text simplification. You simplify text to the CEFR level that perfectly aligns with the target CEFR level.
You only output simplified texts. You do not include anything else in your answer.

User Prompt:

Please simplify the following Complex Text to make it easier to read and understand by {target_cefr} CEFR English learners. To simplify, you may replace difficult words with simpler ones, elaborate or remove them when possible. You may also break down a lengthy sentence into shorter, clear sentences. Ensure the revised sentence is grammatically correct, fluent, and maintains the core message of the original, without changing its meaning.
Only produce the simplified text, do not include anything else.

Complex Text: {text}

Simplified Text:

Prompt 5 (Used during inference for Method 2)

You are an expert in text simplification.

You simplify text to the CEFR level that perfectly aligns with the target CEFR level.
You only output simplified texts. You do not include anything else in your answer.
Please simplify the following Complex Text to make it easier to read and understand by {target_cefr} CEFR English learners. To simplify, you may replace difficult words with simpler ones, elaborate or remove them when possible. You may also break down a lengthy sentence into shorter, clear sentences. Ensure the revised sentence is grammatically correct, fluent, and maintains the core message of the original, without changing its meaning.
Only produce the simplified text, do not include anything else.

Complex Text: {text}

Simplified Text:

A.2 Examples Discussed in Error Analysis

example-id	example-text
original-113	Spider venom is present in most species and serves the purpose of stunning or killing their prey rather than attacking humans. In fact, only 25 of the known spider species produce venom which can actually harm humans, and although spider bites can be painful, they are rarely deadly. Australia's notorious Sydney funnel-web spider has not produced any fatalities since anti-venom was developed in 1981. However, take care not to rub hairy spiders like tarantulas up the wrong way. When they feel in danger, they defend themselves by ejecting a cloud of urticating hairs. These irritating hairs then embed themselves in the skin or eyes of the attacker.
run-1-113-a2	Most spiders have venom that helps them catch their food. [omission] This venom can hurt people, but it is not usually deadly. Only 25 out of many spider species can hurt humans. In Australia, there is a big spider called the Sydney funnel-web spider. It is not deadly since a medicine was made to stop its venom. However, it is still dangerous if you touch it. Tarantulas are also hairy and can hurt you if they feel scared. If they feel threatened, they will throw out irritating hairs to defend themselves. These hairs can get stuck in your skin or eyes. You should be careful around spiders.
run-1-113-b1	Spider venom is present in most spiders. It helps them catch their food by stunning or killing it, not by hurting humans. Most spiders do not harm humans. Only 25 species can hurt humans. Their bites can be painful, but usually not deadly. Australia's Sydney funnel-web spider has not killed anyone since the development of anti-venom in 1981. However, be careful when you touch hairy spiders, like tarantulas. If they feel threatened, they can defend themselves by releasing irritating hairs into the air. These hairs can stick to your skin or eyes if you touch them.
run-2-113-a2	Many spiders have venom. They use it to catch their food. [omission] Most spiders can't hurt people. Only a few species can. If you get bitten by a spider, it can hurt. But it's not usually very bad. In Australia, there's a very dangerous spider called the Sydney funnel-web. It won't kill you. But be careful. Some hairy spiders, like tarantulas, can hurt you if you touch them. When they feel threatened, they spit out hair. These hairs can hurt your skin or eyes.
run-2-113-b1	Most spiders have venom. They use it to catch or kill their food. [omission] Only 25 out of 100 spider species can hurt humans. Spider bites can be painful. But they are not usually deadly. Australia's Sydney funnel-web spider is very dangerous. But it is not deadly since medicine was invented in 1981. If you touch hairy spiders like tarantulas, be careful. When they are scared, they can hurt you by throwing hairs. These hairs can hurt your skin or eyes.
run-3-113-a2	Spider venom is present in most species and serves the purpose of stunning or killing their prey rather than attacking humans. In fact, only 25 of the known spider species produce venom which can actually harm humans, and although spider bites can be painful, they are rarely deadly. Australia's notorious Sydney funnel-web spider has not produced any fatalities since anti-venom was developed in 1981. However, take care not to rub hairy spiders like tarantulas up the wrong way. When they feel in danger, they defend themselves by ejecting a cloud of urticating hairs. These irritating hairs then embed themselves in the skin or eyes of the attacker.
run-3-113-b1	Most spider species have venom that helps them catch prey. They do not usually attack humans. Only 25 spider species have venom that can harm humans. Spider bites can be painful, but they are rarely deadly. The Sydney funnel-web spider in Australia has not caused any deaths since anti-venom was created in 1981. However, be careful not to handle hairy spiders like tarantulas improperly. When they feel threatened, they protect themselves by releasing a cloud of irritating hairs. These hairs can stick to the skin or eyes of the attacker.

Table 3: Examples discussed in the Error Analysis subsection with errors highlighted in red.

Archaeology at TSAR 2025 Shared Task: Teaching Small Models to do CEFR Simplifications

Rareş-Alexandru Roşcan and Sergiu Nisioi*

Human Language Technologies Research Center

Faculty of Mathematics and Computer Science

University of Bucharest

roscanrares@gmail.com

sergiu.nisioi@unibuc.ro

Abstract

Large language models (LLMs) have demonstrated strong performance in text simplification tasks, but their high computational cost and proprietary nature often limit practical use, particularly in educational contexts. In this work, we explore open-source LLMs for CEFR-level text simplification. By reducing the size of the model and the computational requirements, our approach enables greater accessibility and deployment in educational environments. Our results obtain some of the lowest error rates in producing CEFR-compliant texts during TSAR2025, with models of 8 billion and 1 billion parameters. Such approaches have the potential to democratize NLP technologies for real-world applications.

1 Introduction

Text simplification (TS) aims to preserve the original meaning while reducing linguistic complexity, making text more accessible to a wider audience (Alva-Manchego et al., 2020). In educational contexts, this is particularly valuable, as learners benefit from materials adapted to their proficiency levels (Li et al., 2025). In this paper, we explore different prompting methods to Large Language Models (LLMs) and provide a comparative analysis between large closed-source systems and smaller open-weights models. While recent LLMs demonstrate strong language understanding capabilities, their high computational requirements and proprietary nature still limit broad educational deployment, particularly in resource-constrained settings (Bai et al., 2024). One-billion parameter models fine-tuned with synthetic data show promising results in providing accessible solutions for educational use.

While prior work has treated CEFR-level TS as a unified task with controllable outputs (Barayan

et al., 2025; Farajidizaji et al., 2024), we propose decomposing this problem into distinct sub-tasks, where each CEFR level represents an independent simplification objective.

We investigate two approaches: (1) level-specific prompts that are very descriptive individually for the specific target, and (2) specialized fine-tuning where models are trained exclusively for individual CEFR levels. Additionally, we use an iterative prompting technique inspired by Minimum Bayes Risk decoding (Heineman et al., 2024), where multiple attempts with slight prompt variations are used to achieve optimal simplification quality.

2 Simplification Systems

2.1 Prompting

Although prior work used more general prompting strategies (Barayan et al., 2025), we find that highly specific prompts can guide the model more effectively in producing CEFR-aligned simplifications.

Specifically, our prompts make the target CEFR level explicit by including detailed requirements related to vocabulary range, sentence length, and grammatical constructions, as well as by highlighting distinctions between adjacent levels (e.g., A2 vs. B1 and B1 vs B2) while remaining a zero-shot approach. We adopt this method because recent studies have shown that LLMs struggle with reliable CEFR assessment (Barayan et al., 2025; Heineman et al., 2024). The reference to “most common 1500/3000 words” was not based on any external frequency list. Instead, these constraints were communicated in the prompt itself, allowing the LLM to operationalize the notion of common vocabulary internally. The complete prompt templates for each CEFR level are provided in Appendix C and D.

We evaluate our prompting approach across several LLMs to assess their effectiveness in CEFR-aligned text simplification. Specifically, we test GPT-4 Turbo to enable direct comparison with

*Both corresponding authors.

Barayan et al. (2025), Claude Sonnet 4, Kimi K2, and two LLaMA variants: LLaMA 3.1 8B Instruct and LLaMA 3.2 1B Instruct.

Additionally, we implement an iterative simplification approach where each model is given up to 5 attempts per text to achieve better simplifications, retaining the final iteration as our output.

The strategy is:

1. prompt the model to generate a simplification candidate and evaluate it using the CEFR classifier;
2. the generated text is evaluated for its CEFR level using an ensemble of three BERT-based classifiers (**ModernBERT-base-doc_en-CEFR**, **ModernBERT-base-doc_sent_en-Cefr**, and **ModernBERT-base-reference_AllLang2-Cefr2**) the same way it is evaluated in the evaluation script;
3. if the simplification matches the target CEFR level, the process stops; otherwise, corrective feedback is provided in the form of `The text is TOO COMPLEX. Simplify more aggressively.` or `The text is TOO SIMPLE. Add more complexity while staying at target level.`, and the model regenerates a new candidate;
4. the process repeats until the CEFR target is reached or the maximum number of iterations (5) is reached.

Submission 1 represents the iterative prompting applied to Claude Sonnet 4¹ (claude-4-sonnet-20250514). We use the prompts described in Appendix C and D, with decoding parameters set to `top_k = 10` and `temperature = 0.1`.

Submission 2 represents the iterative prompting applied to LLaMA 3.1 8B Instruct², using the same prompt templates from Appendix C and D. For decoding, we set `top_p = 0.8` and `temperature = 0.1`.

We evaluate the simplifications using the official TSAR 2025 test dataset and the corresponding evaluation metrics (Alva-Manchego et al., 2025). The test set consists of 100 independent sentences, which are a subset of the UniversalCEFR dataset,

¹<https://www.anthropic.com/claude>

²<https://huggingface.co/meta-LLaMA/LLaMA-3.1-8B-Instruct>

each paired with two human-authored simplifications: one at the A2 level and one at the B1 level.

2.2 Synthetic Data

UniversalCEFR dataset (Imperial et al., 2025) is a corpus annotated with CEFR levels without parallel versions in simplified form. To address this issue, we use LLaMA 3.1 8b to simplify the texts at the target CEFR levels by prompting the model with the iterative method presented earlier.

We select all texts annotated at B2 level or above, both at the document and sentence level. Sentence-level texts are used directly, while document-level texts are parsed into smaller texts. In total, we obtain 12,000 distinct texts, each paired with two simplified versions: one at A2 level and one at B1 level.

2.3 Finetuning Small models

We use the synthetic dataset described in subsection 2.2 to finetune LLaMA 3.2 1B Instruct for CEFR-based text simplification. We experiment with two finetuning setups, in the first approach, we train a single model on the entire dataset, such that it learns to generate both A2 and B1 level simplifications. In the second approach, we train two separate models, one specialized for A2 simplification and the other for B1. We perform fine-tuning using the Axolotl framework³, which provides training pipelines for instruction-tuned LLaMA models.

Our training configuration uses a learning rate of $5e-5$ with cosine learning rate scheduling and the AdamW optimizer with 8-bit quantization. We set the micro-batch size to 1 with gradient accumulation steps of 32, resulting in an effective batch size of 32. The models are trained for 2 epochs with a maximum sequence length of 4096 tokens, using sample packing to improve training efficiency. We apply a weight decay of 0.001 and use automatic mixed precision (bf16) when available. The validation set comprises 10% of the training data, and we evaluate model performance every 50 steps using perplexity as the primary metric. Early stopping is not employed, and we save model checkpoints once per epoch.

Submission 3 represents the iterative prompting to LLaMA 3.2 1B Instruct⁴, fine-tuned using the configuration mentioned in subsection 2.3. We use

³<https://github.com/OpenAccess-AI-Collective/axolotl>

⁴<https://huggingface.co/meta-LLaMA/LLaMA-3.2-1B-Instruct>

the same iterative prompting process and prompt templates as in previous submissions.

3 Results and Discussion

Table 1 summarizes the performance of all submissions across the evaluation metrics. Overall, the results indicate that Claude Sonnet 4 achieves the best performance, with the lowest RMSE (0.122) and the highest automatic ranking score (1.0). This highlights its strong ability to reliably adjust lexical and syntactic complexity while preserving meaning, this being the reason why we chose it for synthetic data.

While Kimi K2 achieves strong results across automatic metrics, qualitative inspection suggests that its simplifications often lack fluency and cohesion. In particular, the model tends to produce excessively fragmented and repetitive sentences, which explains the gap between metric-based performance and actual readability. This pattern also raises the possibility that Kimi K2 may be exploiting weaknesses in current evaluation metrics, producing outputs that score well numerically. A detailed comparison with human references and outputs from Claude and LLaMA 3.1 is provided in Appendix 2.

In contrast, the results for LLaMA 3.1 8B Instruct illustrate a divergence between automatic evaluation and human judgment. According to Table 1, this model ranks below Kimi K2, with a higher RMSE and slightly lower automatic scores. However, our manual analysis found the outputs of LLaMA 3.1 8B to be more coherent, fluent, and contextually appropriate. This discrepancy underscores the fact that automatic metrics alone do not fully capture aspects such as fluency and cohesion, which are essential in CEFR-based text simplification. Consequently, we regard LLaMA 3.1 8B as overall stronger in practical terms, despite its lower position in the automatic ranking.

When examining the outputs of LLaMA 3.2 1B Instruct without fine-tuning, two issues become evident. First, the model often fails to follow all prompt instructions, even when these are stated multiple times. For example, it frequently returns additional comments or explanations such as “Here is the simplified text...” or even detailed notes about the transformation process, despite being explicitly instructed to output only the simplified text. Second, the base model tends to produce excessively short simplifications, sometimes reducing an entire

paragraph to one or two sentences. This extreme shortening is caused partly by the iterative prompting setup: since the model has no prior knowledge of how a CEFR-target simplification should look (zero-shot), the repeated corrections encourage it to strip down content aggressively rather than restructure it appropriately.

Moreover, the extra notes and process explanations negatively bias the iterative pipeline: because the CEFR classifier evaluates the entire response, the added commentary shifts the predicted level upwards or downwards in unintended ways. For instance, in one case the model returned a step-by-step justification of lexical changes (“I removed ‘common’ from ‘Common dreams’, changed...”), which was treated as part of the text by the classifier and distorted the CEFR label. Examples of both the over-shortening and adherence failures are documented in Appendix 3, demonstrating how iterative prompting without prior fine-tuning leads to unreliable outputs and noisy automatic evaluations.

By contrast, the fine-tuned 1B model overcomes these limitations: it aligns much more closely with the intended structure of CEFR simplifications, avoids adding unwanted commentary, and produces outputs that are considerably richer in content while still controlled in complexity. These improvements are particularly noteworthy given that the fine-tuning relied on synthetic data generated with LLaMA 3.1 8B, itself a relatively modest model, and that the resulting dataset was not manually cleaned or verified. This suggests that the gains observed here represent only a lower bound of what could be achieved; with higher-quality, carefully curated, or state-of-the-art synthetic simplification data, fine-tuning small models would likely provide even greater improvements.

For completeness, we also conducted a small-scale fine-tuning experiment on LLaMA 3.1 8B, using approximately 1000 synthetic simplifications generated with Claude Sonnet 4 and the same fine-tune configuration as the one used for LLaMa 3.2 1B. While the dataset was limited in size and not manually verified, it still allowed us to observe meaningful improvements. Due to the small amount of data, we did not apply iterative prompting during inference and instead evaluated the model in a single-try setting. This setup affected the results in two ways: (1) RMSE scores were higher compared to the iterative experiments, since the model had fewer opportunities to adjust toward the exact CEFR level, and (2) the Mean-

Model	RMSE	MeaningBERT (orig)	MeaningBERT (ref)	Avg Score	Auto Rank
Claude Sonnet 4¹	0.122	0.779	0.804	0.338	1.000
LLaMA 3.1 8B finetune non-iterative	0.324	0.807	0.822	0.195	1.710
LLaMA 3.1 8B²	0.255	0.781	0.789	0.135	2.000
Kimi K2	0.173	0.726	0.771	0.099	2.180
GPT-4 Turbo	0.274	0.730	0.776	0.000	2.670
LLaMA 3.2 1B individual finetune³	0.212	0.706	0.731	-0.104	3.180
LLaMA 3.2 1B finetune	0.300	0.721	0.738	-0.167	3.490
LLaMA 3.1 8B non-iterative	0.561	0.794	0.795	-0.192	3.610
LLaMA 3.2 1B	0.663	0.574	0.588	-1.285	9.000

Table 1: AUTORANK results for all submitted runs under custom weighting. ^{1,2,3} mark our official submissions. “Individual” refers to individually fine-tuned models while “non-iterative” refers to single try simplifications. If not mentioned, all models are iterative/nonfinetuned. The other models are included for comparison.

ingBERT scores improved, consistent with our hypothesis that the first simplification attempt usually yields the highest semantic similarity. Even under these constrained conditions, the one-try RMSE remained competitive, showing that even with a very limited number of training examples, fine-tuning substantially enhanced the model’s ability to perform CEFR simplification.

It is also important to note that our experiments were primarily optimized towards minimizing RMSE, as this metric directly reflects the alignment of the outputs with the target CEFR level. Consequently, aspects related to meaning preservation were not prioritized to the same extent. This explains why some systems, despite achieving lower RMSE, occasionally sacrificed fidelity or omitted information from the original text. A more balanced approach that explicitly weights both accuracy of CEFR control and semantic consistency could represent an important direction for future work.

Furthermore, our focus on minimizing RMSE exposed some of the limitations and biases in the current evaluation metrics. For example, Kimi K2 achieves relatively strong scores despite producing outputs that are qualitatively weak, highlighting that automatic metrics can sometimes reward undesirable behavior. This suggests that more robust, state-of-the-art evaluation methods are necessary to obtain a fair ranking of systems. Under such an evaluation framework, we hypothesize that an approach built on well-balanced Minimum Bayes Risk decoding could result in models that are both reliable for CEFR control and practical enough to be deployed.

In addition, as shown in Table 1, fine-tuning two

separate LLaMA 3.2 1B models individually for A2 and B1 simplification yields lower slightly RMSE compared to a single model trained on both levels. This indicates that task specialization, even at very small model scales, can provide improvements in controlling output complexity.

4 Ethical and Deployment Considerations

Our methods enable practical deployment of CEFR-aware simplification, but operational use raises concrete risks. Synthetic training data and automatic simplifiers may introduce factual inaccuracies, level misclassification, or unequal behaviour across learner groups. Therefore, any real-world deployment should include human validation, continuous monitoring for meaning preservation and fairness, clear documentation of model limitations and licensing, and user-facing transparency about automated processing.

Furthermore, the proposed approach contributes to a more sustainable deployment paradigm: instead of relying on repeated inference from large proprietary LLMs, it enables the generation of synthetic data to train smaller, open models. This reduces computational and financial costs while improving accessibility for educational institutions with limited resources.

Another advantage of our method lies in its applicability to low-resource settings, where high-quality labeled simplification data are scarce. Synthetic data generation through LLMs can bootstrap new CEFR-aligned datasets for such languages.

Future work could incorporate human-in-the-loop feedback from language teachers to refine level control and to assess pedagogical usefulness beyond automatic metrics.

5 Conclusions

This work demonstrates several key findings for CEFR-based text simplification using smaller, open-source language models.

Our iterative prompting approach provides a computationally inexpensive method to significantly improve simplification quality. By allowing models up to 5 attempts per text, we achieved substantial RMSE improvements across all tested models. This technique offers a practical way to enhance performance without additional training or model scaling, making it particularly valuable for resource-constrained educational environments.

Highly structured, CEFR-specific prompts improve model understanding of proficiency level requirements. Our detailed prompts, which explicitly describe vocabulary constraints, sentence complexity, and grammatical structures for A2 and B1 levels, led to more accurate simplifications compared to generic approaches. This finding suggests that explicit linguistic guidance can compensate for models' limited understanding of CEFR frameworks.

Training separate models for each CEFR level yields better results than training a single model to handle multiple target levels. This specialization allows models to develop more focused representations of the linguistic constraints and stylistic requirements specific to each proficiency level, resulting in more appropriate simplifications.

Our fine-tuning experiments further serve as a strong proof of concept. We observed that even when training with synthetic data generated by LLMs (which were not manually verified and may inherently contain too much noise) substantial improvements can still be achieved. This effect was visible both when using large amounts of automatically generated samples (from LLaMA 3.1 8B) and under very constrained conditions with only 1000 examples (from Claude Sonnet 4 for LLaMA 3.1 8B). These results highlight the potential of fine-tuning: with higher-quality and carefully curated datasets, the gains observed here would likely be amplified, suggesting a clear path forward for building robust simplification systems on top of smaller open models for educational use.

Our code is available at <https://github.com/roscaanraes/emnlp-2025-tsar>.

6 Lay Summary

This study explores how large language models (LLMs) can be guided to produce simplified texts that match the Common European Framework of Reference for Languages (CEFR) levels. Our aim is to make reading materials easier to understand by automatically adjusting text difficulty to the reader's language level.

Many previous studies have used general instructions or relied on large datasets labeled by hand. Our approach experiments with CEFR-aware prompting, which gives the model more specific instructions to generate simplified examples that match the language characteristics of each level. These examples can then be used to train or evaluate smaller, open-access models, reducing the need for expensive proprietary LLMs.

We are also studying how the level of detail in these instructions affects the model's ability to produce appropriate simplifications. Early results suggest that more specific prompts can improve control over vocabulary and sentence structure, leading to texts that are easier to read and score better in automatic evaluations.

In addition, we observe that this method enables a more sustainable way of developing simplification systems. Instead of relying on repeated queries to large commercial models, synthetic data produced once can be reused to fine-tune compact open models. This reduces computational costs, encourages reproducibility, and supports educational organizations that may not have access to high-end infrastructure.

This work is still in progress. While our findings are promising, more research is needed before such methods can be turned into practical tools. Ultimately, we hope this research will help teachers and educational institutions create more accessible, level-adjusted learning materials and support the development of CEFR-aligned datasets for languages with limited resources.

Acknowledgments

This work was supported by the Romanian National Research Council (CNCS) through the Executive Agency for Higher Education, Research, Development and Innovation Funding (UEFISCDI) under grant PN-IV-P2-2.1-TE-2023-2007 InstRead.

References

- Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2020. [Data-driven sentence simplification: Survey and benchmark](#). *Computational Linguistics*, 46(1):135–187.
- Fernando Alva-Manchego, Regina Stodden, Joseph Marvin Imperial, Abdullah Barayan, Kai North, and Harish Tayyar Madabushi. 2025. Findings of the TSAR 2025 shared task on readability-controlled text simplification. In *Proceedings of the Fourth Workshop on Text Simplification, Accessibility, and Readability (TSAR 2025)*, Suzhou, China. Association for Computational Linguistics.
- Guangji Bai, Zheng Chai, Chen Ling, Shiyu Wang, Jiaying Lu, Nan Zhang, Tingwei Shi, Ziyang Yu, Mengdan Zhu, Yifei Zhang, Xinyuan Song, Carl Yang, Yue Cheng, and Liang Zhao. 2024. [Beyond efficiency: A systematic survey of resource-efficient large language models](#). *Preprint*, arXiv:2401.00625.
- Abdullah Barayan, Jose Camacho-Collados, and Fernando Alva-Manchego. 2025. [Analysing zero-shot readability-controlled sentence simplification](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6762–6781, Abu Dhabi, UAE. Association for Computational Linguistics.
- Asma Farajidizaji, Vatsal Raina, and Mark Gales. 2024. [Is it possible to modify text to a target readability level? an initial investigation using zero-shot large language models](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9325–9339, Torino, Italia. ELRA and ICCL.
- David Heineman, Yao Dou, and Wei Xu. 2024. [Improving minimum bayes risk decoding with multi-prompt](#). *Preprint*, arXiv:2407.15343.
- Joseph Marvin Imperial, Abdullah Barayan, Regina Stodden, Rodrigo Wilkens, Ricardo Muñoz Sánchez, Lingyun Gao, Melissa Torgbi, Dawn Knight, Gail Forey, Reka R. Jablonkai, Ekaterina Kochmar, Robert Reynolds, Eugénio Ribeiro, Horacio Saggion, Elena Volodina, Sowmya Vajjala, Thomas François, Fernando Alva-Manchego, and Harish Tayyar Madabushi. 2025. [UniversalCEFR: Enabling Open Multilingual Research on Language Proficiency Assessment](#). *arXiv preprint arXiv:2506.01419*.
- Guanlin Li, Yuki Arase, and Noel Crespi. 2025. [Aligning sentence simplification with ESL learner’s proficiency for language acquisition](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 492–507, Albuquerque, New Mexico. Association for Computational Linguistics.

A Things we Thought Would Work, but didn’t

In addition to our successful experiments, we report several approaches that produced suboptimal results, providing valuable insights for future research in CEFR-based text simplification.

A.1 Extended Synthetic Training Data

We initially hypothesized that including intermediate simplification attempts (not just the final outputs) would provide richer training signals. However, expanding our training dataset from 12,000 to 48,000 examples by including all intermediate simplifications introduced excessive noise and degraded model performance. This indicates that quality of training examples is more critical than quantity in this context and that it might be difficult to finetune for iterative simplifications.

A.2 Alternative Model Architectures

Experiments with Qwen 32B yielded considerably weaker results compared to LLaMA 3.1 8B. Both the base version of Qwen and its fine-tuned variant underperformed relative to LLaMA 3.1 8B across metrics such as RMSE and MeaningBERT, indicating that model size alone does not dictate performance.

In addition, we also attempted fine-tuning on several other open-source models, including EXAONE-4.0-1.2B⁵, OLMo-2-0425-1B-Instruct⁶, and LFM2-1.2B⁷. However, none of these models produced competitive results in our setup.

A.3 Sub-1B Parameter Models

We experimented with several ultra-lightweight models, including Gemma variants and LFM2-350M⁸ systems with fewer than 1 billion parameters. While these models were able to return outputs in a one-shot setup, their results were significantly weaker than those of LLaMA 3.2 1B, both in terms of RMSE and semantic preservation. The shortcomings became even more apparent under iterative prompting, where they showed the same type of failures as the unfine-tuned LLaMA 3.2 1B, including hallucinations and unstable outputs. Given that

⁵<https://huggingface.co/LGAI-EXAONE/EXAONE-4.0-1.2B>

⁶<https://huggingface.co/allenai/OLMo-2-0425-1B-Instruct>

⁷<https://huggingface.co/LiquidAI/LFM2-1.2B>

⁸<https://huggingface.co/LiquidAI/LFM2-350M>

even their one-try simplifications were of very low quality, we conclude that such sub-1B models are not a viable choice for CEFR simplification in our setup, especially when compared to the much stronger performance achieved by LLaMA 3-series models.

A.4 Cross-lingual Performance Variations

Models such as Kimi K2 and Qwen, which were likely trained on larger Chinese language corpora, showed different processing patterns and performance characteristics compared to models with more balanced multilingual training. This highlights the importance of pre-training data composition for downstream task performance.

A.5 LLaMA Models: Exceptional Performance-to-Cost Ratio

Remarkably, our experiments reveal that LLaMA variants, particularly LLaMA 3.1 8B and even the compact 3.2 1B model, achieve competitive or superior performance compared to much larger proprietary models like GPT-4 Turbo in CEFR simplification tasks. This finding has significant implications:

- **Cost-effectiveness:** Smaller LLaMA models provide accessible alternatives for educational institutions with limited computational resources
- **Synthetic data generation:** Larger LLaMA variants (e.g., LLaMA 4 17B, LLaMA 3.3 70B) could serve as ideal candidates for generating high-quality synthetic training data for even smaller specialized models
- **Deployment feasibility:** The strong performance of 1B parameter models opens possibilities for on-device deployment in educational applications

A.6 CEFR Assessment Finetune

We also experimented with joint fine-tuning approaches on LLaMA 3.1 8B that combined CEFR assessment with simplification. Our motivation for this experiment was the observation that models generally lack the capacity to reliably perform CEFR assessment on their own, which led us to hypothesize that an explicit assessment signal could help guide the simplification process. Specifically, we tested two strategies: (1) sequential fine-tuning, where we first trained the model for CEFR text

classification and then adapted it for simplification, and (2) multi-task fine-tuning, where we included both simplification data and CEFR assessment data in the same training stage.

For the assessment task, we relied on the same UniversalCEFR dataset used in our synthetic data experiments, but instead of parsing document-level texts into smaller segments, we used the full documents together with the sentence-level samples, resulting in approximately 12,000 texts. For the simplification task, we used the same synthetic simplifications generated by Claude Sonnet 4. However, neither approach produced improvements over standard fine-tuning. These results suggest that directly combining CEFR assessment and simplification is not straightforward, and that more advanced methods such as curriculum learning or more carefully balanced multi-task objectives may be required.

B Models and Synthetic Dataset

All fine-tuned LLaMA models used in this work are publicly available on Hugging Face:

- **LLaMA 3.1 8B fine-tuned for A2 and B1:**
<https://huggingface.co/roscanraes/llama-3.1-8b-finetuned-a2b1>
- **LLaMA 3.2 1B fine-tuned for A2:**
<https://huggingface.co/roscanraes/llama-3.2-1b-finetuned-a2>
- **LLaMA 3.2 1B fine-tuned for B1:**
<https://huggingface.co/roscanraes/llama-3.2-1b-finetuned-b1>

The synthetic dataset used for training and evaluation can be downloaded from:

- **CEFR-Simplifications dataset:**
<https://huggingface.co/datasets/roscanraes/CEFR-Simplifications>

C A2 Prompt

Role

You are a language teacher simplifying texts to A2 CEFR level.

Objective

Transform this text to A2 level while preserving all original meaning and information. {base_feedback}

A2 Language Requirements

- Vocabulary: Most common 1500 English words only
- Sentences: 8–12 words, one clear idea per sentence
- Grammar: Simple present/past, basic future (will), basic modals (can/must/should)
- Connectors: and, but, because, so, when, if, then
- Style: Personal, concrete, everyday language

Strict Level Control

- Above A1: Include personal experiences, feelings, plans, time references
- Below B1: No present perfect, passive voice, or complex connectors (however, although, despite)
- Below B1: No abstract concepts without concrete explanation

Transformation Process

1. Identify all key information and meaning
2. Break complex sentences into simple A2 structures
3. Replace advanced vocabulary with A2 equivalents
4. Convert complex grammar to simple A2 patterns
5. Verify all original meaning is preserved

Critical

Do not omit, summarize, or change any information. Only change **how** it is expressed.

Instruction

Return only the simplified text. Do not include any other comments, notes, or additional information.

Text to Simplify

D B1 Prompt

Role

You are an expert CEFR B1 text simplification specialist with deep understanding of automatic language assessment systems.

Objective

Transform this text to precise B1 level while preserving all original meaning and information. {base_feedback}

B1 Language Requirements

- Vocabulary: 2000–3000 most common English words, avoid academic/formal terms
- Sentences: 15–22 words, can connect 2 related ideas with clear logic
- Grammar: Present perfect (have/has done), simple passive (is/was done), basic conditionals (if...will/would), modals (should, might, could, would)
- Connectors: however, although, while, since, unless, because, so that, even though
- Style: Clear intermediate language that shows reasoning and personal opinions

Strict Level Control

- Above A2: Include abstract concepts with simple explanation, cause-effect relationships, personal opinions with basic justification, intermediate grammar patterns
- Below B2: No academic/formal vocabulary (facilitate → help, demonstrate → show, utilize → use)
- Below B2: No complex conditional structures, no sophisticated argumentation, no specialized terminology without explanation
- Precise B1 Target: Intermediate complexity using everyday vocabulary — never oversimplify to A2, never undersimplify leaving B2+ elements

Critical B1 Differentiators

- From A2: Can handle abstract ideas but explains them simply using common words
- From B2: Uses everyday vocabulary even for complex concepts, avoids formal/academic tone
- B1 Signature: Connects ideas logically but with simple language patterns

Transformation Process

1. Identify all key information and meaning
2. Scan for B2+ vocabulary and replace with B1 common equivalents
3. Convert complex sentences to B1 structures (maximum 2 clauses per sentence)
4. Add simple explanations for any remaining complex concepts
5. Include 2–3 B1 grammar markers per paragraph naturally
6. Verify consistent B1 complexity throughout — no A2 oversimplification, no B2+ elements remaining

Critical

Do not omit, summarize, or change any information. Only change **how** it is expressed to match B1 patterns that automatic CEFR classifiers consistently recognize as B1 level.

Instruction

Return only the simplified text. Do not include any other comments, notes, or additional information.

Text to Simplify

Source	Text
Original	Your dreams might show you all kinds of insights into things that you didn't realise you were thinking about. Common dreams like being able to fly or falling, your teeth falling out or having no clothes on in a public place probably mean something similar in most people. But the key to understanding exactly what they mean to you is to connect them to the events and feelings in your daily life.
Kimi K2	Your dreams can show you things. They show what you think about. Many people have the same dreams. They dream they can fly. They dream they fall. They dream their teeth fall out. They dream they have no clothes in a public place. These dreams mean the same thing for most people. But you must connect the dream to your life. You must connect it to your feelings. Then you will know what it means to you.
Claude Sonnet 4	Your dreams can show you new ideas about things. You did not know you were thinking about these things before. Many people have the same dreams. Some people dream they can fly. Some people dream they are falling. Some people dream their teeth fall out. Some people dream they have no clothes on when other people can see them. These dreams probably mean similar things for most people. But you must think about your own life to understand your dreams. You must think about what happens in your day. You must think about how you feel every day. Then you can understand what your dreams mean for you.
LLaMA 3.1 8B	Your dreams can show you things you didn't think about before. Many people have dreams like flying or falling. These dreams are probably the same for everyone. To understand your dreams, you must think about what you do and feel every day.
Human Reference	People have different dreams when they sleep. Some common dreams are flying, falling, losing teeth, or being outside with no clothes. These dreams often mean the same thing for many people. To know what your dream means, think about what is happening in your life and how you feel.

Table 2: Example simplifications produced by different models compared to the human reference.

Example 1: Oversimplification (very short output)	
Original	Social media, magazines and shop windows bombard people daily with things to buy, and British consumers are buying more clothes and shoes than ever before. Online shopping means it is easy for customers to buy without thinking, while major brands offer such cheap clothes that they can be treated like disposable items – worn two or three times and then thrown away.
Reference (human)	Every day, people see many clothes and shoes on social media, in magazines, and in shops. In the UK, people are buying more than before. Online shopping makes it easy to buy without thinking. Big brands sell clothes that are very cheap. Some people wear them only a few times and then throw them away.
1B No Fine-tune	People buy things on line. They do not want these things.
1B Fine-tuned	People see many ads on social media and in shops every day. They want to buy things. British people buy more clothes and shoes now than ever before. Online shopping is easy. You don't need to think about it. Major brands sell cheap clothes. These clothes can be used once or twice and then thrown away.
Example 2: Added explanations despite prompt restrictions	
Original	A large number of bridge accidents occur during the construction of the bridge itself. These accidents are often due to an error made by the engineers, such as a miscalculation. The bridge collapses under its own weight, and this can be deadly for the workers on it at the time.
Reference (human)	Many accidents with bridges happen while they are being built. These accidents often happen because engineers make mistakes, like wrong numbers. The bridge can fall down from its own weight. This is very dangerous for the workers on the bridge.
1B No Fine-tune	Here is the simplified text: "Many people get hurt when they drive over bridges. Sometimes there is a mistake in how the bridge is built, like making a wrong calculation. This makes the bridge fall down and it can be very bad for those who work on it."
1B Fine-tuned	Many people get hurt in bridge building accidents. This happens because of a mistake made by the builders. They make a wrong calculation, and the bridge falls down. It can be very dangerous for the workers who are working on it.

Table 3: Outputs from LLaMA 3.2 1B base vs. fine-tuned models compared against original and human reference simplifications. Example 1 illustrates oversimplification by the base model, while Example 2 shows unnecessary explanations being added despite task instructions.

HOPE at TSAR 2025 Shared Task: Balancing Control and Complexity in Readability-Controlled Text Simplification

Sujal Maharjan
Taylor's University
Subang Jaya, Malaysia

sujalmaharjan007@gmail.com

Astha Shrestha
Taylor's University
Subang Jaya, Malaysia

aasthashrestha688@gmail.com

Abstract

This paper describes our submissions to the TSAR 2025 Shared Task on Readability-Controlled Text Simplification. We present a comparative study of three architectures: a rule-based *Baseline*, a heuristic-driven *Expert* system, and a zero-shot generative *T5 Pipeline* with a semantic guardrail. Our analysis of the official results shows a clear trade-off between the controllability of rule-based systems and the fluency of generative models. We detect that in this zero-shot instance, our simpler, confined systems achieved superior meaning preservation scores compared to the powerful but less predictable generative model. We present a diagnostic failure analysis centered in our actual system outputs, illustrating how different architectural choices result in distinct error patterns, such as under-simplification, information loss via heuristics, and semantic drift.

1 Introduction

The Shared Task on Readability-Controlled Text Simplification at the Fourth Workshop on Text Simplification, Accessibility, and Readability (TSAR 2025) (Alva-Manchego, Fernando et al., 2025) requires systems to simplify a given text to a specified Common European Framework of Reference (CEFR) level while preserving meaning. This creates a conflict between reducing linguistic complexity and maintaining semantic fidelity.

To evaluate the trade-off between generative power and controllability, we engineered three systems: a deterministic *Baseline*, a heuristic-driven *Expert* system, and a zero-shot *Generative Pipeline* based on T5. Our core empirical observation is that in this zero-shot setting, constrained and interpretable approaches perform better than a powerful generative pipeline on official meaning preservation metrics. We explore the specific failure modes of each system, illustrating a definite trade-off between generative

power and semantic controllability, and propose directions for future hybrid approaches.

2 Related Work

Text Simplification (TS) has transformed from early rule-based systems primarily focused on lexical and syntactic transformations (Siddharthan, 2014) to the current paradigm which is dominated by neural sequence-to-sequence models (Nisioi et al., 2017). Models like T5 (Raffel et al., 2019) and BART (Lewis et al., 2020), pre-trained on vast text corpora, have become the de facto standard, achieving state-of-the-art fluency when fine-tuned on task-specific data.

However, a key challenge in modern TS is **controllability** (Maddela et al., 2021). While large language models are proficient at fluent paraphrasing, guiding them to simplify to a precise readability level without sacrificing semantic fidelity remains an open problem. Researchers have explored techniques like explicit target-level prompting, but models can still "hallucinate" or deviate from the underlying meaning. Our work directly addresses the issue, questioning whether a powerful generative model utilized in a zero-shot setting is the effective tool for a task with strict semantic constraints, echoing findings in other domains where simpler models can be more robust (Rudin, 2019).

3 System Descriptions

We implemented three systems with increasing complexity to explore the trade-off between control and generative power.

3.1 System 1: Baseline

Our baseline is a deterministic pipeline serving as a high-precision, low-recall benchmark. It performs three operations: (1) lowercasing the input text, (2) applying a curated 10-word substi-

tution dictionary (e.g., ‘approximately’ -> ‘about’), and (3) capitalizing the first letter of the output. This system performs minimal, safe edits designed to maximize meaning preservation.

3.2 System 2: Expert

The Expert system extends the baseline with two key features. First, it uses a more intricate, two-tiered substitution lexicon of approximately 100 entries, governed by resources such as the English Vocabulary Profile (EVP) to ensure CEFR-appropriateness. For lower proficiency targets (A1/A2), a more robust set of substitutions is used (e.g., ‘substantial’ -> ‘big’). For higher levels (B1+), a more conservative lexicon is utilized to maintain nuance.

Second, it implements a heuristic for structural simplification: for texts targeted at A1/A2 levels exceeding 50 words, the system truncates the output to the first three sentences. This rule is a direct, interpretable method to regulate output length, a crucial aspect of lower-level texts, though it comes with the risk of information loss.

3.3 System 3: Generative Pipeline

This multi-stage pipeline was used in a zero-shot setting, as the shared task provided no official training data. The stages are:

1. **Lexical/Syntactic Preprocessing:** The input text is first simplified utilizing the same non-truncating rules as the Expert system.
2. **CEFR-Aware Prompting:** A T5-base model is guided by a dynamic instructional prompt. For example, to simplify a text for a B1 target, the prompt is: ‘Simplify this text using clear language for intermediate level: [original text]’.
3. **Semantic Guardrail:** An embedding-based check is performed. We compute the cosine similarity between the original and T5-generated text embeddings using the ‘all-MiniLM-L6-v2’ model. If the similarity is below a threshold of 0.75, the T5 output is rejected, and the system reverts to the pre-processed text from stage (1). This mechanism is a countermeasure against significant semantic drift.

4 Experimental Setup

Data: We present results on the official TSAR 2025 test set (200 instances; targets A2/B1). The

human-simplified ‘reference’ texts were employed in the official scoring and in our diagnostic analysis.

Official Metrics: We present the official AUTORANK composite score and its components as provided by the organizers.

- **AUTORANK:** Official composite metric used by the shared task (lower is better).
- **MB-orig/ref:** MeaningBERT score against the original and reference texts, respectively (higher is better).
- **RMSE:** Root-mean-square error for CEFR level prediction (lower is better).

Implementation: Appendix A contains reproducibility notes.

5 Results and Analysis

Table 1: Official final results on the TSAR 2025 test set. The official AUTORANK score is a composite metric where lower is better. Best scores for each metric are in bold.

System	RMSE ↓	MB-orig ↑	MB-ref ↑	AUTORANK ↓
Baseline	1.428	0.945	0.815	12.230
Expert	1.402	0.919	0.797	13.260
Generative Pipeline	1.600	0.841	0.730	19.030

Table 1 presents the official final scores for our three systems. The results demonstrate a clear pattern: while the Expert system achieved the best CEFR compliance (lowest RMSE), the simpler Baseline system was superior on both meaning preservation metrics (MB-orig, MB-ref) and, consequently, the final AUTORANK composite score. The Generative Pipeline performed worst across all official metrics.

5.1 Diagnostic Failure Analysis

To understand the trade-offs revealed by these scores, we performed a diagnostic failure analysis by comparing system output against the original text and the human-written reference. Table 2 provides a representative example that illustrates the distinct failure modes of each system.

Failure Spectrum: Our analysis confirms different failure modes. The Baseline makes safe but inadequate modifications. The Expert system’s aggressive heuristics improve readability metrics but can severely hamper fidelity. The Generative Pipeline, despite its potential for fluency, is vulnerable to semantic drift through “content omission”, a common failure mode for

Table 2: Diagnostic failure analysis for a representative example (text_id: 22-a2, Target: A2).

System & Input	Output Text (Actual)	Error Type and Analysis
Original	One of the most interesting cases of wild animals living in a city are the wild dogs of Moscow... They have learned that it is safer to cross the street with people...	A descriptive, multi-fact text of 97 words.
Human Reference	There are many wild dogs in Moscow, about 35,000. Some were born wild, and some were pets... They found that the dogs have learned how to live in the city...	Simplifies vocabulary and rephrases for clarity while retaining all key facts. Serves as the gold standard for simplification.
Baseline	One of the most interesting cases... Moscow. In Moscow, there are about 35,000 wild dogs... they have learned that it is safer to cross the street...	Under-simplification: Only performed lexical substitution ('approximately' -> 'about'). Failed to simplify sentence structure. This explains its high MB-orig score (0.945) but lower MB-ref score (0.815).
Expert	One of the most interesting cases... pets that have been abandoned by their owners.	Information Loss via Heuristic: Our sentence truncation rule for long A2 texts deleted all facts about the dogs' adaptation to urban life. This aggressive rule explains the system's improved RMSE (1.402) but also its drop in meaning preservation scores.
Generative Pipeline	One of the most interesting cases of wild animals living in a city are the wild dogs of Moscow. In Moscow, there are around 35,000 wild dogs living on the streets.	Content Omission: The T5 model fluently paraphrased the first two sentences but omitted all subsequent information about the dogs' origins and learned behaviors. This explains why it received the worst scores on all meaning preservation metrics (MB-orig: 0.841, MB-ref: 0.730).

zero-shot models that we incorrectly labeled "hallucination" in a preliminary analysis. This confirms the reviewer's observation and highlights the difficulty of ensuring factual consistency with generative models.

6 Semantic Guardrail: Design and Limitations

We implemented an embedding-cosine guardrail (model: 'all-MiniLM-L6-v2'; cutoff = 0.75). Its advantages are speed and ease of deployment, and it successfully screens out many gross semantic deviations. However, our analysis reveals its limitations: cosine similarity is an imperfect proxy for entailment. As seen in the example in Table 2,

a text can remain topically similar while missing critical facts, a subtlety that the guardrail can miss. This indicates that thresholds are dataset-specific and that more robust verification techniques are needed for high-stakes applications.

7 Conclusion

Our comparative analysis on the TSAR 2025 Shared Task shows a clear trade-off between control and fluency in zero-shot readability-controlled text simplification. Our findings empirically illustrates that in the absence of fine-tuning data, simpler, interpretable approaches can be more robust for semantic fidelity. A simple `Baseline` preserved meaning but was insufficient for structural

simplification. A heuristic-driven `Expert` system enhanced readability metrics but caused significant information loss. Finally, a `Generative Pipeline` offered fluent paraphrasing but was highly vulnerable to content omission, resulting in the lowest meaning preservation scores. This implies that for tasks with stringent semantic constraints, the controllability of simpler systems provides a distinct advantage.

Limitations and Future Work

Limitations: The primary limitation of this study is the lack of a large-scale human evaluation to confirm whether the automated metrics, including the official `AUTORANK`, fully align with human judgments of simplification quality. While our analysis uses human-written references for scoring, it does not include direct human ratings of our systems' outputs. Furthermore, our analysis revealed that the embedding-based guardrail, while effective at catching major deviations, is an imperfect proxy for fine-grained semantic fidelity.

Future Work: We propose two main directions. First, exploring hybrid systems that balance control and fluency, for instance, through lexicon-constrained decoding to guide generative models away from factual errors. Second, developing stronger, automated fidelity checks. Our analysis showed that cosine similarity can be insufficient; future work should investigate using Natural Language Inference (NLI) or Question-Answering (QA) models to verify the preservation of key facts. Validating these more advanced automated metrics against targeted human evaluation will be a critical next step for the field.

Lay Summary

Making complicated text easier to read is important for everything from education to making public information more accessible. This process is called text simplification. The challenge is not just to make text simpler, but to simplify it for a specific reading level like for a beginner versus an intermediate learner—without changing the original meaning.

Scientists use different tools for this task. Some use simple, strict rules, like swapping a hard word for an easy one. Others use powerful Artificial Intelligence (AI) models, similar to ChatGPT, which can fluently rewrite entire sentences. We wanted to find out which approach works best for this con-

trolled simplification task, especially when there is no specific training data available. Is the most powerful AI always the best choice when preserving the original meaning is critical?

To answer this, we built and compared three systems: a `Baseline` system with just a few word-swapping rules, a smarter `Expert` system with more rules (including one to shorten long texts), and a powerful AI `Generative Pipeline`. Our study found that the simpler, rule-based systems were surprisingly better at keeping the original meaning of the text. The powerful AI, while often producing fluent sentences, made significant errors by deleting important information a problem we call 'content omission'. Our `Expert` system also lost information when its rule to shorten long texts was too aggressive. The safest system was the simplest `Baseline`, which made only minor changes but never altered the core message.

Our findings are important for developers building tools for education and accessibility. They show that for tasks where accuracy is crucial, relying on simple, predictable rules can be more reliable than using a complex AI that you can't fully control. The best path forward may be to create hybrid systems that combine the safety of rules with the fluency of modern AI.

References

- Alva-Manchego, Fernando, Stodden, Regina, Imperial, Joseph Marvin, Barayan, Abdullah, North, Kai, and Tayyar Madabushi, Harish. 2025. Findings of the TSAR 2025 shared task on readability-controlled text simplification. In *Proceedings of the Fourth Workshop on Text Simplification, Accessibility, and Readability (TSAR 2025)*, Suzhou, China. Association for Computational Linguistics.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. [BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.
- Mounica Maddela, Fernando Alva-Manchego, and Wei Xu. 2021. [Controllable text simplification with explicit paraphrasing](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3536–3553, Online. Association for Computational Linguistics.

Sergiu Nisioi, Sanja Štajner, Simone Paolo Ponzetto, and Liviu P. Dinu. 2017. [Exploring neural text simplification models](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 85–91, Vancouver, Canada. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. [Exploring the limits of transfer learning with a unified text-to-text transformer](#). *CoRR*, abs/1910.10683.

Cynthia Rudin. 2019. [Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead](#). *Preprint*, arXiv:1811.10154.

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

A Reproducibility Notes

A.1 Code and Data Availability

To facilitate full replication, the complete source code for all three systems, the final system outputs, and the analysis scripts utilized in this paper are publicly available in a GitHub repository under an MIT License. The repository can be accessed at: <https://github.com/SUJAL390/hope-tsar-emnlp2025>

A.2 Dependencies and Hyperparameters

- **General:** Python 3.8+, numpy, pandas.
- **Libraries:** transformers (v4.55+), evaluate, scikit-learn, torch.
- **T5 decoding:** model t5-base; decoding parameters: temperature = 0.7, top- p = 0.9.
- **Semantic guardrail:** model sentence-transformers/all-MiniLM-L6-v2; cutoff = 0.75.

A.3 Recommended Guardrail Validation

To validate a guardrail’s effectiveness, one should sample N accepted and N rejected outputs, annotate them for meaning preservation (binary) and then compute precision and recall. We recommend $N=100$ for an initial check.

Know-AI at TSAR 2025 Shared Task: Text Simplification with Difficulty Critics

Yiheng Wu[†] Jue Hou^{†‡} Anisia Katinskaia[†] Roman Yangarber[‡]

[†]Department of Computer Science

[‡]Department of Digital Humanities

University of Helsinki, Finland

first.last@helsinki.fi

Abstract

Text simplification is an active research topic with applications in many domains. In a simplification pipeline, assessment of text difficulty plays a crucial role as a quality control mechanism: it acts as a “critic,” and guides models to generate text at the difficulty level that is required by the user. In this paper, we present a LLM-based Text Simplification pipeline using several difficulty critics. We evaluate the pipeline using the TSAR shared task dataset and discuss the challenges in building models for assessment of text difficulty and simplification, including the construction of corpora for training difficulty models.

1 Introduction

Text simplification is a widely studied task in natural language processing (NLP), with applications in accessibility, education, and communication. It is important in many applications where the users—e.g., non-native speakers—struggle to understand complex or standard language. The goal is to reduce the linguistic complexity of a text, while maintaining the original text’s core meaning and coherence. Increasingly, official legislation in Europe ([Inclusion Europe](#)) requires government organizations, NGOs and other public agencies to provide information to clients in clear and accessible form, including for readers who may be unable to understand *standard* language. We are motivated especially by applications of simplification in second-language (L2) education, where *personalized* learning is supported by adapting text to the learner’s proficiency level ([Katinskaia and Yangarber, 2018](#); [Hou et al., 2019](#)).

Our simplification pipeline,¹ shown in Figure 1, uses a critic consisting of two parts: (a) *difficulty*—it evaluates the difficulty level of a text simplified by a large language model (LLM), and (b) *semantic*

¹[simplification.py](#)

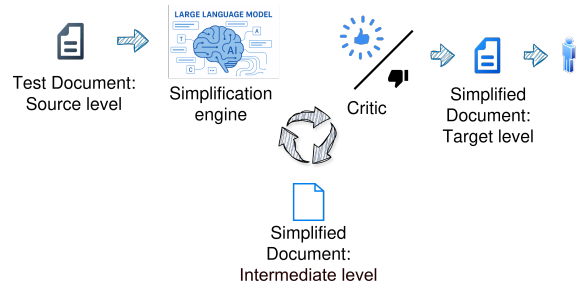


Figure 1: Overview of simplification pipeline.

similarity—it checks how well the simplified text preserves the semantics/meaning of the original text. This framework was introduced in ([Katinskaia et al., 2025](#)), in L2 education. In this paper, we adapt the framework for simplification in English. The pipeline iteratively attempts to generate a “simplified” version of an input text. If the generated text is above the target level of difficulty, then feedback—including the generated text and its currently assessed level—is sent back to the LLM to revise the output. The pipeline makes several attempts at simplification to reach the target difficulty level. We experiment with several critics in the pipeline, including an open-source transformer-based model that classifies text by difficulty level, and a regression model that we train using English-language texts labeled with difficulty levels.

The paper is organized as follows: Section 2 gives a brief overview of related work. Section 3 describes the shared task and the evaluation methods. Section 4 presents the architecture of our simplification pipeline. Section 4.1 describes the experiments with controlling the behavior of a LLM via the difficulty critic. Section 5 presents results and analysis. Section 6 concludes the paper and discusses directions for future work.

2 Related Work

Prior approaches to text simplification relied on assessment of text difficulty to identify sentences requiring simplification. For example, [Gasperin et al. \(2009\)](#) trained a model to detect linguistically complex sentences; [Aluísio et al. \(2010\)](#) developed readability assessment tools to support simplifying texts for low-literacy readers. Readability metrics have also been incorporated directly into rule-based simplifiers: [Woodsend and Lapata \(2011\)](#) integrate the Flesch-Kincaid grade formula ([Flesch, 1948](#)) into optimization-based simplification.

More recent approaches to simplification leverage readability predictors as feedback within generation loops. [Alkaldi and Inkpen \(2023\)](#) use a readability classifier in a reinforcement learning framework to iteratively simplify text until it reaches the desired difficulty. Large-scale neural systems have combined readability prediction with controllable generation techniques to produce text at the target difficulty level ([Agrawal and Carpuat, 2023](#)).

3 Task Description

The Shared Task on Readability-Controlled Text Simplification ([Alva-Manchego et al., 2025](#)) involves simplifying English-language paragraphs written at upper-intermediate or advanced levels. Participants are required to produce simplified versions at a target readability, specified as a CEFR level: Common European Framework of Reference for Languages ([Council of Europe, 2001](#)).

Our experiments are based on the *test* dataset provided by the TSAR shared task. The test set consists of English paragraphs at level B2 or higher, each associated with a target level (A1, A2, or B1). No training data, and no reference simplifications are provided. The evaluation involves measuring multiple aspects of the simplified texts:

- Compliance with target CEFR level is determined using a CEFR-level classifier, which checks whether the generated paragraph meets the specified target proficiency level.
- Meaning preservation is assessed via semantic similarity between the *original* source paragraph and the simplification, ensuring that the essential meaning is retained.
- Similarity to a *reference* simplification is computed, to quantify how closely the system output matches the provided references.

These metrics are calculated using the official

evaluation scripts released by the shared task organizers with the test dataset. The semantic similarity in the evaluation scripts uses meaningbert ([Beauchemin et al., 2023](#)). meaningbert is a BERT-based semantic similarity model that measures how well meaning is preserved between two texts, particularly for tasks such as text simplification and paraphrase assessment.

4 System Overview

We next describe how we use the critic model to guide in LLM-based text simplification pipeline (see Figure 1).

The pipeline begins by determining the difficulty of a source text, either with a difficulty model or manual annotation. The text, together with the target CEFR level and a prompt, is passed to a LLM, which produces a candidate output. The critic model evaluates the candidate’s difficulty; if it matches the target level, the process ends. Otherwise, the LLM is re-prompted with the previous output and the discrepancy from the target. This loop continues for up to N iterations—a predefined maximum, to balance between cost and quality. The system then outputs either a satisfactory simplification, or an error if the target is not reached.

4.1 Methodology

In the context of the shared task, we experiment with two difficulty assessment models in the pipeline critic:

- **Statistical model:** we use the Flesch–Kincaid Reading Ease score ([Flesch, 1948](#); [Kincaid et al., 1975](#)), implemented in the Spacy library.² This model assigns a numeric readability value based on word and sentence length, with higher scores indicating simpler text. To relate these scores to CEFR levels, we apply an approximate mapping³ shown in Table 1. This enables us to interpret Flesch–Kincaid scores within a CEFR framework and use them as difficulty estimates in the simplification pipeline.
- **Transformer-based model:** we use the model AllLang2-Cefr2,⁴ which classifies its input into the 6 CEFR levels: A1–C2. This model is also used in the official evaluation in the Shared Task. We use its prediction on the

²spacy.io/universe/project/spacy_readability

³Flesch-Kincaid readability analysis and CEFR map

⁴ModernBERT-base-reference_AllLang2-Cefr2

LLM-generated text to determine whether to stop the iterative simplification process.

To control the complexity of the generated texts, we use LLM prompts, based on the target CEFR level. Each prompt has three main components, parameterized by the target CEFR level:

- **Role Instruction:** The LLM is instructed to act as an *expert in teaching English*, to adapt English texts for learners to the specified CEFR level (level_target).
- **Output Format:** The LLM must produce a JSON object containing the key "SIMPLIFICATION", to ensure that the result is structured and machine-readable.
- **Adaptation Guidelines:** The LLM is instructed to adapt the input text according to the target CEFR: the prompt contains a description of what the reader can/cannot understand easily (based on the definitions of the CEFR levels). The simplified text should *preserve the meaning* of the original text while matching the target proficiency level.

Detailed prompt templates for all CEFR levels are provided in Appendix A.

To perform the simplification, the pipeline uses GPT-4o (OpenAI, 2024) with the prompts described above. For all test documents, we monitored the simplification process by recording the CEFR level at each iteration and computing the cosine similarity between each intermediate output and the original text.

To measure semantic similarity in the critic—to check how well the simplification preserves meaning—we applied a semantic similarity model all-mpnet-base-v2 (Reimers and Gurevych, 2020), and used a threshold of 0.7 (determined heuristically), retaining only those simplifications that have semantic similarity to the original above this value. Each document was simplified for up to $N = 5$ iterations; the process terminates earlier if the critic judges the text’s difficulty to be at or below the target CEFR level, and its similarity with the original is above the threshold.

5 Results and Analysis

In this section, we examine the results of simplification with different critic models. Beyond exact-match accuracy, we assess how well the predicted difficulty levels match the intended simplification

Flesch-Kincaid	CEFR
90–100	A1
80–89	A2
70–79	B1
60–69	B2
50–59	C1
0–49	C2

Table 1: Mapping from Flesch-Kincaid Reading Ease scores to CEFR levels.

direction. The *Direction Consistency* metric measures whether predictions respect the target level ordering for each input.

Consistent Example			Inconsistent Example		
Target	Pred	Cons.?	Target	Pred	Cons.?
B1	B1	Yes	B1	B1	No
A2	A1		A2	B1	

Table 2: Examples of direction consistency: left = consistent (trend preserved), right = inconsistent (trend violated).

Using the Flesch–Kincaid Reading Ease model as critic in the simplification pipeline, exact match between predicted and target CEFR levels is 38%; only a minority of simplifications reach the target level. Direction consistency measure is 62%, indicating that while the model often misses the exact target, it usually adjusts difficulty in the correct relative direction. Thus, the model offers coarse guidance on text difficulty, but lacks the precision needed for strict level control.

Figure 2 shows that most errors are deviations of ± 1 level, with about 60% of the misclassified samples exceeding the target by one level. Exact matches and predictions one level below are less frequent, and deviations beyond ± 2 levels are rare. Figure 3 illustrates that when the target is A2, outputs often simplify to level B1, while for B1 target, many texts remain at the original difficulty. In rare cases, predictions even drop to A1. These patterns indicate that the model fails to align reliably with CEFR standards, limiting the accuracy of the critic and yielding only modest control over target difficulty in the simplification pipeline.

Using AllLang2-Cefr2 as critic in the simplification pipeline, performance improves. Exact-

Critic model	RMSE	meaningbert	
		Origin.	Refer.
Flesch-Kincaid	0.659	0.801	0.832
AllLang2-Cefr2	0.700	0.821	0.835
Regression	0.600	0.772	0.815

Table 3: Performance of difficulty critics on *simplification*

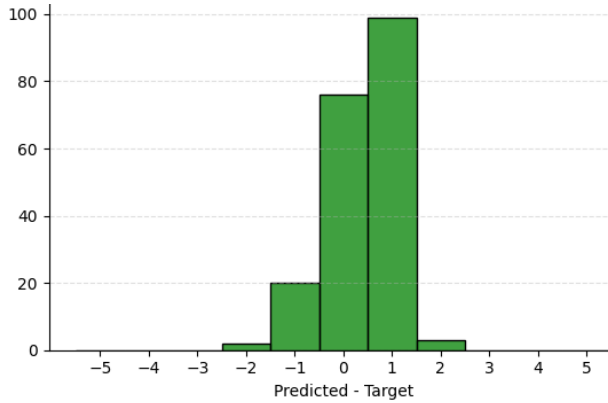


Figure 2: Distribution of the difference between estimated difficulty level of the simplified output and target difficulty level, using Flesch-Kincaid model as critic. X-axis is difference between the estimated level of the output (simplified) text and target level. Y-axis is the *number* of instances in test set.

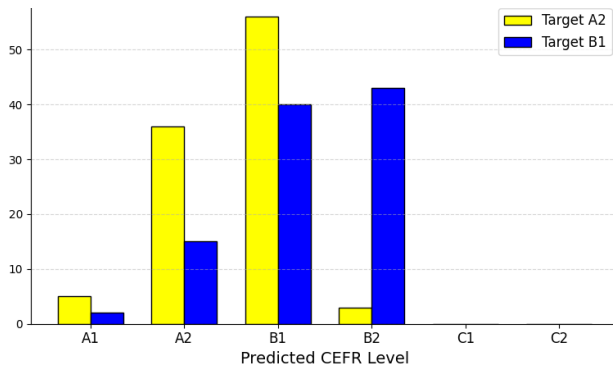


Figure 3: Distribution of estimated difficulty of simplified output texts for different target levels (A2,B1), using Flesch-Kincaid critic. X-axis represents the estimated CEFR levels (A1–C2) of output text; Y-axis indicates the *percentage* of samples at each estimated level. Different target levels are color coded.

match accuracy rises to 57%, well above that of the Spacy critic, while Direction Consistency is 63.5%. As shown in Figure 4, deviations never exceed ± 1 level, with exact matches most common and overshooting by one level less often. Figure 5 shows that when the target is A2, most samples are correctly simplified to A2, with the rest at B1. For target B1, about 60% reach B1, some remain at B2, and the rest overshoot to A2.

The official ranking in the Shared Task is based on (a weighted average of) the three measures reported in Table 3, namely, on correct difficulty and on semantic similarity; *origin* indicates similarity of the simplified text to the original, according to meaningbert; *refer* indicates similarity of the simplified text to reference simplification.

6 Discussion and Future Work

The effectiveness of our proposed pipeline depends on the choice of difficulty assessment model used in the critic, since it guides the simplification pro-

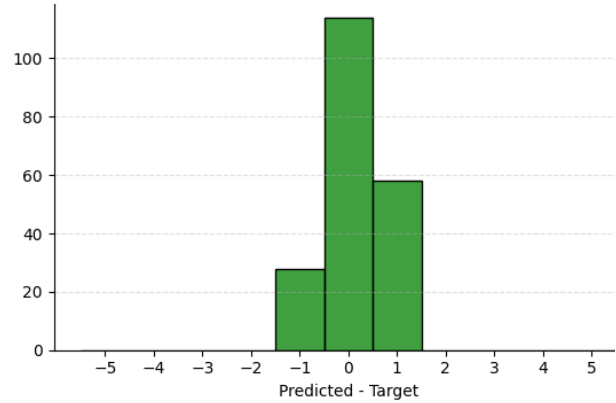


Figure 4: Distribution of the difference between estimated difficulty level of the simplified output and target difficulty level, using AllLang2-Cefr2 as critic.

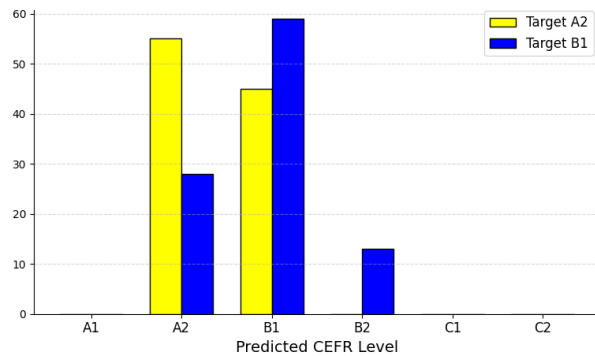


Figure 5: Distribution of estimated difficulty of simplified output texts for different target levels, using AllLang2-Cefr2 as critic.

cess. In addition to the models above, we experimented with training our own difficulty assessment model. Although this approach did not appear in our submissions for the Shared Task, it shows much promise for future work. This section summarizes the lessons learned from this attempt.

First, since no training data were provided for the Shared Task, we construct a training, development and test set—Test Set 1—by taking an existing corpus⁵ described in (Katinskaia et al., 2025) and translating it from Finnish into English, using the OPUS machine translation (MT) toolkit⁶ (Tiedemann et al., 2023). It is crucial to note that we found that the OPUS models are particularly strong at preserving the CEFR levels of the original source text in the MT output text.⁷ We also use the reference set provided by TSAR as a second test

⁵Test Set 1 contains intermediate CEFR levels: A2-B1, B1-B2, etc. For comparability, we applied a special adjustment for AllLang2-Cefr2, in which each intermediate level was “mapped” down to the lower adjacent level.

⁶We use the *sla-eng* MT model (Slavic-to-English).

⁷This property of the OPUS-MT models—that they preserve the CEFR level well from the input to the output text—was confirmed through manual inspection by experts in Simple Language. These findings need to be confirmed more rigorously in quantitative terms, in future work.

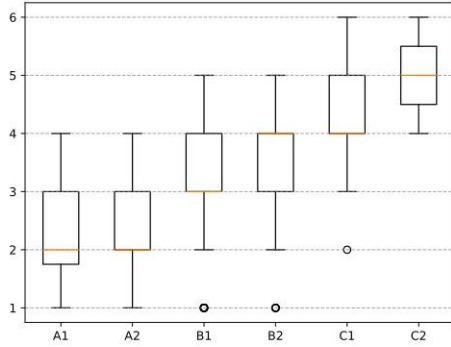


Figure 6: Difficulty estimation using AllLang2-Cefr2 in Test Set 1. Red line means the median of current CEFR level

set—Test Set 2.

Second, following the methodology of [Katin-skaia et al. \(2025\)](#), we train a regression model to predict difficulty. We were unable to gather a sufficient amount of training data and tune our regression model in time for the actual TSAR competition; therefore, as a fallback, we used AllLang2-Cefr2 rather than the regression model as a critic in our submission for the Shared Task.

We next check how well difficulty prediction works—on its own, apart from the simplification task. For Test Set 1, the difficulty prediction results are in Figures 6 and 7. The regression model shows a clear advantage over the AllLang2-Cefr2 model, exhibiting a clear step-wise pattern that aligns well with CEFR levels. It consistently outperforms the baseline across all evaluation metrics. The evaluation metrics for difficulty prediction are shown in the top part of Table 4.

For Test Set 2, the evaluation metrics for difficulty prediction are in the bottom of Table 4. The R^2 values are negative for both models, indicating a limited overall fit to the data. The difficulty prediction results for Set 2 are in Figures 8 and 9.

Several factors may compromise the performance of our regression model. First, the dataset

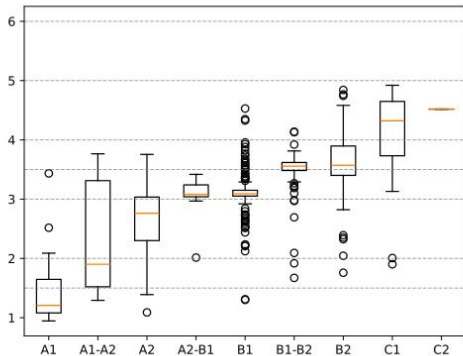


Figure 7: Difficulty estimation error distribution of regression model in Test Set 1.

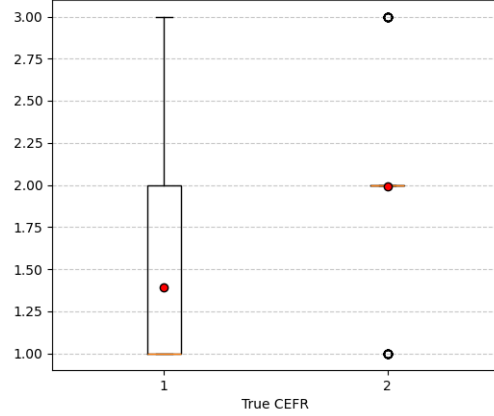


Figure 8: Difficulty estimation using AllLang2-Cefr2 in Test Set 2. Red dot shows the mean score of this CEFR level

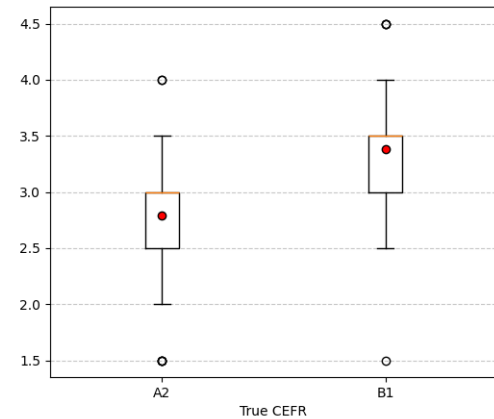


Figure 9: Difficulty estimation using regression model in Test Set 2.

Test	Model	MSE	RMSE	MAE	R^2
1	AllLang2-Cefr2	1.12	1.06	0.82	-0.46
	Regression	0.32	0.57	0.34	0.56
2	AllLang2-Cefr2	0.43	0.66	0.39	-0.72
	Regression	0.64	0.80	0.65	-1.57

Table 4: Performance of difficulty estimation models on two test sets; top section Test Set 1, bottom Test Set 2.

is machine-translated, which may distort the true difficulty of the texts. Ideally, training data is manually annotated for difficulty. However, manual annotation is very complex and time-consuming. Second, the translated dataset is still small, restricting the model’s ability to generalize across different linguistic phenomena.

In future work, we plan to extend the setup relying solely on GPT-4o for text simplification, to consider other models, including smaller models fine-tuned for the simplification task. We will investigate more advanced models to improve the assessment of difficulty, which is central for the simplification pipeline. Larger, more accurate, and more diverse training datasets should further improve performance and generalization.

7 Lay Summary

This study investigates text simplification, in the context of the Shared Task on Text Simplification, Accessibility, and Readability (TSAR).

We present a difficulty-aware simplification pipeline based on large language models (LLMs) and small models for simplification assessment. We use text data in English, of varying levels of difficulty, ranging from A1 to C1 on the CEFR scale. We evaluate performance according to several criteria, including error rates of difficulty assessment models in their assessment of the difficulty of texts in a held-out test set, and the success rates of the simplification pipeline, relative to reference texts provided by the organizers of the shared task.

The paper A. discusses the performance of a number of critic models for assessing difficulty of a text, and B. compares the performance of the simplification pipeline driven by the different critics.

Acknowledgements

This work was supported by Project “*Easy Language for accessible workplace communication*,” funded by BusinessFinland Agency for Technology and Innovation (Grant 4173/31/2024); Project “*Generative AI-Enhanced Knowledge Management in Business*” (GAIK), funded by the European Regional Development Fund (EAKR), and High-performance Computing Project “*Know-AI*,” funded by Academy of Finland (Grant 359285).

References

- Sweta Agrawal and Marine Carpuat. 2023. Controlling pre-trained language models for grade-specific text simplification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12807–12819, Singapore.
- Wejdan Alkaldi and Diana Inkpen. 2023. [Text simplification to specific readability levels](#). *Mathematics*, 11(9):2063.
- Sandra Aluísio, Lucia Specia, Caroline Gasperin, and Carolina Scarton. 2010. Readability assessment for text simplification. In *Proceedings of the 5th Workshop on Innovative Use of NLP for Building Educational Applications*, Los Angeles, California.
- Fernando Alva-Manchego, Regina Stodden, Joseph Marvin Imperial, Abdullah Barayan, Kai North, and Harish Tayyar Madabushi. 2025. Findings of the TSAR 2025 shared task on readability-controlled text simplification. In *Proceedings of the 4th Workshop on Text Simplification, Accessibility, and Readability*, Suzhou, China.
- David Beauchemin, Horacio Saggion, and Richard Khoury. 2023. [Meaningbert: assessing meaning preservation between sentences](#). *Frontiers in Artificial Intelligence*, 6.
- Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge University Press.
- Rudolf Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233.
- Caroline Gasperin, Lucia Specia, Tiago F. Pereira, and Sandra M. Aluísio. 2009. Learning when to simplify sentences for natural text simplification. In *Proceedings of the Encontro Nacional de Inteligência Artificial (ENIA)*, Bento Gonçalves, Brazil.
- Jue Hou, Maximilian W Koppatz, José Maria Hoya Quecedo, Nataliya Stoyanova, Mikhail Kopotev, and Roman Yangarber. 2019. Modeling language learning using specialized Elo ratings. In *BEA: 14th Workshop on Innovative Use of NLP for Building Educational Applications, ACL: 56th annual meeting of Association for Computational Linguistics*.
- Inclusion Europe. Information for all: European standards for making information easy to read and understand. <https://www.inclusion-europe.eu/easy-to-read-standards-guidelines/>. Accessed: 2025-10-12.
- Anisia Katinskaia, Anh-Duc Vu, Jue Hou, Ulla Vanhatalo, Yiheng Wu, and Roman Yangarber. 2025. [Estimation of text difficulty in the context of language learning](#). In *BEA: 20th Workshop on Innovative Use of NLP for Building Educational Applications*, Vienna, Austria.
- Anisia Katinskaia and Roman Yangarber. 2018. Digital cultural heritage and revitalization of endangered Finno-Ugric languages. In *Proceedings of the 3rd Conference on Digital Humanities in the Nordic Countries*, Helsinki, Finland.
- 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 No. RBR875.
- OpenAI. 2024. Gpt-4o system card. <https://arxiv.org/abs/2410.21276>.
- 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 (EMNLP)*, pages 4512–4525.
- Jörg Tiedemann, Mikko Aulamo, Daria Bakshandaeva, Michele Boggia, Stig-Arne Grönroos, Tommi Nieminen, Alessandro Raganato, Yves Scherrer, Raúl Vázquez, and Sami Virpioja. 2023. [Democratizing neural machine translation with opus-mt](#). *Language Resources and Evaluation*, 58:713–755.

Kristian Woodsend and Mirella Lapata. 2011. Learning to simplify sentences with quasi-synchronous grammar and integer programming. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, Edinburgh, Scotland.

A Prompts for CEFR-based Text Simplification

CEFR	Prompt Description
A1	You must always output a JSON object with an "SIMPLIFICATION" key. You are an expert in English language and language teaching. You will be given a text in English. Your task is to read it first and then to provide an adaptation into CEFR level A1. Do not significantly change the meaning of the input text. A1 is the simplest, beginner level with short sentences and easy grammar. Imagine that you are teaching a complete beginner, your adaptation should fit their proficiency level. This is the text to simplify: {text}
A2	You must always output a JSON object with an "SIMPLIFICATION" key. You are an expert in English language and language teaching. You will be given a text in English. Your task is to provide an adaptation into CEFR level A2. Do not significantly change the meaning of the input text. A2 is just above the beginner level and should contain simple grammar and vocabulary. Imagine that you are teaching someone who just started learning the language. This is the text to simplify: {text}
B1	You must always output a JSON object with an "SIMPLIFICATION" key. You are an expert in English language and language teaching. You will be given a text in English. Your task is to provide an adaptation into CEFR level B1. Do not significantly change the meaning of the input text. B1 is an intermediate level. Learners can understand the main points of clear standard input and produce connected text on familiar topics. Adapt the text accordingly. This is the text to simplify: {text}
B2	You must always output a JSON object with an "SIMPLIFICATION" key. You are an expert in English language and language teaching. You will be given a text in English. Your task is to provide an adaptation into CEFR level B2. Do not significantly change the meaning of the input text. B2 corresponds to an upper-intermediate level, allowing complex text understanding and fluent communication. The adapted text should be more advanced than B1 or A2. This is the text to simplify: {text}
C1	You must always output a JSON object with an "SIMPLIFICATION" key. You are an expert in English language and language teaching. You will be given a text in English. Your task is to provide an adaptation into CEFR level C1. Do not significantly change the meaning of the input text. C1 corresponds to an advanced level, capable of producing fluent, well-structured, detailed text with complex grammar and cohesive devices. The adapted text can therefore be more sophisticated and lexically rich. This is the text to simplify: {text}

Table 5: Prompts used for CEFR-based text simplification to target CEFR levels A1–C1. Each prompt instructs the model to adapt the input text to the linguistic characteristics of the target CEFR level.

Author Index

- Alfter, David, 137
Alonso Alemany, Laura, 160
Alva-Manchego, Fernando, 116
Arias Russi, Felipe, 87, 211
Atserias Batalla, Jordi, 160
- Bakker, Jan, 36
Barayan, Abdullah, 116
Barbu, Paul-Gerhard, 173
Bektaş, Fatih, 149
Belem, Catarina, 47
Bezobrazova, Anastasiia, 242
Bian, Kexin, 231
- Cao, Jie, 223
Cheng, Zhaoyi, 19
Cohen Solano, Kevin, 211
Comelles Pujadas, Elisabet, 160
- Dinç, Kutay Arda, 149
- Ebling, Sarah, 1, 19
Eryiğit, Gülşen, 149
- Fujiwara, Yuki, 217
- Gao, Yingqiang, 1
Glenn, Parker, 47
Gombert, Sebastian, 137
- Hayakawa, Akio, 193
Horiguchi, Koki, 217
Hou, Jue, 266
Huynh, Cuong, 223
- Imperial, Joseph Marvin, 116
- Kajiwara, Tomoyuki, 217
Kamps, Jaap, 36
Katinskaia, Anisia, 266
Khallaf, Nouran, 193
Komachi, Mamoru, 231
Kondo, Risa, 217
Kumar, Anoop, 47
- Li, Zihao, 105
Lindner, Lena, 173
Ling, Zhidong, 231
- Lipska-Dieck, Adrianna, 173
Liu, Daben, 47
Liu, Yuhang, 19
Lybarger, Kevin, 70
- Maharjan, Sujal, 261
Manrique, Ruben, 87, 211
Martínez, Paloma, 183
Miyata, Rina, 217
Moreno, Lourdes, 183
- Nisioi, Sergiu, 251
North, Kai, 116
- Orasan, Constantin, 242
Oviedo, Juan Cruz, 160
- Papandreou, Taiki, 36
Przybyła, Piotr, 155
- Rahman, Md Mushfiqur, 70
Romero, Pablo, 105
Roscan, Rares-Alexandru, 251
- Saggion, Horacio, 193
Salazar Lara, Carolina, 87
Samuel, Alf, 47
Sanchez-Escudero, Marco Antonio, 183
Sanchez-Gomez, Jesus M., 183
Shardlow, Matthew, 105
Sharoff, Serge, 193
Shimada, Mao, 231
Shrestha, Astha, 261
Sokova, Daria, 242
Souayed, Belkiss, 1
Stodden, Regina, 116
- Tayyar Madabushi, Harish, 116
- Vajjala, Sowmya, 131
- Wu, Yiheng, 266
- Yangarber, Roman, 266
- Zhang, Mo, 19