StoryLab: Empowering Personalized Learning for Children through Teacher-Guided Multimodal Story Generation

Zhaohui Li[1]*, Feiwen Xiao[2]*, Jiaju Lin[2], Xiaohan Zou[2], Qingxiao Zheng[1], and Jinjun Xiong[1]

[1] University at Buffalo, Getzville, NY 14068, USA
{zli253,qingxiao,jinjun}@buffalo.edu
[2] The Pennsylvania State University, University Park, PA 16802, USA
{ffx5014,jjl7137,xfz5266}@psu.edu

Abstract. Personalized story reading enhances child literacy by aligning content with individual interests, backgrounds, and developmental needs. However, implementing such systems presents challenges, including data privacy concerns, the need for culturally diverse materials, limited resources, and balancing personalization with standardized benchmark objectives. To address these challenges, we introduce **StoryLab**, a multimodal system designed for K-2 teachers and students. The system leverages advanced generative AI to integrate students' personal interests with teacher-defined learning objectives to generate comprehensive learning materials, including story text, illustrated figures, vocabulary support, and a consistent narrative voice. A teacher-in-the-loop design ensures pedagogical alignment and trust. Evaluations demonstrate StoryLab's effectiveness and usability, positioning it as a promising and scalable tool for personalized literacy instruction.

Keywords: Early Literacy · Story Generation · Personalized Learning · Large Language Models · Multimodal Learning Environment

1 Introduction

Personalized reading experiences have been shown to enhance engagement, comprehension, and literacy skill acquisition in young learners [15]. By tailoring content to children's unique interests, cultural backgrounds, and developmental needs, educators can foster intrinsic motivation and bridge gaps in foundational skills development such as phonics, fluency, and vocabulary [7]. However, scalable implementation of personalized literacy tools remains a challenge. Current classroom practices demonstrate that reading decodable text is a common approach to teaching phonics [1,6]. Many existing literacy platforms rely on standardized texts, which may not adequately represent the classrooms' linguistic and cultural diversity in recent years [10]. Children are more likely to engage in reading when they see their own cultures and identities represented in the content [4]. To support diverse learners, reading texts should be personalized to reflect different cultural backgrounds. Educators must navigate competing priorities, such as

* Equal contribution

balancing individualized instruction with standardized learning objectives, and sourcing culturally responsive materials in digital environments [5,18].

To address these challenges, we designed StoryLab, a multimodal generative AI system supporting K-2 teachers and students in personalized story reading and literacy instruction. StoryLab utilizes SOTA generative models GPT-4o [9] and DALLE-3 [12] from OpenAI to produce adaptive learning materials, including customized texts, illustrations, and vocabulary support. Educators can define learning objectives such as phonics patterns, vocabulary, and grammar, while integrating student profiles (interests, cultural backgrounds) to generate engaging, coherent narratives. Unlike traditional platforms, StoryLab employs a teacher-in-the-loop approach, allowing educators to review and refine generated content, thus maintaining pedagogical rigor and data privacy. We evaluated the story's alignment with literacy objectives using a group TF-IDF metric based on an expanded benchmark corpus incorporating words from the DIBELS 8 [17]. Our generated stories outperformed reference corpora in TF-IDF scores, indicating stronger alignment with targeted vocabulary. A user study involving four teachers confirmed StoryLab's pedagogical effectiveness and usability, highlighting its potential as a scalable solution for personalized literacy instruction.

Overall, StoryLab is an AI-powered literacy platform that delivers personalized, story-based reading experiences for early learners and provides teachers with tools for targeted instruction, combining interactive features, customizable content, and a teacher-in-the-loop design to support culturally and linguistically diverse students.

2 Related Works

Recent advancements in AI techniques have driven the development of adaptive reading tools designed to personalize instruction for students. Commercial systems like iReady and Lexia Core5 use machine learning to adjust content difficulty in real-time, emphasizing foundational skills like phonics and fluency [8]. Similarly, DreamBox customizes activities to individual progress, while tools like Natural Reader enhance accessibility via text-to-speech. Amplify Reading effectively boosts literacy outcomes but provides limited teacher customization, focusing primarily on standardized metrics rather than cultural relevance or student interests. The balance between personalization and standardization remains problematic in tools like Amira Learning and Microsoft Reading Coach. Amira Learning offers real-time fluency and comprehension feedback but lacks sufficient teacher customization [5]. Likewise, Ello integrates student interests into phonics learning but restricts content adaptation through closed subscription models. Consequently, current AI-powered systems often overlook teacher input, cultural responsiveness, and curricular integration [18]. These gaps include: (1) limited teacher-involved mechanisms for instructional alignment, (2) inadequate culturally responsive content with personal interest, and (3) reliance on proprietary models restricting adaptability in resource-limited classrooms. StoryLab addresses these challenges by integrating generative AI with educator-guided customization, enabling culturally responsive, dynamic story creation aligned with student interests and instructional goals.

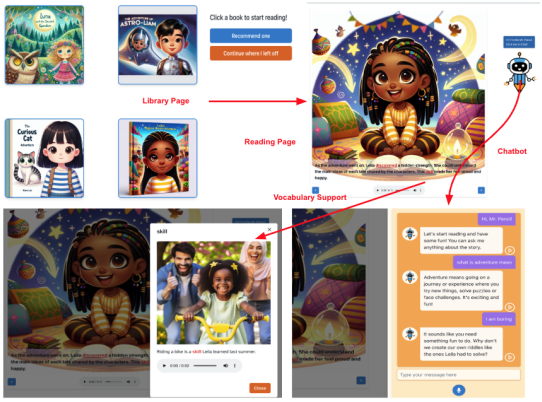


Fig. 1: Overview of the Student Interface in StoryLab.

3 The StoryLab Prototype

The system is designed with two primary user interfaces: Student View and Teacher Dashboard, both accessible through a user-friendly web application.

Student View The student view (Figure 1) offers an engaging, adaptive reading experience through a *Library Page* for browsing teacher-assigned or recommended books, and a *Reading Page* that combines narrations and illustrations to enhance comprehension. Literacy support features include clickable, highlighted words for pictorial definitions and pronunciations, TTS-based *Narration* supports for accessibility, and an interactive *Chatbot* that answers story-related questions via text or voice, using child-friendly prompts to ensure age-appropriate engagement.

Teacher Dashboard As shown in Figure 2, the teacher dashboard enables educators to monitor student progress and generate AI-assisted learning materials. The *Student Profile* section provides key demographic information, such as grade level, ethnicity, and interests. Teachers can also access the *Student Library*, which displays each student's reading history and current books in progress. To further personalize instruction, the *Story Generation Canvas* allows educators to specify learning objectives, including vocabulary, phonics, grammar, and comprehension goals. Based on these inputs, the system produces customized stories aligned with individual student needs. Once a story is generated, teachers can use the *Story Review Page*, a pop-up editor that presents text and illustrations in a book-like format to review, edit, or regenerate specific pages. This feature ensures the content meets pedagogical goals while offering flexibility and control over the final output. By integrating AI-assisted storytelling with teacher supervision and student interactivity, StoryLab provides a dynamic and personalized learning experience. The system enhances literacy development by offering narratives, adaptive reading support, and interactive feedback mechanisms, ensuring a more effective and enjoyable reading journey for students.

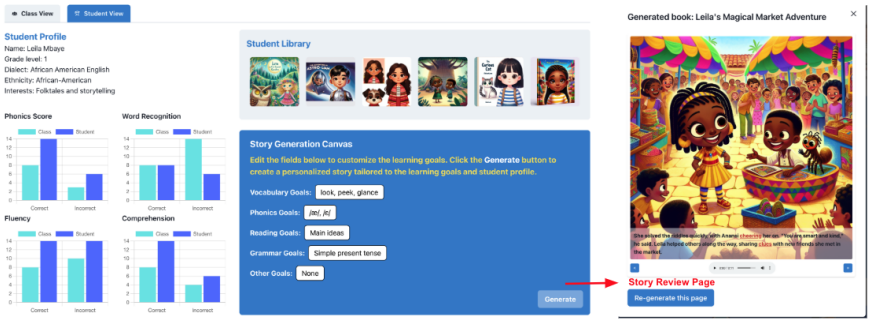


Fig. 2: Overview of the Teacher Dashboard in StoryLab.

Story Generation Canvas The core of the teacher's dashboard is an editable story generation canvas (Figure 2), where teachers create stories aligned with student profiles. Content consistency is ensured via detailed multimodal prompts describing character appearances, available in supplementary materials. The generation pipeline involves three steps: (1) outline drafting, (2) story filling, and (3) multimodal content generation. **Outline drafting**: This initial phase employs LLM planning capabilities to produce structured storylines, a method empirically validated for enhancing narrative quality and consistency [20,11,3]. The outline, generated by a story planner agent, includes the story title, plot points for logical progression, and target vocabulary tailored to learners' reading levels and objectives, thus boosting engagement [13]. **Story filling**: Following the outline, a writer agent develops the narrative by integrating learners' interests, vocabulary, and reading levels, ensuring alignment with educational objectives [19]. The writer agent employs: Dynamic adaptation that personalizes content based on learners' profiles to maintain engagement; Pedagogical alignment that embeds learning objectives within narratives for natural skill development; Complexity gradation that adjusts text difficulty to learners' proficiency to facilitate manageable learning without cognitive overload. **Multimodal content generation**: Multimodal content is created using OpenAI APIs, specifically DALL-E 3 for images and TTS for audio, including: 1. Annotations for target vocabulary: Providing illustrations, sample sentences, and audio pronunciations for new vocabulary, aiding multimodal comprehension. 2. Illustrations for story content: Ensuring visual consistency through detailed character descriptions generated by the story planner agent. This maintains recognizable characters throughout the narrative (Figure 3). These multimodal integrations enhance learner engagement and comprehension by addressing visual consistency challenges.

To ensure the reliability and pedagogical effectiveness of the AI-generated stories, we propose a straightforward teacher-in-the-loop framework. Once the system generates the story and corresponding illustrations, a story review page is presented, displaying the full set of illustrated pages alongside the story text and relevant dictionary entries. Teachers can then review, edit, or refine the story text to ensure that each page aligns with the intended learning objectives and teaching standards.



Fig. 3: Images (a, b) generated with detailed character prompts show consistent visuals, while images (c, d), lacking detailed prompts, show greater variation.

4 Evaluation

Quantitative Evaluation We employed a novel group TF-IDF metric, extending traditional TF-IDF [14] by grouping synonyms and plural word forms to evaluate vocabulary alignment with learning objectives from the DIBELS 8 benchmark dataset [17]. For grades K–2, we selected vocabulary words—such as *job*, *team*, and *time*, and generated ten stories per set, forming our sample set. We compared these with: *Generated Sample Set*: 10 generated stories per DIBELS 8 sample. *StoryLab Corpus*: 30 total generated stories aligned with DIBELS 8. *StoryDB*: English subset of a multilingual narrative dataset [16]. *Children Stories Text Corpus*: 98 public domain children's books from Project Gutenberg [2]. As shown in Figure 4, our stories consistently achieved higher scores across grades K–2 compared to other corpora, indicating strong alignment with target vocabulary. However, further evaluation needs to be conducted for fluency or pedagogical effectiveness.

Teacher Evaluation To better evaluate the pedagogical value of our prototype, we conducted a user study with four educators focusing on: (1) teachers' perceptions of generated stories (language quality, age appropriateness, engagement, pedagogical value, story structure), and (2) potential design improvements. Participants' information, study procedure, and data analysis methods can be found in supplementary materials. Teachers responded positively to the StoryLab pro-

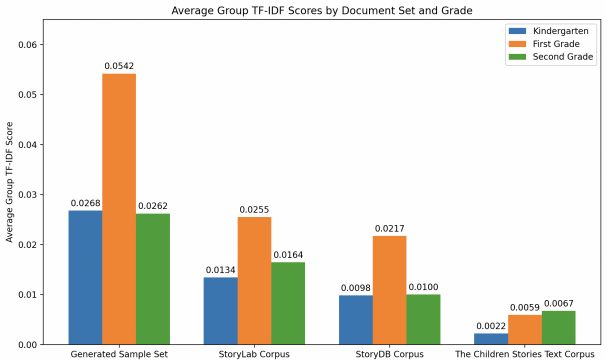


Fig. 4: Comparison of Average Group TF-IDF Scores Across Four Corpora.

totype, highlighting its intuitive dashboard and multimodal tools like read-aloud narration and multimedia dictionaries, which supported word recognition and comprehension. They also praised the story content for its linguistic quality, coherence, and demographic personalization, noting its potential for differentiated instruction (P1–P3). Despite StoryLab's potential, teachers identified key areas for improvement. P4 noted inconsistent character illustrations that disrupted narrative coherence, while others flagged mismatches between story complexity and intended reading levels, citing excessive sight words and advanced vocabulary. P3 recommended a "one sentence per picture" format for Pre-K, and both P3 and P4 emphasized that complex vocabulary without visual support may hinder recognition. Concerns about cultural representation also emerged: while teachers appreciated demographic personalization, P2 and P4 warned against superficial, ethnicity-based depictions and advocated for more authentic, diverse cultural narratives [4]. Additionally, although teachers valued the system's flexibility, P1–P3 found the interface cognitively demanding and called for intelligent scaffolding, such as system-suggested objectives or curated menus to ease use. These findings highlight the need to improve story alignment, cultural sensitivity, and interface usability in future iterations.

5 Discussion and Conclusion

In conclusion, we introduced StoryLab, a multimodal platform supporting early childhood language and literacy development by aligning personalized student experiences with classroom learning objectives. StoryLab integrates student interests, cultural backgrounds, and assessment data, empowering teachers through customizable tools and a teacher-in-the-loop mechanism for reviewing AI-generated content. Our evaluations demonstrate that the generated stories effectively align with educational objectives and student interests, with strong adoption potential among educators. Despite promising outcomes, several limitations must be addressed in future work. Technically, high costs associated with OpenAI's API suggest exploring proprietary LLMs to lower operational expenses. Consistency in character portrayal also requires improvement, achievable through additional training data and model fine-tuning. Furthermore, our user study highlighted inconsistencies in aligning language complexity with grade levels, underscoring the need for detailed classroom benchmark data. Lastly, co-designing interface features with teachers and students, such as recommendation systems and progress tracking, could further enhance usability and instructional effectiveness.

Acknowledgments. This work is supported, in part, by the National Science Foundation under Grant 2229873 (AI4ExceptionalEd) and U.S. Department of Education under Grant R305C240046 (CELaRAI). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the sponsors.

References

1. Cheatham, J.P., Allor, J.H.: The influence of decodability in early reading text on reading achievement: A review of the evidence. Reading and Writing **25**, 2223–2246 (2012)
2. EdenBD: Children stories text corpus (2021), https://www.kaggle.com/datasets/edenbd/children-stories-text-corpus
3. He, T., Liao, L., Cao, Y., Liu, Y., Liu, M., Chen, Z., Qin, B.: Planning like human: A dual-process framework for dialogue planning. In: Ku, L.W., Martins, A., Srikumar, V. (eds.) Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 4768–4791. Association for Computational Linguistics, Bangkok, Thailand (Aug 2024). https://doi.org/10.18653/v1/2024.acl-long.262, https://aclanthology.org/2024.acl-long.262/
4. Koss, M.D.: Diversity in contemporary picturebooks: A content analysis. Journal of Children's Literature **41**(1), 32–42 (2015)
5. Kumar, P.C., Chetty, M., Clegg, T.L., Vitak, J.: Privacy and security considerations for digital technology use in elementary schools. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. pp. 1–13 (2019)
6. Murphy Odo, D.: The use of decodable texts in the teaching of reading in children without reading disabilities: a meta-analysis. Literacy (2024)
7. Neuman, S.B., Celano, D.C.: Giving our children a fighting chance: Poverty, literacy, and the development of information capital. Teachers College Press (2015)
8. Newton, S., Gamble, H., Su, Y., Zoski, J., Damico, D.: Examining the impact of amplify reading on student literacy in grades k-2. 2019 report. Online Submission (2019)
9. OpenAI: Gpt-4 technical report. https://cdn.openai.com/papers/gpt-4.pdf (2023)
10. Paris, D., Alim, H.S.: Culturally sustaining pedagogies: Teaching and learning for justice in a changing world. Teachers College Press (2017)
11. Qiao, S., Zhang, N., Fang, R., Luo, Y., Zhou, W., Jiang, Y., Lv, C., Chen, H.: AutoAct: Automatic agent learning from scratch for QA via self-planning. In: Ku, L.W., Martins, A., Srikumar, V. (eds.) Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 3003–3021. Association for Computational Linguistics, Bangkok, Thailand (Aug 2024). https://doi.org/10.18653/v1/2024.acl-long.165, https://aclanthology.org/2024.acl-long.165/
12. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 (2022)
13. Renninger, K.A., Hidi, S.: The power of interest for motivation and engagement. Routledge (2015)
14. Sparck Jones, K.: A statistical interpretation of term specificity and its application in retrieval. Journal of documentation **28**(1), 11–21 (1972)
15. Tetzlaff, L., Schmiedek, F., Brod, G.: Developing personalized education: A dynamic framework. Educational Psychology Review **33**, 863–882 (2021)
16. Tikhonov, A., Samenko, I., Yamshchikov, I.P.: Storydb: Broad multi-language narrative dataset. arXiv preprint arXiv:2109.14396 (2021)
17. University of Oregon: DIBELS 8th Edition Administration and Scoring Guide. University of Oregon, Eugene, OR (2023), https://dibels.uoregon.edu/sites/default/files/2024-12/dibels8_admin_scoring_guide.pdf

18. Walkington, C., Bernacki, M.L.: Appraising research on personalized learning: Definitions, theoretical alignment, advancements, and future directions (2020)
19. Yelle, L.E.: The learning curve: Historical review and comprehensive survey. Decision sciences **10**(2), 302–328 (1979)
20. Zhao, Y., Chen, L., Cohan, A., Zhao, C.: TaPERA: Enhancing faithfulness and interpretability in long-form table QA by content planning and execution-based reasoning. In: Ku, L.W., Martins, A., Srikumar, V. (eds.) Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 12824–12840. Association for Computational Linguistics, Bangkok, Thailand (Aug 2024). https://doi.org/10.18653/v1/2024.acl-long.692, https://aclanthology.org/2024.acl-long.692/