Evaluation of Linguistic Consistency of LLM-Generated Text Personalization Using Natural Language Processing

Linh Huynh * and Danielle S. McNamara

Posted Date: 13 February 2026

doi: 10.20944/preprints202602.1117.v1

Keywords: natural language processing; large language models; linguistic evaluation; text personalization

 doi:10.20944/preprints202602.1117.v1

Article

Evaluation of Linguistic Consistency of LLM-Generated Text Personalization Using Natural Language Processing

Linh Huynh * and Danielle S. McNamara

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

* Correspondence: lthuynh1@asu.edu

Abstract

This study proposes a Natural Language Processing (NLP)-based evaluation framework to examine the linguistic consistency of Large Language Model (LLM)-generated personalized texts over time. NLP metrics were used to quantify and compare linguistic patterns across repeated generations produced using identical prompts. In Experiment 1, internal reliability was examined across 10 repeated generations from four LLMs (Claude, Llama, Gemini, and ChatGPT) applied to 10 scientific texts tailored for a specific reader profile. Linear mixed-effects models showed no effect of repeated generation on linguistic features (e.g., cohesion, syntactic complexity, lexical sophistication), suggesting short-term consistency across repeatedly generated outputs. Experiment 2 examined linguistic variation across model updates of GPT-4o (October 2024 vs. June 2025) and GPT-4.1 (June 2025). Significant variations were observed across outputs from different model versions. GPT-4o (June 2025) generated more concise but cohesive texts, whereas GPT-4.1 (June 2025) generated outputs that are more academic, lexically sophisticated and complex syntax. Given the rapid evolution of LLMs and the lack of standardized methods for tracking output consistency, the current work demonstrates one of the applications of NLP-based evaluation approaches for monitoring meaningful linguistic shifts across model updates over time.

Keywords: natural language processing; large language models; linguistic evaluation; text personalization

1. Introduction

Recent advances in generative AI (GenAI) and large language models (LLMs) have enabled scalable personalization of educational texts, allowing learning systems to tailor instructional materials and learning pace to support each learner's needs [1–3]. LLM-powered personalized learning has shown improvements in engagement, motivation, and learning outcomes compared to traditional instruction [4,5]. In educational contexts, effective text personalization requires LLMs to tailor linguistic features (i.e., cohesion, lexical sophistication, and syntactic complexity) to be consistent with theories of reading comprehension (e.g., the Construction–Integration model; [6]). These linguistic features affect learners differently depending on their prior knowledge and reading skill. As a result, even small shifts in linguistic patterns across repeated generations or model updates can disrupt the alignment between text characteristics and learners' cognitive needs, thereby undermining comprehension and learning outcomes [7,8].

Despite the widespread application of LLMs for tailoring learning materials [4,9,10], few studies have established standardized evaluation frameworks capable of assessing both the quality and consistency of personalized outputs across model versions and repeated generations [11]. To address these challenges, the present study leverages an NLP-based evaluation framework to examine whether personalized content remain linguistically consistent across repeated generations and model updates over time.

 doi:10.20944/preprints202602.1117.v1

1.1. Variability in LLM-Generated Outputs

Due to differences in model architecture, training data, fine-tuning strategies, and stochasticity, large language models (LLMs) exhibit substantial variability in outputs even under identical prompting conditions, [12–14]. Empirical analyses have shown that no two LLMs produce identical outputs even with the same input instruction, and these differences are observable in measurable lexical, syntactic, and discourse-level features of generated text [15,16]. For instance, an LLM specifically trained using a corpus of scientific or academic text outperforms in scientific tasks whereas a model trained on a coding data corpus would be better at performing code generation and programming tasks [12]. Beyond model architectural differences (e.g., ChatGPT vs. Gemini), output variability also emerges within models over time due to retraining cycles and fine-tuning updates. Longitudinal studies have documented significant behavioral shifts across versions within the same model family (e.g., GPT-4o vs. GPT-4.1) [17,18]. These changes are associated with improvement in model efficiency, leading to systematic shifts in generation style, verbosity, and structural properties of outputs [19]. ChatGPT's behavior evolves rapidly and frequently over time such that even the same model's name can produce significantly different outputs over relatively short time spans.

Additionally, fine-tuning further contributes to within-model variability by altering how models prioritize task constraints. Fine-tuning refers to the process of incorporating domain-specific knowledge to adapt a pre-trained LLM for a specific task, helping the model generates outputs that are more precise and contextually relevant [20]. Fine-tuning also introduces unintended changes in lexical, discourse structure, and reasoning behavior [21–23]. As such, differences in fine-tuning and parameter configurations influence how an LLM responds to prompts [24]. For instance, large-scale reproducibility studies have shown that simpler generation tasks tend to be more stable across repeated runs, whereas complex tasks requiring reasoning or abstraction exhibit greater variability [25].

In addition, stochasticity inherent in generative processes further contributes to output variability [26]. Small variations in prompt structure can result in substantial differences in generated outputs, particularly in the case of simple zero-shot or few-shot prompting [17,27]. Although structured prompting strategies can reduce some sources of variability, stochastic sampling remains a persistent challenge for ensuring reproducibility in deployed NLP systems [28]. Taken together, these sources of variability pose a critical challenge for the development of trustworthy and reproducible LLM-powered systems. This challenge is especially consequential in high-stakes applications such as educational contexts, where consistency and reliability are critical for ensuring quality. Prior work has shown that, when LLMs are applied for assessment tasks, there were variation in grading judgments and feedback quality across repeated evaluations of the same of student's work [18,29]. Without systematic evaluation methods capable of detecting and benchmarking shifts across models, updates, and repeated generations, it is difficult to determine whether observed changes reflect meaningful improvements or unintended stochastic noise. This challenge motivates the need for a theory-aligned evaluation framework that supports continuous monitoring and benchmarking of LLM behavior in real-world applications.

1.2. Natural Language Processing Application as an Evaluation Method

Although traditional human evaluation methods (e.g., expert ratings, comprehension assessments) are informative, they are resource-intensive, time-consuming, and impractical for rapid and iterative evaluation at scale. Existing automatic metrics mainly assess surface-level lexical similarity between model outputs and reference texts [30–32]. However, these approaches fail to capture nuanced linguistic features critical to text readability and comprehension, including cohesion, syntactic complexity, lexical sophistication [7,8,33]. Moreover, because text personalization and other adaptive generation tasks require continuous adjustment over time [34,35], evaluation frameworks must support ongoing validation and monitoring rather than static, one-time assessment.

 doi:10.20944/preprints202602.1117.v1

Natural Language Processing (NLP) analyses emerge as a promising solution for addressing these challenges by enabling systematic assessment of linguistic features in model-generated text. NLP is a computational method that extracts quantifiable linguistic indices related to lexical, syntactic, and discourse-level properties and have been widely used to characterize text complexity and variation across contexts [36,37]. Prior work has also shown that NLP-based analyses can reliably detect systematic linguistic differences across LLMs and prompting conditions among outputs generated by Claude, Llama, ChatGPT, and Gemini [7,8]. Linguistic indices derived from NLP tools are grounded in established models of text processing and coherence building (e.g., the Construction-Integration model; [38]), offering theoretically informed and diagnostic tools for evaluating whether generated outputs exhibit stable behavior. NLP tools support comparative benchmarking of LLM behavior and provide a mechanism for monitoring consistency and shift in deployed NLP systems beyond simple accuracy or similarity metrics.

Personalized text generation is an ideal test case to apply the NLP-based validation framework in assessing linguistic consistency. Unlike other generic text generation tasks, personalization requires LLMs to make controlled adjustments to multiple linguistic properties in response to specific learner profiles while maintaining stability across repeated generations. Even minor shifts in linguistic patterns can signal meaningful changes in model performance. Consequently, personalized generation provides a practically relevant context for evaluating linguistic consistency using NLP-based metrics.

2. Experiment 1: Short-Term Model Consistency

2.1. Introduction Experiment 1

To evaluate the reproducibility of LLM-generated text, Experiment 1 examines the internal linguistic consistency of repeated generations produced by the same model using identical prompts. Although stochastic decoding is an inherent feature of LLMs, reliable deployment of LLM-powered systems requires that stochasticity does not result in variation in critical linguistic properties associated with text readability and comprehension challenges. As such, Experiment 1 examines whether linguistic metrics extracted from an NLP tool [WAT; [39]] can detect meaningful instability attributed to repeated regeneration, independent of model updates or prompt changes.

In Experiment 1, we applied NLP-based evaluation methods to examine linguistic consistency across repeated generations produced by different LLMs under identical prompting conditions. We generated multiple adaptations of the same source texts from four LLMs (i.e., Claude, Llama, Gemini, and ChatGPT) and applied NLP analyses to quantify variability in linguistic features across repeated trials.

2.2. Method Experiemnt 1

2.2.1. Model Selection

Claude 3.5 Sonnet (Anthropic), Llama (Meta), Gemini Pro 1.5 (Google), and ChatGPT-4o (OpenAI) were used in this experiment. All models were accessed via Poe.com using default system configurations, with sampling parameters (e.g., temperature) held constant across repeated generations to ensure comparability. Appendix A provides additional information about each model, including version identifiers, access dates, and information about training size and parameters. A structured prompt template was used to instruct the LLMs modify the input scientific texts for a hypothetical reader profile (see Appendix B). The instruction specified the reader attributes (i.e., age, educational background, reading skill, prior knowledge, and reading goals) and instructed the LLMs to tailor discourse-level and lexical properties accordingly to suit this profile.

 doi:10.20944/preprints202602.1117.v1

2.2.2. Text Corpus

Ten scientific texts were compiled from the iSTART website www.adaptiveliteracy.com/istart (accessed on 10 June 2025). The excerpts varied in level of difficulty and cover a wide range of topics in science domain. Appendix C includes details of the corpus, including domain, text titles, word counts, and Flesch–Kincaid Grade Levels.

2.2.3. Procedure

We prompted the LLMs to tailor complex scientific texts to improve comprehension and engagement for a reader profile. Specifically, this reader was described as an 18-year-old college freshman majoring in Marketing, with below-average ACT scores and limited scientific background knowledge. The description of this reader was held constant across all generations to provide a controlled condition. For each model, the same prompt was applied to all ten texts and repeated ten times, resulting in 100 generated modifications per model and 400 total outputs across four models. Repeated generation under identical conditions allowed us to isolate within-model stochastic variation and assess the internal consistency of linguistic features across runs.

Linguistic features were extracted from each generated text using the Writing Analytics Tool [WAT; [39]]. WAT extracts and provides validated indices of cohesion, syntactic complexity, and lexical sophistication which have been shown to correspond to comprehension difficulties and reading challenges [40,41]. The extracted indices capture interpretable dimensions of linguistic structure and variation that are suitable for benchmarking stability across repeated generations (see Appendix D). These linguistic patterns also provide objective, quantifiable measures of text difficulty and readability and have been validated in prior work [7,8,33].

2.3. Results Experiment 1

2.3.1. RQ1: Do Linguistic Features of Generated Outputs Shift Across Repeated Repetitions?

To examine the stability of linguistic features across repeated generations and differences across models, we analyzed the generated texts using linear mixed-effects models [42] with output text as random factor. The dependent variables were linguistic features extracted using NLP tools, including measures of writing style, cohesion, syntactic complexity, lexical sophistication, and language variety. Across ten repeated generations produced under the same prompting conditions, linguistic features remained stable. No substantial variation was observed across repetitions for any of the extracted indices (all ps > .05). These results suggested that repeated regeneration did not introduce shifts in linguistic properties. Moreover, stability across repetitions was consistent across all four models such that there was no interaction between repetition and model (See Figure 1). These results suggest that under similar prompting conditions, LLM outputs exhibit high internal consistency in linguistic properties across repeated regeneration.

 doi:10.20944/preprints202602.1117.v1

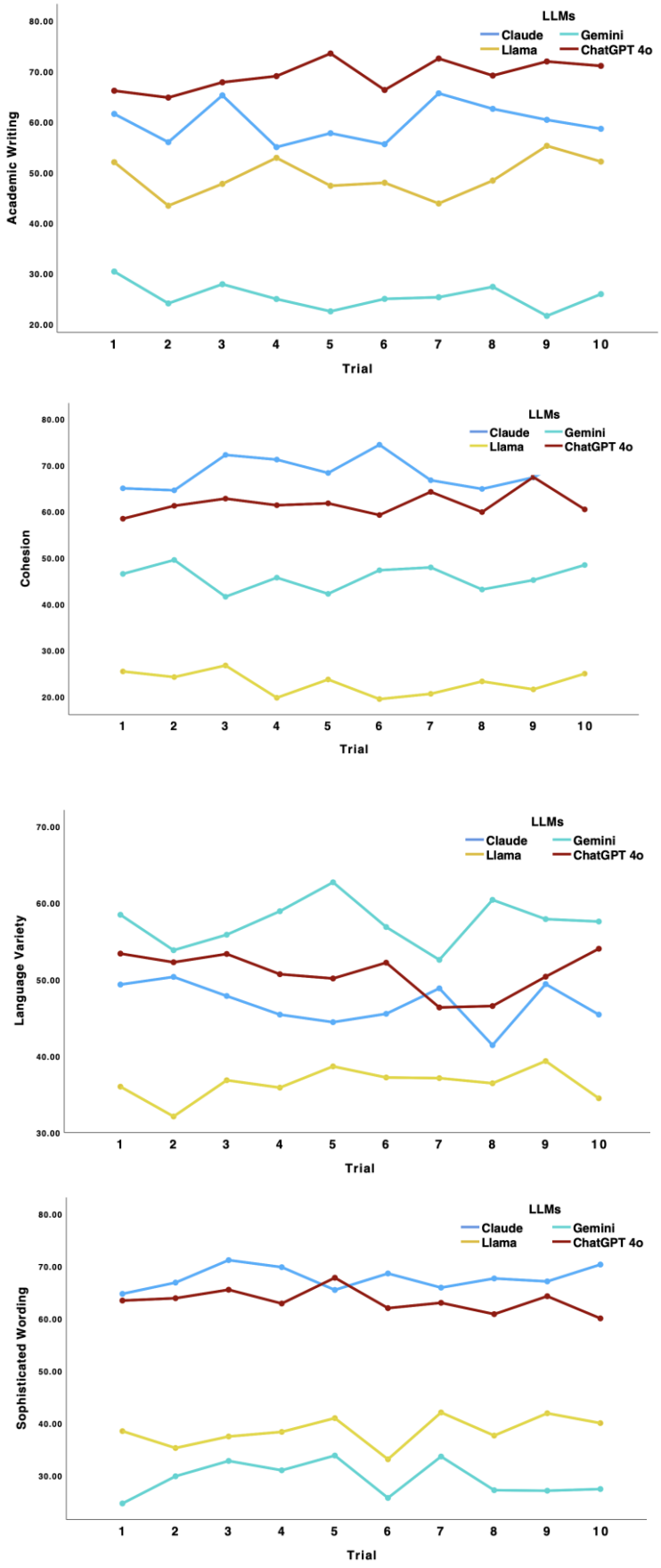

Figure 1. Linguistic features variations across repeated trial and LLMs.

 doi:10.20944/preprints202602.1117.v1

2.3.2. Linguistic Variations in Generated Outputs by Different LLMs

In contrast to the stability observed across repeated regenerations, significant differences emerged across LLMs. The four LLMs produced distinct linguistic profiles across all examined features (see Table 1) which suggested that each model exhibits a characteristic pattern of lexical, syntactic, and discourse-level properties. These results indicated that linguistic variation reflects model-specific linguistic patterns.

Table 1. Descriptive statistics and main effects of LLMs.

Linguistic Features	Claude		Llama		ChatGPT		Gemini		F	p
	M	SD	M	SD	M	SD	M	SD		
Academic Writing	59.98	25.10	52.02	27.76	27.38	19.35	63.98	24.19	8.44	<0.001
Idea Development	64.52	15.54	27.35	14.21	26.32	12.47	79.08	17.21	84.15	<0.001
Sentence Cohesion	13.33	12.95	62.38	24.14	80.24	17.25	37.01	20.55	6.41	<0.001
Noun-to-Verb ratio	6.08	8.29	0.72	1.26	0.71	0.61	2.36	2.24	8.61	<0.001
Sentence Length	11.11	2.86	19.09	4.29	29.12	7.72	17.00	3.25	71.34	<0.001
Language Variety	46.35	24.87	30.78	17.27	53.83	26.00	60.53	20.38	6.76	<0.001
Sophisticated Word	67.86	33.84	40.66	26.42	30.67	16.66	59.46	23.48	9.29	<0.001

To further characterize the nature of these model-specific differences, we conducted a principal component analysis (PCA) on the set of linguistic features. The PCA revealed four underlying dimensions of linguistic variation, together accounting for 85.54% of the total variance (see Table 2). These components provide an interpretable summary of the linguistic dimensions along which LLMs differ, clarifying how model-specific generation strategies manifest in quantifiable linguistic indices.

The first component (accounted for 35.99% of variance) captured variation related to simplified syntax and information density, reflecting shorter sentences with dense conceptual content. The second component (explained 23.61% of variance) corresponded to academic writing style, characterized by sophisticated vocabulary and formal lexical choices. The third component (accounted for 16.09% of variance) reflected discourse-level cohesion and structural density, combining sentence cohesion with high noun-to-verb ratios. The final component (explained 9.85% of variations) associated with language variety, indexing diversity in sentence structures and wording.

Table 2. PCA Results for Linguistic Features.

Linguistic Features	Component 1	Component 2	Component 3	Component 4
Sentence Length	-0.84	-0.28	-0.17	0.20
Idea Development	0.76	0.23	0.16	0.44
Academic Writing	0.34	0.81	-0.15	-0.20
Sophisticated Wording	0.12	0.85	0.16	0.30
Sentence Cohesion	-0.01	0.16	0.89	-0.21
Noun-to-Verb	0.41	-0.17	0.79	0.04
Language Variety	0.01	0.05	-0.18	0.94
Eigenvalue	2.52	1.65	1.13	0.69
Variance Explained	35.99	23.61	16.09	9.85
Cumulative Variance	35.99	59.61	75.70	85.54

 doi:10.20944/preprints202602.1117.v1

2.4. Discussion Experiment 1

There was no significant variability in linguistic features across repetitions. Experiment 1 showed that with similar prompt instruction and testing conditions, linguistic features of LLM-generated outputs remained stable across repeated generations. Across ten regenerations of the same task, there was no significant variation in any of the linguistic features, indicating high internal short-term consistency with the same model version, prompt, and input text. This result suggested that short-term stochastic regeneration does not introduce significant changes in linguistic properties associated with text readability. In contrast to within-model stability, systematic differences were observed across LLMs such that each model exhibited distinct linguistic patterns in their outputs. These differences reflect model-specific generation behavior rather than random variability which further highlights that linguistic variation observed across models is structured and reproducible [7,8].

Despite these findings, a limitation of Experiment 1 is that all outputs were generated within a single day. As these models are subject to ongoing retraining, fine-tuning, and updates, the current research does not fully capture potential long-term performance shifts that may emerge across model versions. Prior work has demonstrated that updates related to training data, architectures, and task optimization can lead to measurable shifts in linguistic patterns over time [17–19]. To address the limitation of Experiment 1, Experiment 2 extends the evaluation across different time points and model versions to examine whether linguistic consistency is preserved across model updates over time.

3. Experiment 2: Linguistic Variability Across Model Versions

3.1. Introduction Experiment 2

A growing body of work has begun to examine model's drift and reproducibility issues across update cycles and deployments. For instance, ChatGPT has shown systematic behavioral changes over a short time span, highlighting how minor updates can alter model responses in ways that complicate reproducibility [17]. More recent work has extended to examine domain-specific tasks in finance and accounting, which shows inconsistency and reproducibility issues when evaluating multiple versions of ChatGPT [16,17]. However, the predominant evaluation paradigm mainly emphasizes task accuracy or end-task metrics rather than the linguistic properties of generated texts [26]. Little research has systematically evaluated how the linguistic properties related to text readability (e.g., lexical sophistication, syntactic complexity, cohesion) evolve across model updates and deployment intervals. The current study addresses this gap by leveraging NLP-based validation framework to assess longitudinal variation of linguistic features. Rather than focusing solely on task accuracy, we extracted interpretable linguistic indices that capture lexical, syntactic, and discourse-level properties and use them to quantify model stability with repeated regenerations (Experiment 1). We then leverage those indices to examine behavioral drifts associated with model updates over time (Experiment 2). This validation approach complements existing research by providing a framework that are sensitive to discourse changes beyond accuracy metrics and surface similarity measures (e.g., BLEU, ROUGE, [31,32]).

In Experiment 2, we investigated linguistic variation across two versions of OpenAI's GPT-4, specifically GPT-4o and GPT-4.1. By examining outputs generated at different deployment intervals (October 2024 and June 2025), we demonstrated how NLP-based metrics can detect behavioral drifts and operationalize longitudinal benchmarking relevant to applied personalization systems. By controlling for the task and prompting instruction constant, this experiment examines whether model updates are associated with systematic shifts in lexical, syntactic, and discourse-level features of generated outputs.

 doi:10.20944/preprints202602.1117.v1

3.2. Method Experiment 2

3.2.1. Model Versions & Deployment Details

This study examined linguistic drift across successive versions of OpenAI's GPT-4 model: ChatGPT 4o introduced in May 2024 (deployed in October 2024), ChatGPT 4o (deployed in June 2025), and GPT-4.1 (deployed in June 2025). Appendix A provides additional technical details, including release dates and accessibility. All three models are built on OpenAI's fourth-generation transformer architecture and are considered comparable in their language capabilities. GPT-4o is a multimodal, latency-optimized version intended for public conversational use, whereas GPT-4.1 is optimized for speed and precision, and can handle long contexts. These versions represent distinct deployment timepoints and update cycles within the same model family, enabling controlled comparison of linguistic properties across model updates.

3.2.2. Task Setup & NLP-Based Evaluation

The generation task involved modifying source texts to align with four distinct reader profiles, each varying in reading proficiency and prior knowledge. Ten science texts and ten history texts were used as inputs. A set of four predefined reader profiles was used to condition generation, varying along dimensions of reading proficiency (high vs. low) and domain-specific prior knowledge (science vs. history). The prompt instructed the LLM to adapt the text to enhance comprehension and engagement while taking into account specific reader needs according to their educational background, reading skill level, prior knowledge, and learning objectives. To isolate linguistic variation attributable to model updates rather than task types or prompt instruction changes, all outputs were generated using the same task specification and prompting structure. The personalized texts were generated at different time points October 2024 and June 2025, using ChatGPT 4o and See Appendix B for more details about the instruction. Each model generated 20 adapted texts per profile (10 science, 10 history), resulting in 80 outputs per model version and 240 total outputs.

Rather than evaluating task accuracy like prior research, we focused the analysis on how these linguistic features systematically differed across model versions. Similar to Experiment 1 procedure, we extracted the linguistic features listed in Appendix D using the Writing Analytics Tool [WAT; [39]]. We applied the same set of NLP features for both science and history passages to examine whether drift patterns were consistent across domains.

3.3. Results Experiment 2

3.3.1. RQ2: Do Linguistic Features in Outputs Shift Due to Model Updates over Time?

To examine whether linguistic properties of generated texts varied across GPT-4 versions and personalization contexts, a two-way MANCOVA (four reader profiles × three model versions), controlling for word count as a covariate.

The main effect of model versions on linguistic features was significant and that linguistic properties differed systematically across GPT-4 versions. Compared to GPT-4o (October 2024), outputs from GPT-4o (June 2025) and GPT-4.1 (June 2025) showed higher alignment with source texts and increased indicators of academic writing and idea development (see Table 3). GPT-4o (June 2025) produced the most cohesive outputs and the lowest language variability, suggesting that new model updates reduced variability but enhanced text cohesion to maintain consistency in generated outputs.

In contrast, GPT-4.1 outputs exhibited higher noun-to-verb ratios, longer sentences, and more sophisticated wording, reflecting a more formal academic writing style. Outputs generated by GPT-4o June 2025 and GPT-4.1 June 2025 were both higher in academic writing metrics compared to outputs generated by Chat GPT-4o Oct 2024 version. Moreover, both GPT-4o June 2025 and GPT-4.1 June 2025 showed significantly enhanced idea development compared to GPT-4o Oct 2024. GPT-4.1 June 2025 outputs contained the highest sophisticated wording compared to both previous model

 doi:10.20944/preprints202602.1117.v1

versions. GPT-4.1 June 2025 also exhibited the highest noun-to-verb ratio which is typical of formal academic writing style.

Table 3. Descriptive statistics and main effects of model versions.

Linguistic Features	ChatGPT 4o Oct 2024		ChatGPT 4o June 2025		ChatGPT 4.1 June 2025				
	M	SD	M	SD	M	SD	F (2, 212)	*p*	η^2
Source Similarity	0.79	0.12	0.90	0.04	0.88	0.06	41.04	<0.001	0.43
Academic Writing	33.72	30.55	61.84	19.72	57.83	22.10	24.83	<0.001	0.32
Idea Development	41.88	26.08	50.15	17.76	58.98	21.15	6.21	0.003	0.10
Language Variety	62.52	24.22	28.32	20.50	48.69	31.19	39.46	<0.001	0.42
Sentence Cohesion	44.27	26.99	65.94	22.28	48.97	25.34	13.44	<0.001	0.20
Noun-to-Verb Ratio	2.07	0.42	2.10	0.35	2.26	0.45	8.39	<0.001	0.14
Sentence Length	19.05	4.80	10.94	2.34	18.22	7.56	54.80	<0.001	0.51
Sophisticated Wording	40.81	33.06	45.35	20.11	53.56	29.14	8.11	<0.001	0.13

3.3.2. RQ3: How Do Newer Models Respond Differently to Personalization Tasks?

There was a significant two-way interaction effect between reader profile and LLM versions, suggesting that model updates impact not only linguistic features of outputs but also how each model generated personalized texts for readers with varying skill and knowledge level.

Earlier versions GPT-4o, October 2024 and GPT-4.1 (June 2025) showed greater differentiation in personalized texts for different reader profiles. Texts modified for advanced readers showed greater language variety, longer sentences, and more sophisticated syntax structures and vocabulary compared to less advanced reader profiles (See Figure 2). Personalized texts intended for low-knowledge readers were more cohesive and less complex. These adaptations align with theoretical expectations for effective personalization such that linguistic complexity increases for readers with greater background knowledge and reading skills [40,41].

However, GPT-4o (June 2025) produced more uniform linguistic patterns across reader profiles, suggesting reduced sensitivity to reader characteristics. Across all reader profiles, GPT-4o (June 2025) produced adaptations with consistently high level of cohesion, similar sentence lengths, lexical and syntax sophistication, and lower language variety. These patterns resulted in less linguistic differentiation between reader profiles, signaling a weaker alignment with intended personalization goals. The adapted texts were not effectively tailored to differences in reader's skill and level of prior knowledge.

In contrast, the newer model GPT-4.1 (June 2025) demonstrated the most effective adaptation performance. This model selectively lowered cohesion, increased sentence length, sophisticated wording and syntax for high-knowledge readers. At the same time, the model effectively lowered text complexity and increased cohesion in modifications intended for low-knowledge readers. These patterns suggest that GPT-4.1 preserved sensitivity to task inputs and generated linguistically differentiated outputs that were more closely aligned with reading comprehension theories.

 doi:10.20944/preprints202602.1117.v1

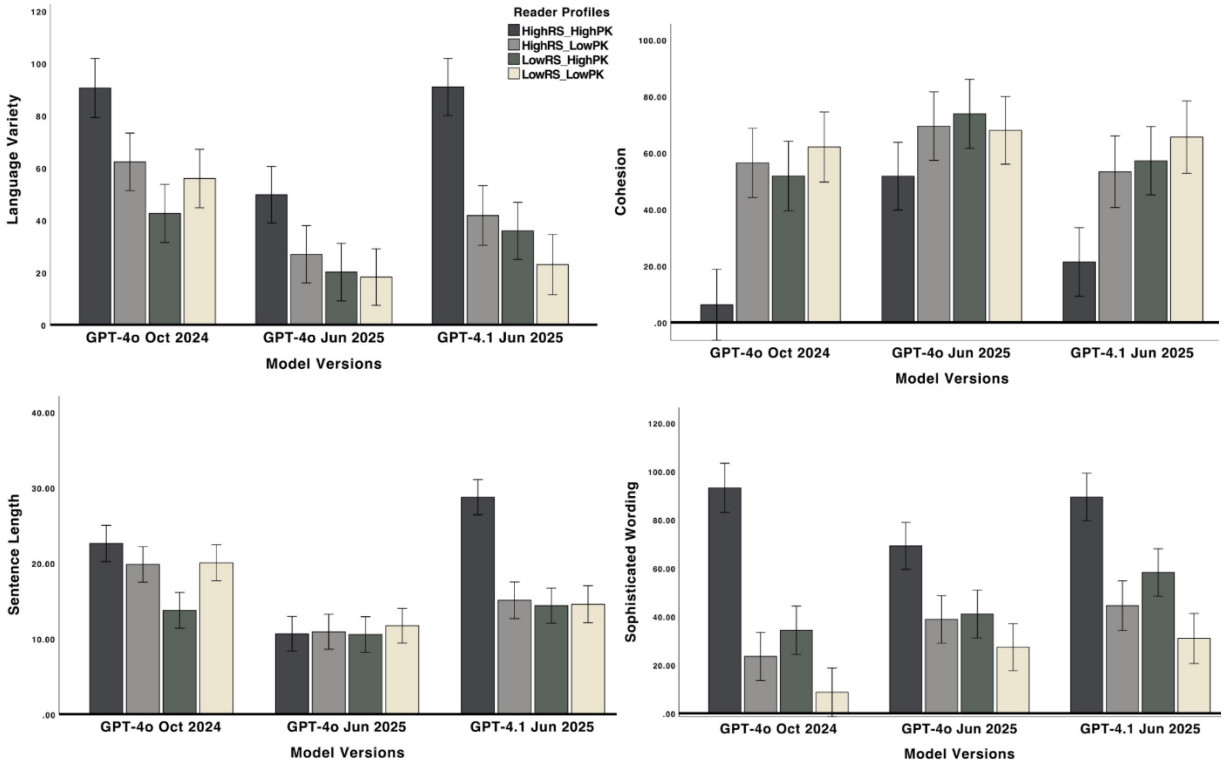

Figure 2. Linguistic complexity patterns across reader profiles shift with GPT-4 model updates. GPT-4o (October 2024) and GPT-4.1 (June 2025) modified texts for skilled readers with greater language variety, more sophisticated vocabulary and syntax, whereas adapted texts for low-knowledge readers demonstrated higher cohesion and reduced complexity. In contrast, GPT-4o (June 2025) generated consistent linguistic patterns across reader profiles, indicating reduced sensitivity in response to cognitive needs of different reader profiles.

4. Discussion

In this paper, we proposed and evaluated an NLP-based framework for tracking linguistic consistency of LLM-generated outputs across repeated regenerations and model updates over time. Experiment 1 established short-term internal reliability within same model such that under controlled prompting instructions, repeated generations produce stable linguistic profiles. Using metrics extracted by the NLP tool [i.e., WAT, [39]], we analyzed features related to text complexity such as cohesion, syntax and lexical sophistication, and language variety in outputs. There were no significant shifts in linguistic measures of outputs related to text complexity. This consistency in linguistic patterns highlighted LLMs' reliability and robustness for producing linguistically stable content across short timeframe [26,43]. Additionally, findings from Experiment 1 also demonstrated variations in linguistic patterns across different LLMs. There were consistent differences across models that revealed structured, model-specific generation behaviors, validating the sensitivity of the NLP-based evaluation framework [44].

In Experiment 2, we demonstrated that linguistic properties of outputs shift across GPT-4 model updates: GPT-4o deployed in October 2024, in June 2025, and GPT-4.1 deployed in June 2025. Specifically, GPT-4.1 June 2025 produced modifications with highest lexical sophistication, containing long sentences with dense noun-to-verb ratios compared to both versions of GPT-4o. These linguistic features are common in advanced academic writing [45,46]. While GPT-4.1 outputs were formal and rigorous, GPT-4o updates over time showed drifts toward more natural and conversational writing style. These findings are consistent with prior research suggesting that model architecture and fine-tuning influence lexical diversity and complexity [47,48]. These changes in linguistic patterns can be attributed to model-specific optimization strategies which align with GPT-4.1's goals to be used for formal contexts as marketed by OpenAI [49]. GPT-4.1 prioritized technical and complex tasks that demand advanced reasoning and rigor outputs. In contrast, updates in GPT-

 doi:10.20944/preprints202602.1117.v1

4o outputs are tailored toward enhancing readability and conversational style, indicated by less complex word use and shorter sentence structures.

Moreover, these findings suggested that model drift affects not only linguistic style but also sensitivity to task inputs. Using the NLP-based evaluation framework to assess linguistic features, we found that NLP indices sensitively capture personalization performance of each GPT-4 model versions. GPT-4o (June 2025) generated outputs that were more for all four readers, this uniformity signaled weaker alignment with intended personalization goals as there was no differences in linguistic profiles of personalized content. In contrast, GPT-4.1 maintained differentiated responses aligned with reader characteristics. As expected from reading comprehension theories, texts modified for skilled and knowledgeable readers exhibited increased syntactic complexity, advanced and diverse vocabulary, and reduced cohesion. These linguistic adaptations align with cognitive theories that skilled readers benefit from challenging and less cohesive texts [40,41]. Conversely, texts tailored for readers with lower reading skills and prior knowledge demonstrated simplified vocabulary and syntax, more cohesive, aligning with best practices in text simplification to enhance readability for less skilled readers [33,36].

Limitations and Future Directions

While the current study contributes to the literature by validating an NLP-based framework for detecting linguistic consistency and drifts in LLM-generated outputs, several considerations should be noted. First, the NLP-based evaluation framework is intended to be a diagnostic tool rather than causal attribution for model drifts over time. The goal of this research is providing a valid method for longitudinal monitoring, identifying whether and how linguistic properties of outputs change across model versions and deployment intervals. Prior research has repeatedly demonstrated LLM performance can shift across model updates due to change in training data, alignment objectives, and fine-tuning strategies [50,51]. Given these expectations that models are subject to change, the challenge is not so much about explaining why these drifts occur but to detect and monitor its effects on model outputs. As such, the current study presents an NLP-based validation framework aimed to focus only on tracking the observable linguistic patterns rather than explaining the internal model changes. By leveraging NLP-based analyses grounded in theories of text difficulty, the framework provides sensitive, interpretable signals of how model outputs change over time. These metrics allow detection of nuanced shifts in lexical use, syntactic structure, and discourse-level cohesion that cannot be captured by task accuracy or surface similarity measures [52]. NLP-based validation framework offers a practical evaluation tool for monitoring and benchmarking, enabling valid assessment of stability and consistency of LLM-generated text in applied contexts such as educational text personalization.

Second, the proposed validation framework aims to assess a critical aspect of personalization which is linguistic properties of LLM-generated outputs rather than comprehensive assessment of system-level personalization quality. Prior research has shown that such linguistic features are closely related to text readability and comprehension processes [33,36]. However, personalization quality encompasses additional properties beyond linguistic adaptation such as materials selection, pacing, content sequencing, feedback timing, and personalized scaffolding [35,53,54]. Therefore, linguistic alignment is a necessary but not sufficient condition for effective personalization, user experience, or pedagogical value. Moreover, hallucinations, which refer to the generation of plausible yet incorrect information, remain difficult to detect using linguistic metrics alone and require complementary evaluation approaches [55–58]. Human-in-the-loop evaluation remains essential for assessing other aspects of personalization quality such as learner perceptions, usability, and instructional impact [59–61]. Extending linguistically grounded monitoring to other components of adaptive systems represents an important direction for future research.

Finally, the evaluation relied on a limited set of predefined simulated reader profiles as a proof-of-concept. While these profiles were designed to represent theoretically meaningful contrasts in reading skill and domain-specific prior knowledge, real-world users exhibit a richer and more diverse

 doi:10.20944/preprints202602.1117.v1

abilities and characteristics [Merino-Campos, 2025; Sharma et al., 2025]. These factors were not fully captured with the current profile descriptions. Extending the framework to more dynamic and data-driven user representations is essential.

5. Conclusions

This work highlights the value of NLP-based linguistic evaluation for longitudinal benchmarking of LLM behavior. In this study, we proposed and applied the NLP-based analysis framework to track drifts in linguistic features of LLM-generated outputs across repeated generations and model updates. By capturing fine-grained lexical, syntactic, and discourse-level properties of generated text, the framework provides sensitive indicators of how models respond to personalization task instructions and how those responses change over time. During the preparation of this manuscript, both the GPT-4o and GPT-4.1 models examined in this study were scheduled for retirement in order to focus more on the development of their current flagship model GPT-5.2 [62]. New models are going to be continuously updated and retired, which further underscores the necessity of a sensitive and reliable evaluation frameworks [63]. The contribution of this work lies not in the analysis of performance of specific model versions, but in presenting an NLP-based validation methodology capable of detecting linguistic drift across update cycles. As models are continuously updated and replaced, longitudinal monitoring becomes critical for ensuring reproducibility and reliability.

The present NLP-based validation framework is designed to support scalable, automated evaluation of linguistic behavior that can flag meaningful linguistic changes in model outputs and guide when more resource-intensive human evaluation is warranted. Unlike task accuracy metrics or lexical overlap measures, NLP-based metrics reveal how drift manifests in interpretable linguistic indices that affect the readability of outputs generated by LLMs. Such rigorous evaluation methods are critical for trustworthy deployment of LLMs in applied settings, where models are continuously updated and content personalization requires rapid validation to support iterative refinements.

Author Contributions: Conceptualization, L.H. and D.S.M.; methodology, L.H. and D.S.M.; formal analysis, L.H.; investigation, L.H.; resources, D.S.M.; data curation, L.H.; writing—original draft preparation, L.H.; writing—review and editing, L.H. and D.S.M.; visualization, L.H.; supervision, D.S.M; project administration, D.S.M; funding acquisition, D.S.M. All authors have read and agreed to the published version of the manuscript.

Institutional Review Board Statement: Not applicable

Informed Consent Statement: Not applicable

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.

Data Availability Statement: The data presented in this study are available in the Open Science Framework at: https://osf.io/49z2s/overview?view_only=d2a0594f4ce44c21a9d9cdb68655eda1

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

Abbreviations

The following abbreviations are used in this manuscript:

PK	Prior knowledge
RS	Reading skills
GenAI	Generative AI
NLP	Natural Language Processing
LLM	Large Language Model
WAT	Writing Analytics Tool

 doi:10.20944/preprints202602.1117.v1

Appendix A

This appendix includes technical details on the LLMs used in the current research, including model versions, training size, and number of parameters. All model was accessed through the Poe.com web platform using default configurations on 10 June 2025.

Table A. Technical details of each LLM. Source: authors' contribution.

Model	Owned by	Release Date	Capabilities	Context Capacities
ChatGPT-4o	OpenAI	May 2024	Better reasoning vs earlier GPT models with stronger multimodal understanding and generative abilities	Not publicly disclosed, but estimate smaller than 1 million tokens
ChatGPT-4.1	OpenAI	April 2025	Faster and more cost-efficient; benchmarks show reduced latency and improved inference vs older models in several tasks Designed for better instruction-following, coding, and long-context comprehension	Not publicly disclosed, but estimate up to 1 million tokens
Claude 3.5	Anthropic	June 2024	Strong reasoning and instruction-following, optimized for analytical writing, summarization, and code understanding with high factual coherence	Up to 200,000 tokens
Llama 3.1	Meta	July 2024	Open-weight LLM optimized for multilingual generation, reasoning, and instruction-tuned tasks, used widely for research and deployment	Up to 128,000 tokens
Gemini Pro 1.5	Google DeepMind	February 2024	Advanced multimodal reasoning with strong long-context performance, effective for document understanding, summarization, and cross-modal tasks	Up to 1 million tokens

Appendix B

This appendix presents detailed prompt that were used to instruct LLM.

Imagine you are a cognitive scientist specializing in reading comprehension and learning science. Your task is to modify this text to enhance text comprehension, engagement, and accessibility for the reader profile while maintaining conceptual depth, scientific rigor, and pedagogical value.

Your goal is to tailor 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. Explain the rationale behind each modification approach and how each change helps the reader grasp the scientific concepts and retain information. Here is information about the readers that you need to tailor this text for:

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 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

[Input scientific text excerpt]

 doi:10.20944/preprints202602.1117.v1

Appendix C

This appendix includes 10 scientific texts. The texts are accessible to users who create an account and can be found under the "Texts Library" section. All materials are licensed under the Creative Commons Attribution 4.0 International License (CC BY 4.0) (https://creativecommons.org/licenses/by/4.0/).

Table C. scientific 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

* Flesch–Kincaid Grade Level.

Appendix D

This appendix presents several linguistic features related to text readability, which is relevant to the assessment of text personalization quality. These indices were extracted using the Writing Analytics Tool [39] and validated in prior work [7,8].

Table D. Linguistic features related to reading comprehension. Source: authors' contribution.

Features	Metrics and Descriptions
Writing Style	**Academic writing:** The extent to which texts include academic wordings and sophisticated sentence structures, typical properties of scientific texts
Conceptual Density	**Development of ideas:** The extent to which ideas and concepts are elaborated in a text, also indicates complex sentences requiring cognitive effort to comprehend **Noun-to-verb ratio:** High noun-to-verb ratio indicates densely packed information in a text **Discourse-level cohesion:** The extent to which the text contains connectives and cohesion cues (e.g., repeating ideas and concepts)
Syntax Structure	**Sentence length:** Longer sentences indicate more complex syntax structure **Language variety:** The extent to which the text contains varied lexical and syntax structures
Lexical Features	**Complex vocabulary:** Lower measures indicate texts contain wordings that are familiar and easy to be recognized. In contrast, high measures indicate texts contain advanced vocabulary

References

 doi:10.20944/preprints202602.1117.v1

1. Ahmed, A.; Aziz, S.; Abd-Alrazaq, A.; AlSaad, R.; Sheikh, J. Leveraging LLMs and wearables to provide personalized recommendations for enhancing student well-being and academic performance through a proof of concept. *Sci. Rep.* **2025**, *15*, 4591.
2. Bhattacharjee, A.; Zeng, Y.; Xu, S.Y.; Kulzhabayeva, D.; Ma, M.; Kornfield, R.; et al. Understanding the role of large language models in personalizing and scaffolding strategies to combat academic procrastination. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, Honolulu, HI, USA, 11–16 May 2024; pp. 1–18.
3. Huang, W.; Abbeel, P.; Pathak, D.; Mordatch, I. Language Models as Zero-Shot Planners: Extracting Actionable Knowledge for Embodied Agents. *arXiv* **2022**, arXiv:2201.07207.
4. Merino-Campos, C. The impact of artificial intelligence on personalized learning in higher education: A systematic review. *Trends High. Educ.* **2025**, *4*, 17.
5. Sharma, S.; Mittal, P.; Kumar, M.; Bhardwaj, V. The role of large language models in personalized learning: A systematic review of educational impact. *Discover Sustain.* **2025**, *6*, 1–24.
6. Kintsch, W. Revisiting the construction–integration model of text comprehension and its implications for instruction. In *Theoretical Models and Processes of Literacy*; Routledge: New York, NY, USA, 2018; pp. 178–203.
7. Huynh, L.; McNamara, D.S. GenAI-Powered Text Personalization: Natural Language Processing Validation of Adaptation Capabilities. *Appl. Sci.* **2025**, *15*, 6791.
8. Huynh, L.; McNamara, D.S. Natural Language Processing as a Scalable Method for Evaluating Educational Text Personalization by LLMs. *Appl. Sci.* **2025**, *15*, 12128.
9. Peláez-Sánchez, I.C.; Velarde-Camaqui, D.; Glasserman-Morales, L.D. The impact of large language models on higher education: Exploring the connection between AI and Education 4.0. *Front. Educ.* **2024**, *9*, 1392091.
10. Yang, Q.; Liang, C. A Second-Classroom Personalized Learning Path Recommendation System Based on Large Language Model Technology. *Appl. Sci.* **2025**, *15*, 7655.
11. Raj, H.; Rosati, D.; Majumdar, S. Measuring Reliability of Large Language Models through Semantic Consistency. *arXiv* **2023**, arXiv:2211.05853.
12. Srivastava, A.; et al. Beyond the Imitation Game: Quantifying and Extrapolating the Capabilities of Language Models. *arXiv* **2022**, arXiv:2206.04615.
13. 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* **2023**, arXiv:2308.07847.
14. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language models are unsupervised multitask learners. *OpenAI Blog* **2019**, *1*, 9.
15. Atil, B.; Chittams, A.; Fu, L.; Ture, F.; Xu, L.; Baldwin, B. LLM Stability: A detailed analysis with some surprises. *arXiv* **2024**, arXiv:2408 (submitted).
16. Wang, J.J.; Wang, V.X. Assessing consistency and reproducibility in the outputs of large language models: Evidence across diverse finance and accounting tasks. *arXiv* **2025**, arXiv:2503.16974.
17. Chen, L.; Zaharia, M.; Zou, J. How is ChatGPT's behavior changing over time?. *Harvard Data Sci. Rev.* **2024**, *6*(2).
18. Tu, S.; Li, C.; Yu, J.; Wang, X.; Hou, L.; Li, J. ChatLog: Carefully evaluating the evolution of ChatGPT across time. *arXiv* **2023**, arXiv:2304.14106.
19. Park, C.; Kim, H. Understanding LLM development through longitudinal study: Insights from the Open Ko-LLM leaderboard. *arXiv* **2024**, arXiv:2409.03257.
20. Anisuzzaman, D.M.; Malins, J.G.; Friedman, P.A.; Attia, Z.I. Fine-tuning large language models for specialized use cases. *Mayo Clin. Proc. Digit. Health* **2025**, *3*, 100184.
21. Durrani, N.; Sajjad, H.; Dalvi, F. How transfer learning impacts linguistic knowledge in deep NLP models?. *arXiv***2021**, arXiv:2105.15179.
22. Lu, W.; Luu, R.K.; Buehler, M.J. Fine-tuning large language models for domain adaptation: Exploration of training strategies, scaling, model merging and synergistic capabilities. *npj Comput. Mater.* **2025**, *11*, 84.
23. Mosbach, M.; Khokhlova, A.; Hedderich, M.A.; Klakow, D. On the interplay between fine-tuning and sentence-level probing for linguistic knowledge in pre-trained transformers. *arXiv* **2020**, arXiv:2010.02616.

 doi:10.20944/preprints202602.1117.v1

24. Luo, Z.; Xie, Q.; Ananiadou, S. Factual consistency evaluation of summarization in the era of large language models. *Expert Syst. Appl.* **2024**, *254*, 124456.
25. Cui, W.; Zhang, J.; Li, Z.; Damien, L.; Das, K.; Malin, B.; Kumar, S. DCR-Consistency: Divide–Conquer–Reasoning for Consistency Evaluation and Improvement of Large Language Models. *arXiv* **2024**, arXiv:2401.02132.
26. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.
27. Sahoo, P.; Singh, A.K.; Saha, S.; Jain, V.; Mondal, S.; Chadha, A. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv* **2024**, arXiv:2402.07927.
28. Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; et al. Chain-of-thought prompting elicits reasoning in large language models. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 24824–24837.
29. Warr, M.; Pivovarova, M.; Mishra, P.; Oster, N.J. Is ChatGPT Racially Biased? The Case of Evaluating Student Writing. In *The Case of Evaluating Student Writing*; Elsevier: Amsterdam, The Netherlands, **2024**.
30. Banerjee, S.; Lavie, A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, Ann Arbor, MI, USA, June 2005; pp. 65–72.
31. Lin, C.-Y. ROUGE: A package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out*, Barcelona, Spain, July 2004; pp. 74–81.
32. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, PA, USA, July 2002; pp. 311–318.
33. Crossley, S.A. Developing linguistic constructs of text readability using natural language processing. *Sci. Stud. Read.* **2025**, *29*, 138–160.
34. Crossley, S.; Salsbury, T.; McNamara, D. Measuring L2 lexical growth using hypernymic relationships. *Lang. Learn.* **2009**, *59*, 307–334.
35. Tetzlaff, L.; Schmiedek, F.; Brod, G. Developing personalized education: A dynamic framework. *Educ. Psychol. Rev.* **2021**, *33*, 863–882.
36. 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.
37. 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.
38. Kintsch, W. The role of knowledge in discourse comprehension: A construction–integration model. *Psychological Review.* **1988**, *95*, 163–182.
39. Potter, A.; Shortt, M.; Goldshtein, M.; Roscoe, R. D. *Assessing academic language in tenth-grade essays using natural language processing. Assess Writing.* **2025**, *64*, 100921.
40. Ozuru, Y.; Dempsey, K.; McNamara, D.S. Prior knowledge, reading skill, and text cohesion in the comprehension of science texts. *Learn. Instr.* **2009**, *19*, 228–242.
41. 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*, 121–152.
42. Baayen, R.H.; Davidson, D.J.; Bates, D.M. Mixed-effects modeling with crossed random effects for subjects and items. *J. Mem. Lang.* **2008**, *59*, 390–412.
43. Bommasani, R.; Hudson, D.A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; et al. On the opportunities and risks of foundation models. *arXiv* **2021**, arXiv:2108.07258.
44. Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F.L.; et al. GPT-4 Technical Report. *arXiv***2023**, arXiv:2303.08774.
45. Biber, D.; Gray, B. Grammatical Complexity in Academic English: Linguistic Change in Writing, 1st ed.; Cambridge University Press: Cambridge, UK, **2016**.
46. Gardner, D.; Davies, M. A new academic vocabulary list. *Appl. Linguist.* **2014**, *35*, 305–327.

 doi:10.20944/preprints202602.1117.v1

47. Reviriego Vasallo, P.; Conde Díaz, J.; Merino-Gómez, E.; Martínez Ruiz, G.; Hernández Gutiérrez, J.A. Playing with Words: Comparing the Vocabulary and Lexical Diversity of ChatGPT and Humans. *Mach. Learn. Appl.* **2024**, *18*, 100602.
48. Gupta, V.; Chowdhury, S.P.; Zouhar, V.; Rooein, D.; Sachan, M. Multilingual performance biases of large language models in education. *arXiv* **2025**, arXiv:2504.17720.
49. OpenAI. Introducing GPT-4.1 in the API. Available online: https://openai.com/index/gpt-4-1/ (accessed on 14 April 2025).
50. Mosbach, M.; Andriushchenko, M.; Klakow, D. On the stability of fine-tuning BERT: Misconceptions, explanations, and strong baselines. *arXiv* **2020**, arXiv:2006.04884.
51. Merchant, A.; Rahimtoroghi, E.; Pavlick, E.; Tenney, I. What happens to BERT embeddings during fine-tuning?. *arXiv* **2020**, arXiv:2004.14448.
52. Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; et al. Survey of hallucination in natural language generation. *ACM Comput. Surv.* **2023**, *55*, 1–38.
53. Koedinger, K.R.; et al. Predictive Analytics in Education. *Int. J. Learn. Analytics & AI in Educ.* **2013**.
54. Aleven, V.; McLaren, B.M.; Sewall, J.; van Velsen, M.; Demi, S. Embedding Intelligent Tutoring Systems in MOOCs and e-learning Platforms. In *Proceedings of IT*, 2016; Springer.
55. Alkaissi, H.; McFarlane, S.I. Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. *Cureus* **2023**, *15*, e35179.
56. Hatem, R.; Simmons, B.; Thornton, J.E. A call to address AI "hallucinations" and how healthcare professionals can mitigate their risks. *Cureus* **2023**, *15*.
57. Wang, Y.; et al. Factuality of Large Language Models: A Survey. *arXiv* **2024**, arXiv:2402.02420.
58. Maleki, N.; Padmanabhan, B.; Dutta, K. AI hallucinations: a misnomer worth clarifying. In *Proceedings of the 2024 IEEE Conference on Artificial Intelligence (CAI)*, Singapore, 24–26 June 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 133–138.
59. Amershi, S.; et al. Guidelines for Human-AI Interaction. *CHI / Microsoft Research Technical Report* **2019**.
60. Mitchell, M.; et al. Model Cards for Model Reporting. In *Proceedings* **2019**.
61. Holstein, K.; et al. (2019). Design / Human-in-the-loop for Education — teacher-AI complementarity (multiple proceedings/examples). **2019**.
62. OpenAI, Retiring GPT-4o, GPT-4.1, GPT-4.1 mini, and OpenAI o4-mini in ChatGPT, OpenAI, Jan. 29, 2026. Available online: https://openai.com/index/retiring-gpt-4o-and-older-models/ (accessed on 11 February 2026).
63. 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* **2023**, arXiv:2305.14540.