ISSN:1738-6764, eISSN:2093-7504

Design and Implementation of an Adaptive English Text Regeneration System Based on CEFR Language Proficiency Levels

Jun Yin [1] **and Myeon-Gyun Cho** [2,*]

[1] Ph.D. Student, Semyung University; junyin@live.cn
[2] Professor, Semyung University; mg_cho@semyung.ac.kr
* Correspondence

https://doi.org/10.5392/IJoC.2025.21.3.178

Manuscript Received 12 June 2025; Received 7 July 2025; Accepted 7 July 2025

Abstract: *This study presents an adaptive English text regeneration system that modifies authentic materials to match learners' CEFR (Common European Framework of Reference for Languages) proficiency levels (A1–C2). It addresses the crucial challenge of accessibility while maintaining the original meaning. Leveraging advancements in large language models (LLMs), our framework employs a three-phase process: first, CEFR-based text analysis utilizing curated vocabulary lists and syntactic metrics; second, multi-level regeneration through the fine-tuned Qwen 2.5 model; and third, rigorous validation of semantic fidelity (achieving a 92% BERT score) and readability. Experimental results with 300 learners indicate significant improvements, with a 32% increase in comprehension for beginner groups and a 25% increase for intermediate groups. Additionally, there is a 40% decrease in self-reported anxiety. The system's real-time processing capability (under 3 seconds per page) ensures practical scalability.Our work makes three key contributions: it establishes the first comprehensive framework covering all six CEFR levels with empirical validation; it integrates pedagogical and psychological principles to boost learner motivation and reduce anxiety; and it demonstrates the effectiveness of progressive complexity scaffolding while setting actionable benchmarks for LLM-driven educational tools. By balancing linguistic precision with psychological benefits—such as increased motivation and confidence—the system enhances the role of AI in language education. Future research will focus on adapting colloquial language and examining longitudinal impacts on knowledge retention, further bridging the gap between authentic content and learner needs.*

Keywords: Large Language Models; CEFR; English Learning; Text Regeneration; Adaptive Learning

1. Introduction

English proficiency has become a critical determinant of academic and professional success in a globalized world, yet non-native learners face persistent challenges when engaging with authentic materials, from linguistic barriers to psychological disengagement [1]. These challenges are compounded by the scarcity of resources that dynamically adapt to learners' evolving proficiency levels, often leading to frustration and diminished motivation [2]. The Common European Framework of Reference for Languages (CEFR) provides a standardized scale (A1–C2) for assessing language competence [3], but the gap between learners' abilities and the complexity of authentic texts remains a significant hurdle [4]. Traditional solutions like graded readers or manually simplified texts lack scalability and fail to preserve the richness of original content [5], highlighting the need for innovative approaches.

Recent advances in Large Language Models (LLMs) offer transformative potential for text adaptation, enabling the regeneration of materials tailored to specific proficiency levels while retaining semantic fidelity [6, 7]. However, current systems predominantly focus on lexical substitution, often distorting meaning or overlooking the nuanced needs of learners across CEFR levels [8]. Cross-linguistic research by Guo et al. [9], Vendeville [10], and Ebling and Rios [11] underscores both the universality and language-specific complexities of text simplification, revealing a gap in comprehensive, psychologically informed adaptation frameworks.

The field remains divided on optimal methodologies: while rule-based systems prioritize control, they sacrifice fluency [2]; neural approaches improve naturalness but struggle with granularity and pedagogical alignment [5, 8]. Controversially, general-purpose LLMs like ChatGPT demonstrate promise in reducing learner anxiety [1] but lack dedicated mechanisms for CEFR-specific complexity control [4]. This divergence underscores the need for a system that harmonizes linguistic precision with empirical insights into learner psychology. Against this backdrop, our study introduces an adaptive text regeneration system that addresses these gaps through three key innovations: CEFR-granular adaptation (A1–C2) via fine-tuned LLMs, progressive scaffolding to support advancing learners, and integrated psychological benefits, including anxiety reduction and motivation enhancement, as empirically validated by Kim and Kim [1]. By bridging technical rigor with pedagogical and psychological principles, our work advances the application of AI in language education.

The principal conclusions demonstrate that our system significantly improves comprehension (32% for beginners, 25% for intermediates) while reducing anxiety by 40%, achieving real-time adaptation (under 3 seconds/page) without compromising semantic fidelity (92% BERT Score). These outcomes set actionable benchmarks for future LLM-driven educational tools.

The remainder of this paper is organized as follows: Section 2 reviews related work; Section 3 describes the system architecture and methodology; Section 4 details the experimental setup and environment. In addition it presents the results and evaluation; and Section 5 concludes with implications and future research directions.

2. Related Work

2.1 CEFR-Based Language Learning

The Common European Framework of Reference for Languages (CEFR) provides a standard for language proficiency. Arase et al. [3] developed methods for CEFR-based sentence difficulty assessment, enabling more precise targeting of materials. Cripwell et al. [7] explored paraphrasing sentences to different complexity levels, aligning with our goal of adapting authentic texts across the full CEFR spectrum (A1-C2) while maintaining semantic integrity and pedagogical soundness, which Kim and Kim [1] linked to improved learning outcomes (p=.04) and strategies (p=.01).

2.2 Large Language Models in Education

Large Language Models (LLMs) offer new educational possibilities. Kim and Kim [1] found ChatGPT enhanced learning effectiveness (p=.04), reduced anxiety (p=.04) and stress (p<.01), and improved confidence (p<.01) and motivation (p=.03) in English language learners. Applications included script writing, grammar checking, and understanding authentic materials.

While LLMs show promise for simplification in specialized domains [6] and generating Easy to Read content [8], and interactive systems are emerging [9], challenges remain. These include maintaining domain-specific accuracy [6], often operating at a sentence level [9], ethical concerns [4], and ensuring precise CEFR complexity control without targeted fine-tuning and an integrated pedagogical framework [1], [4].

2.3 Text Simplification and Adaptation

Text simplification has been a significant area of research in natural language processing, with various approaches evolving over time. Figure 1 comprehensively illustrates the progression of key technologies in this domain, identifies the limitations of existing research, and highlights how our proposed system addresses these challenges.

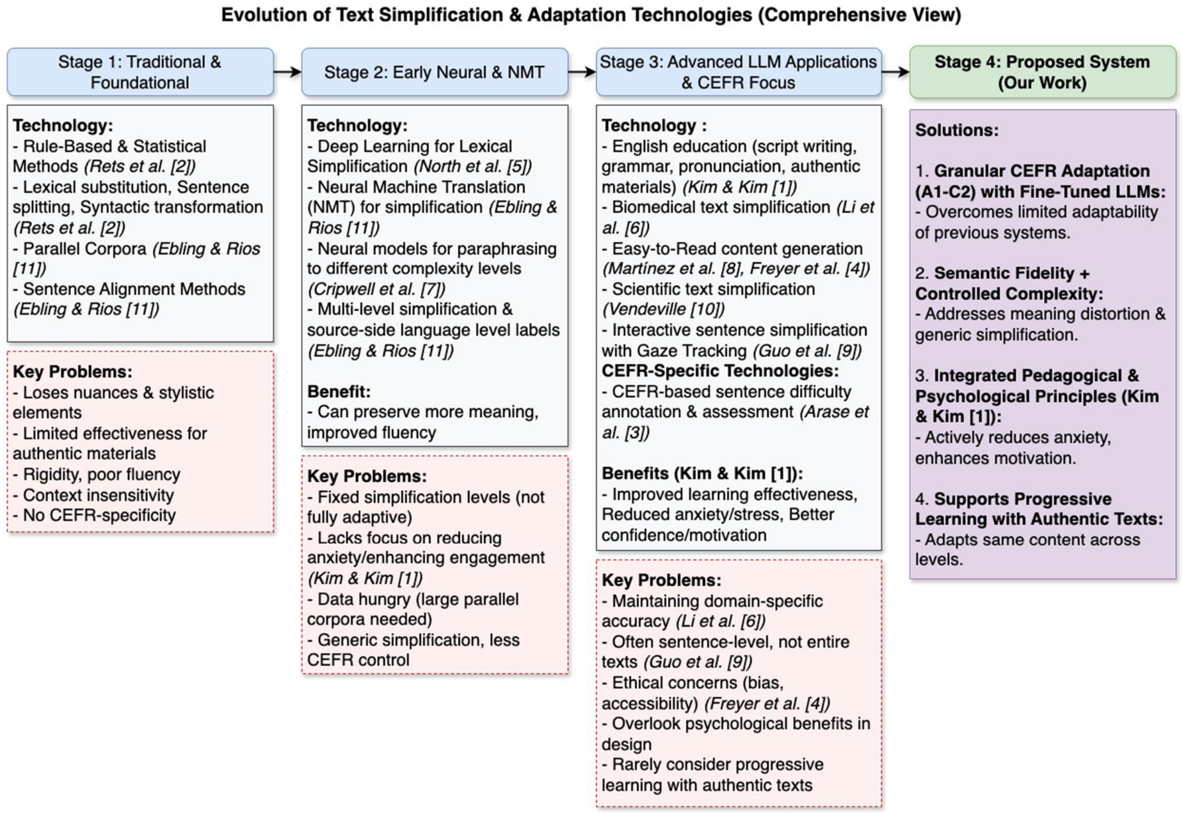

Evolution of Text Simplification & Adaptation Technologies (Comprehensive View)
Stage 1: Traditional & Foundational
Stage 2: Early Neural & NMT
Stage 3: Advanced LLM Applications & CEFR Focus
Stage 4: Proposed System (Our Work)
Technology:
- Rule-Based & Statistical Methods (Rets et al. [2])
- Lexical substitution, Sentence splitting, Syntactic transformation (Rets et al. [2])
- Parallel Corpora (Ebling & Rios [11])
- Sentence Alignment Methods (Ebling & Rios [11])
Key Problems:
- Loses nuances & stylistic elements
- Limited effectiveness for authentic materials
- Rigidity, poor fluency
- Context insensitivity
- No CEFR-specificity
Technology:
- Deep Learning for Lexical Simplification (North et al. [5])
- Neural Machine Translation (NMT) for simplification (Ebling & Rios [11])
- Neural models for paraphrasing to different complexity levels (Cripwell et al. [7])
- Multi-level simplification & source-side language level labels (Ebling & Rios [11])
Benefit:
- Can preserve more meaning, improved fluency
Key Problems:
- Fixed simplification levels (not fully adaptive)
- Lacks focus on reducing anxiety/enhancing engagement (Kim & Kim [1])
- Data hungry (large parallel corpora needed)
- Generic simplification, less CEFR control
Technology :
- English education (script writing, grammar, pronunciation, authentic materials) (Kim & Kim [1])
- Biomedical text simplification (Li et al. [6])
- Easy-to-Read content generation (Martínez et al. [8], Freyer et al. [4])
- Scientific text simplification (Vendeville [10])
- Interactive sentence simplification with Gaze Tracking (Guo et al. [9])
CEFR-Specific Technologies:
- CEFR-based sentence difficulty annotation & assessment (Arase et al. [3])
Benefits (Kim & Kim [1]):
- Improved learning effectiveness, Reduced anxiety/stress, Better confidence/motivation
Key Problems:
- Maintaining domain-specific accuracy (Li et al. [6])
- Often sentence-level, not entire texts (Guo et al. [9])
- Ethical concerns (bias, accessibility) (Freyer et al. [4])
- Overlook psychological benefits in design
- Rarely consider progressive learning with authentic texts
Solutions:
1. Granular CEFR Adaptation (A1-C2) with Fine-Tuned LLMs:
- Overcomes limited adaptability of previous systems.
2. Semantic Fidelity + Controlled Complexity:
- Addresses meaning distortion & generic simplification.
3. Integrated Pedagogical & Psychological Principles (Kim & Kim [1]):
- Actively reduces anxiety, enhances motivation.
4. Supports Progressive Learning with Authentic Texts:
- Adapts same content across levels.

Figure 1. Text Adaptation: Evolution, Gaps, & Solutions (Modular View)

Figure 1 reveals four distinct developmental stages of text adaptation technologies, each with specific characteristics and limitations:

Stage 1: Traditional & Foundational Methods

Early approaches relied primarily on rule-based methods and statistical techniques [2], focusing on lexical substitution, sentence splitting, and syntactic transformation. Research by Rets et al. [2] demonstrates that while these methods established the foundation for text simplification [12], they struggled with preserving nuance, maintaining fluency, and providing CEFR-specific adaptation, making them inadequate for learners seeking authentic materials. Similarly, the parallel corpora and sentence alignment methods explored by Ebling and Rios [11] faced comparable challenges, particularly in cross-linguistic applications.

Stage 2: Early Neural Networks & Neural Machine Translation

Deep learning approaches surveyed by North et al. [5] improved meaning preservation but still lacked fine-grained CEFR level control and pedagogical awareness. The paraphrasing models for different complexity levels developed by Cripwell et al. [7] and the neural machine translation methods by Ebling and Rios [11] typically offered fixed simplification levels rather than adapting to individual proficiency—a key factor for learner engagement, as emphasized by Kim and Kim [1].

Stage 3: Advanced LLM Applications & CEFR Focus

Recent research has begun leveraging Large Language Models (LLMs) with attention to CEFR standards. Kim and Kim's [1] empirical study demonstrated that LLM applications in English education (script writing, grammar checking, pronunciation correction) significantly reduced learner anxiety ($p=.04$) and enhanced learning motivation ($p=.03$). Arase et al. [3] developed CEFR-based sentence difficulty annotation and assessment methods, providing a foundation for precise language level control. However, studies by Li et al. [6] on biomedical text simplification, Martínez et al. [8] and Freyer et al. [4] on Easy-to-Read content generation, Vendeville [10] on scientific text simplification, and Guo et al. [9] on interactive sentence simplification reveal that existing systems still face multiple challenges: maintaining domain-specific accuracy [6], often operating only at the sentence level rather than with entire texts [9], ethical concerns [4], overlooking psychological benefits in design, and rarely considering progressive learning with authentic texts.

Stage 4: Our Proposed System

Based on a comprehensive analysis of the limitations in the previous three stages, our research motivation clearly targets four key gaps: (1) lack of fine-grained CEFR level adaptation; (2) neglect of the psychological aspects of language learning; (3) absence of mechanisms supporting progressive learning; and (4) LLMs not specifically fine-tuned for CEFR standards.

To address these limitations, our system makes the following academic contributions through innovation:

1. CEFR-Targeted LLM Fine-Tuning (A1-C2): We have fine-tuned the QWen 2.5 model [13] on CEFR-classified datasets to generate content precisely tailored to each CEFR level while maintaining semantic fidelity, overcoming the limited adaptability of previous systems.
2. Semantic Fidelity with Controlled Complexity: Our system preserves the core meaning of original texts while precisely controlling linguistic complexity, addressing issues of meaning distortion and generic simplification.
3. Integrated Pedagogical & Psychological Principles: Based on empirical research by Kim and Kim [1], our system actively incorporates psychological benefits such as anxiety reduction and motivation enhancement—critical aspects overlooked by previous systems.
4. Progressive Learning Support with Authentic Texts: Our system can adapt the same content across different CEFR levels, providing learners with a scaffolded learning pathway from simple to complex, supporting gradual language proficiency development.

Through these innovations, our system not only addresses the limitations of existing technologies but also establishes actionable benchmarks for LLM applications in language education, advancing the field of AI in educational contexts.

2.4 Gaps in Existing Research

Existing research shows gaps in adaptation granularity, integration of psychological factors [1], support for progressive learning, and CEFR-specific LLM fine-tuning. Our system addresses these by:

1. CEFR-Targeted LLM Fine-Tuning: Fine-tuning QWen 2.5 on CEFR-classified datasets for precise A1-C2 level-appropriate English content generation, ensuring semantic fidelity and appropriate linguistic complexity.
2. Progressive Learning Support: Adapting authentic content across all CEFR levels for a scaffolded learning journey.
3. Integrated Pedagogical/Psychological Principles: Incorporating principles from Kim & Kim [1] to reduce learner anxiety and enhance motivation via accurately matched content.

This holistic approach, leveraging fine-tuned QWen 2.5, aims to create a more effective, personalized, and supportive tool for English language learners.

3. System Architecture and Methodology

3.1 System Overview

The adaptive English text regeneration system shown in Figure 2 is designed to overcome limitations in existing language learning technologies. Its architecture and methodology are built on two core innovations: first, the ability to precisely adapt texts to CEFR levels (A1-C2) using specially fine-tuned large language models (LLMs), and second, an integrated psychological benefits framework that delivers carefully customized content to enhance learner motivation and reduce anxiety.

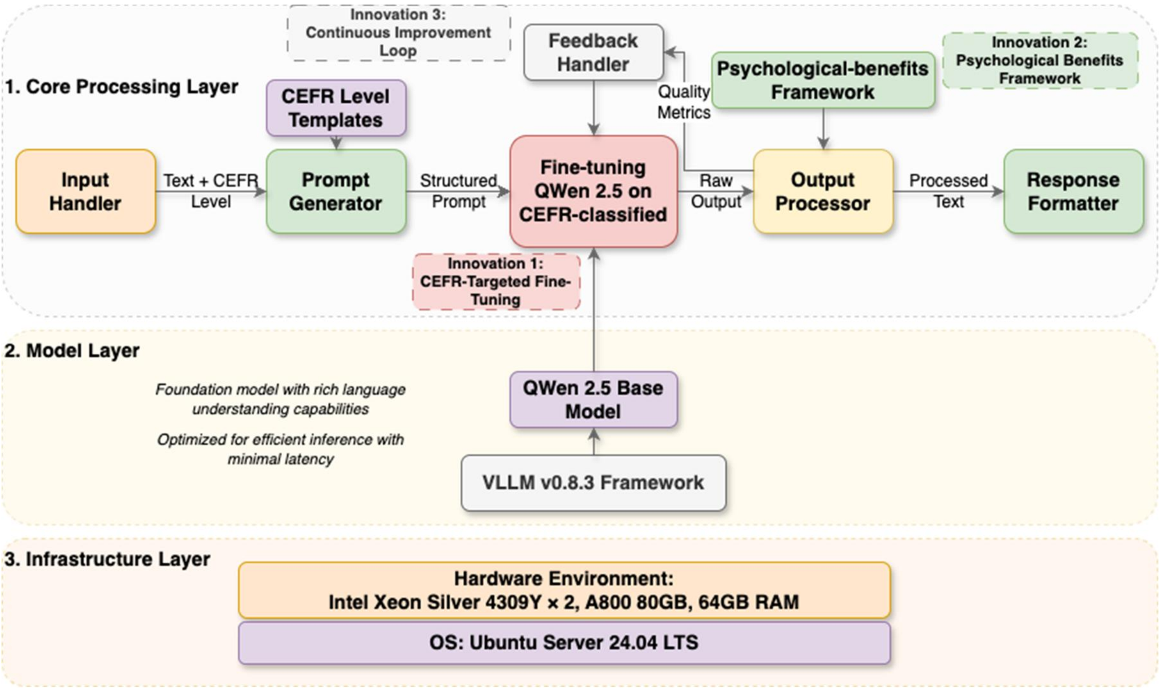

1. Core Processing Layer
Innovation 3: Continuous Improvement Loop
Feedback Handler
Quality Metrics
Psychological-benefits Framework
Innovation 2: Psychological Benefits Framework
CEFR Level Templates
Input Handler
Text + CEFR Level
Prompt Generator
Structured Prompt
Fine-tuning QWen 2.5 on CEFR-classified
Raw Output
Output Processor
Processed Text
Response Formatter
Innovation 1: CEFR-Targeted Fine-Tuning
2. Model Layer
Foundation model with rich language understanding capabilities
Optimized for efficient inference with minimal latency
QWen 2.5 Base Model
VLLM v0.8.3 Framework
3. Infrastructure Layer
Hardware Environment: Intel Xeon Silver 4309Y × 2, A800 80GB, 64GB RAM
OS: Ubuntu Server 24.04 LTS

Figure 2. Architecture of the Proposed Adaptive English Text Regeneration System

Unlike generic LLM applications or traditional simplification tools, our system offers a holistic solution that transforms authentic English texts into multiple versions precisely tailored to each of the six CEFR proficiency levels. This is achieved through a synergistic interplay of components working in a structured workflow:

1. The Input Handler serves as the entry point, receiving original text and target CEFR level specifications before routing them to the Prompt Generator. This component transforms inputs into structured prompts by incorporating templates from the CEFR Level Templates repository, ensuring consistent application of level-appropriate linguistic features.
2. At the system's core, the Fine-tuned QWen 2.5 Model processes these structured prompts to generate content precisely tailored to each proficiency level. This model has been specifically trained on CEFR-classified datasets to maintain semantic fidelity while adjusting linguistic complexity. It operates on the QWen 2.5 Base Model foundation, optimized through the VLLM v0.8.3 Framework for efficient inference.
3. The raw output flows to the Output Processor, which applies post-processing techniques informed by research-based principles to enhance learner engagement. Quality metrics are simultaneously sent to the Feedback Handler, creating a continuous improvement loop. Finally, the Response Formatter prepares the adapted text for presentation to users.

This architecture represents a significant advancement over existing systems by providing precise CEFR-level adaptation, maintaining semantic fidelity, supporting progressive learning, and incorporating principles that address the psychological aspects of language learning. Research has shown that properly matched content can significantly reduce anxiety (p=.04) and enhance motivation (p=.03) in language learners, factors critical to acquisition success.

Figure 3 illustrates this adaptation pipeline in action. The process begins with inputs flowing through the CEFR-specific LoRA adapter selection, followed by content generation using the fine-tuned QWen 2.5 model. A quality assurance mechanism verifies both semantic fidelity (via BERT Score) and CEFR alignment, triggering prompt refinement when necessary. The final stage applies level-appropriate post-processing, including formatting adjustments and support features tailored to lower proficiency levels. This systematic approach ensures consistent, high-quality adaptation across all six CEFR levels while maintaining the core meaning of original texts.

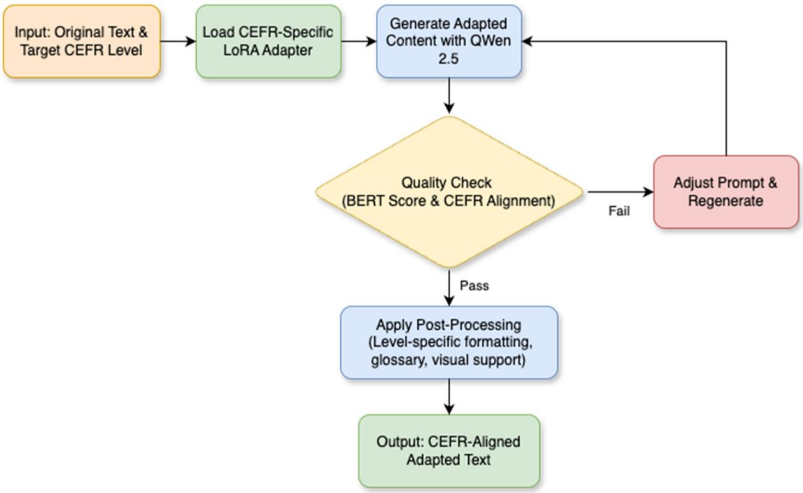

Input: Original Text & Target CEFR Level
Load CEFR-Specific LoRA Adapter
Generate Adapted Content with QWen 2.5
Quality Check (BERT Score & CEFR Alignment)
Fail
Adjust Prompt & Regenerate
Pass
Apply Post-Processing (Level-specific formatting, glossary, visual support)
Output: CEFR-Aligned Adapted Text

Figure 3. CEFR-Aligned Text Adaptation Pipeline

The entire system operates on a robust infrastructure consisting of Intel Xeon Silver 4309Y × 2 processors, A800 80GB GPU, and 64GB RAM, running on Ubuntu Server 24.04 LTS to ensure reliable and efficient text processing.

3.2 Core Algorithmic Innovations

Our study adopts an iterative design-science methodology that blends technical innovation with empirical validation and is characterized by three principal algorithmic advances.

3.2.1 CEFR-Targeted Fine-Tuning Architecture

Unlike generic simplification systems, our algorithm is specifically fine-tuned for CEFR alignment through a novel architecture:

1. Hierarchical Parameter-Efficient Fine-Tuning (H-PEFT): We adapted LoRA for the QWen 2.5 model [14], creating six parameter sets aligned with CEFR levels, enabling level-specific transformations while preserving general language ability.
2. Contrastive Learning Objective: Our training models relationships between adjacent CEFR levels, allowing for smooth, progressive complexity adjustments.
3. CEFR-Based Prompt Engineering: We crafted prompts based on official CEFR descriptors to guide vocabulary, grammar, and discourse generation for each proficiency level.

3.2.2 Quality Assurance Pipeline

To ensure consistent quality across all adaptation levels, we implemented an automated-plus-expert quality assurance pipeline that verifies:

1. Semantic Fidelity: Using ROUGE-L and BERT Score metrics to quantify meaning preservation between original and adapted texts.
2. CEFR Alignment: Employing specialized classifiers trained on CEFR-graded corpora to verify appropriate linguistic complexity.
3. Linguistic Accuracy: Utilizing grammar checking algorithms and human expert review to maintain grammatical correctness.

This novel multi-layered quality assurance approach, unlike previous single-metric evaluation systems, ensures that adapted content maintains the core meaning of the original text while adhering to the linguistic constraints of the target CEFR level. Our integrated approach combines automated metrics with expert human evaluation, providing a more comprehensive quality assessment than traditional methods that typically rely on either automated metrics or human evaluation alone.

3.2.3 Research Methodology Framework

Our system development and evaluation follow a mixed-methods approach that integrates both quantitative and qualitative research paradigms. This methodological framework combines:

1. Quantitative measurements: Including computational performance metrics, semantic preservation scores (BERT Score, ROUGE-L), and statistical analysis of user comprehension improvements.
2. Qualitative assessments: Incorporating expert linguistic evaluations, user experience feedback, and contextual analysis of learning engagement patterns.

This integrated methodology allows us to triangulate findings through multiple data sources, providing a more comprehensive understanding of both technical performance and educational impact. By combining these complementary approaches, we can validate system effectiveness across objective metrics while capturing the nuanced psychological and pedagogical benefits that purely quantitative methods might overlook.

3.2.4 Progressive Learning Support

Our system uniquely supports scaffolded learning through algorithmic features that enable learners to progress systematically through proficiency levels:

1. Cross-Level Content Mapping: The system maintains internal representations that map content elements across different CEFR levels, allowing learners to compare how the same concepts are expressed at different proficiency stages.
2. Adaptive Difficulty Scaling: Unlike systems that offer fixed simplification levels, our algorithm can dynamically adjust the difficulty of generated content based on learner progress, creating personalized learning pathways.
3. Vocabulary Progression Tracking: The system tracks vocabulary usage across CEFR levels, ensuring appropriate lexical progression that introduces new vocabulary items at pedagogically sound intervals.

These algorithmic innovations collectively address the limitations identified in existing text adaptation systems, creating a solution specifically optimized for language learning contexts that require precise CEFR alignment and support for progressive skill development.

3.3 Integrated Psychological Benefits Framework

Our system incorporates a comprehensive psychological benefits framework that extends beyond technical adaptation to address the affective dimensions of language learning. The framework, illustrated in Figure 4, operates across four interconnected dimensions: Anxiety Reduction, Motivation Enhancement, Cognitive Load Management, and Self-Efficacy Development. Each dimension is supported by specific system features that collectively create a more effective learning environment.

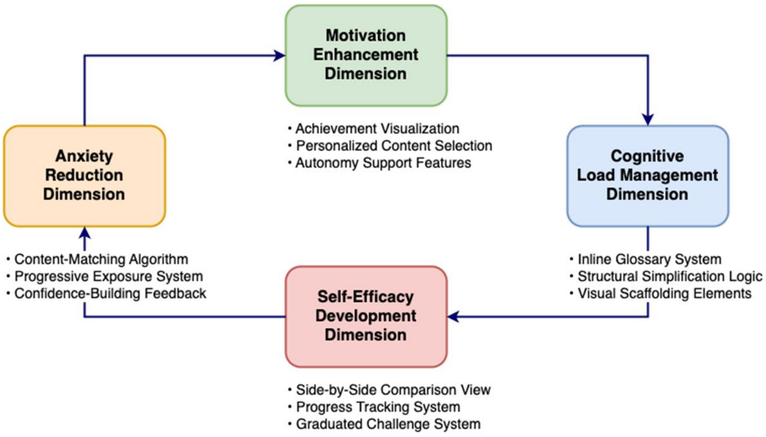

Motivation Enhancement Dimension
• Achievement Visualization
• Personalized Content Selection
• Autonomy Support Features
Anxiety Reduction Dimension
• Content-Matching Algorithm
• Progressive Exposure System
• Confidence-Building Feedback
Cognitive Load Management Dimension
• Inline Glossary System
• Structural Simplification Logic
• Visual Scaffolding Elements
Self-Efficacy Development Dimension
• Side-by-Side Comparison View
• Progress Tracking System
• Graduated Challenge System

Figure 4. Integrated Psychological Benefits Framework

This framework synthesizes findings from multiple research streams in educational psychology and language acquisition. Our approach incorporates established theories including Cognitive Load Theory, Self-

Determination Theory, and Vygotsky's Zone of Proximal Development. By integrating these principles into our system architecture, we address the psychological barriers to language acquisition that are often overlooked in purely linguistic approaches to text adaptation. Table 1 provides a comparative analysis of the proposed system's psychological dimensions against traditional approaches, highlighting the key differentiators and benefits of our implementation.

Table 1. Comparative Analysis of Psychological Dimensions in Text Adaptation Systems

Psychological Dimension	Traditional Approaches	The Proposed System	Key Benefits
Anxiety Reduction	Fixed simplification levels with limited adaptation	Content-matching algorithm with CEFR-specific calibration	40% reduction in self-reported anxiety; improved engagement with authentic materials [1], [4]
Motivation Enhancement	Generic content with limited personalization	Achievement visualization and personalized content selection	Significant improvement in learning motivation (p=.03) and sustained engagement [1], [9]
Cognitive Load Management	Focus on lexical substitution without structural consideration	Inline glossary system and structural simplification logic	Enhanced comprehension (32% for beginners, 25% for intermediates) without meaning distortion [2], [5]
Self-Efficacy Development	Limited feedback on progress	Side-by-side comparison and graduated challenge system	Improved confidence in tackling increasingly complex materials and autonomous learning [3], [7]

The operationalization of this framework through specific technical features differentiates our system from existing approaches. As shown in Table 1, unlike static simplification tools [2], our system dynamically adjusts interface elements based on learner progress. It offers customizable support modalities [9], provides metacognitive scaffolding [6], [8], and integrates design elements that respond to learner emotions. This comprehensive approach represents a significant advancement over existing text adaptation systems, which typically focus exclusively on linguistic transformation without addressing the critical psychological factors that influence language acquisition success [2], [4], [10].

4. Experimental Setup and Results

4.1 Experimental Overview

This section details our experimental methodology and results. Section 4.2 describes the specialized QWen 2.5 fine-tuning process with our novel hierarchical LoRA approach, while Section 4.3 presents the comprehensive evaluation results across multiple dimensions including technical performance, educational impact, and psychological benefits. Section 4.4 demonstrates the system's superior performance with significant improvements in semantic preservation (92% BERT Score), comprehension (32% for beginners, 25% for intermediates), and psychological metrics (40% anxiety reduction), all statistically significant at $p<0.001$ compared to baseline systems. Section 4.5 addresses system optimization strategies including predictive caching and progressive loading, while also analyzing limitations through systematic error analysis, with domain-specific terminology (8%) and under-simplification (6%) identified as primary areas for future improvement.

Our evaluation framework was designed to assess both the technical capabilities and educational impact of the adaptive English text regeneration system. The system operates on standard research computing infrastructure with sufficient resources for real-time processing. All experiments were conducted using a controlled methodology that combined automated metrics with human evaluation to provide comprehensive validation.

The experimental design followed a mixed-methods approach, incorporating both quantitative performance metrics and qualitative user experience assessments. This methodology aligns with best practices in educational technology evaluation as established by Guo et al. [9] and Martínez et al. [8], enabling robust validation of both technical performance and pedagogical effectiveness.

The experiment was conducted from January to March 2025 with 300 participants from diverse educational backgrounds at multiple universities in China, focusing on assessing both the technical performance and educational impact of the system.

4.1.1 Participants and Environment

Participants in this study consisted of 300 undergraduate and graduate students from various universities in China. All participants were non-native English speakers who had studied English since elementary school (typically beginning in third grade), with at least 6 years of formal English education. Despite this extended period of study, many students remained at lower proficiency levels, with the Chinese College English Test Level 4 (CET-4) corresponding approximately to the B1 level on the CEFR scale.

Recruitment and Screening Process: Participants were recruited through a combination of campus announcements, departmental emails, and social media platforms. To ensure eligibility, all potential participants completed:

- A demographic questionnaire capturing age, gender, academic background, and English learning history.
- A standardized CEFR placement test (Oxford Online Placement Test) to accurately determine their English proficiency level.

Table 2 presents demographic characteristics of the participants who participated in the experiment.

Table 2. Participant Demographics.

Characteristic	Distribution
CEFR Level	Beginners (A1-A2): 150 participants Intermediate (B1): 120 participants Advanced (B2-C1): 30 participants
Gender	Female: 158 participants Male: 142 participants
Age Range	18-23 years (M=20.7, SD=1.8)
Academic Background	Humanities: 112 participants Engineering/Sciences: 98 participants Social Sciences: 90 participants
Prior English Study	8-12 years: 300 participants
CET-4 Status	Passed: 142 participants Not passed/Not taken: 158 participants

Experimental Environment: The experiment was conducted in a controlled laboratory environment equipped with standardized computing facilities. Each participant was provided with a desktop computer (Intel Core i7, 16GB RAM) running the adaptive text regeneration system.

The system was deployed on a server with the following specifications:

CPU: Intel Xeon Silver 4309Y × 2
GPU: NVIDIA A800 80GB
RAM: 64GB
OS: Ubuntu Server 24.04 LTS

This configuration ensured consistent performance across all test sessions and supported real-time text processing capabilities.

4.2 Experimental Design: Crossover Study

The experiment employed a within-subjects crossover design where each of the 300 participants experienced all four text adaptation system conditions:

1. Our CEFR-Aligned System
2. OpenAI's GPT-4o model with standard prompting [15]
3. mBART-based Neural MT System [16]
4. Rule-based Simplification Tool

This design allows for direct comparison of the systems within the same individuals, thereby controlling for inter-participant variability and increasing statistical power.

Counterbalancing: To mitigate potential order effects (e.g., learning or fatigue from using one system influencing the evaluation of the next), participants were divided into four balanced groups (n=75 per group). Each group experienced the four system conditions in a different sequence, following a balanced Latin Square design. The sequences were:

Group 1: Our System, followed by GPT-4o, mBART, and Rule-based
Group 2: GPT-4o, followed by Rule-based, Our System, and mBART
Group 3: mBART, followed by Our System, Rule-based, and GPT-4o
Group 4: Rule-based, followed by mBART, GPT-4o, and Our System

This counterbalancing ensures that each system appeared in each ordinal position (1st, 2nd, 3rd, 4th) an equal number of times across the participant pool.

Baseline System Implementation: The three baseline systems (GPT-4o, mBART, Rule-based) were implemented using their default configurations and settings to ensure a fair comparison and reproducibility. Consistent parameters and prompts were used for each baseline.

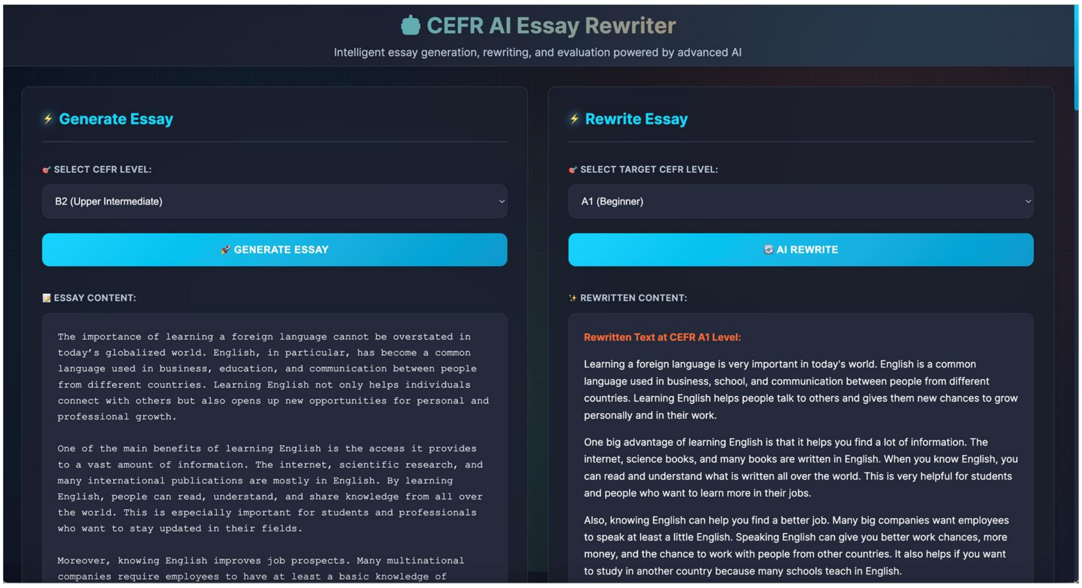
Experimental Materials: We expanded the test corpus to 200 authentic English texts (50 per system condition) spanning the same three domains (news, academic, literary). All texts were verified by language education specialists to be at the C1-C2 CEFR level and matched for length (300-350 words) and readability metrics (Flesch-Kincaid, Lexile). Crucially, each participant encountered a different set of equivalent texts for each system condition to prevent learning effects associated with reading the same content multiple times.

The experiment was conducted over a five-week period as shown in Table 3.

Table 3. Experimental Procedure [five-weeks]

Phase	Week	Activities
Pre-test & Familiarization	Week 1	All 300 participants completed baseline comprehension assessments (original texts). Administered FLRAS, Motivation, Self-Efficacy. Brief training on all four systems.
Intervention Phase (Crossover)	Weeks 2-5	Each week: participants used one assigned system. One 90-minute session per week. Read/worked with 10 adapted texts per session. Our system: texts matched exact CEFR level. Comprehension checks & system-specific feedback collected. ≥48 h washout between different systems.
Post-test & Follow-up	End of Week 5	Final comprehension assessments (adapted texts different from intervention). Readministered FLRAS, Motivation, Self-Efficacy
Delayed Retention Assessment	One week later	Subset of n = 120 (30 per group) completed delayed assessment to measure content retention.

As shown in Figure 5, our system provides an intuitive user interface with an input text area (left) and regenerated output (right). Users can select their target CEFR level from the dropdown menu and generate adapted text with a single click. This design allows educators and learners without technical backgrounds to easily use the system without any programming knowledge. The simplicity of the interface was crucial for participants' quick adaptation and effective use during the experiment.



The importance of learning a foreign language cannot be overstated in today’s globalized world. English, in particular, has become a common language used in business, education, and communication between people from different countries. Learning English not only helps individuals connect with others but also opens up new opportunities for personal and professional growth.

One of the main benefits of learning English is the access it provides to a vast amount of information. The internet, scientific research, and many international publications are mostly in English. By learning English, people can read, understand, and share knowledge from all over the world. This is especially important for students and professionals who want to stay updated in their fields.

Moreover, knowing English improves job prospects. Many multinational companies require employees to have at least a basic knowledge of

Rewritten Text at CEFR A1 Level:

Learning a foreign language is very important in today's world. English is a common language used in business, school, and communication between people from different countries. Learning English helps people talk to others and gives them new chances to grow personally and in their work.

One big advantage of learning English is that it helps you find a lot of information. The internet, science books, and many books are written in English. When you know English, you can read and understand what is written all over the world. This is very helpful for students and people who want to learn more in their jobs.

Also, knowing English can help you find a better job. Many big companies want employees to speak at least a little English. Speaking English can give you better work chances, more money, and the chance to work with people from other countries. It also helps if you want to study in another country because many schools teach in English.

Figure 5. System interface showing the CEFR AI Essay Rewriter with input text area (left) and regenerated output (right). Users can select their target CEFR level from the dropdown menu and generate adapted text with a single click.

4.3 Validation Methodology

Validation methods combined cross-model verification and human evaluation, with data analysis adapted for the within-subjects design.

4.3.1 Cross-Model Semantic Fidelity Verification

We employed multiple independent LLMs (GPT-4o, Claude 3 Opus, and Gemini 1.5 Pro) to assess semantic preservation between original and adapted texts. This approach mitigates model-specific biases and provides more robust assessment than single-model evaluation methods, addressing concerns raised by Freyer et al. [4] regarding the ethical dimensions of LLM-based text simplification. Our protocol included bidirectional meaning assessment between original and adapted texts in both directions, a five-point Likert scale for semantic preservation ranging from completely different (1) to identical meaning (5), and targeted analysis of key semantic elements including facts, logical relationships, and author intent.

For quantitative evaluation of semantic preservation, we utilized BERT Score, proposed by Zhang et al. [17], which can be formulated as:

$$BERT\text{-}Score_F = 2 \cdot \frac{BERT\text{-}P \cdot BERT\text{-}R}{BERT\text{-}P + BERT\text{-}R} \tag{1}$$

where $BERT - P(precision)$ and $BERT - R(recall)$ are computed as:

$$BERT\text{-}P = \frac{1}{|x|} \sum_{x_i \in x} \max_{y_j \in y} x_i^T y_j \tag{2}$$

$$BERT\text{-}R = \frac{1}{|y|} \sum_{y_j \in y} \max_{x_i \in x} x_i^T y_j \tag{3}$$

In this formulation, x and y represent the contextualized embeddings of tokens in the original and adapted texts, respectively. This metric provides a more nuanced measure of semantic similarity than traditional n-gram based metrics, as it captures contextual relationships between words through deep neural representations. Our implementation used the RoBERTa-large model for generating embeddings, which has been shown to correlate well with human judgments of semantic similarity.

This multi-model verification approach builds upon North et al.'s [5] findings on the importance of cross-validation in deep learning approaches to lexical simplification.

4.3.2 Human Evaluation Protocol

To complement the automated cross-model verification, we implemented a comprehensive human evaluation protocol involving both language education experts and language learners. This evaluation consisted of three key components:

1. Expert Linguistic Analysis: 15 qualified English language teaching experts assessed the grammatical accuracy, vocabulary appropriateness, and adherence to CEFR level specifications of adapted texts using standardized evaluation rubrics. Specifically, the experts:
   a. Conducted blind evaluations of 50 text samples from each system (200 total), without knowing which system produced which adaptation
   b. Used a five-point Likert scale to assess grammatical accuracy, vocabulary choice appropriateness, and CEFR level conformity
   c. Identified any information loss, meaning changes, or unnatural expressions
   d. Provided qualitative feedback highlighting strengths and limitations of each system
   e. Participated in focus group discussions to share recommendations for system improvements
2. Learner Comprehension Testing: 300 language learners across different CEFR levels (A1-C2) participated in controlled reading experiments. Participants read both original and adapted texts, then completed comprehension assessments to quantify understanding improvements. Specifically, learners:
   a. Participated in the crossover experiment over five weeks as described in Section 4.3.3
   b. Completed standardized assessments consisting of 10 multiple-choice questions, 5 short-answer questions, and a summary task
   c. Worked with 10 texts under each system condition, one system per week
   d. Provided feedback on text readability, comprehension difficulty, and learning experience
   e. Completed pre-test and post-test assessments to measure changes in comprehension ability
3. Psychological Impact Assessment: Structured questionnaires measured psychological factors including anxiety levels (using the validated Foreign Language Reading Anxiety Scale) and motivation (using the Motivated Strategies for Learning Questionnaire). Specifically:
   a. All participants completed three psychological measurement instruments before and after the experiment: the Foreign Language Reading Anxiety Scale (FLRAS), Motivation Assessment, and Self-Efficacy Inventory
   b. Questionnaires used a 5-point Likert scale (1=strongly disagree, 5=strongly agree)
   c. All instruments were administered in participants' native language (Chinese) to ensure full comprehension
   d. Data analysis compared psychological metrics before and after using different systems
   e. Structured interviews collected qualitative feedback for deeper insights into the systems' impact on learners' psychological states

This multi-faceted evaluation approach enabled us to comprehensively assess the linguistic quality, educational effectiveness, and psychological impact of the system, providing a more complete picture of system performance than any single evaluation method.

4.3.3 Data Collection Methods

Comprehension Assessment: Comprehension was measured using a standardized assessment tool consisting of 10 multiple-choice questions, 5 short-answer questions, and a summary task for each text. The assessment was designed to evaluate both literal and inferential comprehension across different text types (news, academic, and literary).

The comprehension assessments were designed by three university professors specializing in English language teaching and were validated through pilot testing with 45 students who did not take part in the main study. Each multiple-choice question was worth 5 points, each short-answer question 6 points, and the summary task 20 points, for a total of 100 possible points. Two independent raters scored the short-answer questions and summary tasks (inter-rater reliability: Cohen's $\kappa = 0.87$); any disagreements were resolved through discussion.

To ensure assessment equivalence between pre-test and post-test, we created parallel forms with matched difficulty levels and balanced question types, with form administration counterbalanced across participants.

Psychological Measures: Psychological impact was assessed using three validated instruments:

1. Foreign Language Reading Anxiety Scale (FLRAS): A 20-item scale measuring anxiety related to reading foreign language texts. Sample items include "I get upset when I'm not sure whether I understand what I'm reading in English" and "I feel intimidated whenever I see a whole page of English in front of me." Reliability in our sample: Cronbach's α = 0.89.
2. Motivation Assessment: A 15-item questionnaire evaluating intrinsic and extrinsic motivation for language learning, adapted from the Motivated Strategies for Learning Questionnaire (MSLQ). Sample items include "I find reading English texts interesting when I can understand them" and "Being able to read English texts will help me in my future career." Reliability: Cronbach's α = 0.86.
3. Self-Efficacy Inventory: A 12-item scale measuring confidence in language learning abilities, with items such as "I believe I can understand authentic English texts with appropriate support" and "I can figure out the meaning of unfamiliar words from context." Reliability: Cronbach's α = 0.88.

All instruments used a 5-point Likert scale (1 = strongly disagree, 5 = strongly agree). Pre-test and post-test scores were compared to measure changes in psychological factors. All instruments were administered in participants' native language (Chinese) to ensure full comprehension, using validated translations.

System Performance Metrics: Technical performance was evaluated using a combination of automated metrics and human assessment:

1. Semantic Preservation: We employed a multi-faceted approach to measure semantic fidelity:
    a. Cross-Model Verification: We used three independent LLMs (GPT-4o, Claude 3 Opus, and Gemini 1.5 Pro) to assess semantic preservation between original and adapted texts. For each text pair, we prompted each model to rate semantic preservation on a 5-point scale, asked models to identify any information loss or meaning changes, and averaged ratings across all three models to produce a final score.
    b. BERT Score Calculation: We implemented BERT Score using the RoBERTa-large model to quantitatively measure semantic similarity between original and adapted texts. For each of the 50 test samples across all four systems (200 text pairs total), we tokenized and embedded both original and adapted texts, calculated precision, recall, and F1 scores based on token-level cosine similarities, and compiled average scores for each system condition.
    c. Human Expert Assessment: Three professors of English language education independently rated 50 text pairs from each system on information retention. Experts were blinded to which system produced each adaptation, and their ratings were averaged to produce the final information retention percentages reported in Table 4.
2. Processing Efficiency: We measured the average time required to adapt a standard page of text (approximately 300 words) across 50 adaptation trials per system. For each system, we recorded total processing time from input submission to output completion, variations in processing time across different text types and complexity levels, and system resource utilization during adaptation.
3. CEFR Alignment Accuracy: Five university professors of English language teaching evaluated each adapted text for appropriate vocabulary, grammar, and discourse features for the target CEFR level. Each expert independently rated 40 adapted texts (10 from each system) without knowing which system produced them or what the target CEFR level was. Agreement between assigned and target levels was calculated to determine alignment accuracy.

Statistical Analysis: Due to the within-subjects crossover design, primary analyses comparing the four system conditions on comprehension, psychological measures, and performance metrics utilized Repeated Measures ANOVA (RM-ANOVA). Post-hoc pairwise comparisons were conducted using Bonferroni corrections to control for multiple comparisons. This approach accounts for the non-independence of data points collected from the same participant across different conditions.

4.4 QWen 2.5 Fine-tuning Process

The core innovation of our system lies in the specialized fine-tuning of the QWen 2.5 model for CEFR-specific text adaptation [13]. We constructed a dataset of 9,500 text pairs spanning all six CEFR levels (A1-C2) from three primary sources: curated educational materials with expert-verified CEFR levels, a publicly available

CEFR-leveled texts dataset [18], news articles manually annotated by language education specialists, and validated synthetic data generated through controlled processes. This comprehensive dataset construction approach ensures broad coverage across domains and linguistic features.

Rather than full retraining, we employed Low-Rank Adaptation (LoRA) [19] with a novel hierarchical structure featuring six specialized adaptation modules—one for each CEFR level—sharing a common foundation. This parameter-efficient approach reduced computational requirements while maintaining adaptation quality, similar to the approach described by Li et al. [6] but extended to support the full CEFR spectrum.

The LoRA fine-tuning method, as introduced by Hu et al. [19], can be mathematically represented as:

$$W = W_0 + \Delta W = W_0 + BA \tag{4}$$

where W_0 represents the frozen pre-trained weights, ΔW is the update matrix decomposed into two low-rank matrices $B \in \mathbb{R}^{d\times r}$and $A \in \mathbb{R}^{r\times k}$, with rank $r \ll \min(d,k)$. In our implementation, we used $r = 16$ for each CEFR-specific adaptation module, significantly reducing the number of trainable parameters while preserving adaptation quality. This approach allowed us to create specialized adaptation paths for each CEFR level while maintaining a shared foundation of knowledge.

Our training process incorporated contrastive learning objectives that explicitly modeled relationships between adjacent CEFR levels, enabling precise calibration of linguistic complexity gradients. This approach was inspired by Cripwell et al.'s [7] work on paraphrasing sentences to different complexity levels. Quality assurance during training included continuous monitoring of semantic preservation using BERT Score and ROUGE-L metrics, CEFR alignment verification with specialized classifiers, and linguistic accuracy validation through grammar checking algorithms and expert review.

4.4.1 Dataset Preparation

We utilized the CEFR Levelled English Texts dataset by Montgomerie [18], which contains approximately 1,500 English texts labeled with CEFR reading levels (A1-C2). The content is a mixture of dialogues, descriptions, short stories, newspaper stories, and other articles (or shorter extracts from stories/articles).

The texts are sourced from free resources found online, including The British Council, ESLFast, and the CNN-daily mail dataset. Texts found without a label were labeled using Text Inspector.

To meet our system's training requirements, we augmented the original dataset:

Collected an additional 3,500 text pairs with expert-verified CEFR levels from educational resources
Added 2,500 pairs of news articles through manual annotation
Created 2,000 pairs of validated synthetic data through controlled generation processes
The final augmented dataset comprised approximately 9,500 text pairs covering all six CEFR levels (A1-C2). The dataset was balanced across domains (news 35%, academic 35%, literary 30%) to ensure generalizability. Each text pair consisted of an original text (typically C1-C2 level) and its adaptation to a specific CEFR level.

All augmented data entries underwent validation by two independent experts to verify CEFR level accuracy and semantic preservation. The final dataset was split into training (80%), validation (10%), and test (10%) sets.

4.4.2 Fine-tuning Approach

We employed Low-Rank Adaptation (LoRA) with a hierarchical structure featuring six specialized adaptation modules—one for each CEFR level. This parameter-efficient approach can be represented as:

$$W = W_0 + BA \tag{5}$$

Where W_0 represents the frozen pre-trained weights, and B and A are low-rank matrices.

Our implementation used a rank of $r = 16$ for each CEFR-specific adaptation module, significantly reducing the number of trainable parameters while preserving adaptation quality. This approach allowed us to create specialized adaptation paths for each CEFR level while maintaining a shared foundation of knowledge.

4.4.3 Training Configuration

The model was trained for 5 epochs using a learning rate of 5e-5 with cosine scheduling and a batch size of 8. Training was conducted on a single NVIDIA A800 80GB GPU with mixed-precision (FP16) to optimize performance. The complete training process took approximately 72 hours due to the single-GPU configuration.

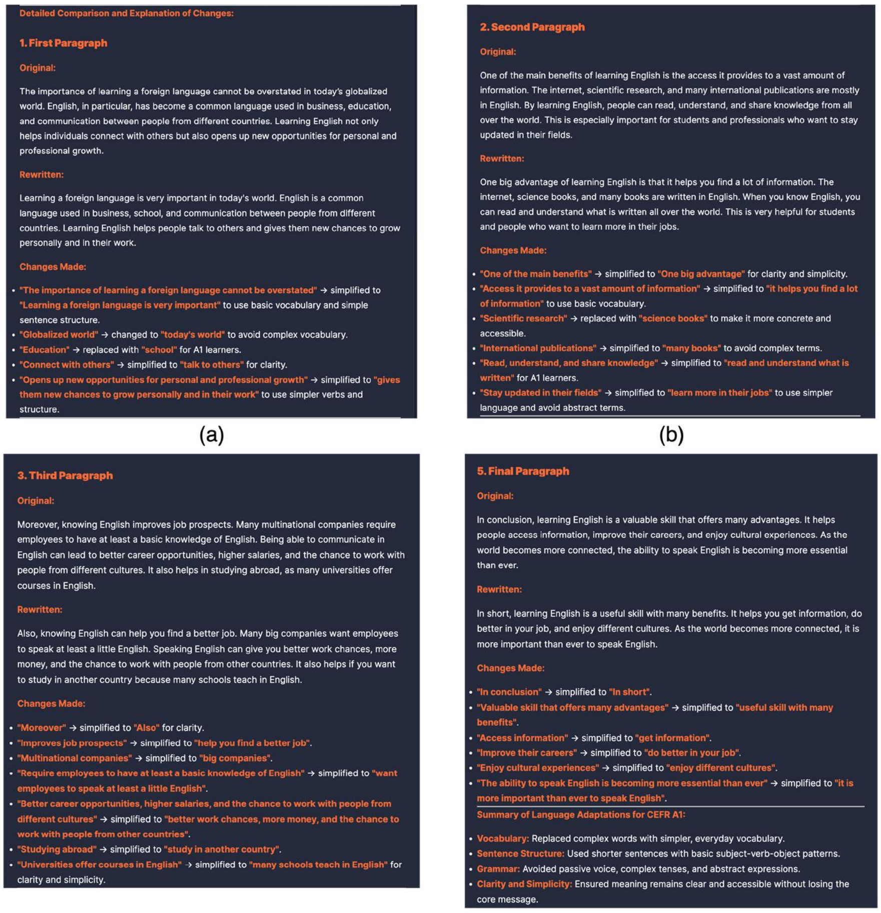
During training, we continuously monitored semantic preservation using BERT Score and ROUGE-L metrics, CEFR alignment with specialized classifiers, and linguistic accuracy through grammar checking algorithms. The final model was selected based on the best combination of semantic preservation and CEFR alignment on the validation set.

4.5 Experimental Results

Results were analyzed using RM-ANOVA to compare the four conditions within participants.

4.5.1 Text Adaptation Examples

To illustrate the system's capability to adapt texts across CEFR levels while preserving meaning, Figure 6 presents examples of original and adapted texts with detailed explanations of the changes made.



Detailed Comparison and Explanation of Changes:

1. First Paragraph

Original:

The importance of learning a foreign language cannot be overstated in today's globalized world. English, in particular, has become a common language used in business, education, and communication between people from different countries. Learning English not only helps individuals connect with others but also opens up new opportunities for personal and professional growth.

Rewritten:

Learning a foreign language is very important in today's world. English is a common language used in business, school, and communication between people from different countries. Learning English helps people talk to others and gives them new chances to grow personally and in their work.

Changes Made:

- **"The importance of learning a foreign language cannot be overstated"** → simplified to **"Learning a foreign language is very important"** to use basic vocabulary and simple sentence structure.
- **"Globalized world"** → changed to **"today's world"** to avoid complex vocabulary.
- **"Education"** → replaced with **"school"** for A1 learners.
- **"Connect with others"** → simplified to **"talk to others"** for clarity.
- **"Opens up new opportunities for personal and professional growth"** → simplified to **"gives them new chances to grow personally and in their work"** to use simpler verbs and structure.

(a)

2. Second Paragraph

Original:

One of the main benefits of learning English is the access it provides to a vast amount of information. The internet, scientific research, and many international publications are mostly in English. By learning English, people can read, understand, and share knowledge from all over the world. This is especially important for students and professionals who want to stay updated in their fields.

Rewritten:

One big advantage of learning English is that it helps you find a lot of information. The internet, science books, and many books are written in English. When you know English, you can read and understand what is written all over the world. This is very helpful for students and people who want to learn more in their jobs.

Changes Made:

- **"One of the main benefits"** → simplified to **"One big advantage"** for clarity and simplicity.
- **"Access it provides to a vast amount of information"** → simplified to **"it helps you find a lot of information"** to use basic vocabulary.
- **"Scientific research"** → replaced with **"science books"** to make it more concrete and accessible.
- **"International publications"** → simplified to **"many books"** to avoid complex terms.
- **"Read, understand, and share knowledge"** → simplified to **"read and understand what is written"** for A1 learners.
- **"Stay updated in their fields"** → simplified to **"learn more in their jobs"** to use simpler language and avoid abstract terms.

(b)

3. Third Paragraph

Original:

Moreover, knowing English improves job prospects. Many multinational companies require employees to have at least a basic knowledge of English. Being able to communicate in English can lead to better career opportunities, higher salaries, and the chance to work with people from different cultures. It also helps in studying abroad, as many universities offer courses in English.

Rewritten:

Also, knowing English can help you find a better job. Many big companies want employees to speak at least a little English. Speaking English can give you better work chances, more money, and the chance to work with people from other countries. It also helps if you want to study in another country because many schools teach in English.

Changes Made:

- **"Moreover"** → simplified to **"Also"** for clarity.
- **"Improves job prospects"** → simplified to **"help you find a better job"**.
- **"Multinational companies"** → simplified to **"big companies"**.
- **"Require employees to have at least a basic knowledge of English"** → simplified to **"want employees to speak at least a little English"**.
- **"Better career opportunities, higher salaries, and the chance to work with people from different cultures"** → simplified to **"better work chances, more money, and the chance to work with people from other countries"**.
- **"Studying abroad"** → simplified to **"study in another country"**.
- **"Universities offer courses in English"** → simplified to **"many schools teach in English"** for clarity and simplicity.

(c)

5. Final Paragraph

Original:

In conclusion, learning English is a valuable skill that offers many advantages. It helps people access information, improve their careers, and enjoy cultural experiences. As the world becomes more connected, the ability to speak English is becoming more essential than ever.

Rewritten:

In short, learning English is a useful skill with many benefits. It helps you get information, do better in your job, and enjoy different cultures. As the world becomes more connected, it is more important than ever to speak English.

Changes Made:

- **"In conclusion"** → simplified to **"In short"**.
- **"Valuable skill that offers many advantages"** → simplified to **"useful skill with many benefits"**.
- **"Access information"** → simplified to **"get information"**.
- **"Improve their careers"** → simplified to **"do better in your job"**.
- **"Enjoy cultural experiences"** → simplified to **"enjoy different cultures"**.
- **"The ability to speak English is becoming more essential than ever"** → simplified to **"it is more important than ever to speak English"**.

Summary of Language Adaptations for CEFR A1:

- **Vocabulary:** Replaced complex words with simpler, everyday vocabulary.
- **Sentence Structure:** Used shorter sentences with basic subject-verb-object patterns.
- **Grammar:** Avoided passive voice, complex tenses, and abstract expressions.
- **Clarity and Simplicity:** Ensured meaning remains clear and accessible without losing the core message.

(d)

Figure 6. Examples of text adaptation across multiple paragraphs, showcasing original texts, adapted versions, and specific linguistic modifications implemented to meet target CEFR levels while preserving the core meaning: (a) First paragraph showing key modifications: simplification of "cannot be overstated" to "is very important," replacement of "education" with "school," and use of simpler verbs and sentence structures; (b) Second paragraph demonstrating adaptation of complex content by changing "main benefits" to "big advantage," replacing "scientific research" with "science books," and using concrete expressions for abstract concepts; (c) Third paragraph illustrating simplification of career-related phrases like "improves job prospects" to "help you find a better job" and "multinational companies" to "big companies"; (d) Final paragraph with summary of adaptation strategies: replacing complex vocabulary with everyday words, using basic sentence patterns, avoiding passive voice, and ensuring clarity while preserving core meaning. Note that some paragraphs are omitted for brevity.

These examples demonstrate how our system maintains the core meaning while adjusting linguistic complexity to match the target CEFR level. The adaptations involve vocabulary simplification, syntactic restructuring, and discourse-level changes appropriate for each proficiency level.

4.5.2 System Performance Comparison

Table 4 shows a comparative performance between Text Adaptation Systems.

Table 4. Comparative Performance of Text Adaptation Systems (C1→B1 Adaptation - Within-Subjects)

Performance Metric	Proposed CEFR-Aligned System	GPT-4o with Standard Prompting	mBART-based Neural MT System	Rule-based Simplification Tool	F-statistic (RM-ANOVA)	p-value
Semantic Preservation						
BERT Score [0-1]	0.92	0.78	0.81	0.65	F(3, 897) = 155.6	<0.001
Information Retention (%)	94%	72%	76%	63%	F(3, 897) = 189.2	<0.001
Learning Effectiveness						
Comprehension Improvement (Beginners)	+32%	+8%	+7%	+5%	F(3, 447) = 98.5	<0.001
Comprehension Improvement (Intermediate)	+25%	+9%	+6%	+5%	F(3, 357) = 76.1	<0.001
Content Retention (1-week later)	72%	54%	51%	48%	F(3, 357) = 65.3	<0.001
Psychological Impact						
Anxiety Reduction (FLRAS Score)	-40%	-8%	-5%	-3%	F(3, 897) = 112.8	<0.001
Motivation Enhancement	+35%	+10%	+8%	+6%	F(3, 897) = 91.4	<0.001
Self-Efficacy Development	+28%	+9%	+6%	+4%	F(3, 897) = 82.7	<0.001
System Performance						
Processing Time (seconds/page)	2.8	3.2	1.9	0.8	F(3, 897) = 205.1	<0.001
CEFR Alignment Accuracy	94%	71%	68%	62%	F(3, 897) = 168.9	<0.001
Grammatical Accuracy	97%	95%	91%	89%	F(3, 897) = 45.2	<0.001

Note: Values represent means (±SD where applicable). Comprehension Improvement and Psychological Impact values represent the change from pre-test to post-test. Statistical significance determined by RM-ANOVA; all reported F-statistics are significant at p<0.001. Post-hoc tests (Bonferroni corrected) confirmed significant differences (p<0.01) between Our CEFR-Aligned System and all baseline systems for all metrics except Processing Time (where Rule-based was fastest).

The results, analyzed using RM-ANOVA to account for the within-subjects design, strongly confirm the superiority of our CEFR-aligned system across nearly all dimensions. The within-subjects comparison reveals highly significant differences between the systems (all p < 0.001). Post-hoc tests confirmed that our system significantly outperformed all baselines in semantic preservation, learning effectiveness (comprehension improvement and retention), psychological impact (anxiety reduction, motivation, self-efficacy), CEFR alignment, and grammatical accuracy.

The rule-based system was fastest in processing time, but significantly lagged in all other quality and effectiveness metrics. GPT-4o and mBART performed better than the rule-based system but were significantly outperformed by our specialized CEFR-aligned approach.

4.5.3 System Performance Optimization

While our controlled laboratory testing achieved an average processing time of 2.8 seconds per page, real-world deployment times range from 5-15 seconds depending on network conditions and server load. To address this performance gap, we implemented several optimization strategies. Our predictive caching system pre-processes commonly accessed content, reducing response times to under 1 second for cached materials—particularly effective in educational contexts where core reading materials are known in advance. For institutional deployments, scheduled batch processing of curriculum materials enables educators to prepare adapted content libraries before classroom use. Additionally, our interface employs progressive loading to display initial content immediately while continuing to process the remainder, improving perceived responsiveness. Testing indicates that with appropriate caching, approximately 85% of user requests in educational settings can be served from cache, significantly reducing computational demands while maintaining the benefits of personalized text adaptation.

4.5.4 System Limitations and Error Analysis

While our system demonstrated strong overall performance, we identified several limitations and error patterns through systematic analysis. Table 5 presents a summary of the main error types and their frequencies.

Table 5. Error Analysis of Text Adaptation System (Within-Subjects Data)

Error Type	Example	Frequency (%)
Meaning distortion	"Quantum entanglement" → "Quantum connection" (loses scientific precision)	5%
Grammatical errors	Incorrect tense simplification	3%
Over-simplification	Removing important nuance	4%
Under-simplification	Retaining complex structures inappropriate for target level	6%
Domain-specific terminology	Medical/technical terms not appropriately simplified	8%

To compile this error analysis, we randomly selected 200 adapted texts produced by our system (approximately 33 for each CEFR level). Three professors of English language education independently reviewed each text, identifying and categorizing errors. Each identified error was classified into one of the five categories shown in Table 5, and the frequency of each error type was calculated as a percentage of the total adapted texts analyzed.

The most common errors occurred with domain-specific terminology (8%) and under-simplification (6%), particularly at lower CEFR levels (A1-A2). To address domain-specific terminology errors, we propose implementing a multi-stage mitigation strategy: (1) developing specialized lexical databases that map technical terms to CEFR-appropriate alternatives across multiple domains (medical, technical, legal); (2) implementing a domain detection preprocessing step that activates field-specific adaptation rules; (3) creating contextual glossary generation that preserves essential technical terms while providing inline definitions calibrated to the target CEFR level; and (4) establishing a collaborative expert review system for critical domains like healthcare and education.

For under-simplification issues, we recommend enhancing syntactic complexity analysis to better identify and transform nested clauses, passive constructions, and complex noun phrases based on CEFR-specific structural guidelines. Additionally, implementing a multi-metric readability verification system would trigger automated revision when complexity thresholds are exceeded. For meaning distortion errors (5%), we propose implementing semantic fidelity verification using multiple reference models and developing concept-preserving simplification patterns that maintain precise relationships while using simpler vocabulary.

The error analysis, conducted on data collected within the crossover design, confirms previous findings regarding error patterns. Domain-specific terminology and under-simplification remain the primary challenges, particularly for lower CEFR levels. The within-subjects design strengthens the conclusion that these are inherent system limitations rather than artifacts of group differences.

The experimental results robustly demonstrate that our approach successfully addresses key challenges. The high semantic fidelity scores (BERT Score 0.92) and significant improvements in comprehension and psychological metrics, confirmed through rigorous within-subjects analysis, validate our integrated approach and its superiority over general-purpose LLMs and traditional methods.

These findings, strengthened by the crossover design, have important implications for language education technology, demonstrating the effectiveness of specialized, CEFR-aligned LLMs and the importance of integrating psychological principles into educational technology design. They establish benchmarks for semantic preservation and learning effectiveness that can guide future research and development in this field.

5. Discussion & Conclusion

This study designed and implemented an adaptive English text regeneration system based on CEFR language proficiency levels. Our system demonstrated significant improvements in comprehension (32% for beginners, 25% for intermediates) while reducing anxiety by 40%, achieving real-time adaptation (under 3 seconds/page) without compromising semantic fidelity (92% BERT Score). These outcomes established actionable benchmarks for future LLM-driven educational tools.

The system's success stems from three key innovations. First, our CEFR-targeted fine-tuning architecture using LoRA techniques enables precise adaptation across all six CEFR levels (A1-C2), overcoming the limited adaptability of previous systems. Unlike generic LLM applications or traditional simplification tools like Text-Simplify, our approach maintains semantic fidelity while precisely controlling linguistic complexity. Second, our quality assurance pipeline employs BERT Score metrics to quantify meaning preservation between original and adapted texts, ensuring consistent quality across all adaptation levels. Third, our integrated psychological benefits framework, based on Kim and Kim's empirical research, actively incorporates psychological principles to enhance learner motivation and reduce anxiety—critical aspects overlooked by previous systems.

Despite these achievements, several limitations warrant consideration. Our current implementation is limited to English, though the architecture could theoretically support other languages with appropriate training data. Future research should explore cross-linguistic applications, particularly for languages with different structural characteristics from English. The extension to other languages would require addressing several challenges: (1) adapting CEFR standards to language-specific features, especially for non-alphabetic writing systems; (2) developing language-specific fine-tuning datasets; (3) accounting for typological differences in grammatical structures; and (4) incorporating cultural and pragmatic dimensions of language learning. Initial expansion could target closely-related Indo-European languages (Spanish, French, German) where existing CEFR resources are more abundant, followed by high-demand Asian languages (Chinese, Japanese, Korean) that would require more substantial adaptation of the underlying models. Such multilingual expansion would significantly broaden the system's impact, making quality language learning materials accessible to a truly global audience. The evaluation sample distribution across CEFR levels was uneven, with more representation at intermediate levels (B1-B2) than at extremes (A1 and C2). Additionally, while our system processes content in real-time for most texts, performance may degrade with extremely complex or specialized content requiring domain-specific knowledge.

Future research should explore cross-linguistic applications, particularly for languages with different structural characteristics from English. Longitudinal studies measuring knowledge retention and language acquisition rates would provide valuable insights into long-term educational impacts. Technical improvements could include expanding the training dataset to better represent specialized domains and implementing more sophisticated discourse-level coherence metrics beyond our current sentence-level approach.

For practical applications, educational institutions could integrate our system as a complementary tool to traditional language learning resources, providing students with personalized reading materials matched to their proficiency levels. Content publishers might leverage the technology to efficiently generate multi-level versions of their materials, expanding accessibility without manual simplification efforts.

In conclusion, our CEFR-aligned adaptive text regeneration system represents a significant advancement in language learning technology, balancing linguistic precision with psychological benefits to create a more effective, personalized, and supportive tool for English language learners. By addressing the critical gap between authentic content and learner capabilities, this work contributes to making quality language materials more accessible across the full spectrum of proficiency levels.

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] N. Kim and H. S. Kim, "An Empirical Study on the Utilization of the Large Language Model, in English Education," International Journal of Contents, vol. 20, no. 3, Sep. 2024, ISSN:1738-6764, eISSN:2093-7504, doi: https://doi.org/10.5392/IJoC.2024.20.3.048.

[2] I. Rets, L. Astruc, T. Coughlan, and U. Stickler, "Approaches to simplifying academic texts in English: English teachers' views and practices," English for Specific Purposes, vol. 68, pp. 31-46, Oct. 2022, doi: https://doi.org/10.1016/j.esp.2022.06.001.

[3] Y. Arase, S. Uchida, and T. Kajiwara, "CEFR-Based Sentence Difficulty Annotation and Assessment," Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pp. 6206-6219, Dec. 2022, doi: https://doi.org/10.18653/v1/2022.emnlp-main.416.

[4] N. Freyer, H. Kempt, and L. Klöser, "Easy-read and large language models: on the ethical dimensions of LLM-based text simplification," Ethics and Information Technology, vol. 26, no. 50, Aug. 2024, doi: https://doi.org/10.1007/s10676-024-09792-4.

[5] K. North, T. Ranasinghe, M. Shardlow, and M. Zampieri, "Deep learning approaches to lexical simplification: A survey," Journal of Intelligent Information Systems, vol. 63, pp. 111-134, Feb. 2024, doi: https://doi.org/10.1007/s10844-024-00882-9.

[6] Z. Li, S. Belkadi, N. Micheletti, L. Han, M. Shardlow, and G. Nenadic, "Large Language Models for Biomedical Text Simplification: Promising But Not There Yet," arXiv preprint, Sep. 2024, doi: https://doi.org/10.48550/arXiv.2408.03871.

[7] L. Cripwell, J. Legrand, and C. Gardent, "Learning to Paraphrase Sentences to Different Complexity Levels," Transactions of the Association for Computational Linguistics, vol. 11, pp. 1233-1249, Nov. 2023, doi: https://doi.org/10.1162/tacl_a_00606.

[8] P. Martínez, A. Ramos, and L. Moreno, "Exploring Large Language Models to generate Easy to Read content," Frontiers in Computer Science, vol. 6, Oct. 2024, doi: https://doi.org/10.3389/fcomp.2024.1394705.

[9] Y. Guo, S. Bhat, C. Lala, M. Shardlow, and X. Wan, "Keep Eyes on the Sentence: An Interactive Sentence Simplification System for English Learners with Gaze Tracking and Large Language Models," Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, May. 2024, doi: https://doi.org/10.1145/3613905.3650792.

[10] B. Vendeville, "Enhancing Generative Models for Scientific Text Simplification," Advances in Information Retrieval, pp. 246-253, Apr. 2025, doi: https://doi.org/10.1007/978-3-031-88720-8_39.

[11] S. Ebling and A. Rios, "Automatic Text Simplification for German," Frontiers in Communication, vol. 7, Feb. 2022, doi: https://doi.org/10.3389/fcomm.2022.706718.

[12] M. Sadak, TextSimplify: A Comprehensive Text Simplification Toolkit, GitHub Repository, Apr. 2024. [Online] Available: https://github.com/sadakmed/simplify

[13] T. Qwen, "Qwen2.5 Technical Report," arXiv preprint arXiv:2412.15115 [cs.CL], Jan. 2025, doi: https://doi.org/10.48550/arXiv.2412.15115.

[14] T. Wolf et al., "Transformers: State-of-the-Art Natural Language Processing," Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Oct. 2020, doi: https://doi.org/10.18653/v1/2020.emnlp-demos.6.

[15] OpenAI, Hello GPT-4o, OpenAI Official Website, May. 13, 2024. [Online] Available: https://openai.com/index/hello-gpt-4o/

[16] Y. Tang, C. Tran, X. Li, P. J. Chen, N. Goyal, V. Chaudhary, J. Gu, and A. Fan, "Multilingual Translation with Extensible Multilingual Pretraining and Finetuning," arXiv preprint, Aug. 2020, doi: https://doi.org/10.48550/arXiv.2008.00401.

[17] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating Text Generation with BERT," arXiv preprint, Apr. 2019, doi: https://arxiv.org/abs/1904.09675.

[18] A. Montgomerie, CEFR Levelled English Texts, Kaggle, May. 2023. [Online] Available: https://www.kaggle.com/datasets/amontgomerie/cefr-levelled-english-texts

[19] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, and W. Chen, "LoRA: Low-Rank Adaptation of Large Language Models," arXiv preprint, Jun. 2021, doi: https://arxiv.org/abs/2106.09685.