Natural Language Processing as a Scalable Method for Evaluating Educational Text Personalization by LLMs

Linh Huynh and Danielle S McNamara *

Posted Date: 24 September 2025

doi: 10.20944/preprints202509.2013.v1

Keywords: natural language processing; large language models; personalized learning; text evaluation

 doi:10.20944/preprints202509.2013.v1

Article

Natural Language Processing as a Scalable Method for Evaluating Educational Text Personalization by LLMs

Linh Huynh and Danielle S. McNamara *

Learning Engineering Institute, Arizona State University, 120 Cady Mall, Tempe, AZ 85281, USA

* Correspondence: dsmcnama@asu.edu

Abstract

Four versions of science and history texts were tailored to diverse hypothetical reader profiles (high and low reading skill and domain knowledge), generated by four Large Language Models (i.e., Claude, Llama, ChatGPT, and Gemini). Natural Language Processing (NLP) technique was applied to examine variations in Large Language Model (LLM) text personalization capabilities. NLP was leveraged to extract and quantify linguistic features of these texts, capturing linguistic variations as a function of LLMs, text genres, and reader profiles. An approach leveraging NLP-based analyses provides an automated and scalable solution for evaluating alignment between LLM-generated personalized texts and readers' needs. Findings indicate that NLP offers a valid and generalizable means of tracking linguistic variation in personalized educational texts, supporting its use as an evaluation framework for text personalization.

Keywords: natural language processing; large language models; personalized learning; text evaluation

1. Introduction

Recent advancements in Generative AI (GenAI) and Large Language Models (LLMs) have transformed various aspects of education. LLMs have shown immense capabilities in text-processing and text-generation tasks such as translation, summarization, and question-answering [1]. LLMs have been applied to automate various tasks such as content creation, reasoning, problem-solving, and tailoring content to individual needs [2–6]. In recent years, LLMs have been widely applied to enable scalable, data-driven personalization in educational contexts [7–10].

1.1. Personalized Learning

Personalized learning refers to tailoring educational experiences and providing individualized support to accommodate students' unique abilities, goals, and preferences [11]. LLMs can dynamically tailor learning content, pace, and feedback to align with their growth and challenges. Each student possesses a complex and unique profile requiring tailored educational experiences [12,13]. Recent work has highlighted the potential of utilizing LLMs to personalize content and demonstrated positive impacts of LLM-driven personalization technologies [14,15]. To create accessible learning materials that meet the needs of learners, particularly those with learning disabilities, LLMs have been used to simplify content and align text complexity to match readers' abilities [16]. This approach is grounded in reading comprehension theories showing that text comprehension depends on both individual characteristics (e.g., prior knowledge, reading skills, cognitive abilities) and text-specific linguistic features (e.g., cohesion, lexical complexity, and syntactic sophistication; [17]).

The urgent need for personalized learning is well-documented due to several factors such as learner diversity, limitations of traditional one-size-fits-all instruction, and the special requirements

 doi:10.20944/preprints202509.2013.v1

of students with disabilities [18–22]. Tailoring text complexity to prior knowledge and reading skills has strong potential to enhance students' motivation, interest, and overall learning outcomes [23]. LLM-powered tutors that tailor explanations, questions, and feedback according to learners' knowledge have been shown to enhance comprehension and engagement [24–26]. For example, in an Adaptive Feedback System (AFS) integrating GPT-4 with deep learning models to tailor feedback, students' online courses demonstrated 12.6 percentage point improvements in performance compared to those in the control condition (Cuéllar et al., 2025). The personalized feedback system also increased student engagement and reduced achievement gaps between students with varying levels of prior knowledge, helping students with low prior knowledge catch up with their peers [27]. These findings supported the value and benefits of integrating LLMs to personalize learning.

While LLMs offer significant benefits and potential to transform education, persistent challenges related to standardized evaluation, data privacy issues, ethical considerations, and effective integration remain unresolved [28–30]. When an LLM is prompted to tailor feedback for students, only high-quality, theoretically grounded prompts consistently generate feedback that are superior to expert human-generated feedback in terms of explanation quality, specificity, and engagement outcomes for the students [31]. These findings highlight the need to establish a rigorous and theory-driven evaluation framework to maximize the potential of LLMs. Rigorously assessing the quality and alignment of personalized educational content, providing reliable and real-time feedback are critical to ensure effective implementation of an LLM-powered personalized learning system [32]. A rigorous validation method is critical to ensure that LLM-driven personalization adapts content to learners' needs in a consistent and effective manner [22,23]. Standardized evaluation frameworks that are theoretically grounded are crucial factors for successful implementation [33–35].

1.2. Text Personalization Evaluation Using Natural Language Processing

A scalable, objective method is essential to validate the extent to which LLM adaptations align with learning theories and evolving student profiles. Human-based evaluation (e.g., expert rating the quality, comprehension assessment, learning outcomes) is time-consuming, resource-intensive, and generally suffers from inherent biases and inter-rater variability. These limitations make it challenging to improve performance rapidly over time and impractical to implement on a large scale. Traditional automated evaluation metrics such as BLEU (BiLingual Evaluation Understudy), ROUGE (Recall-Oriented Understudy for Gisting Evaluation), and METEOR (Metric for Evaluation of Translation with Explicit Ordering) calculate the similarity score between the generated and expected output, focusing only on lexical similarity and word overlap [36–39]. However, adapting text to meet readers' individual needs goes well beyond semantic overlap. Text personalization must evolve according to the learners' changing skills and knowledge over time as learners advance or struggle at different points in their learning trajectories [40,41]. The shortcomings of available evaluation methods are consequential since effective personalized educational content requires rapid and iterative evaluation of text appropriateness tailored to learners' levels of skills and knowledge [42].

To overcome these evaluation challenges, Natural Language Processing (NLP) offers an alternative method, capturing linguistic features that are well-aligned with cognitive theories. NLP can be used to extract and quantify linguistic features that have been shown to be strongly predictive of reading difficulty and comprehension outcomes [43–46]. Text readability refers to the ease that readers process and understand a text, which can be quantified using metrics derived from NLP tools such as the Writing Assessment Tool (WAT; [47]). NLP-based analyses offer a robust evaluation method for assessing text personalization. Unlike traditional evaluation methods, they can be leveraged to assess the linguistic features critical for optimal comprehension and engagement. These metrics allow researchers to differentiate between various types of texts and level of complexity based on linguistic and semantic properties [48]. Specifically, features such as cohesion, language variety, syntax and lexical sophistication significantly influence text comprehension and ease of text processing, particularly in educational contexts [49,50]. These theoretically-driven metrics are

 doi:10.20944/preprints202509.2013.v1

necessary to quickly assess personalized content and iteratively improve performance of the personalization system.

Huynh and McNamara (2025) found that NLP techniques effectively differentiate and evaluate personalized text generated by various LLMs, including Claude, Llama, ChatGPT, and Gemini. They selected theoretical-driven metrics derived from an NLP tool (i.e., WAT; [47]) to assess alignment of tailored scientific texts intended for different reader profiles. Their study highlighted variability in linguistic alignment between reader profiles and outputs generated by different LLMs. NLP analyses successfully captured linguistic variations among texts and provided assessment for cohesion, lexical and syntactic complexity measures consistent with theoretical predictions. These linguistic features varied systematically in alignment with reader profiles and effectively differentiated between the outputs generated by different LLMs. Their findings underscored the importance of measuring fine-grained linguistic alignments between generated texts and specific reader profiles, emphasizing NLP's utility in objectively benchmarking personalization quality.

However, different disciplines (i.e., science and history texts) exhibit unique linguistic patterns due to differences in disciplinary conventions and purposes. Science texts are predominantly explanatory and informative with the purpose of conveying factual information [51–54]. They are characterized by high conceptual density with interconnected abstract concepts and specialized technical terminology [55–57]. Moreover, science texts feature dense nominalization in which nouns or noun phrases are converted from verbs (e.g., oxidation, measurement; [58]). Science texts also exhibit syntactic complexity characterized by longer sentences, passive voice, and embedded clauses [59].

These features highlight the objective nature of science texts but also contribute to comprehension difficulties for readers [60]. Readers need to possess sufficient background knowledge, strong vocabulary, and comprehension strategies to fully grasp the materials [17,46]. In contrast, history texts aim to contextualize and interpret historical events through a blend of descriptive, evaluative and narrative writing style [61,62]. History texts often have varied syntax, fewer nominalizations, and implicit cohesion (e.g., chronological sequencing, storytelling; [63–65]). While also including specialized vocabulary (e.g., terms referring to historical events, institution names, dates, and figures), history texts incorporate concrete and descriptive language that is less complex compared to science texts [66]. These linguistic variations result in comprehension differences such that successful comprehension of science texts relies on vocabulary knowledge while comprehension of history texts depends on understanding the overall context and connections between events [67].

While prior studies have assessed LLM adaptations without explicitly differentiating text domains [42], the inherent linguistic differences between science and history materials suggest that effective personalization must consider these linguistic distinctions [50]. Effective personalization of educational content not only aligns text complexity with readers' unique needs and skills but also retains domain-specific linguistic features. As such, it is imperative to determine whether LLM-generated modifications sufficiently reflect these demands. NLP metrics have primarily been validated within science. Due to the domain-specific nature of science and history, it is necessary to establish the validity and generalizability of NLP-based evaluation methods by assessing text personalization across domains. Cross-domain validation helps strengthen the rigor and validity of NLP-based evaluation methods. Without cross-domain testing, it remains uncertain the extent that these linguistic metrics can be generalized effectively to texts with fundamentally different linguistic structures.

1.3. Current Research

The current research aims to replicate the method and extend prior study [42] by leveraging NLP analyses to assess LLM-generated text personalization across different domains (i.e., science and history). This study examined linguistic variations in LLM-generated personalized texts by analyzing the effects of domain (Science vs. History), LLM (Claude, Llama, ChatGPT, Gemini), and reader

 doi:10.20944/preprints202509.2013.v1

profile (i.e., high and low prior knowledge and reading skill) on linguistic features. Personalization has been primarily examined on math and science subjects, but discourse research shows that texts from history/humanities disciplines differ in connective use, syntax, lexical and nominalization patterns compared to science texts [68,69]. No study has jointly examined reader profile adaptation, LLM variation, and cross-disciplinary text features. Analyzing science and history text adaptations allows us to assess whether LLM-modified texts align with known linguistic complexities inherent to each domain and how tailored content aligns with readers' needs.

We hypothesized that NLP analyses would reveal robust and meaningful variations in linguistic features across text domains, LLMs and reader profiles. Several linguistic metrics are related to syntactic complexity such as noun-to-verb ratio, language variety, sentence length. Lexical sophistication is determined by whether texts include vocabulary commonly used in academic texts and words with low level of concreteness [70,71]. Syntax and lexical complexity present significant challenges for readers with lower reading skills or prior knowledge. These readers often struggle with decoding complex vocabulary and sentence structures since they are less likely to use effective reading strategies, integrate textual information with prior knowledge, or monitor their understanding, all of which can hinder comprehension [72,73]. When readers with limited vocabulary encounter unfamiliar academic terms or phrases, they may misinterpret the text or fail to comprehend it altogether [74]. Therefore, for less skilled readers or those with limited vocabulary knowledge, texts should be modified to use clear and straightforward language, avoiding overly complex vocabulary and sentence structures. Modifications should minimize language variety and avoid complex syntax structures and sophisticated wording.

Based on previous findings [42], we expected alignment between the complexity of personalized texts and the cognitive needs of different reader profiles. High-cohesion texts containing clear referential and causal connections benefit low-knowledge readers who lack the necessary background knowledge [50]. In contrast, low-cohesion texts are more suitable for high-knowledge readers, as they promote deep learning by facilitating inference generation [75]. Skilled readers are able to comprehend complex and low-cohesion texts, which require them to actively generate inferences and engage in deep processing [17]. Examining these linguistic features provides quantifiable insights into alignment between personalized content and readers' needs, allowing researchers to effectively assess personalization quality [50]. Cohesion, lexical and syntactic features are linguistic features grounded in theories of reading comprehension and psycholinguistics and have been shown to be directly related to comprehension and learning difficulties [17,45,76]. These metrics have been validated as predictors of text readability, making them appropriate for objectively assessing LLM-generated personalized texts [48]. Table 1 includes a list of features related to text readability.

Table 1. Linguistic features affecting coherence-building processes and reading comprehension. Source: Authors' contribution.

 doi:10.20944/preprints202509.2013.v1

Features	Metrics and Descriptions
Writing Style	Academic writing*: The extent to which the texts include domain-specific words and sophisticated sentence structures, commonly found in academic writing texts
Conceptual Density and Cohesion	Lexical density: The extent to which text contains sentences with dense and precise information, including complex noun phrases and sophisticated and specific function words Noun-to-verb ratio*: Text with a high noun-to-verb ratio results in dense information and complex sentences that require greater cognitive effort to process Sentence cohesion*: The extent to which the text contains connectives and cohesion cues (e.g., repeating ideas and concepts)
Syntax Complexity	Sentence length*: Longer sentences often have more clauses and complex structure Language variety*: The extent to which the text contains a variety of lexical and syntax structures
Lexical Complexity	Word concreteness: The degree to which words are tangible and refer to concepts that can be experienced by the senses. High measures indicate the texts contain more tangible words, while low measures indicate more abstract concepts Sophisticated wording*: Lower measures indicate the vocabulary familiar and common, whereas higher measures indicate more advanced words Academic frequency*: Indicates the extent of sophisticated vocabulary are used, which are also common in academic texts
Connectives	All connectives: Refers to the overall density of linking words and phrases (e.g., however, therefore, then, in addition). Higher values indicate the text is overtly guiding the reader through logical, additive, contrastive, temporal, or causal relations, increasing cohesion. Lower values imply that relationships must be inferred from context Temporal connectives: Markers that place events on a timeline (e.g., then, meanwhile, during, subsequently) Causal connectives: Markers that signal cause-and-effect or reasoning links (e.g., because, since, therefore, thus, as a result)

* Indices used in Huynh and McNamara (2025).

We also predicted that there would be differences in linguistic features between science and history adaptations based on genre-specific characteristics [51,55,60]. We anticipated measurable differences in lexical, syntactic, and cohesive features when comparing science and history passages using NLP metrics [59]. Regardless of reader profile and LLMs, science modifications would contain more dense and complex syntax, advanced vocabulary commonly used in academic discourse, and more causal connectives compared to history modifications. In contrast, history texts would show higher use of temporal connectives with higher noun-to-verb ratio. Corpus analysis study showed that history texts emphasize temporal relations and nominal references. Nominalizations appear frequently in historical discourse because they often contain narrative accounts of historical figures whereas science texts employ causal reasoning to link abstract concepts [52–54].

Moreover, due to the inherent differences in model design, training corpus and fine-tuning strategies, we anticipated variations in linguistic adaptations across LLMs. This hypothesis is based on previous findings by Huynh and McNamara (2025). Specifically, differences in model architectures also contribute to variability in outputs [80,81]. Moreover, different LLMs are trained on a unique training corpus so they exhibit different capabilities and behaviors even when given

 doi:10.20944/preprints202509.2013.v1

identical prompts [80,81]. As a result, each LLM's unique architectural design impacts how the model processes and generates answers [82,83]. These model differences lead to different interpretations and text generation by different LLMs.

2. Materials and Method

2.1. LLM Selection and Implementation Details

We selected four LLMs: Claude 3.5 Sonnet (Anthropic), Llama 3.1 (Meta), Gemini Pro 1.5 (Google), and ChatGPT 4o (OpenAI). Appendix A provides additional technical details, including version numbers, usage dates, and training specifications. These LLMs have comparable training and parameter sizes which indicates similar capabilities in language understanding and generation. Moreover, these LLMs have a strong track record of high performance on general-purpose natural language processing (NLP) tasks. Although the training sizes differ, all four models are widely recognized for producing coherent, contextually appropriate responses and demonstrating advanced language comprehension skills.

2.2. Text Corpus

Ten science and 10 history texts were compiled from the iSTART website www.adaptiveliteracy.com/istart (accessed on 10 June 2025). The texts are publicly available through the iSTART website (www.adaptiveliteracy.com/istart) (accessed on 10 June 2025). Users may create a free account on the website and access the texts from the "Texts Library" menu. These materials are licensed under the Creative Commons Attribution 4.0 International License (CC BY 4.0), (https://creativecommons.org/licenses/by/4.0/) (e.g., iSTART StairStepper, [77]).

The 20 passages varied in level of difficulty and cover a wide range of subject areas in science and history, making them suitable for evaluating how different LLMs adapt texts to various reader profiles. Table 2 presents details of the corpus, including domain, text titles, word counts, and Flesch–Kincaid Grade Levels.

Table 2. Scientific and history texts. Source: Authors' contribution.

Domain	Topic	Text Titles	Word Count	FKGL*
Science	Biology	Bacteria	468	12.10
Science	Biology	The Cells	426	11.61
Science	Biology	Microbes	407	14.38
Science	Biology	Genetic Equilibrium	441	12.61
Science	Biology	Food Webs	492	12.06
Science	Biology	Patterns of Evolution	341	15.09
Science	Biology	Causes and Effects of Mutations	318	11.35
Science	Biochemistry	Photosynthesis	427	11.44
Science	Chemistry	Chemistry of Life	436	12.71
Science	Physics	What are Gravitational Waves?	359	16.51
History	American History	Battle of Saratoga	424	9.86
History	American History	Battles of New York	445	11.77
History	American History	Battles of Lexington and Concord	483	12.85
History	American History	Emancipation Proclamation	271	13.4
History	American History	House of Burgesses	200	12.8
History	American History	Abraham Lincoln- Rise to Presidency	631	12.15
History	American History	George Washington	260	9.79
History	French & American History	Marquis de Lafayette	356	13.78
History	Dutch & American History	New York (New Amsterdam) Colony	403	12.97

 doi:10.20944/preprints202509.2013.v1

History	World History	Age of Exploration	490	10.49

*Flesch–Kincaid Grade Level.

2.3. Descriptions of Reader Profiles

Reading comprehension is influenced by domain-specific knowledge such that students may comprehend science and history texts differently depending on their prior knowledge in the domain even with comparable reading skills (e.g., [17,46]). A reader with strong prior knowledge in one domain (e.g., history) may not generalize to comprehension of texts in another subject area (e.g., biology, chemistry or physics). As such, we designed four hypothetical reader profiles replicating the profiles described in prior research and paired each with its corresponding domain (e.g., science PK with science texts, history PK with history texts, [42]). Details of the reader profiles are presented in Appendix B. The descriptions varied in terms of domain-specific prior knowledge (science or history) and reading proficiency (high or low). Each profile presented detailed descriptions of the reader's age, education, reading level, background knowledge, interests, and reading goals (see Appendix B for descriptions). While the use of eight profiles does not capture the full complexity of real-world personalization, they serve as a proof-of-concept simulation to assess the extent to which LLMs generate tailored content for each profile and leverage NLP tools to evaluate the alignment of linguistic features for readers' needs.

2.4. Procedure

We prompted the four LLMs to modify 10 science and 10 history texts to align with four distinct reader profiles, each varying in reading proficiency and prior knowledge in science or history domain. The prompt instructed the LLM to adapt the text to enhance comprehension and engagement while considering specific readers' needs according to their educational background, reading skill level, prior knowledge, and learning objectives. We used the theory-aligned RAG augmented prompt that guided linguistic adjustments based on the reader profiles (e.g., higher cohesion/simpler syntax for low-knowledge readers). Prior work showed that adding RAG helps LLM to adhere to these linguistic modifications consistently with established comprehension theories. Details of the prompting strategy is included in Appendix C. For every reader profile listed, an LLM generated 20 adapted texts, 10 in science and 10 in history. After completing a set of prompts for one profile, the conversation history was cleared before beginning the next reader. In total, each LLM generated 80 personalized modifications, resulting in a total of 320 outputs across all four models (Claude 3.5, Llama 3.1, Gemini 1.0, and ChatGPT 4o).

To evaluate the linguistic adaptations, we extracted the linguistic features listed in Table 1 using the Writing Analytics Tool (WAT; [47]). WAT extracts and provides validated indices of cohesion, syntactic complexity, and lexical sophistication that have been shown to correspond to comprehension difficulties faced by students. These linguistic patterns provide objective, quantifiable measures of text difficulty that correlate with expert judgments of readability and text difficulty. WAT metrics provides objective criteria for validating how well each modification corresponds to the cognitive and linguistic needs of distinct reader profiles. We also apply the same NLP features to compare science and history passages. For instance, low-knowledge readers should receive texts with higher cohesion and simpler syntax, whereas high-knowledge readers can handle, and even benefit from lower cohesion and sophisticated syntax and vocabulary [75].

3. Results

3.1. Main Effect of Reader Profile on Variations in Linguistic Features of Modified Texts

The goal of this analysis was to examine the extent to which the linguistic features of LLM-modified texts aligned with different reader profiles. A 4 × 4 × 2 MANCOVA was conducted to examine the effects of reader profiles (High RS/High PK, High RS/Low PK, Low RS/High PK, and

 doi:10.20944/preprints202509.2013.v1

Low RS/Low PK) and LLMs (Claude, Llama, Gemini, and ChatGPT) and text types (Science and History texts) on linguistic features of modifications, with word count included as a covariate.

The main effect of reader profiles on linguistic features was significant. See Table 3 for descriptive statistics and F-values of the main effects of reader profiles. As expected, results showed that linguistic features of modifications significantly differed across reader profiles. Modifications intended for Profile 1 (High RS/High PK) had the highest level of academic writing and text complexity measures, followed by Profile 2, then 3 and 4. Texts modified for Profile 1 had significantly higher language variety (M = 80.73, SD = 19.21) compared to texts modified for Profile 2 (High RS/Low PK) (M = 50.76, SD = 20.57), Profile 3 (Low RS/High PK) (M = 27.72, SD = 17.39), and Profile 4 (Low RS/Low PK) (M = 30.33, SD = 18.44), $p < 0.001$. Lexical density of adaptations for Profile 1 was significantly higher compared to modifications for all other profiles. Profile 1 (High RS/High PK) also had the lowest level of cohesion compared to Profile 2, 3 and 4, $p < 0.001$. Moreover, there were significant effects of reader profiles on word concreteness, sophisticated wording, language variety, and academic vocabulary measures which indicated that lexical complexity varied significantly based on the reading skill and prior knowledge.

Pairwise comparisons also revealed significant differences across reader profiles for measures of text complexity, including language variety, lexical density, noun-to-verb ratio, sentence length, sophisticated wording, and academic vocabulary, aligning with the hypothesized difficulty levels for each profile. As expected, texts modified for Profile 1 had the most complicated syntax and sophisticated vocabulary, followed by Profiles 2, 3, and 4 (all $ps < 0.05$).

Table 3. Descriptive Statistics and Main Effects of Reader Profiles. Source: Authors' contribution.

Linguistic Features	Reader 1 (High RS/ High PK*)		Reader 2 (High RS/Low PK*)		Reader 3 (Low RS/ High PK*)		Reader 4 (Low RS/Low PK*)		Main Effects of Profile		
	M	SD	M	SD	M	SD	M	SD	F (3, 320)	p	η^2
Academic Writing	75.84	24.74	51.66	26.48	33.06	27.15	34.30	22.96	121.25	<0.001	0.38
Language Variety	80.73	19.21	50.76	20.57	27.72	17.39	30.33	18.44	251.32	<0.001	0.55
Lexical Density	.68	.12	.61	.12	.59	.11	.58	.10	226.13	<0.001	0.53
Sentence Cohesion	32.86	28.89	54.75	29.93	55.83	22.68	60.45	26.92	35.11	<.001	.15
Noun-to-Verb Ratio	2.79	.46	2.53	.55	2.54	.72	1.84	.34	119.86	<.001	.37
Sentence Length	18.62	5.97	14.78	5.49	14.59	4.47	13.53	4.11	61.98	<.001	.23
Word Concreteness	29.86	17.79	50.52	25.63	55.18	27.21	60.76	24.96	57.26	<.001	.22
Sophisticated Word	88.85	9.52	51.12	21.09	29.05	17.64	23.42	16.06	603.28	<.001	.75
Academic Frequency	10842.04	1812.02	9471.83	1517.92	9017.30	1812.29	8823.87	1779.10	57.97	<.001	.22
Causal Connectives	.01	.01	.01	.01	.01	.01	.01	.01	3.11	.03	.02
Temporal Connectives	.01	.01	.01	.01	.01	.01	.01	.01	.79	.50	.00

 doi:10.20944/preprints202509.2013.v1

All Connectives	.05	.01	.05	.01	.05	.01	.05	.01	3.54	.02	.02

* RS= Reading Skill, PK= Prior Knowledge.

3.2. Main Effect of LLMs

The main effect of the models was significant. As expected, LLMs significantly differ in how they modify text linguistically regardless of reader characteristics. See Table 4 for descriptive statistics and F-values.

LLMs produced significantly different outputs even when given the same prompt and reader profiles. Llama's modifications (M = 41.55, SD = 27.20) had significantly lower cohesion compared to modifications by Claude (M = 63.55, SD = 28.59), Gemini (M = 47.94, SD = 27.31), and ChatGPT (M = 51.05, SD = 29.52), all $ps < 0.05$. Llama's modifications also contained less concrete, more technical vocabulary compared to modifications by Claude, Gemini, and ChatGPT (all $ps < 0.05$). Modifications by Gemini had the highest language variety (M = 54.42, SD = 28.78) compared to texts generated by Claude (M = 47.16, SD = 29.78), ChatGPT (M = 47.92, SD = 25.37) $p < 0.05$, and Llama (M = 40.04, SD = 27.99), $p < 0.001$. Claude's modifications were the lowest in text difficulty, as reflected by the highest cohesion, shortest sentence length, and lowest academic vocabulary compared to modifications generated by other LLMs.

Table 4. Descriptive statistics and main effects of LLMs. Source: authors' contribution.

Linguistic Features	Claude		Llama		Gemini		ChatGPT		Main Effects of LLMs		
	M	SD	M	SD	M	SD	M	SD	F (3, 320)	p	η^2
Academic Writing	48.42	31.73	53.28	30.50	45.78	30.98	47.37	29.23	1.98	0.12	0.01
Language Variety	47.16	29.78	40.04	27.99	54.42	28.78	47.92	25.37	12.12	<.001	0.06
Lexical Density	.62	.12	.61	.12	.62	.12	.61	.12	5.26	0.001	0.03
Sentence Cohesion	63.35	28.59	41.55	27.20	47.94	27.31	51.05	29.52	21.73	<.001	0.10
Noun-to-Verb Ratio	2.56	.84	2.38	.56	2.44	.57	2.33	.51	7.88	<.001	0.17
Sentence Length	12.46	4.38	16.32	5.15	16.38	5.09	16.35	5.88	42.71	<.001	0.17
Word Concreteness	47.33	26.23	46.25	26.89	51.94	27.85	50.80	26.00	1.60	.189	0.10
Sophisticated Word	47.30	32.20	46.70	27.88	49.38	31.52	49.05	30.82	2.72	.044	0.01
Academic Frequency	8848.43	1915.11	10388.74	1847.94	9592.04	1875.01	9325.83	1638.72	48.392	<.001	0.19
Causal Connectives	.01	.01	.01	.01	.01	.01	.01	.01	.56	.64	0.00
Temporal Connectives	.01	.01	.01	.01	.01	.01	.01	.01	11.87	<.001	0.06

 doi:10.20944/preprints202509.2013.v1

All Connectives	.05	.01	.05	.01	.05	.01	.05	.01	4.03	.007	0.02

3.3. Main Effect of Text Types

The main effect of the text types was significant. See Table 5 for descriptive statistics and F-values. Linguistic features of science and history text modifications differed significantly. Science texts contained longer sentences and higher lexical density ($M = 0.72$, $SD = 0.06$) than history texts ($M = 0.51$, $SD = 0.05$), $p < .001$. Science texts showed significantly more frequent use of sophisticated wording and academic vocabulary ($M = 10557.87$, $SD = 1670.58$) compared to history texts ($M = 8519.65$, $SD = 1540.14$), $p < .001$. These findings suggested that science modifications are more complex in terms of syntax and language used.

Moreover, science modifications contained more connectives compared to history modifications. However, when examining different types of connectives, science texts contained more causal connectives compared to history texts, whereas history text used more temporal connectives, $p < .001$. History texts also contained higher noun-to-verb ratio ($M = 2.51$, $SD = 0.68$) compared to science texts ($M = 2.34$, $SD = 0.58$), $p < .001$. These results suggested that history modifications used more noun phrases (e.g., people names, places, institutions, wars) and temporal connectives to organize and connect the timeline of historical events. In contrast, science texts included more causal connectives which established explicit cause–effect relations, a critical component of scientific writing.

Table 5. Descriptive statistics and main effects of text types. Source: authors' contribution.

Linguistic Features	Science Texts		History Texts		Main Effects of Text Types		
	M	SD	M	SD	F (1, 320)	*p*	η^2
Academic Writing	50.08	27.98	47.34	33.15	2.80	.10	0.01
Language Variety	47.06	29.30	47.71	27.56	.84	.36	0.00
Lexical Density	.72	.06	.51	.05	5743.50	<.001	0.90
Sentence Cohesion	51.26	28.63	50.68	29.80	.09	.76	0.00
Noun-to-Verb Ratio	2.34	.58	2.51	.68	18.31	<.001	0.03
Sentence Length	17.67	5.21	13.09	4.59	254.88	<.001	0.30
Word Concreteness	49.53	28.42	48.62	25.10	.12	.73	0.00
Sophisticated Word	49.27	30.95	46.94	30.25	5.00	.03	0.01
Academic Frequency	10557.87	1670.58	8519.65	1540.14	412.42	<.001	0.41
Causal Connectives	.01	.01	.00	.00	78.44	<.001	0.11
Temporal Connectives	.01	.01	.01	.01	17.01	<.001	0.03
All Connectives	.06	.01	.05	.01	26.50	<.001	0.04

3.4. Interaction Effect Reader x Text Genre

Linguistic adjustments for each reader profile were tailored depending on genre. Science passages are often dense, technical, and include abstract terminology that require strong prior knowledge to decode and derive meanings from text. In contrast, comprehending history texts relied

 doi:10.20944/preprints202509.2013.v1

more on understanding historical context and connections between events [46,50]. As such, cohesion should be enhanced for less skilled and low knowledge readers.

The interaction effects between text genre and profile were significant, F (3, 320) = 23.54, $p < 0.001$, $\eta 2$ = 0.10. As intended, the results showed that cohesion was significantly higher for history texts modified for low-knowledge readers compared to science texts. See Figure 1 for the effect of text genre and reader profiles on cohesion.

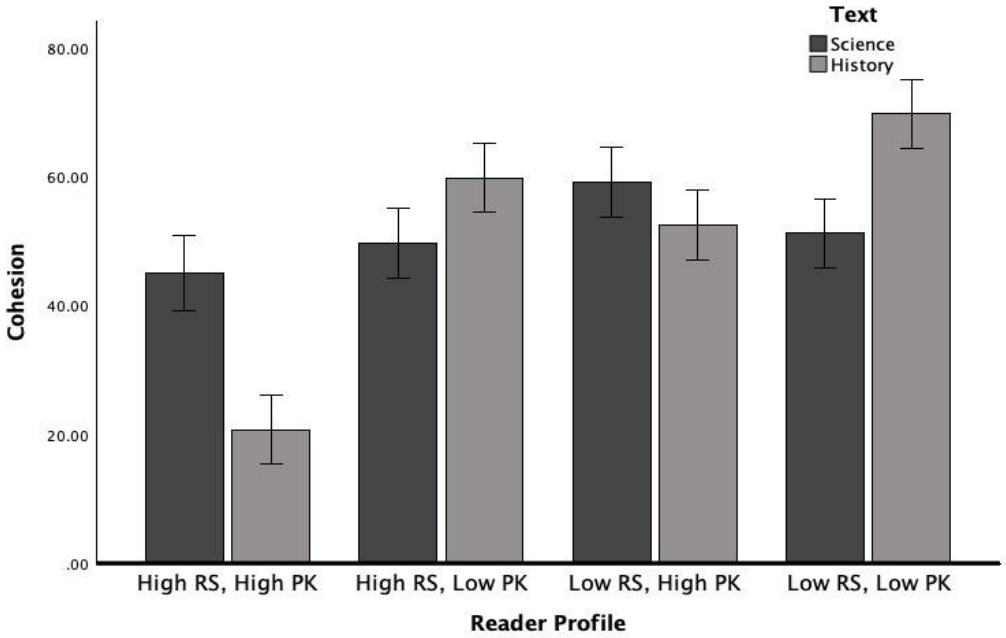

Figure 1. Cohesion as a function of text genre and reader profile. Source: Authors' contribution.

Science texts amplify the comprehension difficulty such that students with limited background struggle more with expository texts due to domain-specific vocabulary and conceptual density (McNamara, 2001). As such, to increase readability, lexical complexity in science text should be simplified for less skilled and low knowledge readers.

The interaction effect between text genre and profile was significant for sophisticated wording, F (3, 320) = 29.14, $p < 0.001$, $\eta 2$ = 0.12, and word concreteness, F (3, 320) = 16.83, $p < 0.001$, $\eta 2$ = 0.08. As intended, science modifications for low knowledge and less skilled profiles significantly decreased sophisticated wording and increased word concreteness. In contrast, modifications for high knowledge and skilled profiles significantly increased word complexity and were less abstract, aligning with the hypothesized difficulty levels for each profile. See Figures 2 and 3 for the effect of genre and reader profiles on sophisticated wording and word concreteness.

 doi:10.20944/preprints202509.2013.v1

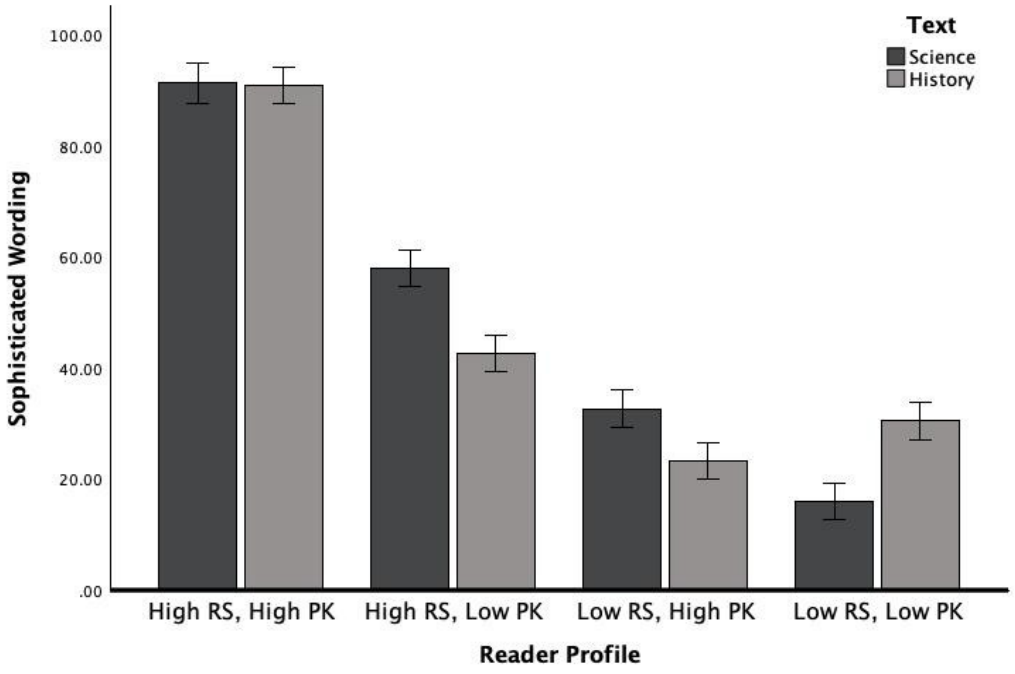

Figure 2.

Sophisticated wording as a function of text genre and reader profile. Source: Authors' contribution.

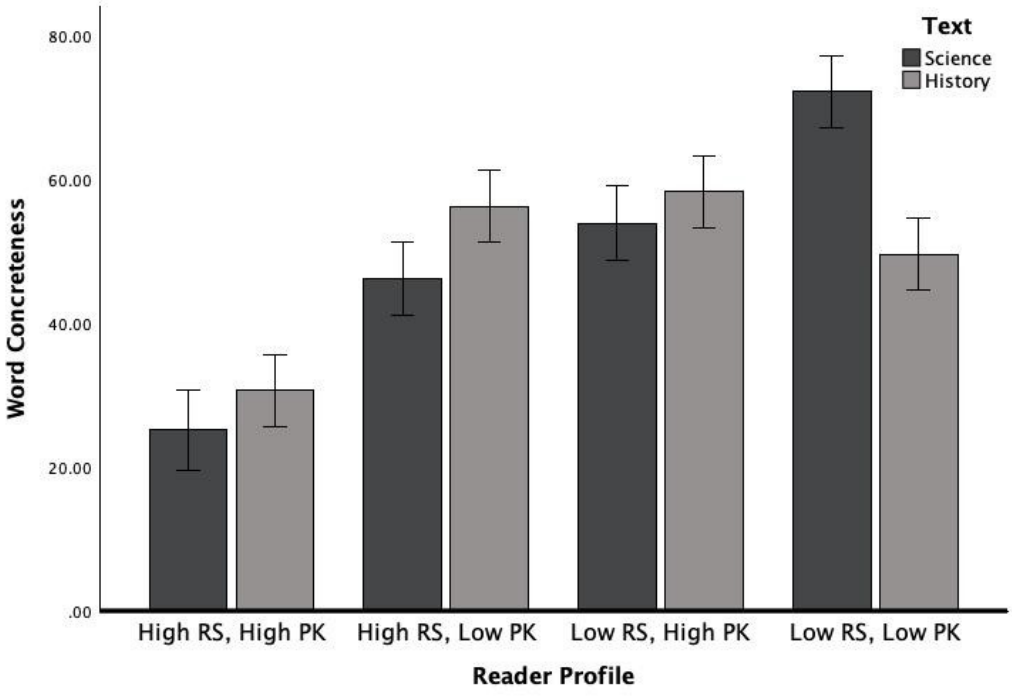

Figure 3. Word concreteness as a function of text genre and reader profile. Source: Authors' contribution.

4. Discussion

Four LLMs (i.e., Claude, Llama, Gemini, ChatGPT) were prompted to tailor educational texts in science and history domains for different reader profiles. NLP analyses were applied to provide insights into how effectively modifications align with readers' needs. Our findings demonstrated that NLP-based evaluation methods can be applied across both science and history domains to assess personalization quality. The results implied that LLMs are sensitive not only to readers' needs but

 doi:10.20944/preprints202509.2013.v1

also to language variation across genres. These linguistic patterns were quantified and demonstrated through NLP analyses showcasing that LLMs retained these domain-specific linguistic features when generating modifications.

4.1. Texts Adapted for Different Reader Profiles

Adaptations intended for less skilled and low-knowledge readers included text features that foster comprehension (i.e., high cohesion, more concrete word, simpler language and sentence structures), whereas modifications for high-knowledge and skilled readers featured sophisticated, less explicit connections to encourage active cognitive processing and deep comprehension. The results suggested that adapted texts for skilled readers with high background knowledge were complex, used more academic language, varied vocabulary, and sophisticated wording. Modifications for high prior knowledge readers had low cohesion and word concreteness which indicate that less explicit explanations were provided to promote deeper cognitive engagement through inference-making [17,45]. Using the theory-aligned prompt with RAG, LLMs modified text features meaningfully based on reader skill and knowledge to promote deep comprehension as informed by theories of text readability and reading comprehension. While less skilled and low knowledge readers received more cohesive, concrete, and syntactically simple modifications, modifications for skilled and high knowledge readers were higher in measures related to academic writing such as low cohesiveness, higher lexical density, more sophisticated syntax structure and vocabulary used. These linguistic modifications facilitate comprehension and engagement, ensuring that the text content is accessible and supportive for the low knowledge readers to fill in knowledge gaps. These findings highlighted that LLM-generated modifications successfully tailored linguistic complexity based on readers' needs by adjusting linguistic features grounded in theories from the reading comprehension literature (when they were prompted to do so and provided a sufficient knowledge base). These linguistic variations captured by NLP analyses align with prior research demonstrating that adjusting features related to text readability is an effective method to support reading comprehension outcomes [78,79].

4.2. LLMs Generated Outputs with Unique Linguistic Patterns

Moreover, LLMs differed in adaptation style due to inherent training data and fine-tuning strategy differences within each model [80,81]. Each model exhibited a unique writing style with variation in linguistic features, even when using the same prompt and personalizing texts for the same reader profile. Claude generated cohesive, shorter, and less academic texts, which lowers text difficulty level. In contrast, Llama generated texts that are more syntactically and lexically more complex but less cohesive, which are similar to texts commonly found in science domains. Modifications generated by Gemini were rich and varied in the language used, and moderately complex. ChatGPT produced outputs with moderate cohesion and complexity. These results highlighted the distinct linguistic pattern of each LLM which were consistent with prior research [42,82,83]. NLP analyses demonstrated how various LLMs approach personalization tasks differently which highlights the importance of understanding each model's unique capabilities and weaknesses. This study reinforced the importance of NLP-based evaluation methods in providing replicable and scalable metrics to rapidly assess personalization quality.

4.3. Linguistic Differences of Adapted Texts from Science Versus History Domain

Prior corpus analyses suggesting that science texts are typically more abstract, information-dense, and contain more complex vocabulary than history texts [84]. Scientific writing also connects propositions using causal connectives that signal explanatory logic [85]. In contrast, history discourse includes more temporal connectives (e.g., then, during, by the time) and noun-dense, highlighting that history texts have a narrative and sequential structure. Historical texts are typically noun-dense

 doi:10.20944/preprints202509.2013.v1

due to frequent references to people, places, and institutions and event sequencing (e.g., George Washington, Industrial Revolution; [86,87]).

In our study, NLP analyses revealed that the adapted texts preserved the same linguistic patterns observed in science and history texts. As expected, science modifications exhibited higher lexical density, longer sentence length, and higher use of domain-specific academic vocabulary (e.g., organism, mutation, photosynthesis). Science modifications also included explicit causal and referential cohesive markers (e.g., because, therefore, thus) to outline processes, explanations, and cause-effect relationships [60,63,88]. In contrast, history modifications featured shorter sentences describing sequences of historical events. They also contained more temporal connectives and higher noun-to-verb ratio compared to science modifications. The modifications preserved genre-typical patterns and thus reflect each discipline's epistemic goals. While science texts aim to explain generalized causal processes, history texts focus on recounting chronology of events.

While text adaptation using LLMs offers a scalable solution to tailor learning content to individual readers' needs, evaluating the effectiveness of personalization can be challenging. The current study contributes to the current literature by demonstrating how NLP analyses can be leveraged to assess personalization across domains. With NLP, meaningful linguistic patterns in modifications were objectively quantified and rendered visible. Rather than relying on intuition, the current study demonstrated how NLP metrics can be used to evaluate whether a text matched the expected profile. NLP-based evaluations enable researchers to extract quantifiable linguistic metrics which are associated with comprehension difficulty. NLP analyses also move beyond readability formulas to align linguistic features with learners' needs based on cognitive and reading comprehension theories. These metrics help to concretely demonstrate how personalized content is aligned for learners with varying needs. Unlike subjective judgments or manual analyses, NLP provides scalable, replicable, and rigorous assessments of linguistic features, facilitating more precise evaluations of text modifications and alignment with reader profiles.

4.4. Limitations and Future Directions

The current study focused on science and history subjects which were chosen due to their distinct linguistic characteristics. Expanding the examination to additional domains such as literature, social studies, or mathematics and non-academic genres would further test the generalizability and robustness of the NLP-based evaluation approach. Moreover, although the current study included a limited set of reader profiles varying reading skills and prior knowledge, which was sufficient for proof-of-concept testing, a more diverse profile set (e.g., cultural background, motivation, interests, and learning disabilities) may capture a more realistic classroom settings [23,29]. Future research might also expand to examine the impacts of LLM-generated personalized text on actual learner performance and engagement, providing further insights into the pedagogical values and effectiveness of personalized LLM modifications.

Moreover, we generated one modification for each reader profile, LLM, and text combination. Since LLMs are stochastic, repeated generations and versioned models would allow in-depth examination of output variance [89,90]. Future insight may be gleaned from examining the consistency and reproducibility by incorporating multiple prompt generation cycles and different model versions.

Finally, hallucinations in which LLM-generated content that is linguistically coherent but semantically incorrect remains a known risk in an LLM system [91,92]. While linguistic metrics from NLP analyses provide insights into readability and cognitive alignment with the reader's ability, evaluation of semantic accuracy remains challenging [93]. Future studies could incorporate multi-agent verification methods that combine automated NLP assessments with human validation in the loop to enhance reliability.

5. Conclusion

 doi:10.20944/preprints202509.2013.v1

In this study, we leveraged an NLP-based validation method to systematically evaluate the personalization capabilities of four LLMs across diverse reader profiles and text genres. Our results demonstrated linguistic variations between science and history texts, highlighting domain-specific adaptations generated by each LLM [51,87]. Additionally, each LLM exhibited unique linguistic characteristics in text generation, providing evidence for the inherent differences attributable to training corpus and model architectures [89,90]. Replicating the NLP evaluation framework from Huynh and McNamara (2025), this study quantified linguistic variations and demonstrated alignment between text modifications and intended reader profiles across different genre. NLP assessment can be automated to provide immediate feedback, enabling real-time assessment and text refinement. As a result, personalized content can be modified based on quantitative feedback, assessed, and adapted iteratively. NLP-based evaluation methods provide a robust framework to continuously assess and improve LLM-generated personalized texts, facilitating adaptive, data-driven personalization across diverse domains.

Funding: The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305T240035 to Arizona State University. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Abbreviations

The following abbreviations are used in this manuscript:

PK	Prior knowledge
RS	Reading skills
GenAI	Generative AI
LLM	Large Language Model
AFS	Adaptive Feedback System
RAG	Retrieval-Augmented Generation
NLP	Natural Language Processing
iSTART	Interactive Strategy Training for Active Reading and Thinking
FKGL	Flesch-Kincaid Grade Level
WAT	Writing Analytics Tool

Appendix A. LLM Descriptions

This appendix includes details on the LLMs used in the submitted research, including model versions, date of access, access method, training size, and number of parameters.

Claude 3.5 Sonnet (Anthropic)

- Version Used: Claude 3.5;
- Date Accessed: 10 June 2025;
- Accessed via https://poe.com web deployment, default configurations were used;
- Training Size: Claude is trained on a large-scale, diverse dataset derived from a broad range of online and curated sources. The exact size of the training data remains proprietary;
- Number of Parameters: The exact number is not disclosed by Anthropic, but it is estimated to be between 70–100 billion parameters.

Llama (Meta)

- Version Used: Llama 3.1;
- Date Accessed: 10 June 2025;
- Accessed via https://poe.com web deployment, default configurations were used;
- Llama 3.1 was trained on 2 trillion tokens sourced from publicly available datasets, including books, websites, and other digital content;
- Number of Parameters: 70 billion parameters.

Gemini Pro 1.5 (Google DeepMind)

 doi:10.20944/preprints202509.2013.v1

- Version Used: Gemini Pro 1.5;
- Date Accessed: 10 June 2025;
- Accessed via https://poe.com web deployment, default configurations were used;
- Training Size: Gemini is trained on 1.5 trillion tokens, sourced from a wide variety of publicly available and curated data, including text from books, websites, and other large corpora;
- Number of Parameters: 100 billion parameters.

ChatGPT 4o (OpenAI)

- Version Used: GPT-4o;
- Date Accessed: 10 June 2025;
- Accessed via https://poe.com web deployment, default configurations were used;
- Training Size: GPT-4 was trained on an estimated 1.8 trillion tokens from diverse sources, including books, web pages, academic papers, and large text corpora;
- Number of Parameters: The exact number is not publicly disclosed but is in the range of 175 billion parameters.

Appendix B. Reader Profile Descriptions

This appendix includes details of the various reader profiles provided to the LLMs.

Table A1. Augmented Prompt. Source: authors' contribution.

	Descriptions of High and Low Knowledge Reader in Science	**Descriptions of High and Low Knowledge Reader in History**
Reader 1 (High RS/High PK*)	Age: 25 Educational level: Senior Major: Chemistry (Pre-med) ACT English composite score: 32/36 (performance is in the 96th percentile) ACT Reading composite score: 32/36 (performance is in the 96th percentile) ACT Math composite score: 28/36 (performance is in the 89th percentile) ACT Science composite score: 30/36 (performance is in the 94th percentile) Science background: Completed eight required biology, physics, and chemistry college-level courses (comprehensive academic background in the sciences, covering advanced topics in biology, chemistry, and physics, well-prepared for higher-level scientific learning and analysis) Reading goal: Understand scientific concepts and principles	Age: 25 Educational level: Senior Major: History and Archeology ACT English: 32/36 (96th percentile) ACT Reading: 33/36 (97th percentile) AP History score: 5 out of 5 History background: Completed 4 years of college-level courses in U.S. and World History (extensive training in historical analysis, primary source evaluation, and historiography) Reading goal: Understand key historical events and their relevance to society
Reader 2 (High RS/Low PK*)	Age: 20 Educational level: Sophomore Major: Psychology ACT English composite score: 32/36 (performance is in the 96th percentile)	Age: 21 Educational level: Junior Major: Biology ACT English: 32/36 (96th percentile) ACT Reading: 31/36 (94th percentile)

 doi:10.20944/preprints202509.2013.v1

	ACT Reading composite score: 31/36 (performance is in the 94th percentile) ACT Math composite score: 18/36 (performance is in the 42th percentile) ACT Science composite score: 19/36 (performance is in the 46th percentile) Science background: Completed one high-school-level chemistry course (no advanced science course). Limited exposure and understanding of scientific concepts Interests/Favorite subjects: arts, literature Reading goal: Understand scientific concepts and principles	AP History score: 2 out of 5 History background: Completed general education high school history; no college-level history courses. Limited interest and knowledge of historical events Interests/Favorite subjects: arts, literature Reading goal: Understand key historical events and their relevance to society
Reader 3 (Low RS/High PK*)	Age: 20 Educational level: Sophomore Major: Health Science ACT English composite score: 19/36 (performance is in the 44th percentile) ACT Reading composite score: 20/36 (performance is in the 47th percentile) ACT Math composite score: 32/36 (performance is in the 97th percentile) ACT Science composite score: 30/36 (performance is in the 94th percentile) Science background: Completed one physics, one astronomy, and two college-level biology courses (substantial prior knowledge in science, having completed multiple college-level courses across several disciplines, strong foundation in scientific principles and concepts) Reading goal: Understand scientific concepts Reading disability: Dyslexia	Age: 22 Educational level: Junior Major: History ACT English: 19/36 (44th percentile) ACT Reading: 20/36 (47th percentile) AP History score: 5 out of 5 History background: Completed 3 years of college-level history courses (specializing in U.S. history and early modern Europe) Reading goal: Understand key historical events and their relevance to society Reading disability: Dyslexia
Reader 4 (Low RS/Low PK*)	Age: 18 Educational level: Freshman Major: Marketing ACT English composite score: 17/36 (performance is in the 33rd percentile) ACT Reading composite score: 18/36 (performance is in the 36th percentile) ACT Math composite score: 19/36 (performance is in the 48th percentile) ACT Science composite score: 17/36 (performance is	Age: 18 Educational level: Freshman Major: Finance ACT English: 18/36 (35th percentile) ACT Reading: 17/36 (32nd percentile) AP History: 1 out of 5 History background: Only completed basic U.S. History in high school; little engagement or interest in history topics Reading goal: Understand key historical

 doi:10.20944/preprints202509.2013.v1

	in the 34th percentile) Science background: Completed one high-school-level biology course (no advanced science course) Limited exposure and understanding of scientific concepts Reading goal: Understand scientific concepts	events and their relevance to society

* RS= Reading Skill, PK= Prior Knowledge.

Appendix C. Prompt Used

This appendix includes the augmented prompt which incorporates advanced prompting strategies such as personification, task objectives, chain-of-thought reasoning, and Retrieval-Augmented Generation (RAG).

Table A2. Augmented Prompt. Source: authors' contribution.

Components	Augmented Prompt
Personification	Imagine you are a cognitive scientist specializing in reading comprehension and learning science
Task objectives	• Modify this text to enhance text comprehension, engagement, and accessibility for the reader profile while maintaining conceptual depth, scientific rigor, and pedagogical value • Adapt the text in a way that supports the readers' understanding of scientific concepts, using strategies that align with empirical findings on text cohesion, reading skills, and prior knowledge • Help the reader retain scientific concepts and reinforce understanding • Ensure that the reader can build meaningful understanding while being challenged at an appropriate level
Chain-of-thought	Explain the rationale behind each modification approach and how each change helps the reader grasp the scientific concepts and retain information
RAG	Refer to the attached pdf files. Apply these empirical findings and theoretical frameworks from these files as guidelines to tailor text • Impact of prior knowledge and reading skills on comprehension of science texts • Impact of prior knowledge on integration of new knowledge according to the Construction-Integration (CI) Model of Text Comprehension • Impact of text cohesion on comprehension the differential effect of cohesion on comprehension depending on level of prior knowledge and reading skills
Reader profile	[Insert Reader Profile Description from Appendix B]
Text input	[Insert Text]

References

1. Lee, J.S. InstructPatentGPT: Training patent language models to follow instructions with human feedback. *Artif. Intell. Law* **2024**, 1–44.
2. Cherian, A.; Peng, K.C.; Lohit, S.; Matthiesen, J.; Smith, K.; Tenenbaum, J. Evaluating large vision-and-language models on children's mathematical olympiads. *Adv. Neural Inf. Process. Syst.* **2024**, 37, 15779–15800.
3. Liu, D.; Hu, X.; Xiao, C.; Bai, J.; Barandouzi, Z.A.; Lee, S.; Lin, Y. Evaluation of large language models in tailoring educational content for cancer survivors and their caregivers: Quality analysis. *JMIR Cancer* **2025**, 11, e67914.

4. Krause, S.; Stolzenburg, F. Commonsense reasoning and explainable artificial intelligence using large language models. In *Proc. Eur. Conf. Artif. Intell.*; Springer: Cham, Switzerland, **2023**; pp. 302–319.
5. Lin, C.-Y. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proc. ACL-04 Workshop*; Assoc. Comput. Linguist.: Barcelona, Spain, **2004**; pp. 74–81.
6. Lee, C.; Porfirio, D.; Wang, X.J.; Zhao, K.; Mutlu, B. VeriPlan: Integrating formal verification and LLMs into end-user planning. *arXiv* **2025**, arXiv:2502.17898.
7. Pesovski, I.; Santos, R.; Henriques, R.; Trajkovik, V. Generative AI for customizable learning experiences. *Sustainability.* **2024**, 16(7), 3034.
8. Laak, K.-J.; Aru, J. AI and personalized learning: Bridging the gap with modern educational goals. *arXiv* **2024**, arXiv:2404.02798.
9. Park, M.; Kim, S.; Lee, S.; Kwon, S.; Kim, K. Empowering personalized learning through a conversation-based tutoring system with student modeling. In *Extended Abstracts of the CHI Conf. on Human Factors in Comput. Syst.*; ACM: New York, NY, USA, **2024**; pp. 1–10.
10. Wen, Q.; Liang, J.; Sierra, C.; Luckin, R.; Tong, R.; Liu, Z.; Cui, P.; Tang, J. AI for education (AI4EDU): Advancing personalized education with LLM and adaptive learning. In *Proc. 30th ACM SIGKDD Conf. on Knowledge Discovery and Data Mining*; ACM: New York, NY, USA, **2024**; pp. 6743–6744.
11. Pane, J.F.; Steiner, E.D.; Baird, M.D.; Hamilton, L.S.; Pane, J.D. Informing Progress: Insights on Personalized Learning Implementation and Effects; Res. Rep. RR-2042-BMGF; Rand Corp.: Santa Monica, CA, USA, 2017.Pane, J.F.; Steiner, E.D.; Baird, M.D.; Hamilton, L.S.; Pane, J.D. Informing progress: Insights on personalized learning implementation and effects. *Rand Corp. Res. Rep.* **2017**.
12. Bernacki, M.L.; Greene, M.J.; Lobczowski, N.G. A systematic review of research on personalized learning: Personalized by whom, to what, how, and for what purpose(s)? *Educ. Psychol. Rev.* **2021**, 33(4), 1675–1715.
13. Kaswan, K.S.; Dhatterwal, J.S.; Ojha, R.P. AI in personalized learning. In *Advances in Technological Innovations in Higher Education*; CRC Press: Boca Raton, FL, USA, **2024**; pp. 103–117.
14. Jian, M.J.K.O. Personalized learning through AI. *Adv. Eng. Innov.* **2023**, 5(1), 16–19.
15. Pratama, M.P.; Sampelolo, R.; Lura, H. Revolutionizing education: Harnessing the power of artificial intelligence for personalized learning. *Klasikal: J. Educ. Lang. Teach. Sci.* **2023**, 5(2), 350–357.
16. Martínez, P.; Moreno, L.; Ramos, A. Exploring large language models to generate easy-to-read content. *Front. Comput. Sci.* **2024**, 6, 1394705.
17. Ozuru, Y.; Dempsey, K.; McNamara, D.S. Prior knowledge, reading skill, and text cohesion in the comprehension of science texts. *Learn. Instr.* **2009**, 19(3), 228–242.
18. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; … Amodei, D. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **2020**, 33, 1877–1901.
19. Hardaker, G.; Glenn, L.E. Artificial intelligence for personalized learning: A systematic literature review. *Int. J. Inf. Learn. Technol.* **2025**, 42(1), 1–14. https://doi.org/10.1108/IJILT-07-2024-0160
20. Ma, X.; Mishra, S.; Liu, A.; Su, S.Y.; Chen, J.; Kulkarni, C.; … Chi, E. Beyond chatbots: ExploreLLM for structured thoughts and personalized model responses. In *Extended Abstracts of the CHI Conf. on Human Factors in Computing Systems*; ACM: New York, NY, USA, **2024**; pp. 1–12.
21. Ng, C.; Fung, Y. Educational personalized learning path planning with large language models. *arXiv Preprint* **2024**, arXiv:2407.11773.
22. Zhang, Y.; Xu, X.; Zhang, M.; Cai, N.; Lei, V.N.L. Personal learning environments and personalized learning in the education field: Challenges and future trends. In *Applied Degree Education and the Shape of Things to Come*; Springer Nature Singapore: Singapore, **2023**; pp. 231–247.
23. Sharma, S.; Mittal, P.; Kumar, M.; Bhardwaj, V. The role of large language models in personalized learning: A systematic review of educational impact. *Discov. Sustain.* **2025**, 6(1), 1–24.
24. Lyu, W.; Wang, Y.; Chung, T.; Sun, Y.; Zhang, Y. Evaluating the effectiveness of LLMs in introductory computer science education: A semester-long field study. In *Proc. 11th ACM Conf. on Learning@Scale*; ACM: New York, NY, USA, **2024**; pp. 63–74.
25. Létourneau, A.; Deslandes Martineau, M.; Charland, P.; Karran, J.A.; Boasen, J.; Léger, P.M. A systematic review of AI-driven intelligent tutoring systems (ITS) in K–12 education. *npj Sci. Learn.* **2025**, 10(1), 29.

 doi:10.20944/preprints202509.2013.v1

26. Xiao, R.; Hou, X.; Ye, R.; Kazemitabaar, M.; Diana, N.; Liut, M.; Stamper, J. Improving student–AI interaction through pedagogical prompting: An example in computer science education. *arXiv Preprint* **2025**, arXiv:2506.19107.
27. Cuéllar, Ó.; Contero, M.; Hincapié, M. Personalized and timely feedback in online education: Enhancing learning with deep learning and large language models. *Multimodal Technol. Interact.* **2025**, 9(5), 45.
28. Wang, S.; Xu, T.; Li, H.; Zhang, C.; Liang, J.; Tang, J.; … Wen, Q. Large language models for education: A survey and outlook. *arXiv Preprint* **2024**, arXiv:2403.18105.
29. Merino-Campos, C. The impact of artificial intelligence on personalized learning in higher education: A systematic review. *Trends High. Educ.* **2025**, 4(2), 17.
30. Yan, L.; Sha, L.; Zhao, L.; Li, Y.; Martinez-Maldonado, R.; Chen, G.; … Gašević, D. Practical and ethical challenges of large language models in education: A systematic scoping review. *Br. J. Educ. Technol.* **2024**, 55(1), 90–112.
31. Jacobsen, L.J.; Weber, K.E. The promises and pitfalls of large language models as feedback providers: A study of prompt engineering and the quality of AI-driven feedback. *AI* **2025**, 6(2), 35. https://doi.org/10.3390/ai6020035
32. Murtaza, M.; Ahmed, Y.; Shamsi, J.A.; Sherwani, F.; Usman, M. AI-based personalized e-learning systems: Issues, challenges, and solutions. *IEEE Access* **2022**, 10, 81323–81342.
33. Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K.Q.; Artzi, Y. BERTScore: Evaluating text generation with BERT. In *Proc. Int. Conf. on Learning Representations (ICLR)*, **2020**.
34. Basham, J.D.; Hall, T.E.; Carter, R.A. Jr.; Stahl, W.M. An operationalized understanding of personalized learning. *J. Spec. Educ. Technol.* **2016**, 31(3), 126–135. https://doi.org/10.1177/0162643416660835
35. Bray, B.; McClaskey, K. A step-by-step guide to personalize learning. *Learn. Lead. Technol.* **2013**, 40(7), 12–19.
36. Banerjee, S.; Lavie, A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proc. ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, Ann Arbor, MI, USA, **2005**; pp. 65–72.
37. Lin, C.-Y. ROUGE: A package for automatic evaluation of summaries. In Text Summarization Branches Out; Proceedings of the ACL-04 Workshop, Barcelona, Spain, July **2004**; pp. 74–81.
38. Novikova, J.; Dušek, O.; Curry, A.C.; Rieser, V. Why we need new evaluation metrics for NLG. *arXiv* **2017**, arXiv:1707.06875.
39. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. BLEU: A method for automatic evaluation of machine translation. In *Proc. 40th Annu. Meeting Assoc. Comput. Linguist.*; Philadelphia, PA, USA, **2002**; pp. 311–318.
40. Crossley, S.; Salsbury, T.; McNamara, D. Measuring L2 lexical growth using hypernymic relationships. *Lang. Learn.***2009**, 59(2), 307–334.
41. Tetzlaff, L.; Schmiedek, F.; Brod, G. Developing personalized education: A dynamic framework. *Educ. Psychol. Rev.***2021**, 33(3), 863–882.
42. Huynh, L.; McNamara, D.S. GenAI-powered text personalization: Natural language processing validation of adaptation capabilities. *Appl. Sci.* **2025**, 15(12), 6791.
43. Cromley, J.G.; Snyder-Hogan, L.E.; Luciw-Dubas, U.A. Reading comprehension of scientific text: A domain-specific test of the direct and inferential mediation model of reading comprehension. *J. Educ. Psychol.* **2010**, 102(3), 687–700. https://doi.org/10.1037/a0019452
44. Frantz, R.S.; Starr, L.E.; Bailey, A.L. Syntactic complexity as an aspect of text complexity. *Educ. Res.* **2015**, 44(7), 387–393. https://doi.org/10.3102/0013189X15603980
45. O'Reilly, T.; McNamara, D.S. Reversing the reverse cohesion effect: Good texts can be better for strategic, high-knowledge readers. *Discourse Process.* **2007**, 43(2), 121–152.
46. McNamara, D.S.; Ozuru, Y.; Floyd, R.G. Comprehension challenges in the fourth grade: The roles of text cohesion, text genre, and readers' prior knowledge. *Int. Electron. J. Elem. Educ.* **2011**, 4(1), 229–257.
47. Potter, A.; Shortt, M.; Goldshtein, M.; Roscoe, R.D. Assessing academic language in tenth-grade essays using natural language processing. *Assess. Writ.* **2025**, in press.
48. Crossley, S.A. Developing linguistic constructs of text readability using natural language processing. *Sci. Stud. Read.***2025**, 29(2), 138–160.

 doi:10.20944/preprints202509.2013.v1

49. Crossley, S.A.; Skalicky, S.; Dascalu, M.; McNamara, D.S.; Kyle, K. Predicting text comprehension, processing, and familiarity in adult readers: New approaches to readability formulas. *Discourse Process.* **2017**, 54, 340–359.
50. Smith, R.; Snow, P.; Serry, T.; Hammond, L. The role of background knowledge in reading comprehension: A critical review. *Read. Psychol.* **2021**, 42(3), 214–240.
51. Biber, D.; Gray, B. Nominalizing the verb phrase in academic scientific writing. In *The Verb Phrase in English: Investigating Recent Language Change with Corpora*; Aarts, B.; Close, J.; Leech, G.; Wallis, S., Eds.; Cambridge University Press: Cambridge, UK, **2013**; pp. 99–132. https://doi.org/10.1017/CBO9781139060998.006
52. Fang, Z.; Schleppegrell, M.J. *Reading in Secondary Content Areas: A Language-Based Pedagogy*; University of Michigan Press: Ann Arbor, MI, USA, **2008**.
53. Halliday, M.A.K.; Martin, J.R. *Writing Science: Literacy and Discursive Power*; Routledge: London, UK, **2003**.
54. Schleppegrell, M.J.; Achugar, M.; Oteíza, T. The grammar of history: Enhancing content-based instruction through a functional focus on language. *TESOL Q.* **2004**, 38(1), 67–93.
55. Biber, D.; Gray, B.; Poonpon, K. Lexical density and structural elaboration in academic writing over time: A multidimensional corpus analysis. *J. English Acad. Purp.* **2021**, 50, 100968.
56. Graesser, A.C.; McNamara, D.S.; Kulikowich, J.M. Coh-Metrix: Providing multilevel analyses of text characteristics. *Educ. Res.* **2011**, 40(5), 223–234. https://doi.org/10.3102/0013189X11413260
57. Nagy, W.E.; Townsend, D. Words as tools: Learning academic vocabulary as language acquisition. *Read. Res. Q.* **2012**, 47(1), 91–108. https://doi.org/10.1002/RRQ.011
58. Biber, D.; Gray, B. Nominalizing the verb phrase in academic science writing. In The Verb Phrase in English: Investigating Recent Language Change with Corpora; Aarts, B., Close, J., Leech, G., Wallis, S., Eds.; Cambridge University Press: Cambridge, UK, **2013**; pp. 99–132. https://doi.org/10.1017/CBO9781139060998.006
59. Dong, J.; Wang, H.; Buckingham, L. Mapping out the disciplinary variation of syntactic complexity in student academic writing. *System* **2023**, 113, 102974.
60. Fang, Z. The language demands of science reading in middle school. *Int. J. Sci. Educ.* **2006**, 28(5), 491–520.
61. Grever, M.; Van der Vlies, T. Why national narratives are perpetuated: A literature review on new insights from history textbook research. *London Rev. Educ.* **2017**, 15(2), 1–16. https://doi.org/10.18546/LRE.15.2.03
62. Huijgen, T.; Van Boxtel, C.; Van de Grift, W.; Holthuis, P. Toward historical perspective taking: Students' reasoning when contextualizing the actions of people in the past. *Theory Res. Soc. Educ.* **2017**, 45(1), 110–144. https://doi.org/10.1080/00933104.2016.1208597
63. Duran, N.D.; McCarthy, P.M.; Graesser, A.C.; McNamara, D.S. Using temporal cohesion to predict temporal coherence in narrative and expository texts. *Behav. Res. Methods* **2007**, 39(2), 212–223.
64. Van Drie, J.; Van Boxtel, C. Historical reasoning: Towards a framework for analyzing students' reasoning about the past. *Educ. Psychol. Rev.* **2008**, 20(2), 87–110.
65. Wineburg, S.S.; Martin, D.; Monte-Sano, C. *Reading like a historian: Teaching literacy in middle and high school history classrooms*; Teachers College Press: New York, NY, USA, **2012**.
66. Shanahan, T.; Shanahan, C. Teaching disciplinary literacy to adolescents: Rethinking content-area literacy. *Harv. Educ. Rev.* **2008**, 78(1), 40–59.
67. Blevins, B.; Magill, K.; Salinas, C. Critical historical inquiry: The intersection of ideological clarity and pedagogical content knowledge. *J. Soc. Stud. Res.* **2020**, 44(1), 35–50. https://doi.org/10.1016/j.jssr.2019.03.001
68. Biber, D.; Conrad, S.; Cortes, V. If you look at…: Lexical bundles in university teaching and textbooks. *Appl. Linguist.***2004**, 25(3), 371–405. https://doi.org/10.1093/applin/25.3.371
69. Hyland, K. As can be seen: Lexical bundles and disciplinary variation. *English Spec. Purp.* **2008**, 27(1), 4–21. https://doi.org/10.1016/j.esp.2007.06.001
70. Malvern, D.; Richards, B.; Chipere, N.; Durán, P. *Lexical diversity and language development*; Palgrave Macmillan UK: London, UK, **2004**; pp. 16–30.
71. McCarthy, P.M.; Jarvis, S. MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behav. Res. Methods* **2010**, 42(2), 381–392.
72. Cain, K.; Oakhill, J.V.; Barnes, M.A.; Bryant, P.E. Comprehension skill, inference-making ability, and their relation to knowledge. *Mem. Cognit.* **2001**, 29(6), 850–859.

 doi:10.20944/preprints202509.2013.v1

73. Magliano, J.P.; Millis, K.K.; RSAT Development Team; Levinstein, I.; Boonthum, C. Assessing comprehension during reading with the Reading Strategy Assessment Tool (RSAT). *Metacogn. Learn.* **2011**, 6(2), 131–154.
74. Cruz Neri, N.; Guill, K.; Retelsdorf, J. Language in science performance: Do good readers perform better? *Eur. J. Psychol. Educ.* **2021**, 36(1), 45–61.
75. McNamara, D.S. Reading both high-coherence and low-coherence texts: Effects of text sequence and prior knowledge. *Can. J. Exp. Psychol.* **2001**, 55(1), 51–62.
76. Pickren, S.E.; Stacy, M.; Del Tufo, S.N.; Spencer, M.; Cutting, L.E. The contribution of text characteristics to reading comprehension: Investigating the influence of text emotionality. *Read. Res. Q.* **2022**, 57(2), 649–667.
77. Arner, T.; McCarthy, K.S.; McNamara, D.S. iSTART StairStepper—Using comprehension strategy training to game the test. *Comput.* **2021**, 10, 48.
78. Viera, R.T. Syntactic complexity in journal research article abstracts written in English. *MEXTESOL J.* **2022**, 46(2).
79. Wu, J.; Zhao, H.; Wu, X.; Liu, Q.; Su, J.; Ji, Y.; Wang, Q. Word concreteness modulates bilingual language control during reading comprehension. *J. Exp. Psychol. Learn. Mem. Cogn.* **2024**, Advance online publication.
80. Chen, L.; Zaharia, M.; Zou, J. How is ChatGPT's behavior changing over time? *Harv. Data Sci. Rev.* **2024**, 6(2).
81. Luo, Z.; Xie, Q.; Ananiadou, S. Factual consistency evaluation of summarization in the era of large language models. *Expert Syst. Appl.* **2024**, 254, 124456.
82. Liu, Y.; Cong, T.; Zhao, Z.; Backes, M.; Shen, Y.; Zhang, Y. Robustness over time: Understanding adversarial examples' effectiveness on longitudinal versions of large language models. *arXiv Preprint* **2023**, arXiv:2308.07847.
83. Rosenfeld, A.; Lazebnik, T. Whose LLM is it anyway? Linguistic comparison and LLM attribution for GPT-3.5, GPT-4, and Bard. *arXiv Preprint* **2024**, arXiv:2402.14533.
84. McNamara, D.S.; Graesser, A.C.; Louwerse, M.M. Sources of text difficulty: Across genres and grades. In *Measuring Up: Advances in How We Assess Reading Ability*; Sabatini, J.P.; Albro, E.; O'Reilly, T., Eds.; Rowman & Littlefield: Lanham, MD, USA, **2012**; pp. 89–116.
85. Achugar, M.; Schleppegrell, M.J. Beyond connectors: The construction of cause in history textbooks. *Linguist. Educ.***2005**, 16(3), 298–318.
86. Gatiyatullina, G.M.; Solnyshkina, M.I.; Kupriyanov, R.V.; Ziganshina, C.R. Lexical density as a complexity predictor: The case of science and social studies textbooks. *Res. Result. Theor. Appl. Linguist.* **2023**, 9(1), 11–26. https://doi.org/10.18413/2313-8912-2023-9-1-0-2
87. de Oliveira, L.C. Nouns in history: Packaging information, expanding explanations, and structuring reasoning. *Hist. Teach.* **2010**, 43(2), 191–203.
88. Follmer, D.J.; Li, P.; Clariana, R. Predicting expository text processing: Causal content density as a critical expository text metric. *Read. Psychol.* **2021**, 42(6), 625–662. https://doi.org/10.1080/02702711.2021.1935786
89. Atil, B.; Chittams, A.; Fu, L.; Ture, F.; Xu, L.; Baldwin, B. LLM Stability: A detailed analysis with some surprises. *arXiv e-prints* **2024**, arXiv:2408.
90. Zhou, H.; Savova, G.; Wang, L. Assessing the macro and micro effects of random seeds on fine-tuning large language models. *arXiv Preprint* **2025**, arXiv:2503.07329.
91. Alkaissi, H.; McFarlane, S.I. Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. *Cureus* **2023**, 15(2).
92. Hatem, R.; Simmons, B.; Thornton, J.E. A call to address AI "hallucinations" and how healthcare professionals can mitigate their risks. *Cureus* **2023**, 15(9).
93. Laban, P.; Kryściński, W.; Agarwal, D.; Fabbri, A.R.; Xiong, C.; Joty, S.; Wu, C.S. LLMs as factual reasoners: Insights from existing benchmarks and beyond. *arXiv Preprint* **2023**, arXiv:2305.14540.