CAT-AI: Supporting Teacher Workflows with AI-Assisted Exercise Creation

ANONYMOUS AUTHOR(S)

Creating classroom exercises consumes substantial teacher time, requiring educators to balance pedagogical soundness with student engagement while adapting materials for diverse learners. Through formative interviews with ten K-12 teachers, we observed that educators increasingly turn to AI tools yet struggle with prompt design, fragmented workflows across multiple applications, and significant verification overhead. From these interviews we synthesized a three-phase workflow model describing how teachers conceptualize, transform, and finalize educational content. Building on these insights, we present CAT-AI, an AI-assisted authoring system that embeds structured parameter specification, unified WYSIWYG editing, and transparent confidence indicators to support teachers throughout exercise creation without distrupting workflow or requiring prompt engineering expertise.

CCS Concepts: • **Human-centered computing → Human computer interaction (HCI)**; **Interactive systems and tools**; • **Applied computing → Education**.

Additional Key Words and Phrases: Education and Learning Technologies, Generative AI, Teacher-AI Collaboration, Human-Centered AI Design

ACM Reference Format:
Anonymous Author(s). 2026. CAT-AI: Supporting Teacher Workflows with AI-Assisted Exercise Creation. 1, 1 (January 2026), 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Educational content creation consumes substantial teacher time, with educators spending hours adapting materials for diverse learners while balancing pedagogical objectives, difficulty calibration, and student engagement [14, 20]. Studies reveal that teachers employ transformation-based approaches to exercise creation, using techniques like addition of supplementary materials, deletion of inappropriate content, and modification of difficulty levels [15, 17]. Traditionally, digital efforts to support teachers have focused on Intelligent Tutoring Systems (ITS) and related authoring environments, which enabled educators to design and train digital tutors without programming expertise [3, 4]. However, their primary focus was on the digitization of tutoring practices rather than on pedagogical content creation.

More recently, the emergence of AI-powered tools has shifted attention directly to educational content creation, supporting teachers in generating and adapting classroom materials. Early neural network approaches demonstrated technical feasibility for generating questions from textbooks [10, 18], and LLMs promised greater flexibility through prompt engineering [11], with validated capabilities now integrated into commercial platforms [6, 13].

Han et al. [8] provide empirical evidence that while teachers appreciate AI's ability to accelerate material preparation, they struggle with limited transparency and alignment with pedagogical intent. The majority of existing Generative AI systems treat trust as a post-generation validation step, leaving teachers uncertain about reliability and forcing them to adopt elaborate verification strategies [5, 7]. Beyond transparency, misalignment with pedagogical goals often derive from prompt engineering difficulties, as teachers must translate nuanced instructional intentions into textual instructions [16, 19]. Template-based systems like REDEEM helped non-programmer teachers create interactive

Manuscript submitted to ACM

content through structured interfaces [1, 2]. However, these systems assume teachers start from scratch, offering limited possibilities to reuse or adapt existing resources. On the contrary, field studies show teachers using AI-generated content as starting points, extensively modifying outputs for their specific contexts [8, 9]. This transformation-based approach represents teachers' natural workflow but cannot be easily expressed through current interfaces, which fragment generation, editing, and formatting into disconnected stages [12].

In this work, we examine how teachers engage with AI throughout their authoring workflow. We conducted formative interviews with 10 K-12 educators using semi-structured protocols and think-aloud observations. Teachers demonstrated their typical exercise creation workflows, revealing recurring challenges that cluster around four areas: difficulty translating pedagogical requirements into effective prompts; use of multiple disconnected tools creating substantial context-switching overhead; effort spent verifying AI-generated content without transparency about reliability; and inability to easily specify what to preserve versus modify when adapting existing materials.

These observations led us to formalize a three-phase workflow model. During conceptualization, teachers specify pedagogical parameters through structured fields rather than open-ended prompts, eliminating the need to translate teaching expertise into prompt engineering. During transformation, a unified WYSIWYG editor enables both direct text manipulation and targeted AI-assisted regeneration, with confidence indicators flagging potentially problematic content. During finalization, the system handles export formatting automatically, producing classroom-ready materials without requiring external tools. We present CAT-AI, a web-based system that operationalizes teachers' workflow through integrated support for each phase.

CAT-AI is available as open-source software with a live demonstration.[1]

2 Formative Study

To ground our design in observed pedagogical practices, we conducted formative interviews with ten K-12 teachers to understand their current workflows for creating educational exercises. We recruited teachers from primary and secondary schools covering grades 1 through 8, with teaching experience ranging from 3 to 40 years and varying levels of AI tool proficiency, balanced between primary (n=5) and middle school (n=5) teachers. Interviews followed a semi-structured protocol lasting approximately 20 minutes each, and participants demonstrated their typical workflows while sharing recently created materials.

We identified a common multi-tool workflow across participants (Figure 1). Teachers begin by identifying learning objectives from their curriculum and determining target difficulty levels based on their knowledge of student abilities. All interviewed teachers reported that what drive exercise design are learning objectives. Clarity in instructions is critical since ambiguous wording undermines the entire purpose of an exercise. Coherence emerged as equally important across: terminological coherence, requires consistent use of the same vocabulary introduced in class, structural coherence means following familiar exercise patterns, and visual coherence which demands consistent formatting conventions.

Difficulty calibration was universally cited as critical but context-dependent: primary teachers emphasized that all students should complete exercises successfully, while middle school teachers focused on testing thinking skills rather than mere knowledge recall.

The content generation phase exposes critical challenges in translating pedagogical intent into effective prompts. Teachers navigate to ChatGPT or similar AI tools and construct prompts through iterative refinement, yet the gap between pedagogical expertise and prompt engineering skills creates repeated cycles of adjustment. As one teacher

[1]Platform demo: https://catai.paperbackwriters.club:8443/; Code: https://repo.paperbackwriters.club:8443/code/catai.

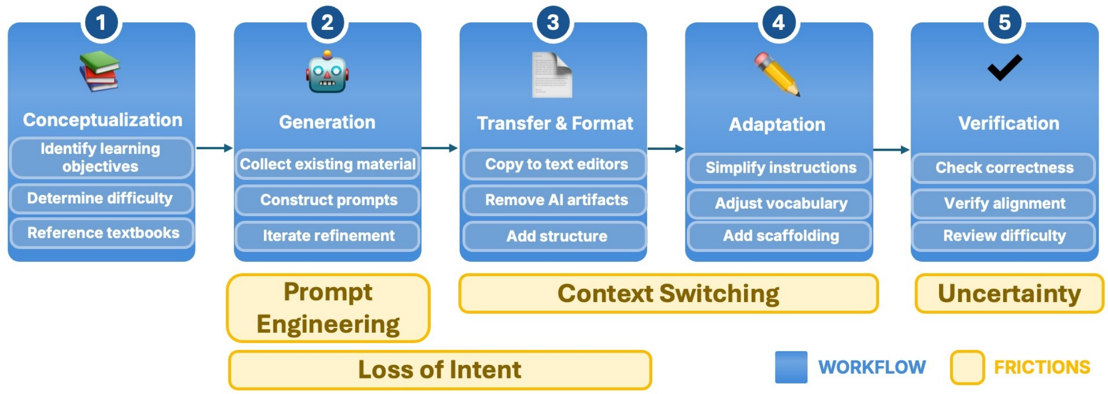

Fig. 1. Typical teacher workflow for AI-assisted exercise creation, showing fragmentation across multiple tools.

explained, “I have to be very specific about what I want, but even then the system frequently ignores these constraints” (T6).

Once content is generated, teachers manually transfer it into word processors, introducing workflow discontinuity. The adaptation phase reveals the inadequacy of unmodified AI-generated content, as teachers extensively modify materials by simplifying instructions, adjusting vocabulary, correcting errors, and adding visual elements. One participant captured this burden: “I spend nearly as much time checking and fixing generated content as I would creating from scratch” (T8).

Throughout the study teachers rarely create exercises from scratch but instead take existing materials and selectively modify them through vocabulary adjustment, content substitution, and difficulty modification. As T6 described, “I find an exercise that’s close to what I need, then change the parts that don’t fit.” Current AI tools fail to support this transformation-based workflow as they require teachers to articulate pedagogical requirements through prompt engineering, they operate through holistic regeneration that produces entirely new content, they provide no transparency about confidence or potential errors, forcing extensive manual verification; and they exist as disconnected applications that fragment the creation process across generation, editing, and formatting tools.

3 Design Goals

While generative models offer powerful content creation capabilities, formative interviews revealed evident frictions in how teachers currently use AI for exercise creation. The cognitive overhead of prompt engineering, workflow fragmentation, and verification burden often negate potential efficiency gains. These findings informed three design goals for CAT-AI.

Goal 1: Minimize Prompt Engineering Burden. Teachers possess deep pedagogical expertise, yet current tools require them to develop prompt engineering skills to leverage AI assistance. Rather than asking teachers to translate requirements into natural language prompts, the system should directly capture dimensions teachers naturally consider: learning objectives, prerequisites, school level, and difficulty calibration. The dual-mode generation approach, allowing either file-based reference or manual specification, emerged from interviews where teachers rarely create from scratch but work from existing materials.

Goal 2: Integrate Generation, Customization, and Export. The observed workflow fragmentation creates significant cognitive overhead as teachers move between ChatGPT for generation, Word for formatting, and various verification steps. A unified environment should support WYSIWYG editing where teachers see content in its final printed form, provide granular control over AI-assisted modifications allowing selective regeneration while preserving context, and handle export formatting automatically to produce classroom-ready materials without manual reformatting.

Goal 3: Provide Transparency and Control. Teachers' ambivalence toward AI-generated content stems from opacity: they cannot assess reliability without extensive verification, and regeneration produces entirely different content rather than targeted changes. Confidence indicators should flag potentially problematic content, directing teacher attention to elements requiring scrutiny. Multiple regeneration modalities, from complete exercise regeneration to element-level adjustments, support teachers' transformation-based workflow by distinguishing what should change from what must remain constant.

4 CAT-AI System

We implemented CAT-AI as a full-stack web application that translates our design goals into a functional system supporting teachers' three-phase workflow (Figure 2). The frontend uses React 18 with Material-UI components, while the backend uses Node.js with Express, integrating with OpenAI's GPT-4o for content generation and verification.

The workflow begins with a **conceptualization phase** where teachers specify pedagogical parameters without requiring prompt construction (Figure 2, top row). From the welcome screen, teachers select between file-based reference or manual entry modes. In file mode, the system accepts PDF or image uploads as shown in the file upload panel, including screenshots or photos of existing exercises, extracts text content server-side using pdf-parse for PDFs or OCR for images, and automatically infers candidate learning objectives, prerequisites, school level, and grade through an initial LLM API call. Teachers review and modify these extracted parameters before generation proceeds. Manual mode presents the structured input forms visible in the manual input panel, directly requesting pedagogical dimensions such as school level, grade, learning objectives as tagged entries, prerequisites, and optional features like worked examples or student reminders. Both modes converge to a common generation pipeline through the pedagogical settings interface.

Generation settings including difficulty level, number of exercises, vocabulary tone, language, and visual style apply globally across sessions through a persistent settings modal. This separation between exercise-specific parameters, representing what to teach, and stylistic preferences, representing how to present, reduces cognitive load during individual exercise creation. The generation process involves two sequential API calls. The first generates exercise and solution content structured as JSON arrays of text elements, each with content, positioning, dimensions, and styling attributes; the second call analyzes this generated content for logical or mathematical errors, assigning confidence scores and explanatory notes to each element. This dual-pass approach is designed to mitigate the reliability concerns raised by teachers, as the system explicitly flags uncertain content rather than presenting all output with equal authority.

Once content is generated, teachers enter the **transformation phase**. Through a WYSIWYG editor they directly interact with exercise content in its final printed form (Figure 2, middle row). The direct manipulation interface eliminates the abstraction gap between editing and output that characterizes the ChatGPT and Microsoft Word workflows observed in our formative study. As visible in the editor interface, exercise and solution occupy separate A4-sized canvases where all content appears as draggable, resizable text or image elements.

The editor supports three distinct modification modalities. Manual editing allows direct text modification through double-click interaction, drag-and-drop repositioning with alignment guides, element resizing within page boundaries,

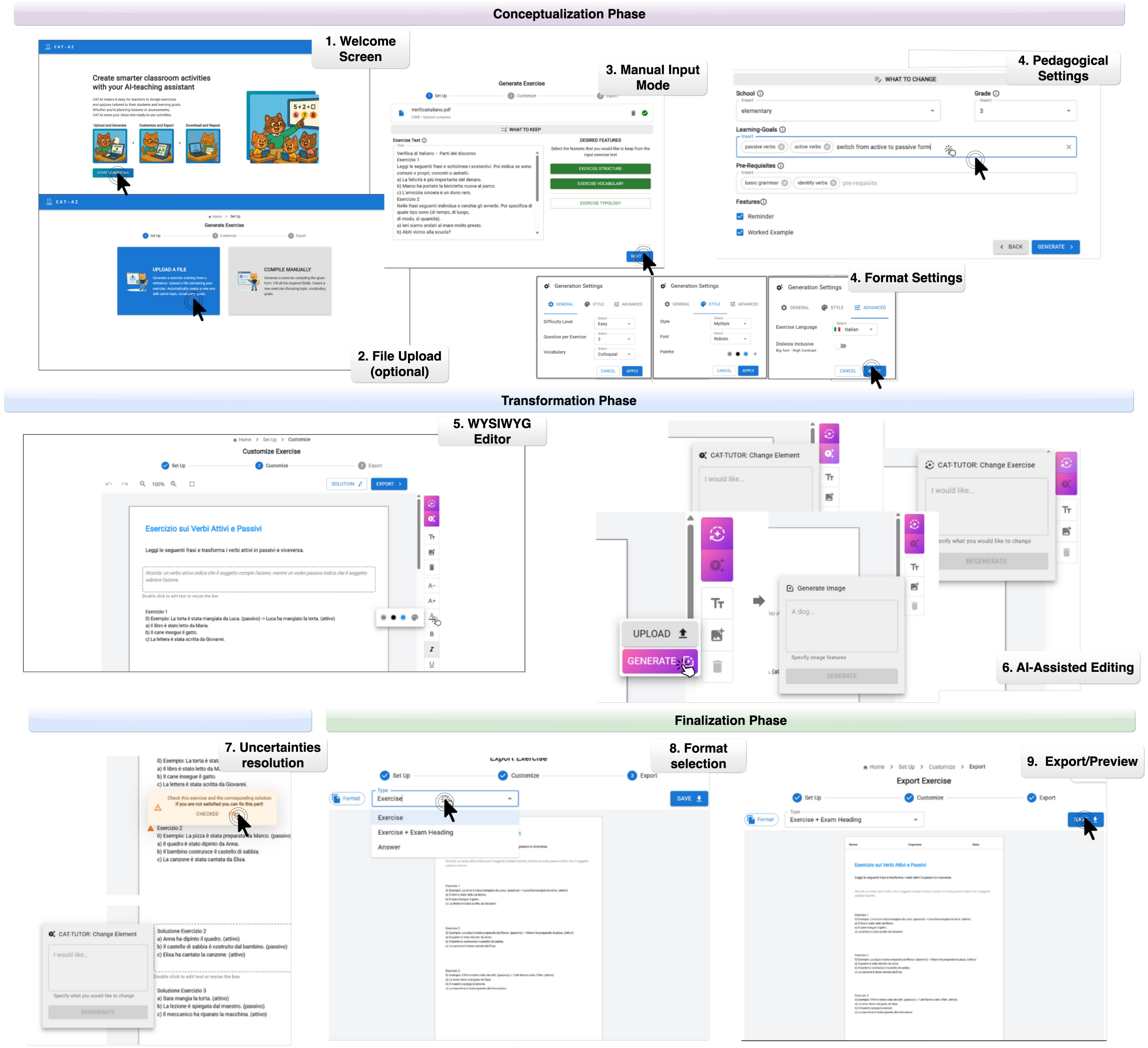

Fig. 2. Complete user interaction flow through CAT-AI's three-phase workflow. The conceptualization phase (top) guides structured parameter specification through welcome screen, optional file upload, manual input, and pedagogical settings. The transformation phase (middle) provides uncertainty resolution, AI-assisted editing, and a WYSIWYG editor interface. The finalization phase (bottom) handles format selection and export with preview.

and styling adjustments including font size, color, and emphasis. The AI-assisted editing panel enables element-level regeneration where teachers select any text element and request modifications through natural language prompts, with the system preserving all other content while regenerating only the specified element. Exercise-level AI regeneration processes natural language requests to add, modify, or remove content across the entire exercise or solution, supporting the transformation-based workflow where teachers make targeted changes to mostly-satisfactory content. The uncertainty resolution interface, shown in the middle row, presents flagged content with confidence indicators and one-click "Fix" options that trigger targeted regeneration.

Additional AI-assisted features include solution recalculation that regenerates solutions when exercise content changes to maintain consistency, and image generation through DALL-E 2 with text prompts. Each AI operation maintains context from the current exercise state, ensuring modifications remain coherent with existing content rather than producing disconnected alternatives. The implementation stores all element data in React Context as structured objects containing position, dimensions, content, styling, confidence scores, and unique identifiers, enabling immediate UI updates and undo/redo functionality.

The **finalization phase** addresses the formatting burden observed in current workflows where teachers spend considerable time reformatting content for distribution (Figure 2, bottom row). Through the format selection interface, teachers preview their complete exercise in standard A4 layout and select from predefined formats: exercise only, exercise with exam header including name, date, and class fields, or solution only. The export panel converts the React component representation to PDF server-side, handling page margins, element positioning, and formatting automatically. Generated PDFs become immediately available for download without additional processing, eliminating the manual reformatting work that fragments teachers' current workflows.

The system architecture separates presentation logic from content generation and file management. The frontend employs React 18[2] for component architecture, Material-UI[3] for interface components, and react-draggable[4] for element positioning. PDF generation uses html2canvas[5] and jsPDF[6]. The backend runs on Node.js[7] with Express[8] for routing, Multer[9] for PDF text extraction, and Tesseract.js[10] for OCR on image uploads. Content generation and verification use OpenAI's GPT-4o[11], while image generation uses DALL-E 2[12].

5 Conclusions

CAT-AI represents an attempt to align AI assistance with teachers' natural exercise creation practices instead of requiring educators to adapt to general-purpose tools. Several directions for future work emerge from this design. We plan to conduct comparative evaluations with teachers using both CAT-AI and their current practices to measure whether the structured approach reduces cognitive load and improves content quality. Longitudinal deployment would reveal how teachers integrate the system into regular practice and whether initial benefits persist over time. The confidence indicator mechanism raises broader questions about AI transparency in educational contexts; future versions might incorporate formal knowledge representations such as curriculum standards to provide more pedagogically grounded validation. Detecting and correcting for potential biases in AI-generated educational content represents another important direction, as exercises should reflect diverse perspectives and avoid perpetuating stereotypes. Cross-cultural evaluation spanning different countries and school systems would clarify which design principles generalize and which require contextual adaptation. Finding the right balance between structured guidance and creative exploration remains an open challenge, as highly structured interfaces may constrain the discovery of novel exercise formats while still honoring teacher specifications.

[2]https://react.dev
[3]https://mui.com
[4]https://github.com/react-grid-layout/react-draggable
[5]https://html2canvas.hertzen.com
[6]https://github.com/parallax/jsPDF
[7]https://nodejs.org
[8]https://expressjs.com
[9]https://github.com/expressjs/multer
[10]https://tesseract.projectnaptha.com
[11]https://openai.com/index/hello-gpt-4o/
[12]https://openai.com/dall-e-2

References

[1] Shaaron Ainsworth, Nigel Major, Shirley Grimshaw, Mary Hayes, Jean Underwood, Ben Williams, and David Wood. 2003. REDEEM: Simple Intelligent Tutoring Systems from Usable Tools. In *Authoring Tools for Advanced Technology Learning Environments*, Tom Murray, Stephen B. Blessing, and Shaaron Ainsworth (Eds.). Springer, Dordrecht, 205–232. https://doi.org/10.1007/978-94-017-0819-7_8

[2] Shaaron E. Ainsworth, Shirley K. Grimshaw, and D. Jean Underwood. 1999. Teachers as designers: Using REDEEM to create ITSs for the classroom. *Computers & Education* 33, 2-3 (1999), 171–188.

[3] Vincent Aleven, Bruce M. McLaren, Jonathan Sewall, and Kenneth R. Koedinger. 2006. The Cognitive Tutor Authoring Tools (CTAT): Preliminary Evaluation of Efficiency Gains. In *Intelligent Tutoring Systems*. Springer, 61–70. https://doi.org/10.1007/11774303_7

[4] Vincent Aleven, Bruce M. McLaren, Jonathan Sewall, Martin van Velsen, Octav Popescu, Sandra Demi, Michael Ringenberg, and Kenneth R. Koedinger. 2016. Example-Tracing Tutors: Intelligent Tutor Development for Non-Programmers. *Int. J. Artif. Intell. Educ.* 26, 1 (2016), 224–269. https://doi.org/10.1007/s40593-015-0088-2

[5] Deepak Varuvel Dennison, Bakhtawar Ahtisham, Kavyansh Chourasia, Nirmit Arora, Rahul Singh, Rene F. Kizilcec, Akshay Nambi, Tanuja Ganu, and Aditya Vashistha. 2025. Teacher–AI Collaboration for Curating and Customizing Lesson Plans in Low-Resource Schools. arXiv:2507.00456 [cs.CY].

[6] Kristen DiCerbo. 2023. Khan Academy explores the potential for GPT-4 in a limited pilot program. (2023). https://openai.com/customer-stories/khan-academy.

[7] Jaimie Drozdal, Justin D Weisz, Dakuo Wang, Gaurav Dass, Bingsheng Yao, Changruo Zhao, Michael Muller, Lin Ju, and Hui Su. 2020. Trust in AutoML: Exploring Information Needs for Establishing Trust in Automated Machine Learning Systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 297–307. https://doi.org/10.1145/3377325.3377501

[8] Yueqi Han, Yuling Zhou, Hanqi Cai, et al. 2024. Teachers, Parents, and Students' Perspectives on Integrating Generative AI into Elementary Literacy Education. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3613904.3642438

[9] Tae Soo Kim, Yoonjoo Lee, Jamin Shin, Young-Ho Kim, and Juho Kim. 2024. EvalLM: Interactive Evaluation of Large Language Model Prompts on User-Defined Criteria. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–21. https://doi.org/10.1145/3613904.3642216

[10] Ekaterina Kochmar, Dung Do Vu, Robert Belfer, Varun Gupta, Iulian Vlad Serban, and Joelle Pineau. 2021. Automated Data-Driven Generation of Personalized Pedagogical Interventions in Intelligent Tutoring Systems. *International Journal of Artificial Intelligence in Education* 32 (2021), 323–349. https://doi.org/10.1007/s40593-021-00267-x

[11] Jaewook Lee and Sanghoon Kim. 2024. Leveraging ChatGPT for Adaptive Learning through Personalized Prompt-based Instruction: A CS1 Education Case Study. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3613905.3637148

[12] Xinyi Lu, Simin Fan, Jessica Houghton, Lu Wang, and Xu Wang. 2023. ReadingQuizMaker: A Human-NLP Collaborative System that Supports Instructors to Design High-Quality Reading Quiz Questions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–18. https://doi.org/10.1145/3544548.3580957

[13] MagicSchool ai. 2023. AI for teachers - lesson planning and more! (2023). https://www.magicschool.ai/.

[14] Jason K McDonald. 2021. The Everydayness of Instructional Design and the Pursuit of Quality in Online Courses. *Online Learning* 25, 4 (2021), 156–173.

[15] Ni Nyoman Padmadewi and Luh Putu Artini. 2024. Textbook Adaptation Techniques in a Technology-Integrated Environment by an Indonesian EFL Teacher. *TEFLIN Journal* 35, 1 (2024), 89–107.

[16] Auste Simkute, Viktor Kewenig, Abigail Sellen, Sean Rintel, and Lev Tankelevitch. 2025. The New Calculator? Practices, Norms, and Implications of Generative AI in Higher Education. *arXiv preprint arXiv:2501.08864* (2025).

[17] Brian Tomlinson. 2022. Theorising Textbook Adaptation in English Language Teaching. *Innovation in Language Learning and Teaching* 16, 3 (2022), 218–231.

[18] Zichao Wang, Andrew S Lan, Weili Nie, Andrew E Waters, Phillip J Grimaldi, and Richard G Baraniuk. 2018. QG-net: A Data-Driven Question Generation Model for Educational Content. In *Proceedings of the Fifth Annual ACM Conference on Learning at Scale*. 1–10. https://doi.org/10.1145/3231644.3231654

[19] J.D. Zamfirescu-Pereira, Richmond Y. Wong, Bjoern Hartmann, and Qian Yang. 2023. Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–21. https://doi.org/10.1145/3544548.3581388

[20] Xiaofei Zhou, Jingwan Tang, Beilei Guo, Hanjia Lyu, and Zhen Bai. 2022. Challenges and Design Opportunities in Data Analysis for ML-Empowered Scientific Inquiry - Insights from a Teacher Professional Development Study. In *Proceedings of the International Society of the Learning Sciences Annual Meeting*. 847–854.