

Evaluation of AI-generated Passages for Reading Practice: An application of causal forests

Walter Leite, Qian Shen, Wonchae Lee, Sierra Evans,
Sienna Perez, Wei Li, and Jinnie Shin

University of Florida

Gainesville, FL, USA

{walter.leite, wei.li, jinnie.shin}@coe.ufl.edu

{qian.shen1, wonchaelee, sierra.evans, perezsienna}@ufl.edu

Abstract

This study presents a causal approach for evaluating automated generation of reading passages for early readers. It consists of estimating the average treatment effect of a candidate model, then conditional average treatment effects with causal forests. We apply it to evaluate a fine-tuned model for a digital literacy platform.

1 Introduction

Early reading proficiency is a crucial predictor of long-term academic success and lifelong literacy (Snow et al., 2007). One of the most effective approaches to foundational reading instruction is systematic phonics, in which students are explicitly taught to decode words by mapping letters (graphemes) to sounds (phonemes) (Ehri, 2015). To support this learning process, beginning readers require access to decodable text composed primarily of words that match the phonics patterns and sight words students have already learned (Cheatham and Allor, 2012). These texts allow cumulative skill-building, reduce cognitive load, and let children practice new sound patterns in context. Elementary reading curricula (e.g., Lane and Con-tesse (2022)) typically provide teachers with decodable passages. These passages also align with the curricular scope and sequence, allowing teachers to select a passage that includes phonemes taught in specific lessons for practice and formative assessment.

Despite the importance of decodable text in reading instruction, creating decodable and curriculum-aligned passages is labor-intensive and requires expert knowledge of phonics progression, word selection, and narrative construction. As a result, the quantity and topical variety of decodable passages available to teachers can be limited, potentially reducing student engagement, opportunities for differentiated instruction, and the ability to

match reading passages to student interests (Mesmer, 2005; Cheatham and Allor, 2012).

Educators and researchers are increasingly exploring LLMs to create customized instructional materials, including leveled readers and decodable texts (Kasneci et al., 2023). Currently available LLMs can generate decodable passages for reading instruction aligned with lessons in a specific reading curriculum, given appropriate prompts. However, their performance may vary widely depending on model choice (off-the-shelf or fine-tuned), inputs, constraints, and contextual factors. Causal inference methods to evaluate the effects of LLMs and potential moderation by implementation variables are important because they allow unfounded inference about effect heterogeneity. However, their application in this context remains limited. Thus, this methodological study adapts and demonstrates a causal inference method to evaluate LLM performance in generating decodable passages for elementary reading instruction tied to lessons in a reading curriculum. We illustrate the method with a causal comparison of a fine-tuned model and an off-the-shelf model for generating decodable children's stories for a reading app.

2 Background and Related Work

2.1 Large Language Models to Generate Decodable Text

Despite their instructional value, creating high-quality decodable reading passages is time-consuming and expertise-intensive. Authors must carefully select words and sentence structures to match each lesson's phonics scope, while also crafting engaging, coherent narratives within strict constraints. In recent years, teachers can turn to large language models (LLMs), such as GPT-4o mini and Llama-3.3 70B (Dubey et al., 2024), to create decodable text.

Educational researchers have investigated LLMs

for creating personalized learning materials. Preliminary studies suggest that LLMs can rapidly generate age-appropriate passages at varying reading levels and on diverse topics, offering new opportunities for supporting literacy instruction (Kasneci et al., 2023; Manlises and Ong, 2026). However, research on using LLMs to generate text that meets specific demands for phonics alignment and decodability remains limited. The ability of LLMs to adhere to strict phonics patterns, exclude untaught graphemes, and match the nuanced structure of professional decodable stories remains underexplored.

2.2 Automated Text Analysis of Reading Passages

Recent advances in automated text analysis now make it possible to compute reading-passage properties at scale, enabling metrics-based evaluation of LLM-generated texts. Tools that leverage natural language processing and custom phonics-alignment algorithms can rapidly compute metrics related to text readability, coherence, and complexity. Metrics-based evaluation can provide a fast way to examine LLM-generated texts and discard problematic ones.

To operationalize text evaluation, researchers employ specific linguistic indices to capture the multidimensional nature of text quality. Traditional metrics like Flesch Reading Ease (Flesch, 1948) provide a baseline for surface readability, but are increasingly supplemented by Multisyllabic Percentage (Kearns, 2015) to monitor specific decoding demands. A high density of multisyllabic words can create a "decoding bottleneck," where cognitive resources are consumed by word recognition rather than meaning-making (Kearns and Hiebert, 2022). Beyond the word level, Syntactic Complexity indices (e.g., T-unit length) predict grade-level difficulty more accurately than vocabulary alone (Nelson et al., 2012), ensuring structural accessibility for novice readers.

Local cohesion in text can be measured by Entity Overlap (Graesser et al., 2004). Reading passages with Entity Overlap may act as a critical cognitive scaffold, facilitating deep comprehension (McNamara et al., 1996). Texts with high local cohesion help early learners build coherent "situation models" despite limited background knowledge. Finally, for LLM-generated materials, Perplexity serves as a robust proxy for linguistic naturalness. By measuring the surprisal of generated passages, Perplexity ensures that AI-generated texts main-

tain the idiomaticity and prosodic flow required for authentic language acquisition, filtering out stilted or atypical linguistic inputs (Gauthier et al., 2020; Martinc et al., 2021).

The need for automated text analysis for reading passage evaluation becomes clear when considering the limitations of conventional readability metrics. Common metrics, such as Lexile level, emphasize surface-level difficulty but overlook critical aspects such as developmental appropriateness. As a result, texts deemed "appropriate" by these measures may not actually align with what students have been taught. Research warns that this misalignment, especially when text difficulty increases too rapidly, can undermine reading proficiency and engagement, potentially widening achievement gaps (Hiebert and Mesmer, 2013). Automated metrics-based evaluation tools can help prevent such misalignments by providing more thorough, phonetics-based analysis.

2.3 Causal Evaluation of LLMs for Reading Passage Generation

Causal inference methods can benefit multiple areas of LLMs research. In a review of related applications, Liu et al. (2025) identified six areas of application of causal methods to LLMs: 1) reasoning capacity, 2) fairness and bias, 3) safety, 4) explainability, 5) multimodality, 6) evaluation and benchmark. This final application area shapes LLM performance across the other areas. Causal evaluation of decodable passages for reading instruction can focus on several sources of evidence, including fairness, bias, and safety. In this study, we selected a few sources of evidence related to readability, coherence, and complexity of reading passages.

While human and LLM-as-a-judge evaluations of reading passages aim to rate them as pass/fail or on a Likert scale, causal evaluation focuses on estimating the effect of a condition that may lead to variation in reading passage quality. Examples include the choice of LLM, the specific prompt strategy used, or whether examples were provided. In general, causal evaluation of LLMs requires three steps (Keller and Branson, 2024): 1) define a causal estimand of interest, 2) identify an appropriate estimator for the estimand, after reviewing whether the estimator's assumptions are tenable with the available research design and data, 3) apply the steps to obtain an estimate from the selected estimator. Popular causal estimands include the average treatment effect (ATE) and the average treatment effect

on the treated (ATT) (Leite et al., 2025a).

Causal estimands are commonly defined using Rubin’s causal model, also known as the potential outcomes framework or counterfactual framework for causal inference (Holland, 1986). In Rubin’s causal model, a study participant has potential outcomes under both treated and control conditions, but only one condition is observed. The ATE is the expected difference between potential outcomes under treatment and control for all participants. Common assumptions include unconfoundedness, overlap (also known as positivity), the stable unit treatment value assumption (SUTVA), perfect compliance with treatment, and no attrition from the post-test (Rubin, 2008).

The use of causal inference methods, combined with text metrics, for evaluating LLM-generated text remains in its early stages. One recent example is a study in which (Ji et al., 2025) adapt Pearl’s causal graph framework (Pearl, 1998) to evaluate prompts for generating computer code. They define the treatment condition as prompt rephrasing and the control condition as the original prompt without rephrasing. Their analysis included lexical, syntactic, and semantic characteristics of the prompt. However, rather than using a fixed set of confounders for all ATE estimates, the authors used a gradient-based causal discovery algorithm to create a different causal graph for each estimate. For each causal graph, they then estimated ATEs using a double machine learning algorithm. However, the authors estimate a very large number of ATEs, and there is no indication that they controlled the false discovery rate or employed any multiple-comparison correction to mitigate inflated Type I error rates.

2.4 Evaluation of heterogeneity of LLM effects

In applications of LLMs to generating passages for reading instruction, effect heterogeneity can be substantial, as LLMs can interact with input characteristics, constraints, and context, resulting in wide variation in passage quality. A traditional statistical approach for investigating these interaction effects would involve a regression model with product terms of the condition under evaluation and potential moderating variables, then plotting simple slopes (Aiken and West, 1991). However, it is important to distinguish effect heterogeneity from causal moderation. If causal identification of the ATE was achieved by either random assignment

or analysis methods (e.g., conditioning, weighting), and there is a significant interaction between the condition and a covariate, the subgroup effects of the condition within levels of the covariate are still a causal estimate, even if there are confounders of the relationship between the covariate and the outcome. In this case, we could refer to the subgroup effects as effect heterogeneity, but not causal moderation. However, if there are no confounders of the covariate (e.g., the levels of the covariate were randomly assigned), then the heterogeneity could be characterized as causal moderation (VanderWeele, 2015).

Moderation hypotheses can be difficult to specify in LLM evaluation because theory is limited and many covariates may modify model effects. Causal machine-learning methods provide an exploratory alternative by estimating conditional average treatment effects (CATEs) without manually specifying all interaction terms (Caron et al., 2022). We selected causal forests (Athey et al., 2019) because they directly estimate treatment-effect heterogeneity and can account for clustering in multilevel data, which is important in educational applications (Morbeeck, 2004).

2.5 Research Questions

Comparing off-the-shelf LLMs with fine-tuned models for reading passage generation requires evaluating them on the same task using the same evaluation protocol, enabling direct performance comparisons (Howcroft et al., 2020). In our application (Leite et al., 2025b), we selected Granite-3.3 8B Instruct as the treatment because, under Rewarded SFT, it showed a favorable trade-off across the target metrics in our prior development study, including the highest coherence and lowest LM-PPL among the three 8B models evaluated (Shen et al., 2026). We selected GPT-4o mini as the control because it is the platform’s existing off-the-shelf story-generation model. The comparison therefore tests whether a compact, task-adapted model can serve as a viable replacement for the existing platform model. However, because this study is methodological, the choice of LLM may be arbitrary, and the causal evaluation method should generalize to other pairwise LLM comparisons. The research questions addressed in this study are:

1. What is the average treatment effect (ATE) of a fine-tuned model (treatment) compared to an off-the-shelf model (control)?
2. Is there substantial heterogeneity of treatment effects (HTE) of the

fine-tuned model? 3. Which covariates predict the conditional average treatment effects (CATE) of the fine-tuned model?

3 Methods

3.1 Sample

We collected data for this study using Storiza (Leite et al., 2025b), a digital literacy platform that enables students to create and illustrate their own stories for reading practice. We recruited parent/child pairs through the official Facebook group of the UFLI Foundations phonics program (Lane and Contesse, 2022). The study instructions asked parents to use the Storiza digital literacy platform with their child at least once per week for four weeks. This produced a dataset of 760 stories created by 177 children from kindergarten to second grade, aligned with 90 curriculum lessons.

Given that, in the collected data, LLMs generated only a single story for each child request, and given the large difference in sample sizes between conditions, we applied data augmentation to the collected data. We retained all 760 story-generation configurations selected by the children: grade level, keywords, theme, story length, and corresponding curriculum. We used the treated and control LLMs to generate 20 stories for every condition and 10 per prompt type, yielding a total sample size of 30,400.¹

3.2 Measures

The treatment model was IBM Granite-3.3 8B Instruct, further fine-tuned for curriculum-aligned children’s story generation using the Rewarded SFT procedure described by (Shen et al., 2026). Briefly, we applied QLoRA with 4-bit NF4 quantization and LoRA adapters to 2,580 stories generated by GPT-4o and Llama-3.3 70B from 129 K-2 curriculum lessons. In Rewarded SFT, each training example’s cross-entropy loss was weighted by a scalar reward defined as the unweighted mean of five normalized automatic metrics: Spache readability, LM-PPL, coherence, syntactic complexity, and toxicity. The control model was the off-the-shelf GPT-4o mini. To obtain the analysis’ outcome, we calculated five NLP-based evaluation metrics for each generated story and for curriculum stories aligned with its configuration: Flesch

readability score (Flesch, 1948), syntactic complexity, entity overlap, polysyllabic word proportion, and GPT-2-derived language model perplexity (LM-PPL) (Meister and Cotterell, 2021). We then computed the outcome variable as the Mahalanobis distance between the metric vectors of the LLM-generated stories and those of the corresponding expert-written stories provided by the UFLI Foundations phonics program (Lane and Contesse, 2022). A smaller Mahalanobis distance indicates that the LLM-generated story more closely matches its corresponding expert-written curriculum story across the five linguistic metrics.

The covariates of interest are LLM inputs, because they precede the model’s operations and may moderate its performance. For operational reasons, we used two different prompts during data collection: one for the first 18 days, yielding 557 stories, and a second for the remaining 18 days, yielding 203 stories.² We coded prompt type as a binary covariate. We coded the curriculum requirements as 89 dummy-coded lesson indicators. In the curriculum, lessons are clustered into 15 topics (e.g., alphabet, digraphs, diphthongs); therefore, we created 14 dummy topic indicators. We coded story length as a numeric covariate with seven possible values: 50, 75, 100, 200, 250, 300, 400 words. Grade had three levels (K, 1, and 2), represented by two dummy indicators for kindergarten and second grade, with first grade as the reference category. We included the same five metrics mentioned above for the expert-written example stories as covariates. We also included a binary schema-version indicator as a covariate. In total, we included 113 covariates.

3.3 Analysis

Causal forests estimate the ATE and CATEs by growing an ensemble of “honest” trees that recursively partition the covariate space to capture systematic variation in treatment effects (Breiman, 2001; Wager and Athey, 2018; Athey et al., 2019). Because the data contain repeated observations nested within higher-level units (multiple generated stories per child/user and per story configuration), outcomes may be correlated within clusters, violating the independent-and-identically-distributed assumption. To account for this multilevel structure, we fit a cluster-robust causal forest in which each tree is trained on a subsample of clusters (rather than individual children), and out-of-bag (OOB)

¹An anonymized dataset and code sufficient to reproduce the analyses are available at: https://osf.io/qvy5t/overview?view_only=75b75b3ae0ba43258972a44fb2ca90f7

²Full details about the prompts are omitted to protect intellectual property.

predictions treat an observation as OOB only when its entire cluster is excluded from the tree’s subsample (Athey and Wager, 2019). All analyses were conducted in R using the *grf* (Generalized Random Forests) package. We trained a causal forest with 1,000 trees and enabled honest splitting (Wager and Athey, 2018; Athey et al., 2019). At each split, we considered 4 candidate covariates and used a larger minimum leaf size to stabilize effect estimates given the augmented sample and within-cluster dependence. We first estimated ATE and obtained OOB CATE predictions (Wager and Athey, 2018; Athey et al., 2019). Next, we assessed treatment-effect heterogeneity by (a) inspecting the distribution of predicted CATEs, (b) computing forest-based variable-importance to identify covariates most strongly associated with variation in the estimated effects (Breiman, 2001; Tibshirani et al., b), and (c) performing a calibration-based omnibus test for heterogeneity using a linear calibration regression fitted to held-out forest predictions. Finally, we summarized effect heterogeneity with an interpretable best linear projection (BLP) of the CATE on key moderators, estimated via doubly robust scores from the causal forest (Semenova and Chernozhukov, 2021; Tibshirani et al., a). We report heteroskedasticity-robust standard errors with clustering handled via the forest’s cluster specification.

4 Results

4.1 Research Question 1

The causal forest analysis revealed a significant ATE of 0.264 ($SE = 0.016$, $p < .001$), indicating that Granite-3.3 8B Instruct produced stories with larger Mahalanobis distances from the corresponding expert-written curriculum stories than GPT-4o mini. The standardized effect size was Hedges’ $g = 0.213$ ($p < .05$).

4.2 Research Question 2

The calibration test suggested the forest’s average effect was well calibrated (estimate = 0.997, $SE = 0.059$; an ideal value is 1). The heterogeneity component was positive and highly significant (estimate = 2.519, $SE = 0.117$, $p < .001$), providing an omnibus rejection of constant treatment effects and indicating that LLM effects vary across configurations. The distribution of CATEs ranged from 0.197 to 0.310 (median = 0.265), with all individuals experiencing increased distance under Granite,

though the magnitude varied by approximately 70% between the least and most affected users.

4.3 Research Question 3

We calculated variable importance as the weighted sum of the number of times each covariate was split at each depth in the forest. We selected covariates with importance above the median as potential moderators. These included 31 covariates, such as prompt type, story word count, all five NLP metrics calculated from the stories obtained from UFLI phonics program (Lane and Contesse, 2022), and second-grade status. The best linear projection results (Table 1) showed that prompt type was the strongest moderator; users who received prompt type 2 experienced an additional 0.408 unit increase in the Mahalanobis distance ($p < 0.001$), indicating particularly poor performance by Granite under this prompt configuration. Second-grade children also showed a significant moderating effect, with an additional 0.233-unit increase ($p = .030$). Pedagogical content dummy indicators (i.e., specific lessons and topics) did not significantly moderate LLM effects, suggesting that Granite’s quality disadvantage was relatively consistent across educational content.

Table 1: Best linear projection results

Variable	Estimate	SE	p
Number of words	0.001	0.001	.283
Schema version	−0.011	0.065	.871
Flesch reading ease	0.053	0.029	.064
Syntactic complexity	−0.042	0.055	.442
Entity overlap	0.075	0.241	.754
Multisyllabic percentage	4.958	3.760	.187
Perplexity	−0.009	0.015	.538
Grade (2nd)	0.233	0.107	.030*
Grade (Kindergarten)	0.293	0.262	.263
Type of Prompt (2)	0.408	0.054	< .001***

Note. Non-significant estimates for dummy indicators of topic and lesson are not shown. The intercept is omitted. Cluster- and heteroskedasticity-robust (HC3) standard errors. † means $p < .10$, * means $p < .05$, *** means $p < .001$.

5 Discussion

The positive treatment effect indicates a task-specific performance difference between the Rewarded-SFT Granite-3.3 8B Instruct model and GPT-4o mini under our evaluation protocol. Because OpenAI does not report GPT-4o mini’s parameter count or detailed architecture in its public model description, we do not characterize this comparison as size-matched or attribute the ob-

served difference to a specific architectural mechanism. Instead, the comparison is deployment-oriented: it evaluates whether a compact, task-adapted model can replace the platform’s existing off-the-shelf model while preserving the targeted passage-quality characteristics. Most importantly, the analysis shows that causal machine learning methods can provide a principled way to compare any pair of LLMs on passage-generation metrics.

All five NLP-based metrics for the phonics reading curriculum had above-median variable importance and were therefore selected as potential moderators. However, none of their BLP coefficients reached $p < .05$, so these results do not identify any single curriculum metric as a statistically significant moderator. The pattern nevertheless suggests that linguistic characteristics of the target passages may contribute to treatment-effect heterogeneity and warrants further investigation.

6 Conclusion

6.1 Implications for Practice

The causal evaluation of the fine-tuned Granite-3.3 8B Instruct versus the off-the-shelf GPT-4o mini supports keeping the latter as the story-generating model in the platform. If the fine-tuned model had outperformed the off-the-shelf model, adopting it would have led to substantial cost reductions at scale. Instead, the current results indicate that another fine-tuning study is necessary to achieve the goal of replacing GPT-4o mini.

The Causal Forest results indicate that prompt type moderated the effects of LLMs. Although the prompt change was not randomly assigned, the switch was unrelated to model performance or other contextual factors, so the unconfoundedness assumption likely holds. Given this assumption, the interaction between LLM and prompt is an example of causal moderation. However, we do not know what specific characteristic of the prompt changed that contributed to the difference in performance of LLMs, as the way the specific tokens of a prompt interact with model components is part of the "black box" of the model (Ji et al., 2025). Further experimentation with small prompt changes may help clarify the issue.

6.2 Future Research

Developers of AI-enabled digital literacy platforms have a set of tools that provide different sources of evidence about the validity of LLM outputs for

intended purposes. These tools include metrics, benchmarks, human raters, LLM-as-a-judge, and causal inference analyses. However, these tools are typically used independently and do not build on each other as part of an integrated evaluation framework. Casabianca (2025) recently proposed combining and extending the evidence-centered design framework for test development (Mislevy et al., 2003) together with the validity framework widely used in educational and psychological measurement (American Educational Research Association et al., 2014) to create a new framework that combines evidence from metrics, human raters, and LLM-as-a-judge. While this proposed framework has strong potential, it does not include causal inference methods. We propose that future research into a unified framework for LLM evaluation should also incorporate causal frameworks such as Rubin’s causal model (Holland, 1986), Pearl’s causal graphs (Pearl, 1998), and Campbell’s design validity threat taxonomy (Shadish et al., 2002).

Limitations

This study has a few limitations regarding the set of covariates included in the causal forest analysis. The heterogeneity in effect associated with grade level may not meet assumptions for causal moderation because confounders related to both grade level and story quality may be present. The outcome measure also has limitations. To address the multidimensionality of passage metrics, we used Mahalanobis distance to create a single score for each story across five metrics. This yields a convenient scalar summary of how closely a generated story matches an expert-written reference in a joint-feature space. However, this operationalization embeds several assumptions that may not fully align with how decodable passages are judged in classroom use. Moreover, Mahalanobis distance weights dimensions according to the estimated covariance among metrics, which can unintentionally down-weight features that are instructionally important but less variable in the reference set. Hence, additional metrics, such as average sentence length and sentence diversity, could have provided different results.

Acknowledgments

References

Leona S. Aiken and Stephen G. West. 1991. *Multiple Regression: Testing and Interpreting Interactions*.

- Sage Publications, Newbury Park, CA.
- American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, editors. 2014. *Standards for Educational and Psychological Testing*. American Educational Research Association, Washington, DC.
- Susan Athey, Julie Tibshirani, and Stefan Wager. 2019. [Generalized random forests](#). *The Annals of Statistics*, 47(2):1148–1178.
- Susan Athey and Stefan Wager. 2019. [Estimating treatment effects with causal forests: An application](#). *Observational Studies*, 5(2):37–51.
- Leo Breiman. 2001. [Random forests](#). *Machine Learning*, 45:5–32.
- Alberto Caron, Gianluca Baio, and Ioanna Manolopoulou. 2022. [Estimating individual treatment effects using non-parametric regression models: a review](#). *Journal of the Royal Statistical Society Series A: Statistics in Society*, 185(3):1115–1149.
- Jodi M Casabianca. 2025. Psychometrics is all you need.
- Jennifer P Cheatham and Jill H Allor. 2012. The influence of decodability in early reading text on reading achievement: A review of the evidence. *Reading and Writing*, 25(9):2223–2246.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.
- Linnea C Ehri. 2015. How children learn to read words linnea c. ehri. *The Oxford handbook of reading*, page 293.
- Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.
- Jon Gauthier, Jennifer Hu, Ethan Wilcox, Peng Qian, and Roger Levy. 2020. Syntaxgym: An online platform for targeted evaluation of language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 70–76.
- Arthur C Graesser, Danielle S McNamara, Max M Louwerse, and Zhiqiang Cai. 2004. Coh-matrix: Analysis of text on cohesion and language. *Behavior research methods, instruments, & computers*, 36(2):193–202.
- Elfrieda H Hiebert and Heidi Anne E Mesmer. 2013. Upping the ante of text complexity in the common core state standards: Examining its potential impact on young readers. *Educational Researcher*, 42(1):44–51.
- Paul W. Holland. 1986. [Statistics and causal inference](#). *Journal of the American Statistical Association*, 81(396):945–960.
- David Howcroft, Anya Belz, Miruna-Adriana Clinciu, and Dimitra Gkatzia. 2020. Evaluation practices for data-to-text, conditional, and open-ended generation. *Computational Linguistics*.
- Zhenlan Ji, Pingchuan Ma, Zongjie Li, Zhaoyu Wang, and Shuai Wang. 2025. [Causality-aided evaluation and explanation of large language model-based code generation](#). *Proceedings of the ACM on Software Engineering*, 2(ISSTA):1374–1397.
- Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, and 1 others. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274.
- Devin M Kearns. 2015. How elementary-age children read polysyllabic polymorphemic words. *Journal of Educational Psychology*, 107(2):364.
- Devin M Kearns and Elfrieda H Hiebert. 2022. The word complexity of primary-level texts: Differences between first and third grade in widely used curricula. *Reading Research Quarterly*, 57(1):255–285.
- Bryan Keller and Zach Branson. 2024. [Defining, identifying, and estimating causal effects with the potential outcomes framework: a review for education research](#). *Asia Pacific Education Review*, 25(3):575–594.
- H. Lane and V. Contesse. 2022. *UFLI Foundations: An explicit and systematic phonics program*. Ventriss Learning, Gainesville, FL.
- Walter Leite, Huibin Zhang, Zachary Collier, Kamal Chawla, Lingchen Kong, YongSeok Lee, Jia Quan, and Olushola Soyoye. 2025a. Machine learning for propensity score estimation: A systematic review and reporting guidelines. *Psychological Methods*.
- Walter L Leite, Stephanie Hammerschmidt-Snidarich, A Corinne Huggins-Manley, Hatice K Karakis, Holly Lane, Seyedahmad Rahimi, Qian Shen, Xinyi Tai, and Matthew Schmidt. 2025b. Storiza: A platform to support children’s oral reading fluency development with generative ai. In *Proceedings of the 19th International Conference of the Learning Sciences-ICLS 2025*, pp. 1574–1578. International Society of the Learning Sciences.
- Xiaoyu Liu, Paiheng Xu, Junda Wu, Jiabin Yuan, Yifan Yang, Yuhang Zhou, Fuxiao Liu, Tianrui Guan, Hao-liang Wang, Tong Yu, Julian McAuley, Wei Ai, and Furong Huang. 2025. [Large language models and causal inference in collaboration: A comprehensive survey](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7668–7684, Albuquerque, New Mexico. Association for Computational Linguistics.

- Maria Monica Manlises and Ethel Ong. 2026. [Zero shot phonics: Evaluating constraint-adherent phonics story generation in large language models](#). In *Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2026)*, pages 914–932, San Diego, California, USA. Association for Computational Linguistics.
- Matej Martinc, Senja Pollak, and Marko Robnik-Šikonja. 2021. Supervised and unsupervised neural approaches to text readability. *Computational Linguistics*, 47(1):141–179.
- Danielle S McNamara, Eileen Kintsch, Nancy Butler Songer, and Walter Kintsch. 1996. Are good texts always better? interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and instruction*, 14(1):1–43.
- Clara Meister and Ryan Cotterell. 2021. Language model evaluation beyond perplexity. *arXiv preprint arXiv:2106.00085*.
- Heidi Anne E Mesmer. 2005. Text decodability and the first-grade reader. *Reading & Writing Quarterly*, 21(1):61–86.
- Robert J Mislevy, Russell G Almond, and Janice F Lukas. 2003. A brief introduction to evidence-centered design. *ETS Research Report Series*, 2003(1):i–29.
- Mirjam Moerbeek. 2004. [The consequence of ignoring a level of nesting in multilevel analysis](#). *Multivariate Behavioral Research*, 39(1):129–149.
- Jessica Nelson, Charles Perfetti, David Liben, and Meredith Liben. 2012. Measures of text difficulty: Testing their predictive value for grade levels and student performance. *Council of Chief State School Officers, Washington, DC*.
- Judea Pearl. 1998. [Graphs, causality, and structural equation models](#). *Sociological Methods & Research*, 27(2):226–284.
- Donald B. Rubin. 2008. [For objective causal inference, design trumps analysis](#). *The Annals of Applied Statistics*, 2(3):808–840.
- Vira Semenova and Victor Chernozhukov. 2021. [Debiased machine learning of conditional average treatment effects and other causal functions](#). *The Econometrics Journal*, 24(2):264–289.
- William R. Shadish, Thomas D. Cook, and Donald T. Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Houghton Mifflin Company, Boston, MA.
- Qian Shen, Fanghua Cao, Min Yao, Shlok Gilda, Bonnie J Dorr, and Walter L Leite. 2026. Children’s english reading story generation via supervised fine-tuning of compact llms with controllable difficulty and safety. *arXiv preprint arXiv:2605.13709*.
- Catherine E Snow, Michelle V Porche, Patton O Tabors, and Stephanie Ross Harris. 2007. Is literacy enough? pathways to academic success for adolescents.
- Julie Tibshirani, Susan Athey, Erik Sverdrup, and Stefan Wager. a. grf documentation: best_linear_projection. https://grf-labs.github.io/grf/reference/best_linear_projection.html. Accessed 2026-02-04.
- Julie Tibshirani, Susan Athey, Erik Sverdrup, and Stefan Wager. b. *grf: Generalized Random Forests*. R package. DOI: <https://doi.org/10.32614/CRAN.package.grf>. Accessed 2026-02-04.
- Tyler VanderWeele. 2015. *Explanation in causal inference: methods for mediation and interaction*. Oxford University Press.
- Stefan Wager and Susan Athey. 2018. [Estimation and inference of heterogeneous treatment effects using random forests](#). *Journal of the American Statistical Association*, 113(523):1228–1242.