Forum for Linguistic Studies

https://journals.bilpubgroup.com/index.php/fls

ARTICLE

Decrypting Complexity: A Tri-Metric Evaluation of Readability and Fidelity in AI-Simplified Scientific Texts for ESL University Learners

Musharraf Aziz [1*], ***Omar Al-Jamili*** [2], ***Nur Rasyidah Mohd Nordin*** [1], ***Shamim Akhter*** [3]

[1] *School of Language, Civilisation and Philosophy, University Utara Malaysia (UUM), Sintok 06010, Malaysia*

[2] *School of Computing, University Utara Malaysia (UUM), Sintok 06010, Malaysia*

[3] *Faculty of Education & Liberal Arts (FELA), INTI International University, Nilai 71800, Malaysia*

ABSTRACT

Undergraduate university students in ESL contexts often need enhanced readability of complex scientific articles in research journals. This study aimed to assess the efficacy and "*toolability*" of AI-based Chat-GPT in readability amplification of research abstracts in language and linguistics journals, indexed in Scopus and Web of Science. Robust latent semantic analysis (LSA), with vectorial space document-embedding, was performed to evaluate co-occurrence and notional preservation. One hundred abstracts (n = 100), extracted from four journals, were prompted into an open Chat-GPT 4.o session for simplification at undergraduate level ESL users. Three metrics, Flesch-Kincaid Grade Level, Flesch Reading Ease and McAlpine *EFLAW* were used for readability measurement at pre-transformation and post-transformation stages. The content fidelity in the input and output models were determined by latent semantic analysis recorded from 0 to 1 of the fidelity range. To rule out bias, objective evaluation by field experts was performed on a randomly extracted subgroup (n = 50). Further, *t*-tests and correlation analysis were conducted for comparing estimations and accuracy evaluation. The findings showed adequate semantic similarity and fidelity, almost overruling post-simplification semantic disruption. The readability increased, with a low Flesch-Kincaid Grade, high Flesch- Kincaid Ease and representative *EFLAW* score. However, weak correlation of LSA and field experts' estimations warranted caution and human-AI contra-estimations. The study offers micro-, meso- and macro-implications for incorporating AI in scientific reading comprehension, given caution

*CORRESPONDING AUTHOR:

Musharraf Aziz, School of Language, Civilisation and Philosophy, University Utara Malaysia (UUM), Sintok 06010, Malaysia;
Email: musharrafazizkaifi@gmail.com

ARTICLE INFO

Received: 26 July 2025 | Revised: 5 August 2025 | Accepted: 14 August 2025 | Published Online: 17 October 2025
DOI: https://doi.org/10.30564/fls.v7i11.11269

CITATION

Aziz, M., Al-Jamili, O., Mohd Nordin, N.R., et al., 2025. Decrypting Complexity: A Tri-Metric Evaluation of Readability and Fidelity in AI-Simplified Scientific Texts for ESL University Learners. Forum for Linguistic Studies. 7(11): 86–103. DOI: https://doi.org/10.30564/fls.v7i11.11269

is practiced with unsupervised dependence. Future research may involve other metrics like BERTScore, robust mixed research designs, comparative cognitive protocols evaluation of texts and other AI models.

Keywords: AI-Assisted Text Simplification; Scientific Text Readability; Latent Semantic Analysis; Reading Comprehension; Esl University Learners

1. Introduction

AI language models can be trained to provide the closest possible reading input interpretations that can enhance the reader's comprehension of the input data. This can inform the model performance while serving educational purposes. Undergraduate students need to navigate through a number of research articles for fulfilling academic demands however, in ESL/EFL contexts, comprehending scientific research may be challenging for young researchers. Students have started using AI tools, based on Generative Language Models (GLM), such as Chat-GPT and DeepSeek, for increasing readability of complex texts. Effective comprehension of scientific text in research journals is vital in the modern-day academia while most undergraduate students focus on extracting the gist from abstracts. AI readability-amplifying paraphrasing, with a purpose to access scientific texts, can provide epistemological and intellectual exposure to learners. It is an AI-reading comprehension interface that necessitates examining the critical functions computed during the simplification process, such as the extent to which AI can preserve the intended ideas at the post-simplification stage, adequate readability level for ESL undergraduates, and possible semantic loss. Research, in this area, can help in improving semantic expanse and transfiguration functions of GLM. Moreover, AI text simplification can decrease textual complexity making concepts accessible for students. Additionally, research has shown that reading comprehension and scientific thinking among undergraduates, especially in developing ESL contexts, varies widely and, in that, marked deficits may exist in socio-economically disadvantaged students[1,2]. Easily available AI software can enhance research accessibility for such learners, ultimately leading to better understanding of scientific research and reading comprehension. Indirectly, this may enhance the learner performance in required academic reading and writing.

In undergraduate studies, introducing research in related fields requires students to adequately comprehend scientific information provided in research journals. Most undergraduates initiate absorbing research through abstracts in articles. However, scientific journals do not universally apply any specific readability standards for different readers so low-ability ESL/EFL comprehenders may struggle with understanding scientific research. It is noted that target audience of research journals are mainly researchers and academics, and the script is technically complex, specialized for presenting complex concepts[3]. It may affect the prospective landscape of learning through research, and requires knowledge recontextualization. Thus, improving its accessibility is crucial for undergraduate learners in general, and struggling readers in particular. In this regard, Artificial Intelligence (AI) can be utilized for enhancing the accessibility of research abstracts, and simultaneously promote AI in education (AIED). This study involves scientific text in language and linguistics research journals on account that several scholars[4,5] have noted decreasing readability in scientific articles. Researchers have also noted the phenomenon of readability decrease with reference to undergraduate university learners. Based on this observation, we assume that investigating the use of generative language models (GLM) for amplifying readability may be beneficial with reference to ESL learners. The potential benefit therefore warrants empirical investigation of GLM as a scientific research transformer.

In text simplification, investigations have been conducted on AI models, such as Chat-GPT, exploring how they enhance readability of medical texts that can provide necessary information to the stakeholders such as patients[6]. Further, researchers have investigated the role of AI in simplifying specialized text in broader perspectives, that may lead to AI-based learning[7]. However, these researchers explore, in the context of legal education, how certain steps can be enforced in AI-based learning for descriptive simplification, whereas we note that three of these steps, that perform summarizing, simplifying and terminological convenience creation, in fact, are the model computations of the software. We contend that Chat-GPT, to a certain level, is trained at simplification, employing specific steps. It analyzes the text complexity to locate challenging morphological, syntactic and semantic "portions", proceeding to simplification strategies that are extended to these particular locations. Subsequently, it generates the output. However, certain concerns, such as semantic similarity preservation at post-simplification stage, remain relatively unexplored. The present study aims to address this issue along with the use of AI's potential scaffolding in scientific reading. Theoretically, this study critically engages two theoretical lenses; Zone of Proximal Development (ZPD) and Augmented Intelligence[8,9]. ZPD favours cognitive scaffolding for the learner performance, that can be coupled with AI intelligence, functioning as a cognitive enhancement partner. Together these two theories offer a convincing ground for the case of AI, such as Chat-GPT, as a simplification assistant that provided cognitive scaffolding for enhancing comprehension. Thus, this theoretical integration legitimizes the pedagogical role of AI by which linguistic complexity may be decreased without jeopardizing semantic similarity. Thus, ESL learners' proximal zones can be widened by ethically using AI's cognitive partnership in learning.

Provided that AI models can contribute to text simplification, there is a literature gap in how AI models and tools can be utilized to increase readability of research journals, benefitting ESL undergraduates. These students generally struggle with reading comprehension, so digesting specialized scientific texts, largely produced by L1 users, for assisting research projects can demotivate several learners. Students with lower self-efficacy may be easily demotivated because of intense cognitive load. Ethically acceptable AI support can increase learning opportunities however, further research is needed to understand how software, like Chat-GPT, behave when prompted to perform criterion-based research text simplification, and how the output affects readability of scientific texts for ESL learners. Considering the experimental stage and functionality of GPT 4.0, its computational proneness to overfitting input data patterns raises understandable concerns because in this case, simplifying text for ESL learners may create bi-dimensional contextual misunderstanding in the cognitive *AI-learner* space leading to miscomprehension. Research on hallucinatory misattributions is in progress however, the tendency may lead to loss of semantic accuracy during simplification[10]. Hence, it is crucial to explore the software's semantic efficiency to determine if it offers a satisfactory and appropriate readability at undergraduate level. For this, methods related to natural language processing (NLP) are a logical choice of which latent semantic analysis (LSA) is a powerful candidate. It analyzes semantic structures and associations in a given corpus, and compares the input for reducing dimensionality[11]. Preserving key ideas, the method aims to filter extraneous information for producing contextually and semantically coherent data. In the previous studies[12,13], it has been used for detecting data patterns however, to the best of the researchers' knowledge, this is the first study to leverage LSA for identifying data transformation, in Chat-GPT 4.0's simplification of the scientific research text published in language and linguistics journals.

Extant literature review indicates the need of further inquiry in the area of AI text simplification for students in ESL contexts, with reference to assistance in comprehending scientific texts. Earlier studies have tackled automatic text simplification for learning in either broader or different perspectives. For instance, researchers recently performed a study on AI text simplification of legal content in a translation class, specifically for students from areas of external expertise[7]. They utilized a Chat-GPT-based penta-partite method applying which they summarized, simplified, extracted, and mapped content while generating multimodal data. This method integrated the generative efficacy of AI with strategic pedagogy that enhanced the samples' learning activated by multimodal data. Further, another recent research explored how automatic text simplification (ATS) and automatic text complexity evaluation (ATCE) can be utilized to provide inclusive education to children aged 6–11 years[14]. They experimented with personalizing learning through AI use for reading proficiency.

Another study leveraged priorly trained AI simplification models for increasing the readability of scientific text comprising broad data at a symposium[15]. The findings demonstrated that GPT and BERTScore can provide simplified summaries but information erosion may frequently occur. These researchers warranted for further research on AI model's (e.g., Chat-GPT) ability to coordinate between preserving semantic integrity and text simplification, specifically in medical, engineering and language studies that need technical preservation of data. This indicates the need to further research on AI- based text simplification in defined domains to reach focused findings.

Unlike above investigations, researchers recently conducted a quasi-experiment on undergraduates. They investigated how an AI-driven text simplification (TS) software called "CoAST" can be utilized to produce integrated learning in which both the teacher and learner engage with theoretically sophisticated academic texts[16]. However, their focus was on determining the impact of collaborative teaching-learning synergy that incorporated AI-based TS for comprehension augmentation. The findings showed that AI-human collaboration provided an effective digitized learning space to students. However, the texts involved in this study were not domain-specific and the AI software utilized was CoAST that supplies digitized learning opportunities along with TS. Though such an approach tends to align with curricular goals, it may run scalability problems in face of diverse educational requirements, hence needing investigations based on certain readability scales.

In the above regard, lack of text specificity was observed because earlier studies did not essentially work on scientific text, with an exception of the study that analyzed scientific content through using broad data[15]. Later, researchers analyzed scientific text simplification using AI-systems, but in multilingual context concentrating on Spanish and Basque other than English[17]. This study presented a robust case of AI-usage for scientific text simplification in a globally relevant context, including both English and non-English languages yet, a substantial scope remained for investigating the ways AI text simplification handles semantic challenges using a readability scale and standard specifically designed for adult ESL learners.

While the literature review demonstrated substantial interest in AI integration in educational text simplification, certain marked gaps were observed. For instance, AI-driven simplification in educational setting necessitates evaluating simplification against the validated readability standards. Most studies have relied on general simplification without involving standard parameters into readability, hence the resulting configuration remaining largely unmeasured. Further, there is scarce research on domain-specific text simplification (e.g., in ESL contexts), with prior investigations focusing on either non-specialized texts or multilingual contexts. Thus, usability of AI software in domain-specific text simplification involving specialized text (e.g., scientific literature in applied sciences, linguistics etc.) and the way this AI-based simplification can be tailored to meet specific academic goals, remains largely unexplored. While cross-domain relevance is mostly facilitating, its functionality reduces when simplification operations are required for technically specific purposes, because effective simplification needs application guidelines such as readability scores or categorization. Further, without these criteria, AI-dependent text simplification may suffer lack of scalability and automation criteria, particularly in a university context that has learners from diverse cultural and linguistic backgrounds.

AI-driven text clarification has substantial significance in educating undergraduate learners for research-oriented learning however, certain challenges need to be addressed for satisfactory and standard utilization of AI in this regard. Notably, text simplification requires standard integration for ESL learners and validation of the designed protocols through empirical methods (e.g., latent semantic analysis). Based on this, we formulated following research questions. Each question logically furthers the critical inquiry, which is grounded in the aforementioned theories:

(i) How can an integrated approach to readability criteria (i.e., Flesch-Kincaid Grade Level, Flesch-Kincaid Reading Ease & McAlpine *EFLAW* Readability Score) be utilized to assess AI-based text fidelity?

(ii) To what extent can the AI output be considered input-representative in latent semantic analysis?

(iii) What implications do the criterion-based AI text simplification have for ESL undergraduates in relation to scientific text accessibility?

Overall, this study intended to discover in depth the semantic model capability of Chat-GPT in elevating readability of scientific abstracts in the areas of linguistics and language learning. The next section presents a critical review of recent related works while highlighting the gaps. The theoretical grounds are also given, elaborating how they interrelated to support our research aims.

2. Materials and Methods

This study adopted a quantitative research design with positivist parameters that required statistical data for analysis.

The material and data preparation are elaborated below:

2.1. Textual Data

Four language and linguistics journals were selected based on Scimago ranking and standards; these were (a) Applied Linguistics, (b) Computer Assisted Language Learning, (c) Language Teaching Research, and (d) RELC Journal. From their recent volumes, a total of 100 abstracts were selected with 25 each journal.

2.2. Data Preprocessing

The data was carefully preprocessed at the preliminary stage so that a comprehensive data analysis could later be conducted. The steps-wise data preprocessing is given below:

2.2.1. Initial Readability Assessment

The initial tripartite readability metrics were gauged applying the Flesch-Kincaid Reading Ease (FKRE) score, Flesch- Kincaid Grade Level (FKGL) score and McAlpine *EFLAW* readability score (MRS). These criteria were adopted to measure readability through both native and ESL readability criteria as the first two standards are not specific to native or non-native learners, while the third readability criteria were specifically designed for ESL learners by McAlpine[18]. The formulae are respectively below:

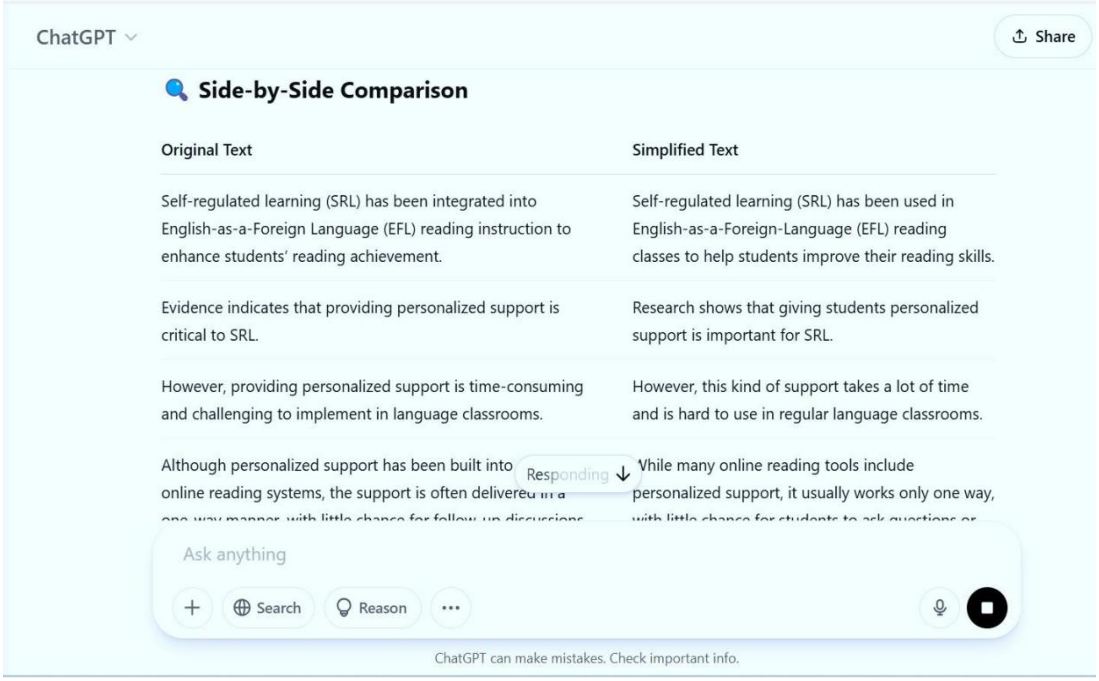
$$\begin{gathered}\text{FKRE} = 206.835 - 1.015 \\ \times(\text{Total Sentences/Total Words}) \\ -84.6\times(\text{Total Words/Total Syllables})\end{gathered} \tag{1}$$

$$\begin{gathered}\text{FKGL} = 0.39\times(\text{Total Sentences/Total Words}) \\ +11.8\times(\text{Total Words/Total Syllables}) - 15.59\end{gathered} \tag{2}$$

$$\begin{gathered}\text{MRS} = 0.4\times(\text{Total Sentences/Total Words}) \\ +3.8\times(\text{Total Words/Total Syllables})\end{gathered} \tag{3}$$

2.2.2. Chat-GPT Simplification

Initially, the text corpus (i.e., the abstracts) was configured into Chat-GPT 4.0 for simplification, storing each dataset. GPT 4.0 software was used because of transformer architecture capability that is more nuanced, as compared to that of the other tools, in adapting linguistics abstractions and discourse-level comprehension in text simplification (TS). Further, this system is more request-oriented, uses large data-driven language models that are textually and contextually more exposed to authentic and diverse language patterns, allowing it to function beyond the prescriptive algorithms, generally used by other software. The input query for TS was, "*Please simplify the following text to be readable at an undergraduate ESL level, without adding or deleting any information*." GPT 4.0, using TextStat model, generated the simplified data as the output, that was criteria-evaluated using FKRE, FKGL and MRS. The resulting quantifications were stored as separate datasets. **Figure 1** shows an actual abstract, transformed in this study:

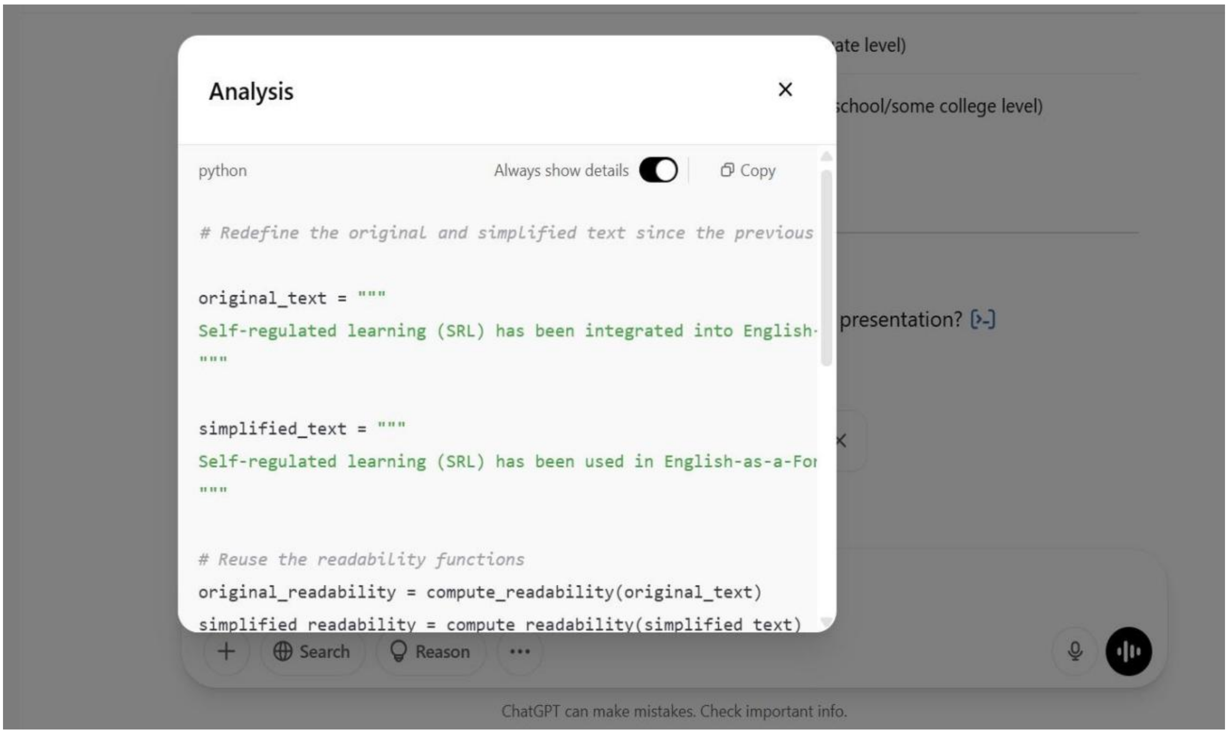


🔍 Side-by-Side Comparison

Original Text	Simplified Text
Self-regulated learning (SRL) has been integrated into English-as-a-Foreign Language (EFL) reading instruction to enhance students' reading achievement.	Self-regulated learning (SRL) has been used in English-as-a-Foreign-Language (EFL) reading classes to help students improve their reading skills.
Evidence indicates that providing personalized support is critical to SRL.	Research shows that giving students personalized support is important for SRL.
However, providing personalized support is time-consuming and challenging to implement in language classrooms.	However, this kind of support takes a lot of time and is hard to use in regular language classrooms.
Although personalized support has been built into online reading systems, the support is often delivered in a one-way manner, with little chance for follow-up discussions	While many online reading tools include personalized support, it usually works only one way, with little chance for students to ask questions or

Figure 1. A sample transformation from the data: The original version (on the left) *from* Using Artificial Intelligence Chatbots to Support EFL Students' Self-Regulated Reading[19]; and the simplified version (on the right) in Chat-GPT 4.0.

During the TextStat model's internal computational analysis of the above data sample, the Python results was captured real-time using the "analysis view". **Figures 2** and **3** show the computations and results respectively:


ate level)
school/some college level)
presentation?
Analysis
python
Always show details
Copy
Search
Reason
ChatGPT can make mistakes. Check important info.

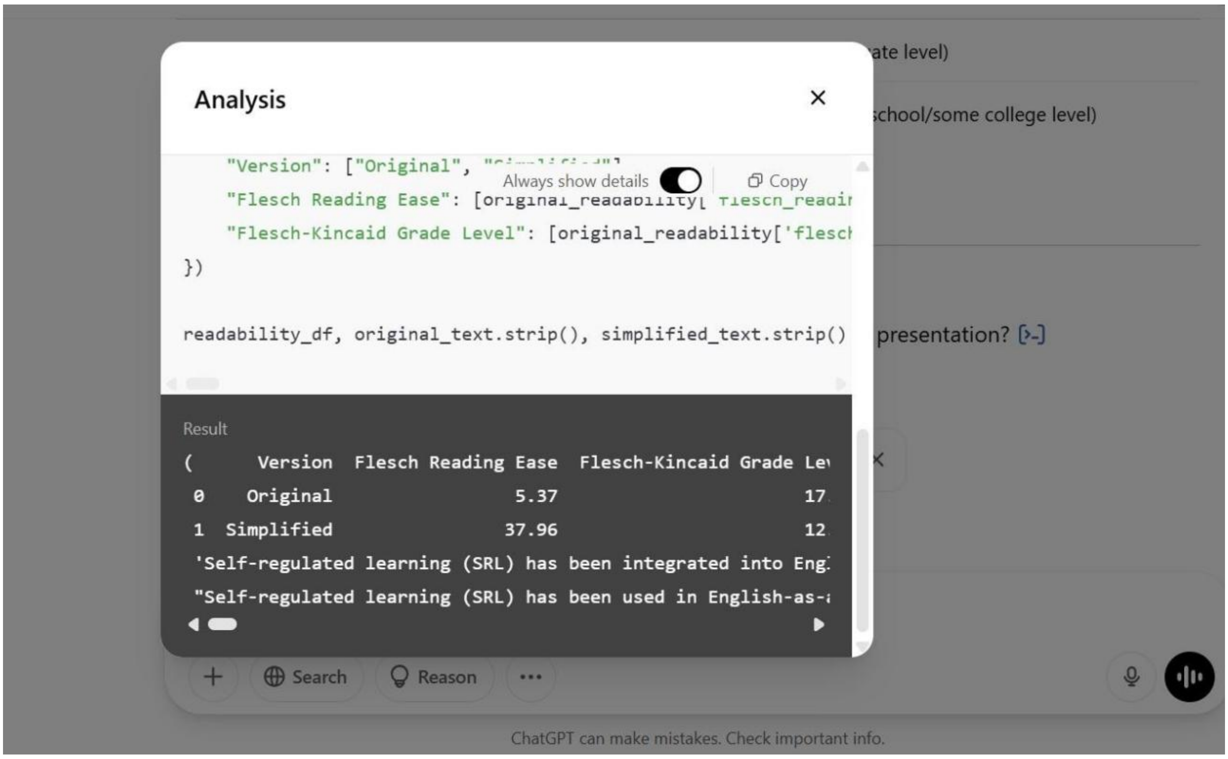
```
# Redefine the original and simplified text since the previous

original_text = """
Self-regulated learning (SRL) has been integrated into English-
"""

simplified_text = """
Self-regulated learning (SRL) has been used in English-as-a-For
"""

# Reuse the readability functions
original_readability = compute_readability(original_text)
simplified_readability = compute_readability(simplified_text)
```

Figure 2. Python's real-time capture of computational analysis of the abstracts.


ate level)
school/some college level)
presentation?
Analysis
Always show details
Copy
Result
Search
Reason
ChatGPT can make mistakes. Check important info.

```
    "Version": ["Original", "Simplified"],
    "Flesch Reading Ease": [original_readability['flesch_readi
    "Flesch-Kincaid Grade Level": [original_readability['flesch
})

readability_df, original_text.strip(), simplified_text.strip()
```

```
(      Version  Flesch Reading Ease  Flesch-Kincaid Grade Le
 0    Original                 5.37                       17
 1  Simplified                37.96                       12
 'Self-regulated learning (SRL) has been integrated into Eng
 "Self-regulated learning (SRL) has been used in English-as-a
```

Figure 3. Readability analysis of the original and simplified versions.

2.2.3. Content Fidelity Assessment

The post-configuration content fidelity was analyzed by pair-wise latent semantic analysis (LSA) using Python SciKit- Learning library. The LSA modeling was selected to generate a document-term matrix, indicating term-frequency, with columns representing textual segments of the selected abstracts, and rows depicting concepts and labels.

Firstly, 100 abstracts were prepared with each entry as a string, resulting into the data as:

$$\text{“abstracts} = [\text{“Abstract 1 text”, “Abstract 2 text”,} \\ \ldots, \text{“Abstract 100 text”].”}$$

Secondly, vectorizer parameters were configured to perform TF-IDF vectorization for obtaining the term-document matrix of the abstracts' data. The parameters were set to:

$$\text{“vectorizer} = \text{TfidfVectorizer(max\_features} \\ = 1000, \text{max\_df} = 0.5, \text{min\_df} = 1\text{”}$$

The above parameters were configured because “max_features = 100” limited the matrix to the highest weighted 1000 notional terms on TF-IDF scores; max_df = 0.5 excluded, from the collected data of 100 abstracts, the terms that appeared more than 50%, hence, reducing over-collection of common terms, and min_df = 1 included the terms that appeared in a minimum of one abstract, to ensure inclusion of terms unique to the data. Next, “X_tfidf = vectorizer.fit_transform(abstracts)” was applied for vectorization. It converted the textual data into numerical form, as term-document matrix in shape (100, 1000), with rows representing the abstracts while columns corresponding to unique terms in these abstracts.

Thirdly, Singular Value Decomposition (SVD) was applied for reducing the dimensionality, using A = $U\Sigma V^T$*[Wherein, A = Matrix; U = Left matrix (Terms), and Σ = Diagonal matrix (Concepts) and* V^T *= Right matrix (Documents)]* determining the latent semantic structures in the data. By morphological reduction, crucial information was preserved glossing the semantic contribution of content words in (n_components=100). Thus, the semantic noise of the content-neutral words neutral words (e.g., “over”, “the”) was eliminated. The initialized SVD model was applied to TF-IDF matrix that resulted in the reduced matrix, “X_svd”, of 100 abstracts, carrying the core semantic features.

The max_features were incrementally tested (*n* = 50/increment) until 99% explained SVD variance was achieved with max_features=1000, (*n* = 100 abstracts) implying that crucial information was preserved.

Lastly, after truncating the matrices, the latent semantic representation was obtained for 100 abstracts, that represented:

(i) Term by Topic matrix (**U_k**) corresponding to the abstracts’ “terms × latent topics”.

(ii) Topic Strength Matrix (**$\sum_k$**) representing the identified “latent topics”.

(iii) Document by Topic Matrix (**V_k^T**) representing “the documents × latent topics”.

Figure 4 illustrates the above LSA process:

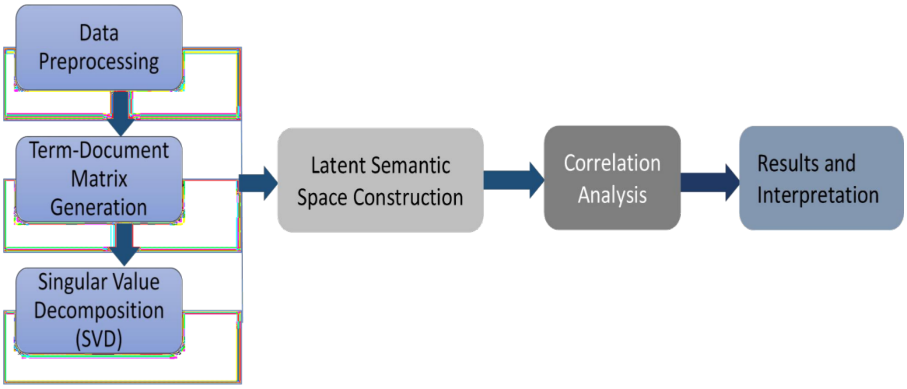

Data Preprocessing
Term-Document Matrix Generation
Singular Value Decomposition (SVD)
Latent Semantic Space Construction
Correlation Analysis
Results and Interpretation

Figure 4. Readability analysis of the original and simplified version.

Resulting from the initial rational querying, the latent topics were derived in the course of LSA. **Table 1** shows the yielded topics (i.e., the matrix terms).

Table 1 depicts the LSA topic distribution in the journals by integrating the mean difference spread for each topic, while the *t* and *p* values shows the significance of difference of the values across the selected journals. There is patternicity of difference indicating how well topics relate to the journals. "CD" indicated a statistically robust difference ($t = 5.76$, $p < 0.001$), with the journals "AL" (ΔM = 0.31 ± 0.07) and "CALT" (ΔM = 0.43 ± 0.09), implying that AL and CALT demonstrate a greater emphasis on the domain of Curriculum Design (CD), as compared to the other two journals, "RELCJ" and "LTR". Likewise, L2A exhibited significant varying mean difference distribution ($t = 4.46$, $p < 0.001$), with "AL" (ΔM = 0.40 ± 0.05) that implied a greater orientation of this journal towards the topic of L2 assessment, as compared to "RELCJ" and "LTR". Contrarily, the journal "LTR" showed lower mean difference (ΔM = 0.24 ± 0.08) in this regard. In addition, "LCT" had a significant mean difference ($t = 8.32$, $p < 0.001$) with "LTR" (ΔM = 0.42 ± 0.09) and "CALT" (ΔM = 0.37 ± 0.07), underscoring a greater focus on learned centered pedagogical strategies, as compared to "AL" (ΔM = 0.23 ± 0.06). and "RELCJ" (ΔM = 0.28 ± 0.06). Similarly, the topic, "LP" demonstrated a significant mean difference ($t = 6.92$, $p < 0.001$) with the journal, "RELCJ" (ΔM= 0.42 ± 0.10) that indicated the highest mean value. This underscored the higher significance that "RELCJ" assigns to L2 policymaking, as compared to "CALT" (ΔM = 0.39 ± 0.05) and "AL" (ΔM= 0.26 ± 0.09). Additionally, the topics, "ML" and "SLA", also demonstrated significant differences. "ML" had a significant difference ($t = 5.53$, $p < 0.001$) with "CALT" (ΔM = 0.43 ± 0.07), while SLA also received a significant emphasis ($t = 2.78$, $p = 0.000$) in CALT (ΔM = 0.46 ± 0.04), as compared to RELCJ (ΔM = 0.23 ± 0.04) and LTR (ΔM = 0.23 ± 0.09).

However, the topic, "L2T" demonstrated an insignificant difference ($t = 1.29$, $p = 0.202$), similar to "TLL" ($t = 1.19$, $p = 0.240$) across the selected journals indicating an even distribution. These findings indicated that these topics did not demonstrate any marked preference across the journals. Overall, the findings in **Table 1** showed a normal SLA topic distribution across the journals.

Table 1. LSA Topic Distribution of Journals.

Topic	AL (M±SD)	CALT (M±SD)	LTR (M±SD)	RELCJ (M±SD)	*t*	*p*
CD	0.31 ± 0.07	0.43 ± 0.09	0.20 ± 0.10	0.24 ± 0.06	5.76	0.000
L2A	0.40 ± 0.05	0.33 ± 0.07	0.24 ± 0.08	0.31 ± 0.08	4.46	0.000
L2L	0.44 ± 0.09	0.36 ± 0.10	0.40 ± 0.04	0.29 ± 0.04	3.26	0.002
L2T	0.23 ± 0.06	0.25 ± 0.06	0.24 ± 0.06	0.41 ± 0.07	1.29	0.202
LCT	0.23 ± 0.06	0.37 ± 0.07	0.42 ± 0.09	0.28 ± 0.06	8.32	0.000
LP	0.26 ± 0.09	0.39 ± 0.05	0.31 ± 0.09	0.42 ± 0.10	6.92	0.000
ML	0.33 ± 0.07	0.43 ± 0.07	0.40 ± 0.04	0.42 ± 0.05	5.53	0.000
SLA	0.41 ± 0.09	0.46 ± 0.04	0.23 ± 0.09	0.23 ± 0.04	2.78	0.007
TLL	0.42 ± 0.06	0.44 ± 0.07	0.34 ± 0.06	0.43 ± 0.06	1.19	0.240
ToT	0.39 ± 0.07	0.34 ± 0.07	0.39 ± 0.06	0.23 ± 0.07	2.77	0.008

Note. [**Journals**:
AL = Applied Linguistics;
CALT = Computer Assisted Language Learning; LTR = Language Teaching Research;
RELCJ = RELC Journal].
[**Topics**:
CD = Curriculum Design; L2A = L2 Assessment; L2L = L2 Learning;
L2T = L2 Teaching;
LCT = Learner-Centered Teaching; LP = Language Policy;
ML = Multilingualism;
SLA = Second Language Acquisition;
TLL = Technology in Language Learning;
ToT = Training of Trainers].

2.2.4. Expert Evaluation

Despite careful and rigorous preprocessing, the resulting transfigurations were submitted to two field experts for review. A randomized dataset of 20 abstracts was provided for independent rating using five evaluation criteria, (a) Topic Relevance, (b) Topic Accuracy, (c) Topic Completeness, (d) Topic Coherence and (e) Result Agreement, along the scale of 1–5 (i.e., 1 = *Very Poor*; 2 = *Poor*; 3 = *Fair*; 4 = *Good*; 5 = *Excellent*). The Cohen's kappa parametric analysis of inter-rater reliability showed a well-preserved threshold ($\kappa = 0.82$). The experts' data simplification rating and the data's LSA scores were compared by Pearson's correlation analysis. Further, segmented data analysis was conducted to examine the effect of the text length on the extracted data, with a configured threshold <450 words and >450 words. This threshold was reasonable because the abstracts of the review subset ranged between 200 to 600 words thus furnishing 450 as around the median.

2.2.5. Normality Test

The normality of the data was assessed through Shapiro-Wilk test. The following section presents the results and interpretations.

3. Results

This section presents the results of descriptive and inferential data analyses. The descriptive analysis consists of readability and semantic similarity measure, and non-parametric normality tests. The inferential analysis comprises the *t*- test and correlation analysis.

3.1. Readability and Semantic Similarity Analysis and Normality Testing

Extracted from four research journals, a total of 100 abstracts from research articles were transfigured performing Flesch-Kincaid Grade Level, Flesch-Kincaid Reading Ease and McAlpine *EFLAW* Readability tests. Their semantic fidelity was assessed during latent semantic analysis. Further, text length and expert similarity were estimated. **Table 2** depicts these elements with regard to the selected journals:

Table 2. Readability, Text Length and Semantic Similarity.

Measure	AL (n = 25)	CALT (n = 25)	LTR (n = 25)	RELCJ (n = 25)	Aggregated (n = 100)
FKGL (Original)	11.31 ± 1.78	11.64 ± 2.62	14.33 ± 1.49	11.96 ± 1.63	12.31 ± 2.27
FKGL (Chat-GPT)	4.76 ± 1.00	4.58 ± 0.75	5.74 ± 1.22	4.45 ± 0.87	4.88 ± 1.10
FKGL Change	6.55 ± 2.31	7.06 ± 2.76	8.59 ± 1.57	7.51 ± 2.04	7.43 ± 2.34
FKRE (Original)	42.65 ± 9.16	40.04 ± 13.46	36.92 ± 6.75	37.06 ± 8.30	39.17 ± 10.02
FKRE (Chat-GPT)	84.24 ± 5.83	85.81 ± 5.37	79.72 ± 6.84	86.58 ± 5.35	84.09 ± 6.45
FKRE Change	41.59 ± 11.86	45.78 ± 14.30	42.80 ± 7.70	49.52 ± 9.84	44.92 ± 11.61
EFLAW (Original)	45.32 ± 4.21	44.89 ± 5.02	42.13 ± 4.87	43.75 ± 3.94	44.52 ± 4.56
EFLAW (Chat-GPT)	81.55 ± 3.98	83.72 ± 4.11	78.94 ± 5.06	84.61 ± 4.20	82.21 ± 4.34
EFLAW Change	36.23 ± 4.65	38.83 ± 5.18	36.81 ± 4.91	40.86 ± 4.78	37.69 ± 4.88
Text Length (Original)	408.80 ± 64.00	344.32 ± 53.06	491.96 ± 62.94	265.04 ± 34.16	377.53 ± 99.84
Text Length (Chat-GPT)	226.00 ± 31.11	251.12 ± 29.57	251.36 ± 35.82	230.28 ± 38.86	239.72 ± 35.84
LSA Similarity	0.60 ± 0.30	0.65 ± 0.28	0.91 ± 0.15	0.72 ± 0.21	0.71 ± 0.26
Expert Similarity	0.90 ± 0.14	0.67 ± 0.10	0.80 ± 0.21	0.73 ± 0.13	0.78 ± 0.16
Accuracy Score (1-5)	4.2 ± 0.75	3.4 ± 0.49	3.6 ± 1.02	3.6 ± 0.49	3.89 ± 1.1

Note. [**Journals**:
AL = Applied Linguistics;
CALT = Computer Assisted Language Learning; LTR = Language Teaching Research;
RELCJ = RELC Journal].

Table 2 shows the effect of AI-simplification of the research abstracts across the involved journals, "Applied Linguistics" (AL), "Computer Assisted Language Testing" (CALT), "Language Teaching Research" (LTR) and "RELC Journal" (RELCJ). The simplified version demonstrated a marked increase in text readability in all the applied measures. **Table 1** indicates that the aggregated FKGL was significantly decreased from 12.31 (SD = 2.27) to 4.88 (SD = 1.10) indicating improved text readability. Similarly, FKRE mean score increased from 39.17 (SD = 10.02) to 84.09 (SD = 6.45) at the post-simplification stage. Further, McAlpine EFLAW score followed the corresponding trend, rising from 44.52 (SD = 4.56) to 82.21 (SD = 4.34), showing that the abstracts became substantially easier to read.

Moreover, a marked decrease in text length was observed at the post-simplification stage at which the original text consisted of 377.53 words (SD = 99.84), while AI-based simplified text comprised 239.72 words (SD = 35.84) after. However, LSA revealed a marked semantic similarity at 0.71 (SD = 0.26) between the pre- and post-simplified texts. This implied that semantic fidelity was intact, despite the reduced text length. Correspondingly, the experts rating validated the fidelity with the mean similarity score at 0.78 (SD = 0.16), while the accuracy score was 3.89/5 (SD = 1.10) that suggested that key concepts and meanings were substantially preserved after the AI-simplification. These findings endorse the ability of chat-GPT's AI-models in improving readability of scientific text, while preserving the semantic fidelity, particularly suiting the learning needs of undergraduates.

To justify the assumptions of the suitable parametric tests, each readability measure's distribution was evaluated through Shapiro-Wilk test. This measure was selected because of its rigour in detecting non-normal distributions. **Table 3** presents the results:

Table 3. Shapiro-Wilk Statistics of the Readability Measures across the Journals.

Measure	Journal	Shapiro-Wilk Statistics (W)	*p*
FKGL (Original)	AL	0.97	0.241
FKGL (Original)	CALT	0.94	0.051
FKGL (Original)	LTR	0.98	0.445
FKGL (Original)	RELCJ	0.97	0.226
FKGL (Chat-GPT)	AL	0.97	0.298
FKGL (Chat-GPT)	CALT	0.96	0.164
FKGL (Chat-GPT)	LTR	0.97	0.255
FKGL (Chat-GPT)	RELCJ	0.97	0.282
FKRE (Original)	AL	0.96	0.117
FKRE (Original)	CALT	0.95	0.080
FKRE (Original)	LTR	0.97	0.268
FKRE (Original)	RELCJ	0.96	0.142
FKRE (Chat-GPT)	AL	0.97	0.216
FKRE (Chat-GPT)	CALT	0.95	0.056
FKRE (Chat-GPT)	LTR	0.97	0.299
FKRE (Chat-GPT)	RELCJ	0.96	0.187
EFLAW (Original)	AL	0.95	0.078
EFLAW (Original)	CALT	0.94	0.053
EFLAW (Original)	LTR	0.97	0.211
EFLAW (Original)	RELCJ	0.96	0.143
EFLAW (Chat-GPT)	AL	0.97	0.318
EFLAW (Chat-GPT)	CALT	0.96	0.120
EFLAW (Chat-GPT)	LTR	0.98	0.390
EFLAW (Chat-GPT)	RELCJ	0.97	0.266
TextLength (Original)	AL	0.98	0.467
TextLength (Original)	CALT	0.94	0.033
TextLength (Original)	LTR	0.96	0.141
TextLength (Original)	RELCJ	0.97	0.248
TextLength (Chat-GPT)	AL	0.97	0.277
TextLength (Chat-GPT)	CALT	0.97	0.346
TextLength (Chat-GPT)	LTR	0.97	0.292
TextLength (Chat-GPT)	RELCJ	0.97	0.310
LSA Similarity	AL	0.97	0.313
LSA Similarity	CALT	0.95	0.076
LSA Similarity	LTR	0.98	0.498
LSA Similarity	RELCJ	0.97	0.263
Expert Similarity	AL	0.96	0.142
Expert Similarity	CALT	0.96	0.167
Expert Similarity	LTR	0.97	0.288
Expert Similarity	RELCJ	0.96	0.154
Accuracy Score	AL	0.98	0.413
Accuracy Score	CALT	0.97	0.295
Accuracy Score	LTR	0.97	0.309
Accuracy Score	RELCJ	0.97	0.321

Table 3 shows that the data distribution was normal, with only a single value at $p < 0.05$. The majority of values met the normality assumption, that supported performing parametric test (*t*-test) for comparing the pre- and post-simplification means scores of readabilities. In view of the robustness of *t*-tests in the face of minor normality issues, the overall distributional pattern of the data allowed parametric paired-sample *t*-testing.

3.2. Pre- and Post-Simplification Readability Differences across the Measures

Paired-sample *t*-tests were performed for comparing the mean scores of the readability criteria to estimate the possible post-simplification readability increase. **Table 4** depicts the results of FKGL *t*-test.

A paired-sample *t*-test was performed to compare pre- and post-simplification FKGL readability mean scores. The findings showed a statistically significant readability increase after textual simplification, t (99) = 32.80, $p < 0.001$, d = 1.16. The large effect size (d) demonstrated a substantial reduction in the reading grade level required.

Table 5 presents the results of FKRE *t*-test.

Next, a paired-sample *t*-test was conducted to compare the pre- and post-simplification FKRE mean scores. A statistically significant increase was observed in reading, t (99) = 20.89, $p < 0.001$, $d = -0.63$. To note, FKRE scale is opposite to FKGL and other descending scales, thus the negative effect size (d) indicates the increased readability at the post-simplification stage. In FKRE scale, higher scores indicate greater readability.

Table 6 depicts the results of *EFLAW t*-test.

Lastly, a paired sample *t*-test was conducted to estimate the effect of text simplification on linguistic complexity and writing errors through using *EFLAW* mean scores. The results indicated a statistically significant complexity reduction between pre- and post-simplification means, t (99) = 18.17, p < 0.001, d = 1.63. The large effect size (d) demonstrated that AI-simplification markedly increased the clarity of the abstracts.

Overall, the results indicated a significant increase in readability after AI's intervention in the simplification process, with large effect sizes, differentiating pre- and post-simplification readability differences. Cohen's (1988) categorization (d) was used to interpret the effect size[20]. Thus, the effect sizes (FKGL = 3.28; FKRE =2.09; *EFLAW* = 1.82) indicated a substantial increase in readability after ChatGPT's textual simplification.

Table 4. Descriptive statistics and Paired-Sample *t*-Test Results for FKGL Pre- and Post-Simplification Readability Scores.

Paired Sample *t*-Test				**95% Interval of Confidence the Difference**				
Groups	**Mean**	**Std. Deviation**	**Std. Error Mean**	**Lower**	**Upper**	***t***	***df***	**Sig. (2-Tailed)**
Pre- and Post-Simplification	5.84	1.78	0.18	3.49	6.19	32.80	99	0.001

Table 5. Descriptive Statistics and Paired-Sample *t*-Test Results for FKRE Pre- and Post-Simplification Readability Scores.

Paired Sample *t*-Test				**95% Interval Confidence of the Difference**					
Groups	**Mean**	**Std. Deviation**	**Std. Error Mean**	**Lower**	**Upper**	***t***	***df***	**Sig. (2-Tailed)**	***d***
Pre- and Post-Simplification	32.59	15.59	1.56	29.50	35.68	20.89	99	0.001*	−0.63

Table 6. Descriptive Statistics and Paired-Sample *t*-Test Results for *EFLAW* Pre- and Post-Simplification Readability Scores.

Paired Sample *t*-Test				**95% Interval Confidence of the Difference**					
Groups	**Mean**	**Std. Deviation**	**Std. Error Mean**	**Lower**	**Upper**	***t***	***df***	**Sig. (2-Tailed)**	***d***
Pre- and Post-Simplification	8.63	4.74	0.47	7.69	9.57	18.17	99	0.001*	1.63

3.3. Comparison of the Readability Metrics across the Journals

Based on the findings of the *t*-tests, the post-hoc Brown-Forsythe test was conducted to compare the readability metrics across the journals. In this test, unequal variances were observed therefore Welch's ANOVA test was performed, followed by Games-Howell post-hoc analysis to assess the intergroup differences. This was to ensure the validity of the observed differences and enhanced reliability. Moreover, these tests determined if the readability metrics significantly differed across the four journals. **Table 7** provides the results:

Table 7. Brown-Forsythe, Welch's ANOVA and Games-Howell Post-hoc Tests.

Measure	Test Type	F-Statistic	*p*	Post-Hoc Comparison (Games-Howell)	*p* for Pairwise Comparisons
FKGL (Original)	Brown-Forsythe	5.76	0.000	AL vs CALT AL vs LTR AL vs RELCJ	0.010 0.004 0.150
	Welch's ANOVA	6.12	0.000	CALT vs LTR CALT vs RELCJ	0.020 0.070
FKRE (Original)	Brown-Forsythe	4.46	0.000	AL vs CALT AL vs LTR AL vs RELCJ	0.030 0.010 0.080
	Welch's ANOVA	4.65	0.001	CALT vs LTR CALT vs RELCJ	0.060 0.040
EFLAW (Original)	Brown-Forsythe	5.92	0.000	AL vs CALT AL vs LTR AL vs RELCJ	0.020 0.003 0.140
	Welch's ANOVA	5.91	0.000	CALT vs LTR CALT vs RELCJ	0.050 0.090

Note. [**Journals**:
AL = Applied Linguistics;
CALT = Computer Assisted Language Learning; LTR = Language Teaching Research;
RELCJ = RELC Journal].

Firstly, the findings for FKGL were observed. Brown-Forsythe test demonstrated a statistically robust difference in FKGL for the selected journals, F (3, 89.52) = 5.76, *p* < 0.001. Correspondingly, Welch's ANOVA corroborated this difference, F (3, 84.67) = 6.12, *p* < 0.001. Further, post-hoc Games-Howell analysis indicated that FKGL were statistically significant for "AL" relative to "CALT" (*p* = 0.010) and "LTR" (*p* = 0.004). However, FKGL were not significantly different for "RELCJ" (*p* = 0.150). The findings further showed that "CALT" was significantly different from "LTR" (*p* = 0.020), like "CALT" and "RELCJ" that also were significantly different (*p* = 0.070).

Secondly, the FKRE scores, in Brown-Forsythe test, were also statistically different for the selected journal, F (3, 88.63) = 4.46, *p* < 0.001. Further, Welch's ANOVA showed a statistical difference across the journals F (3, 86.51) = 4.65, *p* = 0.001. Specifically, for each journal, Games-Howell analysis demonstrated that "AL" was significantly different from "CALT" (*p* = 0.030) and "LTR" (*p* = 0.010). However, similar to FKGL, FKRE scores for RELCJ were not significantly different (*p* = 0.080). Additionally, significant differences were observed for "CALT" and "LTR" (*p* = 0.060), and "CALT" and "RELCJ" (*p* = 0.040) with respect to FKRE scores.

Lastly, the results of Brown-Forsythe test for *EFLAW* scores and journals indicated a statistically robust difference, F (3, 90.37) = 5.92, *p* < 0.001. It was confirmed by Welch's ANOVA, F (3, 85.44) = 5.91, *p* < 0.001. Subsequently, post-hoc Games-Howell comparison revealed that "AL" had a significant difference from "CALT" (*p* = 0.020) and "LTR" (*p* = 0.003). For "RELCJ", the earlier pattern of insignificant difference was demonstrated (*p* = 0.140). Specifically, the *EFLAW* difference between "CALT" and "LTR" was statistically different (*p* = 0.050), while that for "RELCJ" was insignificant (*p* = 0.090).

3.4. Correlational Assessment of Post-Simplification Metrics, Semantic Similarity and Expert Evaluation (Accuracy Rating)

For further inferential exploration, Pearson's correlation analysis was performed to evaluate the associations among readability changes, semantic similarity, text length and expert evaluation. **Table 8** presents the results:

A strong positive correlation was observed between FKGL change and LSA similarity (r = 0.70, *p* = 0.000), indicating that improvements in readability via AI retained the intended textual meaning. This is the highest correlation observed as compared to the others. Next, the correlation between the expert similarity and FKRE change (r = 0.52, *p* = 0.000) and EFLAW change (r = 0.50, *p* = 0.000) were moderate yet statistically significant. This, in terms of meaning preservation, implied exercising caution. Additionally, a weak correlation was observed between LSA similarity assessment and expert accuracy rating (r = 0.25, $p = 0.012$). This indicates a difference between human and machine assessment of content fidelity. Its plausible explanation is discussed in the next section.

Table 8. Correlation of FKGL, FKRE, EFLAW, Accuracy and LSA.

Variable	N	r	Sig. (2-Tailed)	Description
Post-Simplification FKGL	100	0.70	0.000	Strong
Post-Simplification FKRE	100	0.52	0.000	Moderate
Post-Simplification EFLAW	100	0.50	0.000	Moderate
Expert Evaluation (Accuracy)	100	0.25	0.012	Weak

$p < 0.01$ (**), $p < 0.05$ (*).

4. Discussion

This study aimed to explore AI-based Chat-GPT 4.o efficacy in simplifying scientific text, while preserving the semantic fidelity, for ESL undergraduates. The study, thus, simultaneously premised in ESL education, artificial intelligent (AI) and natural language processing (NLP). The findings, to a greater extent, present convincing evidence that transformed-based language models of AI can mediate between complexity and understanding of academic comprehension for undergraduates.

The findings showed a significant decrease in the scores of FKGL, FKRE and *EFLAW* scores, at the post- simplification stage. This implied that the scientific text that Chat-GPT transfigured and simplified became linguistically more comprehensible for undergraduate ESL learners. It indicates the efficacy of the applied readability model in making texts accessible for learner. This is in tandem with a very recent work on the efficacy of the language models, that concluded readability increase after text simplification[21]. It also supports another recent research, that proved that using AI-transfiguring tools may enhance reading comprehension, at different learning levels[22]. However, the strength of the present study, as compared to the existing studies, lies in two aspects: *First*, our study explored AI's simplification capability using three readability metrics, in which *EFLAW* referred specifically to assess the semantic preservation of meanings of the mini-words, idioms and catch phrases, in the context of ESL learners. This study paid special attention to leveraging a metric, for assessing textual simplification enhancement, that was meant more for non-native speakers. Nevertheless, other two metrics (i.e., FKGL & FKRE) have been used for all learners without any obvious specificity. *Second*, our study aimed to explore if significant content fidelity was observed at the post-simplification stage. In this regard, a strong correspondence was found between LSA and FKGL, that indicates that Chat-GPT can be utilized as a scientific research text simplifier AI at the undergraduate level of education, without worrying about meaning loss to a great extent. However, reliance on AI-simplification must be taken with caution because the findings across other two metrics (i.e., FKRE and EFLAW), revealed much lower significance. This stands in contrast with an earlier study that found a value of $p < 0.05$ for FRKE scores at post-simplification stage of medical texts[6]. Further, the present study contributes to and expands notably limited existing literature on AI simplification measured through *EFLAW* metrics for ESL learners. Presumably, no earlier study has tackled AI-simplification of scientific research abstracts in the field of applied linguistics and ESL education, using FKGL, FKRE and *EFLAW* criteria comparison in a single investigation. Collectively, a marked increase was observed in the texts at the post-simplification stage that indicated ample potential of utilizing AI abilities in language learning.

This study is grounded in the critical premise of cognitive load theory, which proposes that overloaded linguistic complexity can overwhelm learners' memory[23]. Based on this, young L2 learners, that generally perceive foreign language learning tasks taxing, may sense complexity as extraneous cognitive overload. AI (e.g., Chat-GPT) reduces this load by simplifying textual channels to better comprehension, thereby motivating learners to utilize cognitive resources for relatively more representative and closely wired information processing. In our study, the significant correlation between FKGL, FKRE and *EFLAW* post-simplification readability scores and semantic similarity showed that the textual meaning was undiluted to a significant extent, thus highlighting that AI can be dependably utilized as a meaningful cognitive resource maximizer for L2 learners. However, AI's capacity of content fidelity preservation, in a controlled environment, should be approached with utmost caution, because it does not, in any case, imply a constant grip on undiluted meaning preservation in all cases. This warrants further research involving different cognitive ad machine models in Chat-GPT and other relevant AI manifestations. Notably, while LSA and FKGL were strongly correlated, FKRE and *EFLAW* had respectively threshold and approximately moderate correlations. A plausible explanation of these results is that FKGL and FKRE indices employ inverse algorithms for scoring, and different readability metric are likely to utilize unique computational and linguistic parameters, manifesting in varied assessments. Thus, the findings support a recent research that noted the phenomenon while assessing readability of a large-scale corpus[24]. Despite variability, the metrics showed significant correlations with semantic preservation that substantiates adequate text simplification ability of AI. However, the variability in semantic similarity may be critically viewed, questioning (i) the robustness of the traditional readability metrics, and (ii) AI's optimal meaning *reconstruction* potential, in relation with scientific research text. This supports the scholars who have recommended to be cautious with traditional readability metrics' limitations, as they generally reply on surface structures[25]. In the current study, using McAlpine *EFLAW* was rationalized to neutralize similar possible effects while capturing sentence complexity viewing the content through ESL lens. Hence, the present study is different from earlier investigations on this behalf.

Besides, this study extends the theoretical stance of Vygotsky's theory of zone of proximal development, that recommends developing ESL learners' linguistic abilities through an exposure to cognitive load, made accessible through cognitive bridging such as scaffolding[8]. The present study illustrates how AI-based GPT can serve as an automated scaffolding, adequately reconstructing textual meanings, to be digestible for undergraduates. Given that Chat-GPT is not precisely conditioned for academic textual simplification, the significant correlations suggest faithful simplification of the research abstracts, that can develop learners' proximal zones. Further, both semantic similarity and accuracy ratings indicate meaning retention despite significant text simplification. The findings are in tandem with recent investigations that showed the ability of AI-based models to retain content fidelity, using BERTScore, natural language query design and model responsiveness assessment[26–28].

4.1. Micro, Meso and Macro Implications

This study offers significant micro-, meso- and macro-implications for incorporating AI tools in ESL instructional designs, learning accessibility and dissemination of scientific knowledge, especially in relation to undergraduate education. The findings demonstrated that Chat-GPT possesses sufficient capability to enhance readability of scientific content while preserving content fidelity during meaning reconstruction by simplifying the text. It suggests that AI-based simplification can increase comprehension of scientific texts, positioning AI-based GPT models from *agentic language relay* tool to *assistive pedagogical resource*, that can mediate between complex scientific text and reading comprehension. Thus, the micro-implications of this study are learner-focused, relating to pedagogy. The findings imply that ESL university learners can utilize AI tools for increasing comprehension of scholarly discourse, and may be trained, through the process, to develop metacognitive reading control. This study, incorporating multiple readability metrics (i.e., FKGL, FKRE & *EFLAW*) and latent semantic analysis (LSA), provides a convincing validation on AI-based text simplification to enhance reading comprehension. Likewise, the meso-implications offer the innovative aspect in that AI simplification tool and strategies can be incorporated in curriculum frameworks for upscaling university learners' reading comprehension of scientific research and access to existing stock of knowledge. This can transform curricular strategies and pedagogical approaches in modern classrooms, leading to provide transformational approach in pedagogy for teachers.

The macro-implications relate to theoretical premise. The findings relate to the cognitive load theory, by providing evidence that textual simplification, mediated by AI models, can efficiently decrease the cognitive load without jeopardizing the essential semantic load of involved concepts[23]. Moreover, the results emphasize formation of the Zone of Proximal Development (ZPD), proposed in the theory of learning and development, indicating how GPT model can serve as an "substitution scaffold"[8]. This can support learner comprehension beyond the current level, thereby providing a motivational load stimulating the learner to learn the inaccessible. This, to a great extent, theoretically functionalize amplified readability and comprehension of dense scientific texts, wherein AI assists learners to engage with texts at a relatively more familiar proximal zone. Additionally, the significant relationship between readability metrics, particularly FKGL, and semantic similarity offers valuable methodological implications for future research on AI-enabled text simplification. The integration of readability metrics and semantic vector modeling illustrates a doubly-validated mechanism, that assures quality in conceptual and structural outcomes.

The macro-implications also encompass epistemological and ethical aspects. The significant semantic similarity indicates preserved content fidelity however, comparatively weaker correlation between expert accuracy and LSA warrants further research on the interpretation of AI-based text simplification by humans in learning landscape. This pinpoints the requisite of relatively an adaptive model of simplification, that may indicate the simplification nuances through cognitive- computational algorithms. This may provide a consistent feedback for refinement.

4.2. Limitations and Recommendations

This study has a few limitations, that simultaneously indicate directions for future research. For conceptual clarity and efficiency, we combine limitations and recommendations in this section, as the latter stem from the former.

Firstly, the study included only the research abstracts from well-known journals in applied linguistics and ESL domain. While full-length texts were not involved, future researchers may extend the data, populating it using articles. They may investigate similar studies in disciplines other than applied linguistics. Additionally, denser data can be utilized for further inquiry involving other strategic topics such as Artificial Intelligence-Based Education (AIED), gamification in L2 learning and STEM.

Secondly, the study investigated the AI text simplification using a concentrated but non-comparative approach, focusing only on one area of study. Future researchers may compare AI text configuration and remodeling, in different areas to measure possible differences that can enlighten further on AI text conversion. Further, researchers may investigate computational text simplification involving other AI platforms such as DeepSeek etc. This may lead to varied results on readability metrics, depending on the internal computational algorithms of the different models. Likewise, comparative studies may involve comparing simplification in different domains such as, AI-based education in IT etc. Such empirical comparisons may assist in determining if text simplification patterns of Chat-GPT remain similar across disciplines and conceptual structures.

Thirdly, this study employed readability metrics of FKGL, FKRE and *EFLAW*, mainly to focus on ESL undergraduates, however, future studies may apply deeper indices such as "System output Against References and against the Input sentence" (SARI) or "Bidirectional Encoder Representations for Transformers Score" (BERTScore), that may yield varied findings, functioning at different semantic levels.

Fourthly, this study followed a quantitative design, relying on statistical data. Qualitative methodology was not utilized in measure readability amplification. Future studies may incorporate qualitative design or mixed method studies, such as combining quantitative metrics with reading-based think-aloud protocols, involving human ESL learners, for real- time semantic and comprehension feedback. This may provide a cognitive feedback after obtaining AI-based computational semantic response to the text.

Lastly, the present study worked with 100 abstracts; the future research may include a larger sample for more reliable results. Moreover, this study used only one prompt for simplification while future researchers may compare multi-prompt simplification results that may lead to a different output.

5. Conclusions

This study investigated the efficacy of AI-based models, such as Chat-GPT 4.0, in simplifying research abstracts, to undergraduate level, extracted from four leading journals in the discipline of applied linguistics and ESL education. The descriptive and inferential results provided empirical evidence that AI textual facilitation, measured by FKGL, FKRE and *EFLAW* metrics, substantially increased content readability. Moreover, latent semantic analysis, expert evaluation and accuracy score illustrated that content fidelity was sufficiently preserved, confirming AI tools' ability to facilitate ESL learners' scientific comprehension by enhancing readability in instructional context.

Theoretically, this study promotes the principles of Vygotsky's Zones of Proximal Development and Sweller's cognitive load theory, framing AI generally and its simplification platforms specifically, as a cognitive scaffold in learning[8,23]. From a methodological perspective, this study presents a novel operationalization of readability metrics (FKGL, FKRE & *EFLAW*), latent semantic analysis and expert evaluation. This provides an empirical reference point for further research in AI-facilitated comprehension, especially at undergraduate level of education. This may help in timely research on enhancing ESL learners' reading abilities that are the need of the time, especially in developing countries[29]. The methodology can also be used to prepare materials for enhancing other language skills, such as speaking, in developing contexts, as indicated in earlier studies[30].

However, modest relationship between automated and cognitive evaluation of content fidelity, denoted by semantic similarity, warrants caution in unmonitored reliance on AI textual outcomes. Yet, it is shown that AI's Chat-GPT model can be utilized for enhancing the learner comprehension in a supervised and reflective environment. In this way, the study affirms that AI text simplification can sufficiently mediate between scientific text complexity and knowledge accessibility for the learner. This has future implications for designing pedagogical goals utilizing the related abilities of generative models, to advance learning.

Author Contributions

Conceptualization, methodology, formal analysis and writing, M.A.; IT software, data curation and analysis validation, O.A.J.; resources and review, N.R.M.N.; resources and visualization, S.A. All authors have read and agreed to this version of the manuscript.

Funding

This work received no external funding.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

This study is based on a successfully defended doctoral project. The raw data files can only be shared after the mandatory embargo period passes. However, basic data may be made available on a reasonable request.

Acknowledgments

The corresponding author extends her sincere thanks to Professor Jerome F. Keating (University of Syracuse, USA) for his genuine interest in validating the author's conceptualization and line of model inquiry.

Conflicts of Interest

The authors declare no conflict of interest engaging this research.

References

[1] Cáceres-Serrano, P., Alvarado-Izquierdo, J.M., 2017. The effect of contextual and socioeconomic factors on reading comprehension levels. Modern Journal of Language Teaching Methods. 7(8), 76–85.

[2] Corso, H.V., Cromley, J.G., Sperb, T., et al., 2016. Modeling the relationship among reading comprehension, intelligence, socioeconomic status, and neuropsychological functions: The mediating role of executive functions. Psychology & Neuroscience. 9(1), 32. DOI: https://doi.org/10.1037/pne0000036

[3] Velilla Sánchez, M.Á., 2025. Recontextualizing knowledge in academic video publications: A discourse analysis of multimodal science dissemination. Pragmatics and Society. 16(5), 676–700. DOI: https://doi.org/10.1075/ps.23124.vel

[4] Wang, S., Liu, X., Zhou, J., 2022. Readability is decreasing in language and linguistics. Scientometrics. 127(8), 4697–4729. DOI: https://doi.org/10.1007/s11192-022-04427-1

[5] Plavén-Sigray, P., Matheson, G.J., Schiffler, B.C., et al., 2017. The readability of scientific texts is decreasing over time. Elife. 6, e27725. DOI: https://doi.org/10.7554/elife.27725

[6] Picton, B., Andalib, S., Spina, A., et al., 2025. Assessing AI simplification of medical texts: readability and

content fidelity. International Journal of Medical Informatics. 195, 105743. DOI: https://doi.org/10.1016/j.ijmedinf.2024.105743

[7] Araújo, S., Aguiar, M., 2023. Simplifying specialized texts with AI: a ChatGPT-based learning scenario. In Proceedings of the International Conference in Information Technology and Education, Singapore, June 2023; pp. 599–609. DOI: https://doi.org/10.1007/978-981-99-5414-8_55

[8] Vygotsky, L.S., 1978. Mind in society: The development of higher psychological processes. Harvard University Press: Cambridge, MA, USA.

[9] Raisch, S., Krakowski, S., 2021. Artificial intelligence and management: The automation–augmentation paradox. Academy of Management Review. 46(1), 192–210. DOI: https://doi.org/10.5465/amr.2018.0072

[10] Siontis, K.C., Attia, Z.I., Asirvatham, S.J., et al., 2024. ChatGPT hallucinating: can it get any more human-like? European Heart Journal. 45(5), 321–323. DOI: https://doi.org/10.1093/eurheartj/ehad766

[11] Li, S., 2024. A Cross Language Information Retrieval Model Based on Latent Semantic Analysis. In: Intelligent Computing Technology and Automation. IOS Press: Amsterdam, Netherlands. pp. 1082–1089. DOI: https://doi.org/10.3233/atde231290

[12] Egger, R., Gokce, E., 2022. Natural language processing (NLP): An introduction: making sense of textual data. In: Applied Data Science in Tourism: Interdisciplinary Approaches, Methodologies, and Applications. Springer International Publishing: Cham, Switzerland. pp. 307–334. DOI: https://doi.org/10.1007/978-3-030-88389-8_15

[13] Jeon, C.H., Shin, J.Y., Ryu, S., 2025. Analyzing Student Communication Patterns in Science Classes Using Machine Learning and Natural Language Processing: A Case Study on High School Science Education. Journal of Science Education and Technology. 1–21. DOI: https://doi.org/10.1007/s10956-025-10226-z

[14] Schicchi, D., Taibi, D., 2023. AI-driven inclusion: Exploring automatic text simplification and complexity evaluation for enhanced educational accessibility. In Proceedings of the International Conference on Higher Education Learning Methodologies and Technologies Online, Cham, 2023; pp. 359–371. DOI: https://doi.org/10.1007/978-3-031-67351-1_24

[15] Anjum, A., Lieberum, N., 2023. Automatic Simplification of Scientific Texts using Pre-trained Language Models: A Comparative Study at CLEF Symposium 2023. In Proceedings of the CLEF 2023: Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, 18–21 September 2023; pp. 2899–2907. Available from: https://ceur-ws.org/Vol-3497/paper-242.pdf

[16] Shardlow, M., Sellar, S., Rousell, D., 2022. Collaborative augmentation and simplification of text (CoAST): Pedagogical applications of natural language processing in digital learning environments. Learning Environments Research. 25(2), 399–421. DOI: https://doi.org/10.1007/s10984-021-09368-9

[17] Uçar, S.Ş., Aldabe, I., Aranberri, N., et al., 2024. Exploring automatic readability assessment for science documents within a multilingual educational context. International Journal of Artificial Intelligence in Education. 34(4), 1417–1459. DOI: https://doi.org/10.1007/s40593-024-00393-2

[18] McAlpine, R., 2012. From Plain English to Global English. CC Press: Wellington, New Zealand.

[19] Pan, M., Guo, K., Lai, C., 2024. Using Artificial Intelligence Chatbots to Support English-as-a-Foreign Language Students' Self-Regulated Reading. RELC Journal. DOI: https://doi.org/10.1177/00336882241264030

[20] Cohen, J., 1988. Statistical Power Analysis for the Behavioral Sciences, 2nd ed. Erlbaum: Hillsdale, NJ, USA.

[21] Qiang, J., Huang, M., Zhu, Y., et al., 2025. Redefining Simplicity: Benchmarking Large Language Models from Lexical to Document Simplification. arXiv preprint. arXiv:2502.08281.

[22] Tessensohn, T.C., Yunus, M.M., Ismail, H.H., 2025. Using AI-Powered Tools in Enhancing Reading Skills in the ESL Classroom: A Systematic Review (2020–2024). International Journal of Academic Research in Progressive Education and Development. 14(2), 57–70. DOI: https://doi.org/10.6007/IJARPED/v14-i2/24959

[23] Sweller, J., 1988. Cognitive load during problem solving: Effects on learning. Cognitive S. 12, 25.

[24] Crossley, S., Choi, J.S., Scherber, Y., et al., 2023. Using Large Language Models to Develop Readability Formulas for Educational Settings. In: Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky: 24th International Conference, AIED 2023, Tokyo, Japan, 3–7 July 2023. DOI: https://doi.org/10.1007/978-3-031-36336-8_66

[25] Han, Y., Ceross, A., Bergmann, J.H., 2024. The use of readability metrics in legal text: A systematic literature review. arXiv preprint. arXiv:2411.09497.

[26] Alsulami, M.M., 2025. Evaluating ChatGPT's semantic alignment with community answers: A topic-aware analysis using BERTScore and BERTopic. Preprints. Available from: https://www.preprints.org/manuscript/202504.2000/v1 (cited 20 July 2025).

[27] Liu, Y., Han, T., Ma, S., et al., 2023. Summary of ChatGPT-related research and perspective towards the future of large language models. Meta-Radiology. 1(1–2), 100017. DOI: https://doi.org/10.1016/j.metrad.2023.100017

[28] Nahatame, S., Yamaguchi, K., 2025. Revisiting Text Readability and Processing Effort in Second Language

Reading: Bayesian Analysis of Eye-Tracking Data. OSF Preprints. DOI: https://doi.org/10.31219/osf.io/5wksq_v3

[29] Aziz, M., Rawian, R., 2022. Modeling higher order thinking skills and metacognitive awareness in English reading comprehension among university learners. In: Frontiers in Education. Frontiers Media SA: Lausanne, Switzerland. DOI: https://doi.org/10.3389/feduc.2022.991015

[30] Al-Jamili, O., Aziz, M., Mohammed, F., et al., 2024. Evaluating the efficacy of computer games-based learning intervention in enhancing English speaking proficiency. Heliyon. 10(16). DOI: https://doi.org/10.1016/j.heliyon.2024.e36440