NLP Validation of Prompt Strategies for Theory-Aligned LLM-Generated Personalization

Linh Huynh & Danielle S. McNamara

Large Language Models
natural language processing
Personalized Learning

This study applies an NLP-based validation framework to examine how Large Language Models (LLMs) can be iteratively refined for theory-aligned text personalization. Building on prior work, we extend the evaluation method to history texts and focus on prompt design as a key factor in personalization quality. Four LLMs (Claude, Llama, Gemini, ChatGPT-4) were prompted to adapt ten history passages for four reader profiles varying in reading skill and prior knowledge using one-shot and task-specific instruction prompts. Linguistic indices were extracted using the Writing Analytics Tool to assess the alignment of linguistic features with students' needs. Although LLMs appropriately tailored text complexity, cohesion patterns failed to match theoretical expectations even under explicit guidance. This iteration highlights the limits of current prompting strategies and the importance of theory-augmented refinement. Through iterative prompt evaluation, the study demonstrates how NLP provides a scalable, real-time framework for validating and improving theory-driven personalization across multiple domains.

Linh Huynh1 and Danielle S. McNamara2

1 Learning Engineering Institute, Arizona State University; lthuynh1@asu.edu

2 Learning Engineering Institute, Arizona State University; dsmcnama@asu.edu

Abstract.

Keywords: Large Language Models; Natural Language Processing; Personalized Learning.

Introduction

Personalized learning refers to tailoring materials to students' unique abilities and needs which enables learners to engage with content that is most relevant, meaningful, and effective to them (Pesovski et al., 2024). Building on this idea, generative AI (GenAI) has the potential to advance personalized learning by dynamically modifying materials to match students' abilities and learning goals. In particular, Large Language Models (LLMs) have also been widely applied in education to adapt writing style, simplify content, and generate materials that are aligned with individual learners (Abbes et al., 2024; Martínez et al., 2024). Providing students with appropriately challenging materials enhances comprehension, engagement, and vocabulary acquisition while supporting skill development and academic growth (Dong et al., 2025; Leong et al., 2024). However, a key

challenge is determining the extent that these personalizations align with learning theory and students' needs. Effective text personalization therefore requires continuous validation to ensure that LLM-generated texts adapt to students' evolving skills and learning needs. As adaptive learning technologies expand, scalable and theory-aligned validation frameworks are needed to evaluate personalization quality in real time and at scale.

Natural Language Processing (NLP) provides a scalable method for evaluating the extent to which personalized texts align with readers' cognitive needs by examining linguistic features. As a computational method, NLP quantifies text difficulty and readability based on linguistic features that influence comprehension (McNamara, 2021; Ozuru et al., 2009). Building on comprehension theory, text readability refers to how the impact of textual features depends on reader characteristics and can be assessed using indices such as cohesion, syntactic complexity, and lexical sophistication (Crossley et al., 2017). While surface-level features (e.g., sentence length, vocabulary, and syntax) influence the ease of processing, discourse-level features (e.g., cohesion) determine how ideas are connected to support the construction of a coherent mental model and learning from text. These linguistic features correspond to comprehension processes and indicate how well texts align with readers' cognitive needs (Ozuru et al., 2009). For instance, domain-specific vocabulary and complex syntax require readers to possess sufficient background knowledge and reading skills to infer meaning (Frantz et al., 2015; Nagy & Townsend, 2012). Cohesion features play an important role in comprehension but they differentially affect understanding depending on prior knowledge of the reader. While cohesive texts support low-knowledge readers by bridging conceptual gaps, low-cohesion texts are more beneficial for high-knowledge readers by promoting inference generation and deeper processing (O'Reilly & McNamara, 2007). These findings demonstrate how NLP operationalizes comprehension theory by analyzing measurable linguistic features that predict how effectively adapted texts support comprehension for each reader.

Prior work by Huynh and McNamara (2025a) introduced the NLP validation framework as a theory-aligned method for evaluating LLM-generated personalization, later extending it across domains to test generalizability (Huynh & McNamara, 2025b). Building on prior work, the current study extends the NLP validation framework to examine how different prompt instructions influence the quality of LLM-generated adaptations. Because prompting strategies influence how models adapt linguistic features to different readers (He et al., 2024; Ye et al., 2024), NLP analyses were used to evaluate how well these adaptations align with comprehension theory and readers' cognitive needs. The NLP analysis provides quantitative feedback that identifies where outputs diverge from theoretical expectations and informs prompt refinement. In doing so, this work demonstrates how NLP-based evaluation can be integrated into adaptive learning systems as part of an iterative design cycle, advancing learning engineering by supporting continuous, theory-informed improvement in AI-driven text personalization.

This work is situated within a learning engineering framework that emphasizes iterative, evidence-based improvement of learning systems through tightly coupled cycles of design, implementation, measurement, and refinement. Learning engineering views educational technologies not as static artifacts but as adaptive systems that must be continuously evaluated and improved using theoretically grounded evidence (Baker et al., 2022; Goodell & Kolodner, 2023). In this context, LLM-driven text personalization represents a complex learning system component whose behavior must be systematically aligned with learning theory to support comprehension across diverse learners. The present study operationalizes a nested learning engineering cycle within this larger personalization system by focusing specifically on the prompt design and evaluation layers. Comprehension theory informs prompt construction, LLM-generated adaptations serve as design instantiations, and NLP-based linguistic analyses function as embedded measurement mechanisms. The resulting evidence is used to diagnose theory misalignment and guide subsequent refinement. By embedding theory-driven validation directly into the design loop, this study demonstrates how learning engineering principles can be applied to iteratively improve AI-mediated personalization at scale.

Method

Four LLMs Claude 3.5 Sonnet (Anthropic), Llama (Meta), Gemini Pro 1.5 (Google), and ChatGPT 4 (OpenAI) were prompted to modify 10 historical texts with topics including American History and World History. The texts were selected and compiled

from the iSTART website (www.adaptiveliteracy.com/istart) (McNamara, 2004).

Four hypothetical reader profiles were developed to simulate combinations of high versus low reading skill (RS) and prior knowledge (PK) in history domain: (1) high RS/high PK, (2) high RS/low PK, (3) low RS/high PK, and (4) low RS/low PK. These profiles served as simplified learner models to test the adaptability of LLM-generated texts through evaluating whether linguistic adaptation patterns followed theoretical expectations.

The LLMs were prompted using two different prompting techniques. The one-shot prompt provided only an example demonstrating the desired adaptation, the reader characteristics and input text. In contrast, the task-specific prompt provided explicit task goals and instruction for modifying texts to suit reader characteristics. For example:

Modify this text to improve comprehension and engagement for a reader. The goal of this personalization is to tailor a reading experience that aligns with the unique characteristics of the reader (age, background knowledge, reading skills, reading goal and preferences, interests) while maintaining content coverage and important concepts from the original text. Analyze the input text and reader profiles, identify key information about the readers, then modify the syntax, vocabulary, and tone to suit the reader's characteristics.

These prompts were designed to evaluate how the level of instructional detail affects the quality of personalization and its alignment. The aim of this analysis is not to examine comprehension outcomes but to diagnose how prompt structures influence linguistic alignment within a scalable personalization system. Prior research has shown that specific, structured prompts can improve task performance such that task-specific prompts often elicit more complex and precise outputs due to clearer goal alignment (Brown et al., 2020; Sahoo et al., 2024). Huynh and McNamara (2025a) also showed that augmenting prompts using Retrieval-Augmented Generation (RAG) informed by comprehension theory can enhance linguistic alignment with reader needs. As such, we compared prompt designs to identify weaknesses in existing prompting techniques and guide future refinement toward theory-aligned personalization.

After generating the adaptations using LLMs, we used the Writing Analytics Tool (WAT; Potter et al., 2025) to extract and analyze linguistic features related to text readability (e.g., academic writing, sentence length, noun-to-verb ratios, cohesion, lexical sophistication, academic wording). These metrics, validated in Huynh and McNamara (2025a, 2025b), were used to assess the extent that each adaptation aligned with comprehension theory predictions. Specifically, texts for advanced readers were expected to have lower cohesion and higher complexity, whereas texts for low-knowledge readers were expected to exhibit simpler syntax, concrete vocabulary, and higher cohesion.

Results

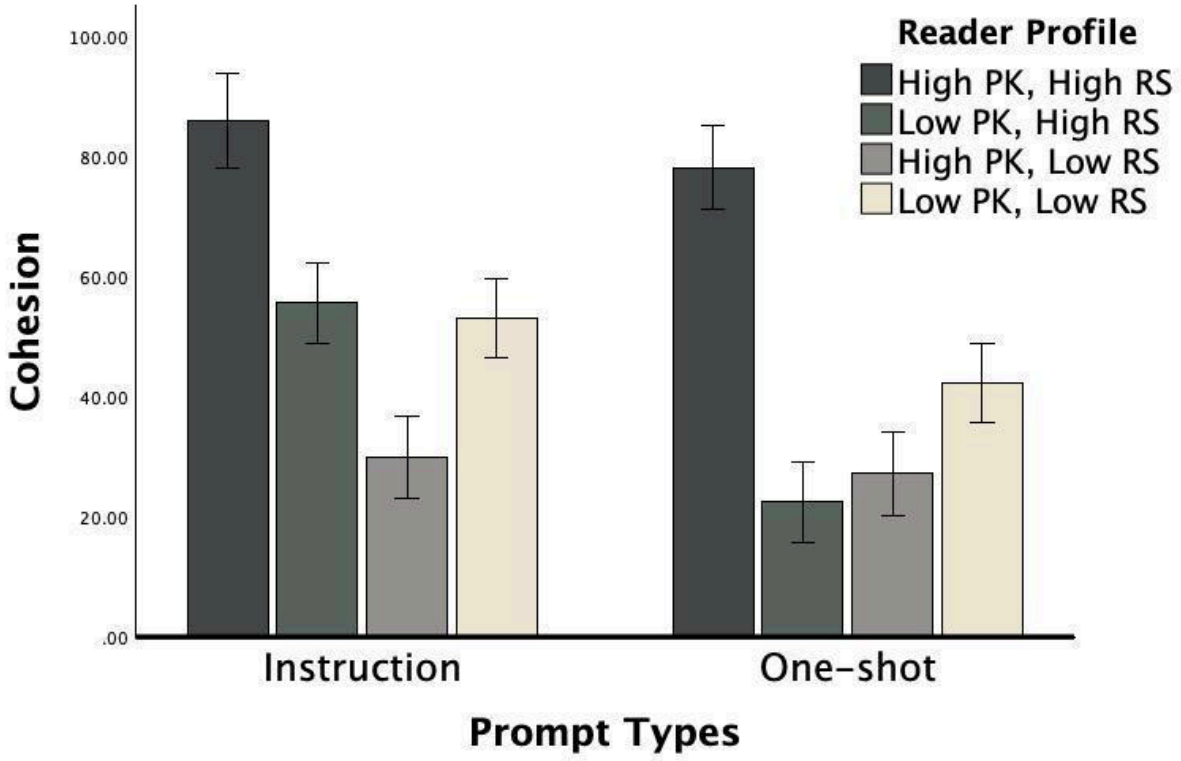
A 4 (Reader Profile: High RS/High PK, High RS/Low PK, Low RS/High PK, Low RS/Low PK) × 4 (LLM: ChatGPT-4, Gemini Pro 1.5, Claude 3.5 Sonnet, Llama 3.1) MANCOVA, controlling for word count, examined how linguistic features of LLM-modified history texts varied across reader profiles and prompting types.

Consistent with prior findings (Huynh & McNamara, 2025a; 2025b), LLMs systematically adjusted surface-level linguistic features to match reader ability. Texts adapted for skilled, high-knowledge readers (Profile 1) featured the highest syntactic and lexical sophistication, using academic writing style. Adaptations for less skilled or low-knowledge readers (Profiles 2, 3, and 4) featured shorter sentences, simpler vocabulary, and more concrete wording, indicating that LLMs successfully tailored text complexity to suit the readers' needs. These adjustments confirm that LLMs can scale personalization.

In contrast to the theory-aligned patterns observed for surface-level features (e.g., shorter sentences, simpler vocabulary, and more concrete wording for less skilled and low-knowledge readers), cohesion patterns did not align with comprehension theory (see Figure 1). Prior research suggests that low-knowledge readers benefit from cohesive texts, whereas high-

knowledge readers can learn more effectively from less cohesive texts (Ozuru et al., 2009). Contrary to this expectation, texts generated for lower-knowledge readers were less cohesive than those tailored for higher-knowledge readers. This reversal effect indicates that while LLMs can tailor surface features, they lack sensitivity to readers' discourse needs due to limited theoretical and contextual understanding. It is critical to provide additional scaffolding such as RAG-augmented technique to enhance LLM's adaptation.

Figure 1. Cohesion features as a function of reader profile and prompt types.



Discussion

LLMs systematically adjusted linguistic features to match reader characteristics, producing modifications that were generally appropriate for readers' needs. Texts for high-skill, high-knowledge readers exhibited greater syntactic and lexical complexity and a more academic writing style, whereas adaptations for low-knowledge or less skilled readers simplified sentence structures and vocabulary to enhance readability. These patterns confirm that LLMs can tailor surface-level complexity effectively and that NLP reliably quantifies these differences. However, cohesion patterns diverged from theoretical expectations. Although task-specific prompts improved linguistic sophistication, neither prompt type successfully aligned cohesion with comprehension theory. As a result, personalized texts seem to be more readable but they are not aligned with the readers' background knowledge in ways that are predicted to support comprehension. This mismatch highlights the need for knowledge-augmented prompting that integrates theoretical principles directly into prompt design. Embedding comprehension theory could enable adaptive systems to control both surface and discourse-level features more effectively (Huynh & McNamara, 2025a).

From a learning engineering perspective, this study represents a focused sub-cycle embedded within a broader adaptive personalization pipeline. The learning problem—ensuring that LLM-generated adaptations align with comprehension theory across reader profiles—motivated the design of theory-informed prompts. Implementation occurred through LLM-based text

generation, while NLP-based linguistic indices provided fine-grained, scalable measurements of alignment at both surface and discourse levels. Critically, the observed cohesion misalignment constitutes actionable design evidence rather than a terminal evaluation outcome. Within a learning engineering cycle, such evidence directly informs redesign decisions, including the integration of theory-augmented prompting strategies (e.g., RAG-based or self-reflective prompts) in subsequent iterations. This positioning distinguishes the framework from standalone validation or analytics approaches by embedding measurement within an ongoing improvement loop. As such, NLP functions not merely as an assessment tool but as an engineering control mechanism that enables rapid, data-driven refinement of AI behavior. This nested-cycle approach illustrates how learning engineering can support the development of adaptive, theory-aligned personalization systems that are both scalable and responsive to learner diversity (Baker et al., 2022; Goodell & Kolodner, 2023).

Extending prior studies, this work demonstrates the framework's generalizability to the history domain and identifies prompt design as a key factor in theory-aligned personalization. This study advances learning engineering by demonstrating how NLP serves as a scalable evaluation layer within adaptive learning systems. The framework enables real-time assessment of personalization quality and supports continuous, data-driven refinement across domains.

Although NLP can potentially provide a scalable method for validating LLM personalization, several open questions remain. How can these metrics be dynamically linked to learner performance data to determine whether linguistic alignment truly improves comprehension? What kinds of prompting architectures (e.g., RAG-based, self-reflective, or multi-agent systems) best translate theoretical constructs into generative behavior? Finally, how might these validation frameworks be generalized across languages and educational domains? Addressing these questions will advance the scientific foundations needed for reliable, explainable, and equitable AI-driven personalization.

Applying NLP validation in real educational systems also requires a robust infrastructure that connects text analytics with learner modeling and content generation, embedding NLP analytics directly into the design and delivery cycle. A scalable pipeline would (1) generate personalized content through LLMs, (2) apply NLP analytics in real time to evaluate linguistic alignment with learner profiles, and (3) feed these metrics back into the prompt or content generation process. Achieving this vision requires developing interoperable tools that connect LLM APIs, learner modeling systems, and NLP analytics dashboards capable of interpreting theory-based linguistic indices. Such integration will enable a transition from static, offline analyses to dynamic, real-time validation pipelines that ensure theory-aligned personalization at scale. Integrating these components transforms NLP validation from a research method into an operational mechanism for continuous system improvement—an essential step for learning engineering in practice.

Acknowledgments

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 the views of the Institute or the U.S. Department of Education.

References

Abbes, F., Bennani, S., & Maalel, A. (2024). Generative AI and gamification for personalized learning: Literature review and future challenges. SN Computer Science, 5(8), 1–12.

Baker, R. S., Boser, U., & Snow, E. L. (2022). Learning engineering: A view on where the field is at, where it's going, and the research needed. Technology, Mind, and Behavior.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., et al. (2020). Language models are few-shot learners. arXiv preprint arXiv:2005.14165.

Crossley, S. A., Skalicky, S., Dascalu, M., McNamara, D., & Kyle, K. (2017). Predicting text comprehension, processing, and familiarity in adult readers: New approaches to readability formulas. Discourse Processes, 54(1), 1–20.

Dong, L., Tang, X., & Wang, X. (2025). Examining the effect of artificial intelligence in relation to students' academic achievement in classroom: A meta-analysis. Computers and Education: Artificial Intelligence, 8, 100400.

Frantz, R. S., Starr, L. E., & Bailey, A. L. (2015). Syntactic complexity as an aspect of text complexity. Educational Researcher, 44(7), 387–393.

Goodell, J., & Kolodner, J. (Eds.). (2023). Learning engineering toolkit: Evidence-based practices from the learning sciences, instructional design, and beyond. Taylor & Francis.

He, J., Rungta, M., Koleczek, D., Sekhon, A., Wang, F. X., & Hasan, S. (2024). Does prompt formatting have any impact on LLM performance? arXiv preprint. arXiv:2411.10541. https://doi.org/10.48550/arXiv.2411.10541

Huynh, L., & McNamara, D. S. (2025a). GenAI-Powered Text Personalization: Natural Language Processing Validation of Adaptation Capabilities. Applied Sciences, 15(12), 6791.

Huynh, L., & McNamara, D. S. (2025b). Natural Language Processing as a Scalable Method for Evaluating Educational Text Personalization by LLMs. Applied Sciences, 15(22), 12128.

Leong, J., Pataranutaporn, P., Danry, V., Perteneder, F., Mao, Y., & Maes, P. (2024). Putting things into context: Generative AI-enabled context personalization for vocabulary learning improves learning motivation. In Proceedings of the CHI Conference on Human Factors in Computing Systems (pp. 1–15).

Magliano, J. P., & Millis, K. K. (2004). Assessing reading skill with a think-aloud procedure. Cognition and Instruction, 21, 251–283.

Martínez, P., Ramos, A., & Moreno, L. (2024). Exploring large language models to generate easy-to-read content. Frontiers in Computer Science, 6, 1394705.

McNamara, D. S. (2021). Chasing theory with technology: A quest to understand understanding. Discourse Processes, 58(5–6), 422–448.

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). Automated evaluation of text and discourse with Coh-Metrix. Cambridge University Press.

McNamara, D. S., Levinstein, I. B., & Boonthum, C. (2004). iSTART: Interactive strategy training for active reading and thinking. Behavior Research Methods, Instruments, & Computers, 36(2), 222–233.

Nagy, W., & Townsend, D. (2012). Words as tools: Learning academic vocabulary as language acquisition. Reading Research Quarterly, 47(1), 91–108.

O'Reilly, T., & McNamara, D. S. (2007). Reversing the reverse cohesion effect: Good texts can be better for strategic, high-knowledge readers. Discourse Processes, 43(2), 121–152.

Ozuru, Y., Dempsey, K., & McNamara, D. S. (2009). Prior knowledge, reading skill, and text cohesion in the comprehension of science texts. Learning and Instruction, 19(3), 228–242.

Pesovski, I., Santos, R., Henriques, R., & Trajkovik, V. (2024). Generative AI for customizable learning experiences. Sustainability, 16(7), 3034.

Potter, A., Shortt, M., Goldshtein, M., & Roscoe, R. D. (in press). Assessing academic language in tenth-grade essays using natural language processing. Assessing Writing.

Sahoo, P., Singh, A. K., Saha, S., Jain, V., Mondal, S., & Chadha, A. (2024). A systematic survey of prompt engineering in large language models: Techniques and applications. arXiv preprint arXiv:2402.07927.

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., & Wang, G. (2022). Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. arXiv preprint arXiv:2206.04615.

Ye, Q., Axmed, M., Pryzant, R., & Khani, F. (2023). Prompt engineering a prompt engineer. arXiv preprint arXiv:2311.05661.

Comments and Action Items for Authors