Open Access Library Journal
2026, Volume 13, e16029
ISSN Online: 2333-9721
ISSN Print: 2333-9705

Core Narrative Information Retention and Readability in AI-Generated Reading Materials: A Comparative Evaluation of Generative Systems

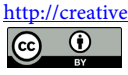
Zhihan Fu[1], Yaowei Xu[2]

[1]School of English Studies, Zhejiang Yuexiu University, Shaoxing, China
[2]Foreign Language Experimental Teaching Center, Zhejiang Yuexiu University, Shaoxing, China
Email: fzh_0813@qq.com, 20231027@zyufl.edu.cn

How to cite this paper: Fu, Z.H. and Xu, Y.W. (2026) Core Narrative Information Retention and Readability in AI-Generated Reading Materials: A Comparative Evaluation of Generative Systems. *Open Access Library Journal*, **13**: e16029.
https://doi.org/10.4236/oalib.1116029

Received: September 14, 2026
Accepted: October 26, 2026
Published: October 29, 2026

Copyright © 2026 by author(s) and Open Access Library Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/



Open Access

Abstract

Generative artificial intelligence (AI) systems are increasingly used to create and adapt educational reading materials, but the quality of these materials requires careful evaluation before they can be used in learning contexts. This study examines two questions: whether different generative AI systems can preserve the narrative information of source materials, and whether these systems produce different levels of reading difficulty in their generated texts. Three generative AI systems, GPT-5.6 Sol, Gemini 3.7 Flash, and Lingualite, were compared using three selected chapters from Journey to the West. Before generation, the source texts underwent limited preprocessing. Extended poetic passages and highly literary descriptions were excluded because the present study focused on AI-generated reading materials for general educational use rather than advanced literary analysis or translation studies. Core narrative events, character relationships, and causal connections were retained. Each system generated five versions of each chapter. Core Narrative Information Retention (CNIR) was defined as a chapter-specific, eight-item binary measure of whether each generated passage preserved eight predefined central events, character relationships, or narrative outcomes. Each content point received a score of 1 when its central content was clearly preserved and 0 when it was absent, materially altered, or assigned to the wrong character. The retention rate was calculated as the number of preserved points divided by eight. CNIR was used as a baseline measure of narrative content preservation; it was not intended to measure semantic equivalence, factual completeness, cultural fidelity, literary quality, or translation accuracy, while reading difficulty was examined using Lexile and Flesch-Kincaid Grade Level (FKGL), supported by

linguistic indicators including sentence length and lexical diversity. The results showed that all generated texts retained the predefined narrative information. However, the three systems produced significantly different readability profiles. GPT-5.6 Sol generated texts with the lowest average Lexile and FKGL scores, while Gemini 3.7 Flash and Lingualite showed different relative patterns across the two measures. Repeated generation analysis further indicated no systematic shift across the five runs within the current dataset. These findings suggest that evaluating AI-generated educational materials requires separating Core Narrative Information Retention from readability rather than relying on a single quality indicator. The study provides a text-level evaluation approach for examining the suitability of AI-generated reading materials in educational contexts.

Subject Areas

Educational Technology

Keywords

Generative AI, Educational Technology, Reading Materials, Readability, Multi-Agent Systems, Core Narrative Information Retention

1. Introduction

Large language models are increasingly used to draft and adapt reading materials. Rather than just fielding student queries in chatbot interfaces, instructors and platform developers now prompt large language models to draft extensive reading passages and reformat dense source materials in seconds. Educational technologists have studied these outputs for reading assessment, custom storytelling, and proficiency-calibrated materials [1]-[3]. Large language models can now produce or adapt long reading passages with little manual effort, which is attractive for educational use, but the resulting text still needs evaluation. A fluent passage may remain difficult for its intended readers, and reducing difficulty may sometimes remove information that matters for understanding.

Lower reading difficulty is useful only when the main narrative information is retained. Recent empirical evaluations confirm that automated readability improvements frequently decouple from factual retention [2] [4]. The problem becomes harder to interpret when systems use different generation workflows. Some models produce a response in one conversational turn, while other platforms revise the text through several stages. Multi-stage architectures process source text through specialized drafting, editing, and quality-checking gates. Each processing layer modifies vocabulary, merges syntactic clauses, and shifts overall document length.

Existing literature focuses heavily on prompted simplification, forcing models toward fixed CEFR bands or explicit grade targets. That setup tests prompt adher-

ence. It overlooks what happens when an educator simply requests an English translation without feeding the model an arbitrary reading formula. In this setting, readability is measured after generation rather than specified in the prompt.

We examine that case with GPT-5.6 Sol, Gemini 3.7 Flash, and Lingualite. The first two are general-purpose models; Lingualite uses a staged workflow with ten documented processing nodes. We use three preprocessed chapters from *Journey to the West* and five separate runs for each system-chapter condition. The analysis first checks Core Narrative Information Retention and then compares Lexile and FKGL, with several linguistic measures used for interpretation. We therefore compare the outputs rather than trying to rank the systems as translation tools. The analysis asks whether they preserve the same narrative information and how their reading difficulty differs across the selected chapters.

2. Literature Review

2.1. Generative AI for Educational Content Generation

Early educational AI research prioritized automated feedback and conversational tutoring. Recent work has also examined language models as tools for generating instructional texts. In assessment authoring, models successfully draft reading passages but require automated filtering and human review to maintain grade appropriateness [1]. Similarly, automated children's stories maintain global narrative coherence while routinely violating target age-band vocabulary boundaries [3].

Controlled text adaptation presents a similar problem. Huang *et al.* [2] found factual and local editing errors even when models were prompted toward specific Lexile levels. Benedetto and Buttery [4] approached the issue differently by testing whether simplified passages still contained the information needed to answer comprehension questions. Together, these studies show why readability and narrative information retention need to be evaluated separately.

2.2. Readability and Text Complexity in Educational Materials

Readability measures describe text demand in numerical terms, although no single score captures every source of difficulty. The Lexile Framework places reader ability and text demand on the same scale [5]. Fan *et al.* [6] applied Lexile to 800 English passages used in Chinese educational contexts, from junior secondary materials to CET-6. In their dataset, Lexile values generally rose with the educational level of the materials. We use Lexile here because it offers an established educational scale for English text demand.

A second measure helps expose differences hidden by Lexile alone. Models do not always match requested grade levels even when prompts name a readability standard [7]. README++ likewise reports wide differences in semantic difficulty among texts with similar sentence lengths [8]. Trott and Rivière [9] and Ünaldi and Ateş [10] caution against reducing text suitability to one automated score. These findings support the use of more than one readability measure when com-

paring AI-generated texts.

2.3. Generation Workflows for AI-Generated Educational Materials

Generative systems can differ not only in their underlying models but also in how they organize generation and revision. General-purpose conversational systems usually present users with a single final response, while some agentic systems divide the task across several visible stages. ExpertEase, for example, uses expert, teacher, and student agents to support grade-specific rewriting and revision [11]. AgentSimp adopts a similar staged approach for document-level simplification, with separate roles for planning, terminology handling, restructuring, and proofreading [12].

A staged workflow does not necessarily improve every aspect of the final text. ExpertEase reported gains in grade-level control, but some single-model baselines performed better on semantic preservation measures [11]. The TSAR 2025 shared task likewise included iterative and agent-based systems with different balances between readability and semantic similarity [13]. These findings make the final output, rather than the number of processing stages alone, the more useful basis for comparison. Because Lingualite does not disclose its underlying foundation models, the present study treats it as a complete system condition and does not attribute observed differences to any individual internal node.

2.4. Core Narrative Information Retention in AI-Generated Educational Materials

Lower readability is not necessarily desirable if important source information is lost. Previous studies have therefore evaluated simplification together with semantic preservation or discourse-level information [14] [15]. Benedetto and Buttery [4], for example, used question answerability to determine whether information needed for comprehension remained available after simplification.

The present study uses a narrower form of narrative information evaluation. Eight core narrative points were defined for each chapter and checked in every generated output. The measure is not intended to capture every stylistic or cultural detail; it provides a basic narrative information baseline before readability differences are interpreted.

3. Methodology

3.1. Research Questions

The analysis focuses on two aspects of the generated texts: reading difficulty and preservation of the main narrative information. Three generative systems were given the same source chapters, and their outputs were evaluated with the same set of measures. We also collected repeated generations to check whether the observed patterns were stable across runs.

RQ1. To what extent do the three generative systems preserve the core narrative

information of the source chapters?

RQ2. How do the three generative systems differ in the reading difficulty of their outputs, as measured by Lexile and FKGL?

3.2. Source Materials

The Chinese source texts were Chapters 14, 23, and 27 of Journey to the West by Wu Cheng'en, obtained from Chinese Wikisource [16], a Chinese-language digital edition of the text (accessed September 30, 2026). All three chapters come from the same work but differ in narrative focus: Chapter 14 develops the early relationship between the pilgrims, Chapter 23 centers on temptation and character testing, and Chapter 27 follows repeated deception and conflict.

Before generation, the three chapters were preprocessed according to a fixed rule set. Extended verse, poetic descriptions, authorial commentary, and other stylistically elaborate passages that were not necessary for the central narrative were removed. The remaining prose was not modernized or simplified. Where deletion would otherwise interrupt narrative continuity, only minimal neutral prose bridges were inserted.

3.3. Generative Systems and Generation Procedure

GPT-5.6 Sol and Gemini 3.7 Flash were run in separate conversations with the same prompt: "Translate the attached Chinese text into English. Return only the English translation." A new conversation was opened for every run, and no conversational context was carried over from the preceding run.

Lingualite was accessed through its website using the Chinese-to-English Standard mode [17]. At the time of data collection, the production backend was version v4.31.0, built on August 17, 2026. The Standard workflow exposed ten functional processing nodes for background analysis, preprocessing, terminology preparation, literal translation, free translation, professional translation, post-editing, quality checking, final polishing, and sentence-level alignment. These nodes operated sequentially on shared intermediate outputs rather than producing ten independent translations. The specific foundation models, sampling parameters, and internal system prompts used by the platform were not publicly disclosed. For each run, the same cleaned Chinese chapter was submitted as a new Standard-mode translation task with Chinese as the source language and English as the target language. No Lexile, CEFR, grade-level, or additional readability constraint was supplied.

Each system generated five versions of each chapter, producing 45 outputs in total. The outputs were evaluated without manual rewriting.

3.4. Readability and Linguistic Characteristics

FKGL and Dale-Chall scores were calculated in Python 3.12.6 using textstat 0.7.8 with US English settings. Sentence counts followed the textstat sentence-counting procedure, which identifies sentence candidates using periods, exclamation marks,

and question marks and excludes candidates containing two or fewer words.

For word count, mean sentence length, and lexical-diversity measures, words were extracted using the regular expression [A-Za-z]+(?:['']][A-Za-z]+)? and converted to lowercase. Contractions were treated as single tokens, hyphenated compounds were split, and no stemming or lemmatization was applied. Mean sentence length was calculated as the regex-based word count divided by the textstat sentence count.

MATTR was calculated as the mean type-token ratio across overlapping 50-token windows, with the window advancing by one token at a time.

Lexile measures were obtained directly from the MetaMetrics Lexile® Analyzer integrated into ChatGPT, with one official Lexile® text measure obtained for each English output.

3.5. Core Narrative Information Retention

Eight core narrative information points defined each chapter before evaluation. Fully preserved key events, character relationships, and narrative outcomes earned one point. Missing or substantially distorted information earned zero points. Acceptable stylistic variations included minor wording differences, syntactic restructuring, and standard character-name variants.

The same standardized scoring rubric governed all generated texts. Every individual evaluation record paired a final binary score with a supporting textual excerpt and a concise analytical justification. One evaluator, the first author, coded all 360 output-by-content-point decisions. The evaluator was blinded to system identity during coding. A predefined codebook was applied to each output, and every decision was linked to an exact textual excerpt and its location in the generated passage. The evaluator reviewed all 360 records rather than reviewing only a selected sample. Because the data were coded by a single evaluator, independent inter-rater agreement could not be calculated. This single-coder design is therefore acknowledged as a limitation. For each output, the Core Narrative Information Retention Rate was calculated as: preserved content points/8 × 100%.

3.6. Data Analysis

Lexile and FKGL were analyzed using two-way ANOVAs with System and Chapter as factors. For the significant Lexile System × Chapter interaction, pairwise comparisons among the three systems were conducted separately within each chapter using Tukey's HSD. The procedure adjusted for the three pairwise comparisons within each chapter, and mean differences, multiplicity-adjusted p values, and 95% simultaneous confidence intervals were reported. The five generation runs were also examined for systematic run effects. As an exploratory robustness check, Run (1 - 5) was entered as a continuous predictor in linear models controlling for System, Chapter, and their interaction to test for a common directional trend across generation order. Because run number indexed generation order rather than an experimental treatment, this analysis was interpreted only as a

check for directional drift. The slope coefficient (b), 95% confidence interval, and p value were reported. Core Narrative Information Retention Rate was reported descriptively because all outputs reached the same score. Because the sample size had already been fixed before the present statistical review, a sensitivity power analysis was conducted rather than an a priori sample-size calculation. With $\alpha = 0.05$ and a target power of 0.80, the current 3 × 3 factorial design could detect approximately Cohen's $f = 0.48$ for the System main effect and $f = 0.55$ for the System × Chapter interaction. These values describe the sensitivity of the present fixed-effects analysis and do not imply equivalent power for generalization across source texts [18].

4. Results

4.1. Core Narrative Information Retention Results

All 45 outputs retained all eight predefined core narrative information points in their respective chapters. Core Narrative Information Retention was therefore 100% for every system, chapter, and run. Because the measure showed no variation, no inferential test was applied. This provided a common narrative information baseline for the readability comparison.

4.2. Readability Profiles

The two primary readability measures produced clear but not identical system profiles. GPT-5.6 Sol had the lowest average Lexile measure (844 L), followed by Gemini 3.7 Flash (927 L) and Lingualite (965 L). FKGL again placed GPT lowest (5.59), but Gemini was highest (6.63), with Lingualite between them (6.10). Lexile and FKGL were moderately correlated across the 45 outputs ($r = 0.654$, $p < 0.001$), but their different system ordering shows that the two measures were not interchangeable (see Table 1).

Table 1. Descriptive statistics by system.

System	Core Narrative Information Retention	Lexile M (SD)	FKGL M (SD)	MSL M (SD)	MATTR M (SD)	Word count M (SD)
GPT-5.6 Sol	100%	844.0 (28.5)	5.59 (0.23)	12.86 (0.38)	0.824 (0.009)	5899.8 (636.2)
Gemini 3.7 Flash	100%	927.3 (14.9)	6.63 (0.22)	15.51 (0.61)	0.827 (0.011)	5598.9 (683.1)
Lingualite	100%	964.7 (26.7)	6.10 (0.30)	15.89 (0.70)	0.807 (0.012)	5889.0 (762.4)

A two-way ANOVA confirmed a significant system effect for both primary measures: Lexile, $F(2, 36) = 142.72$, $p < 0.001$, and FKGL, $F(2, 36) = 72.77$, $p < 0.001$. The main ANOVA results are summarized in Table 2.

4.3. Variation across Chapters

Chapter did not show a significant overall effect on Lexile ($p = 0.087$), but the System × Chapter interaction was significant ($p = 0.003$), indicating that the mag-

nitude of system differences varied across the three source chapters. Tukey-adjusted pairwise comparisons were therefore conducted separately within each chapter (Table 3). GPT-5.6 Sol produced significantly lower Lexile measures than both Gemini 3.7 Flash and Lingualite in all three chapters. The difference between Gemini and Lingualite, however, varied by chapter: it was 12 L in Chapter 14 and was not statistically significant, increased to 36 L in Chapter 23, and reached 64 L in Chapter 27.

Table 2. Two-way ANOVA for the primary readability outcomes.

Outcome	Effect	F	p	partial η^2
Lexile	System	142.72	<0.001	0.888
Lexile	Chapter	2.61	0.087	0.127
Lexile	System × Chapter	4.92	0.003	0.353
FKGL	System	72.77	<0.001	0.802
FKGL	Chapter	5.80	0.007	0.244
FKGL	System × Chapter	0.29	0.883	0.031

Table 3. Tukey-adjusted pairwise system comparisons for Lexile within each chapter.

Chapter	Contrast	Mean difference (L)	95% CI (L)	Adjusted p
14	Gemini - GPT	96	[65.50, 126.50]	<0.001
14	Lingualite - GPT	108	[77.50, 138.50]	<0.001
14	Lingualite - Gemini	12	[−18.50, 42.50]	0.561
23	Gemini - GPT	58	[22.21, 93.79]	0.003
23	Lingualite - GPT	94	[58.21, 129.79]	<0.001
23	Lingualite - Gemini	36	[0.21, 71.79]	0.049
27	Gemini - GPT	96	[61.15, 130.85]	<0.001
27	Lingualite - GPT	160	[125.15, 194.85]	<0.001
27	Lingualite - Gemini	64	[29.15, 98.85]	0.001

4.4. Repeated-Generation Stability

The five generation runs showed modest variation but no systematic directional change. After controlling for System, Chapter, and their interaction, generation order was not associated with a significant linear change in Lexile (b = 1.44 L per run, 95% CI [−2.87, 5.76], p = 0.502) or FKGL (b = 0.0023, 95% CI [−0.0488, 0.0534], p = 0.927). The first and fifth runs were also similar, with mean differences of 1.1 L for Lexile and −0.011 for FKGL (Table 4).

4.5. Supporting Linguistic Patterns

Sentence length followed the main readability pattern. GPT produced the shortest

sentences on average (12.86 words), compared with 15.51 for Gemini and 15.89 for Lingualite. Mean sentence length was strongly correlated with Lexile ($r = 0.934$, $p < 0.001$) and FKGL ($r = 0.798$, $p < 0.001$).

Table 4. First-to-fifth run comparison.

Outcome	Run 1 mean	Run 5 mean	Difference	95% CI	p
Lexile	905.6 L	906.7 L	+1.1 L	−14.94 to +17.16 L	**0.877**
FKGL	6.047	6.036	−0.011	−0.281 to +0.260	**0.931**

MATTR showed a different pattern: Gemini averaged 0.827, GPT 0.824, and Lingualite 0.807. Output length also varied. Gemini produced the shortest texts on average (5599 words), while GPT and Lingualite averaged 5900 and 5889 words. The shorter Gemini outputs did not correspond to lower difficulty on the primary measures, especially FKGL.

5. Discussion

5.1. Core Narrative Information Retention as a Minimum Requirement

All 45 outputs preserved the eight predefined core narrative information points. This does not mean that every detail of the source was retained, because the rubric covered only the main events, relationships, and outcomes. In this study, preservation therefore acts as a minimum requirement rather than a ranking criterion. The readability differences are more informative once that baseline has been met.

5.2. Reading Difficulty and Educational Use

The three systems produced noticeably different readability profiles within the selected chapters. GPT received the lowest scores on both Lexile and FKGL. Gemini and Lingualite were less consistent across the two measures: Lingualite had the higher Lexile score, while Gemini had the higher FKGL score.

Lexile and FKGL are related, but they do not describe reading difficulty in exactly the same way. The disagreement between the two measures shows why one readability score should not be treated as a complete description of learner difficulty. Automated measures can help screen or compare candidate texts, but they should not be treated as interchangeable indicators of learner suitability.

The supporting measures point in the same direction. GPT used shorter sentences, while Gemini and Lingualite produced longer sentence structures. Gemini also generated the shortest texts overall, yet this did not result in the lowest FKGL score. Shorter output, shorter sentences, and lower reading demand are therefore not necessarily the same property.

Because only three chapters were tested, these patterns should be treated as system-level tendencies within the present materials, not stable characteristics of the models in general. The significant variation across chapters in some measures also suggests that the source text remains part of the final reading profile.

5.3. Implications for AI-Generated Reading Materials

The results also have a practical implication for AI-generated reading materials. A text can preserve its core narrative information and still vary substantially in measured reading difficulty. For this reason, preservation and readability should not be collapsed into a single score. In the present dataset, repeated generation did not change the overall pattern, although individual outputs still varied slightly across runs.

These results do not show which system is best for learners. They only show that the same source material can lead to different reading profiles across systems. A later learner-based study would be needed to determine whether those differences matter for comprehension or material selection. Instead, they highlight the need for evaluation procedures that can identify whether AI-generated materials meet different instructional purposes before being introduced into learning environments.

6. Conclusions

This study compared the reading characteristics and core narrative information retention of English texts produced by three generative AI systems from the same source chapters. All 45 outputs retained the eight predefined narrative information points. The systems nevertheless differed in measured reading difficulty: GPT-5.6 Sol had the lowest mean Lexile and FKGL scores, while Gemini 3.7 Flash and Lingualite changed position across the two measures. Repeated runs showed no systematic first-to-fifth shift.

These results support treating core narrative information retention as a minimum requirement rather than a ranking measure. Once the core narrative information is retained, Lexile and FKGL can provide complementary views of the reading demand of generated texts. The disagreement between the two measures also shows why one readability score should not be treated as a complete description of educational suitability.

The study has several limitations. It used only three chapters from one literary work, so the observed patterns may not extend to other genres or subjects. *Journey to the West* is widely available in both Chinese and English, and prior model exposure to the source or existing translations cannot be excluded. The narrative information rubric covered eight core narrative points per chapter rather than every factual, stylistic, or cultural detail. Finally, Lexile and FKGL describe textual characteristics rather than actual learner comprehension.

Future learner-based studies should define a minimum educationally meaningful effect before data collection and use an a priori power analysis to determine the required numbers of learners and source texts.

Author Contributions

Zhihan Fu designed the study, conducted the analyses, and drafted the manu-

script. Yaowei Xu provided research supervision and reviewed and revised the manuscript.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Bezirhan, U. and von Davier, M. (2023) Automated Reading Passage Generation with Openai's Large Language Model. *Computers and Education*: *Artificial Intelligence*, **5**, Article ID: 100161. https://doi.org/10.1016/j.caeai.2023.100161

[2] Huang, C., Wei, J. and Huang, T.K. (2024). Generating Educational Materials with Different Levels of Readability Using LLMs. *Proceedings of the Third Workshop on Intelligent and Interactive Writing Assistants*, Honolulu, 11 May 2024, 16-22. https://doi.org/10.1145/3690712.3690718

[3] Valentini, M., Weber, J., Salcido, J., Wright, T., Colunga, E. and von der Wense, K. (2023) On the Automatic Generation and Simplification of Children's Stories. *Proceedings of the* 2023 *Conference on Empirical Methods in Natural Language Processing*, Singapore, 6-10 December 2023, 3588-3598. https://doi.org/10.18653/v1/2023.emnlp-main.218

[4] Benedetto, L. and Buttery, P. (2025) Towards CEFR-Targeted Text Simplification for Question Adaptation. *International Conference Recent Advances in Natural Language Processing*, Varna, 8-10 September 2025, 150-157. https://doi.org/10.26615/978-954-452-098-4-018

[5] Stenner, A.J. (2023) Measuring Reading Comprehension with the Lexile Framework. In: Fisher Jr., W.P. And Massengill, P.J., EDS., *Explanatory Models*, *Unit Standards*, *and Personalized Learning in Educational Measurement*, Springer, 63-88. https://doi.org/10.1007/978-981-19-3747-7_6

[6] Fan, J., Jin, T. and Chen, I. (2022) Applying the Lexile Framework to Evaluate Text Complexity in the Chinese Context: A Corpus-Based Study. British Council & MetaMetrics, Research into Reading Grants, Report 1/2022.

[7] Imperial, J.M. and Tayyar Madabushi, H. (2023) Flesch or Fumble? Evaluating Readability Standard Alignment of Instruction-Tuned Language Models. *Proceedings of the Third Workshop on Natural Language Generation*, *Evaluation*, *and Metrics* (*GEM*), Singapore, 6 December 2023, 205-223. https://aclanthology.org/2023.gem-1.18/

[8] Naous, T., Ryan, M.J., Lavrouk, A., Chandra, M. and Xu, W. (2024) ReadMe++: Benchmarking Multilingual Language Models for Multi-Domain Readability Assessment. *Proceedings of the* 2024 *Conference on Empirical Methods in Natural Language Processing*, Miami, 12-16 November 2024, 12230-12266. https://doi.org/10.18653/v1/2024.emnlp-main.682

[9] Trott, S. and Rivière, P. (2024) Measuring and Modifying the Readability of English Texts with GPT-4. *Proceedings of the Third Workshop on Text Simplification*, *Accessibility and Readability* (*TSAR* 2024), Miami, November 2024, 126-134. https://doi.org/10.18653/v1/2024.tsar-1.13

[10] Ünaldi, A. and Ateş, B. (2025) Evaluating Text Suitability for EAP Reading Assessment: Teachers versus Lexile, Coh-Metrix and ChatGPT. British Council. https://doi.org/10.57884/zass-nw57

[11] Mo, K. and Hu, R. (2024) ExpertEase: A Multi-Agent Framework for Grade-Specific

Document Simplification with Large Language Models. *Findings of the Association for Computational Linguistics*: *EMNLP* 2024, Miami, November 2024, 9080-9099. https://doi.org/10.18653/v1/2024.findings-emnlp.530

[12] Fang, D., Qiang, J., Ouyang, X., Zhu, Y., Yuan, Y. and Li, Y. (2025) Collaborative Document Simplification Using Multi-Agent Systems. *Proceedings of the* 31*st International Conference on Computational Linguistics*, Abu Dhabi, 19-24 January 2025, 897-912.

[13] Alva-Manchego, F., Stodden, R., Imperial, J.M., Barayan, A., North, K. and Tayyar Madabushi, H. (2025) Findings of the TSAR 2025 Shared Task on Readability-Controlled Text Simplification. *Proceedings of the Fourth Workshop on Text Simplification*, *Accessibility and Readability* (*TSAR* 2025), Suzhou, 4-9 November 2025, 116-130. https://doi.org/10.18653/v1/2025.tsar-1.8

[14] Jeong, J., Park, J. and Yu, Y. (2026) Right at My Level: A Unified Multilingual Framework for Proficiency-Aware Text Simplification. *Proceedings of the* 64*th Annual Meeting of the Association for Computational Linguistics* (*Volume* 1: *Long Papers*), San Diego, 2-7 July 2026, 23680-23706. https://doi.org/10.18653/v1/2026.acl-long.1086

[15] Zhu, J., Yang, E., Zhu, H. and Yang, L. (2026) Discourse Structure Variation for Large Language Model Evaluation in Text Simplification. *Journal of Chinese Information Processing*, **40**, 41-51. (In Chinese) https://www.sohu.com/a/1014585802_121124334

[16] Wu, C. (n.d.) Journey to the West. Chapters 14, 23, and 27. Chinese Wikisource. (In Chinese) https://zh.wikisource.org/wiki/%E8%A5%BF%E9%81%8A%E8%A8%98

[17] (2026) Lingualite: Multi-Agent Translation Platform. https://lingualite.cn/

[18] Lakens, D. (2022) Sample Size Justification. *Collabra*: *Psychology*, **8**, Article ID: 33267. https://doi.org/10.1525/collabra.33267