More than Model Documentation: Uncovering Teachers' Bespoke Information Needs for Informed Classroom Integration of ChatGPT

Mei Tan
Stanford University
USA
mxtan@stanford.edu

Hari Subramonyam
Stanford University
USA
harihars@stanford.edu

ABSTRACT

ChatGPT has entered classrooms, circumventing typical training and vetting procedures. Unlike other educational technologies, it placed teachers in direct contact with the versatility of generative AI. Consequently, teachers are urgently tasked to assess its capabilities to inform their use of ChatGPT. However, it is unclear what support teachers have and need and whether existing documentation, such as model cards, provides adequate direction for educators in this new paradigm. By interviewing 22 middle- and high-school ELA and Social Studies teachers, we connect the discourse on AI transparency and documentation with educational technology integration, highlighting the information needs of teachers. Our findings reveal that teachers confront significant information gaps, lacking clarity on exploring ChatGPT's capabilities for bespoke learning tasks and ensuring its fit with the needs of diverse learners. As a solution, we propose a framework for interactive model documentation that empowers teachers to navigate the interplay between pedagogical and technical knowledge.

CCS CONCEPTS

• **Human-centered computing** → *HCI theory, concepts and models*; • **Applied computing** → *Education*; • **Computing methodologies** → *Natural language processing*;

KEYWORDS

large language models, chatgpt, AI in education, machine learning documentation

1 INTRODUCTION

ChatGPT [82], a conversational artificial intelligence interface using the latest advancements in natural language processing, has presented *immediate* and *significant* impacts on the education domain. Initial concerns with ChatGPT centered around issues of academic integrity and plagiarism and their threat to conventional forms of educational assessment [104]. Consequently, school districts responded by blocking access to the website from their schools' networks [94]. However, the ineffectiveness of such bans and their subsequent repeals [100] pushed the narrative among educators to evolve and consider opportunities to positively apply it to teaching and learning. Across teacher-facing websites, there is a surge in ad-hoc high-level suggestions on leveraging ChatGPT to assist teachers by generating lesson plans, reading passages, assessment questions, models of student work, etc. [103]. At the same time, researchers, administrators, and media narratives advise teachers of ChatGPT's risks and limitations, cautioning them to adopt a thoughtful, well-researched, and purposeful approach to integrating ChatGPT in educational settings–i.e., *informed classroom integration.*

Educators are *tasked* to test ChatGPT's capabilities to determine potential effects on student learning [72, 92, 100], devote time to oversight and prevention of misuse, and devise a clear strategy for facilitating student use of the tool [42]. Such an evaluation necessitates that teachers develop competencies to understand the technology, recognize its vulnerabilities, and adeptly leverage its capabilities in alignment with pedagogical goals [53]. However, this is easier said than done. Unfortunately, educators and educational institutions often lack the knowledge or expertise to effectively integrate new technologies in their teaching [11, 43]. Teachers lack training in technology and technical support [35, 48]. The act of integrating technology is time-consuming and interrupts instruction, requiring additional planning and new routines to implement technology-integrated lessons [62].

Compared with the integration of conventional educational technology, that of ChatGPT introduces more technical complexities and fewer facilitating resources. To integrate conventional tools, educators and school administrators may work with the product or service companies to assess infrastructural needs, compatibility with existing systems, and cost [91]. They may even engage in a pilot testing phase to ensure its alignment with educational objectives while emphasizing ethical considerations, particularly concerning student data and privacy [34]. During this process, some companies provide formal training, documentation, continued technology support, and established processes for troubleshooting failures. Though

such technical supports are not universal across products, the sudden accessibility of ChatGPT circumvented any typical training and vetting procedures.

Further, while conventional educational software is designed to perform well-defined tasks in specified use contexts, ChatGPT is general-purpose; it can "write stories, give life advice, even compose poems and code computer programs [28]." Teachers and students are placed in direct contact with the versatility and customizability of generative large language models. Emergent behaviors of these models present new challenges for fairness, accountability, and transparency [115]. Documentation from OpenAI[1] has disclosed the risks of ChatGPT, including its propensity to produce false information and content that perpetuates harmful biases and stereotypes [81, 83]. However, established frameworks for AI transparency [76] cater predominantly to technical experts, emphasizing the technology's capabilities and overlooking the practical challenges and contexts faced by practitioners [17, 60]. As a result, documentation may fail to support teachers' abilities to grasp ChatGPT's underlying intricacies, discern valuable uses for its features, and navigate its potential pitfalls.

In this work, we investigate teachers' information needs regarding the integration of ChatGPT and ask the following research questions:

- RQ1: How do teachers seek and evaluate available information about ChatGPT to motivate integration decisions?
- RQ2: How do teachers assess the technical capabilities of ChatGPT for pedagogical utility?
- RQ3: How do existing AI transparency and documentation frameworks fall short in addressing the information needs of teachers for the effective integration of ChatGPT into classroom settings?

We interview experienced teachers to understand their evaluation of available information, their approach to information seeking, and their strategies for capability testing in the ChatGPT interface. In doing so, we identify critical information gaps limiting effective integration. Our findings reveal that teachers face challenges discovering applications of ChatGPT relevant to their pedagogical goals, struggling to contextualize its technical affordances and limitations in classroom practice. Teachers lack competencies in prompt engineering and approach experimentation with faulty mental models and narrowly defined goals, resulting in flawed outputs that limit perceptions of ChatGPT's utility in education. We highlight the gaps in existing documentation practices that inadequately address these challenges and present arguments for experimentation-based documentation tools that support dynamic and interactive understanding. Our key contributions include taxonomies of teachers' information needs and experimentation behaviors, as well as a proposed framework to support teachers in exploring—rather than reading about—the technical capabilities of ChatGPT.

2 RELATED WORK

2.1 Information Needs in Educational Technology Adoption

Though the notion of technology integration is multi-dimensional and operationalized in research through numerous instruments [27, 39], perhaps the most prominent model of technology integration in education is the TPACK framework [44]. TPACK provides an explicit mechanism for discussing technology integration in service of teaching and learning [10]. It models the complex interactions and intersections of technological (TK), pedagogical (PK), and content knowledge (CK) [59]. Teachers' knowledge of teaching with technology involves knowledge about standard technologies and how to operate them, how teaching may change as a result of using a particular technology, how technology and content are reciprocally related, and how technologies can support and help overcome problems in the process of teaching subject matter [58]. Measures of teachers' TPACK knowledge assess their abilities to "play around" with a variety of technologies relevant to their subject field, choose technologies that enhance the content of a lesson, and adapt the use of technologies to different teaching activities [98]. Extensions of TPACK additionally address the role of contextual knowledge for adapting technology use to the needs of individual students and classrooms [72].

Studies applying TPACK have found that teachers struggle to develop sufficient pedagogical content knowledge prior to technology integration [86], leading to inadequate consideration of curricular issues. Teachers' rationales for technology adoption are often disconnected from improving content understanding [38]. This discrepancy constrains teachers' design of technology-integrated activities and perpetuates barriers to integration [52].

New waves of technological developments, including the emergence of artificial intelligence, place additional demands on teachers' technology-related knowledge, skills, and attitudes [6, 67, 99]. Researchers have called for teachers to develop new literacies to understand the affordances of AI-based tools [53] and pedagogical strategies for their integration [23, 119]. Such technical competencies also involve evaluating AI in terms of its limitations and ethical issues [23, 55, 75], enabling practitioners to interpret documents explaining the behaviors of systems, discover and prevent bias, and demand accountability for fairness and transparency [16]. Extensions of TPACK in light of advances in generative AI underscore technical knowledge about how generative AI can be used to address common pedagogical challenges, enable new approaches to assessment, and develop appropriate supports for learners [75]. To this end, teachers require a working understanding of how algorithmic factors impact generated outputs, skills to expertly prompt models to generate the most helpful content, and the ability to corroborate the accuracy of generated content [78, 108].

2.2 Large Language Models in Education

Large language models, including ChatGPT [109], have presented novel opportunities and disruptions for practitioners and researchers in education. These models have shown versatility in a range of language tasks, from translation and question answering to writing

---

[1]Since the release of ChatGPT in November 2022, available documentation consisted of the GPT-3 model card [81] and a since-deprecated page of notes for educators that introduced potential opportunities for applications in education and documented similar disclosures to those in the model card. In August 2023 OpenAI replaced this page with the release of its Teaching with AI guide [85] and FAQ [83].

essays and computer programs [53]. Researchers have applied large language models to specific pedagogical tasks, including generating assessment questions [32], code explanations [68], and feedback [51]. Though long-term impacts on student learning remain to be seen, a growing body of work discusses the potential for ChatGPT to facilitate teaching by generating assessments [54] and activity templates [110], scaffolding students' writing assignments [13], and serving as a Socratic think-aloud partner for content understanding [33] or conversational partner for language practice [50].

At the same time, studies have underscored the pedagogical shortcomings of ChatGPT by comparing its outputs with human-authored items in remediating math mistakes [113], providing students with algebra tutoring hints [87] and formative feedback [89], and providing teachers with instructional coaching insights [41, 112]. In math education, language models have been found to ineffectively identify misconceptions [114], make numeric mistakes [79], and offer incorrect solution steps [87]. For writing instruction, they have been found to provide abstract and generic essay feedback without concrete suggestions for improvement [118]. Students have found ChatGPT's writing suggestions lackluster - grammatical but generic, and suppressing students' own voices in writing [95].

Indeed, language models were not designed to supplant the roles of teachers [26, 40]. Support from educators is critical to adapting, contextualizing, and authenticating the content produced by ChatGPT [40]. Yet, designing tools and classroom practices for complementarity between teachers and AI remains a persistent challenge [47, 96]. As AI-based technologies increasingly perform academic and instructional tasks [107], they recalibrate classroom roles and expectations. Specific to ChatGPT, it remains unclear how teachers navigate and make pedagogical advantage of such open-ended generative systems [67, 119]. Though research continues to update the mapping of capabilities and applications of language models in education, the cognitive challenges that underlay teachers' own sense-making of whether and how to integrate ChatGPT persist.

2.3 Technology Documentation

Despite calls for equipping teachers with technical knowledge about AI [16, 46, 78], teachers are provided limited information and continue to face difficulties understanding AI behaviors [22]. Though documentation is a central mechanism for ensuring the fair and accountable application of AI in practice, recent research has shown that current documentation guidelines do not provide sufficient guidance for practitioners [60]. Insufficient documentation is an essential barrier to the adoption of AI [61]. While some non-adopters are cautious of the risks of limited explainability, others perceive documentation as a burden to consume.

Researchers have argued that documentation for software and AI are subject to different requirements [60]. Software documentation describes what a software system does, how it operates, and how it should be used [4]. Documents include technical reference guides for developers and auditors and user manuals for practitioners. Though such documentation is essential to validate the proper functioning of software tools, studies have found problems negatively impacting their quality and usefulness, including incomplete

or out-of-date information, missing definitions, low readability, and technical jargon [3, 5, 120].

A smaller body of work is dedicated to the development of AI documentation and reporting frameworks to make AI systems more transparent and to help users establish trust. Researchers have proposed factsheets [8], datasheets [14, 36, 45], and model cards [76] to specify the content of AI documentation, including purpose, performance metrics, training data composition, and disclosures of limitations. Though model cards were originally proposed to accommodate a spectrum of stakeholders [76], their highly technical text-based format leaves many without AI expertise underserved [30]. As studies have emphasized the importance of practitioner interaction with AI models and their data [7, 30], companies such as HuggingFace [1] and Google Cloud [2] have implemented interactive modalities for assessing model performance, though still targeting a technical audience.

The lack of research evaluating the degree to which existing AI documentation is understandable and applicable for end-users [88] is expected. Conventionally, end-users interact with software documentation, while AI documentation is commonly considered the domain of technical developers. However, the release of ChatGPT as an open-ended interface that functions as a thin wrapper around a model API situates end-users, in this case, as direct consumers of an AI model. To support the effective integration of ChatGPT tools in education, researchers must investigate the delta between available AI documentation frameworks and teachers' contextual information needs.

3 METHOD

To investigate how teachers approach information gathering and assessment of ChatGPT for classroom integration, we conducted semi-structured interviews with experienced teachers. Each interview consisted of discussions around participants' experiences with general technology adoption and the appearance of ChatGPT in education contexts, an evaluation of available transparency documentation regarding the technology behind ChatGPT, and experimentation with the ChatGPT interface. Interviews were held virtually via Zoom and lasted between 90 and 120 minutes. Participation was voluntary, and all participants were compensated with $100 for their involvement. Our institution's IRB approved the study.

3.1 Participants

We recruited experienced teachers through the alumni networks of the Stanford Teacher Education Program and the Hollyhock Fellowship Program. We also reached out to school districts engaged in long-standing partnerships with our university. Due to the central functionality of text generation in the ChatGPT interface, we recruited middle- and high-school teachers for whom student writing is central to their instructional practice. We selected 22 participants meeting our selection criteria by order of sign-up. The participants included 9 teachers with experience teaching at the middle school level and 19 teachers with experience teaching at both middle and high school levels. Ten teachers had between two and five years of experience, nine teachers had between six and ten years of experience, and three teachers had more than ten years of experience.

Phase 1: Open	**Experience Sharing**	**Open Exploration**	**Model Card Evaluation**
	Semi-Structured Interview Teachers describe their adoption experiences for ChatGPT and other educational technologies. Teachers answered questions about their information seeking and tool-evaluation practices and the role of administrators and students in technology integration processes.	*Think-Aloud Activity* Teachers discuss their reasoning while exploring ChatGPT capabilities to inform integration decision-making. Teachers share their browser screens and were allowed to search for new information and/or interact with the ChatGPT interface.	*Semi-Structured Interview* Teachers read the GPT-3 model card and reflect on the degree to which the disclosures are applicable, understandable, or actionable in classroom practice. Teachers discuss outstanding information needs unaddressed by the model card.

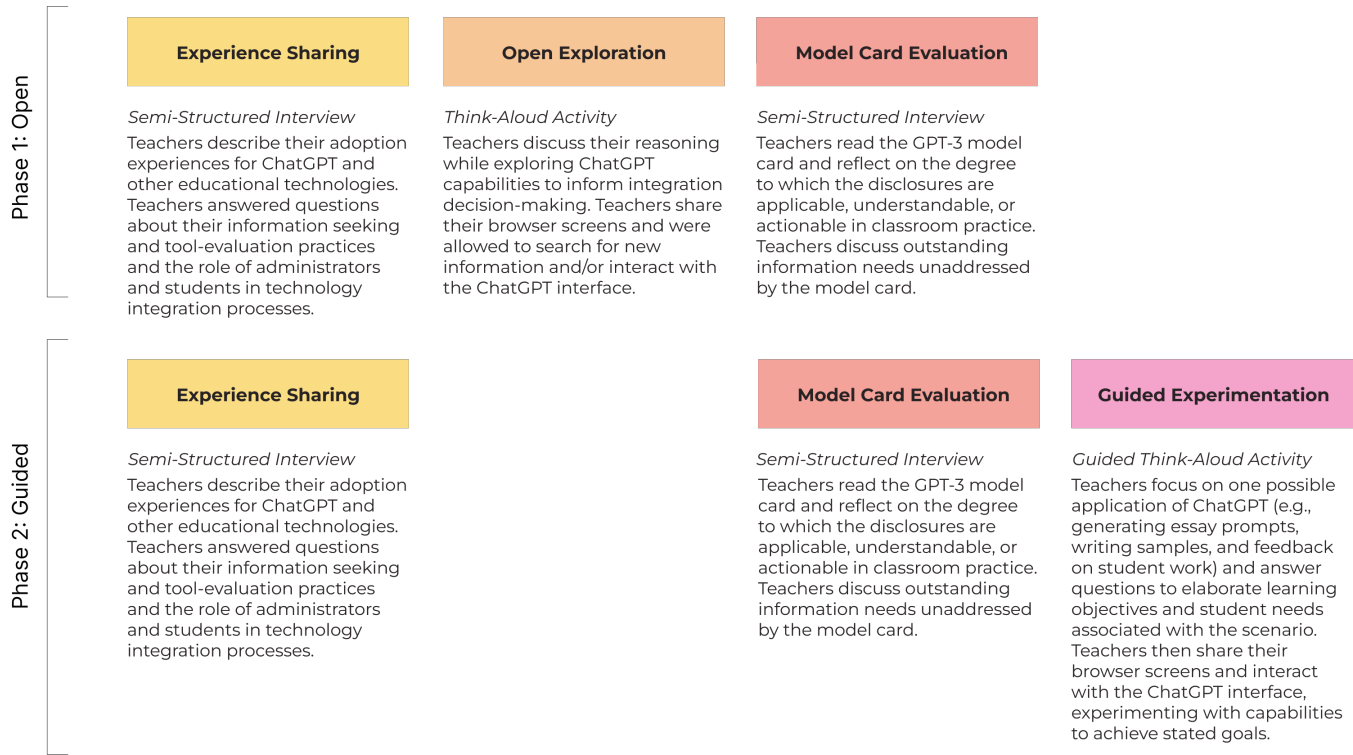| Phase 2: Guided	**Experience Sharing**		**Model Card Evaluation**	**Guided Experimentation**
	Semi-Structured Interview Teachers describe their adoption experiences for ChatGPT and other educational technologies. Teachers answered questions about their information seeking and tool-evaluation practices and the role of administrators and students in technology integration processes.		*Semi-Structured Interview* Teachers read the GPT-3 model card and reflect on the degree to which the disclosures are applicable, understandable, or actionable in classroom practice. Teachers discuss outstanding information needs unaddressed by the model card.	*Guided Think-Aloud Activity* Teachers focus on one possible application of ChatGPT (e.g., generating essay prompts, writing samples, and feedback on student work) and answer questions to elaborate learning objectives and student needs associated with the scenario. Teachers then share their browser screens and interact with the ChatGPT interface, experimenting with capabilities to achieve stated goals.

Figure 1: Components of our interview protocols across two phases. Both phases engaged teachers in experience sharing and model card evaluation, but the second phase replaced open exploration with a guided experimentation think-aloud activity.

Their teaching experiences involved a variety of writing-intensive subjects, including English Language Arts, Social Studies, History, and Humanities.

3.2 Protocol

In line with the iterative nature of qualitative research, our interview protocol underwent periodic revisions as the study progressed [25, 29]. This was essential to ensure that the questions being posed were not only reflective of our evolving understanding of the research context but also responsive to the emerging needs and insights from the participants. Initial interviews were guided by a preliminary set of questions developed from literature and involved open-ended exploration (phase 1, with 14 participants). However, as patterns began to emerge and new themes were identified, the protocol was adjusted to delve deeper into specific areas of interest and to better support participant engagement (phase 2, with 8 participants). All participants engaged in sharing their initial experiences with ChatGPT and evaluating model card information. Here we discuss each of these steps in detail.

3.2.1 Experience Sharing. The semi-structured interviews started with a discussion of participants' experiences with classroom technology integration. The interviewer first asked participants to share contextual factors, such as hardware distribution and access, administrative oversight and decision-making, information channels, and professional development opportunities. Then, teachers were asked to reflect on the information they seek and the criteria they

use to assess whether a technology application is supportive of student learning and how they would approach incorporating the application into their instructional practice. The interview questions then shifted to focus on AI applications and ChatGPT. The interviewer asked teachers to discuss student, teacher, and administrative responses to the emergence of ChatGPT and probed about perceived opportunities and difficulties. Teachers answered questions about the information and narratives they have consumed regarding ChatGPT and their degree of personal interaction with the tool.

3.2.2 Model Card Evaluation. The interviewer introduced the model card for GPT-3 [20, 81] as a technical document for the technology that ChatGPT is based on and explained that model cards [76] address transparency needs for machine learning models by documenting training data composition, performance accuracy, and limitations. We chose to evaluate model cards, a standardized documentation framework for AI models, due to ChatGPT's open-ended interface that positions users as direct consumers of the model, creating its specific functionality. In this way, the teacher directly using ChatGPT is no different from the developer incorporating its model into a software tool. Though end-users are not conventionally the primary audience of AI documentation, the design of ChatGPT has disrupted AI information standards and made model cards relevant for teachers. In other segments of the structured interviews, teachers were encouraged to seek their own conventionally user-facing information sources. Note that educator-specific documentation

Phase	Teacher ID	Years of Experience	Grade Levels	Subjects
Phase 1: Open	1	4	10-11	English Language Arts
	2	2	9	English Language and Composition
	3	2	7	English Language Arts, Social Studies
	4	3	9-12	Social Studies
	5	3	7-8, 12	English Language Arts
	6	19	7-12	English Language Arts
	7	9	7-12	English Language Arts
	8	6	9-12	English Language Arts, Humanities
	9	7	9-12	English Language Arts, AVID
	10	13	7-12	AP Government, World History
	11	7	7-12	English Language Arts
	14	9	7-8	English Language Arts
	15	4	7-8	English Language Arts
	16	9	9-12	Social Studies, Human Geography, U.S. History
Phase 2: Guided	12	3	9-11	English Language Arts
	13	5	9-12	AP Language and Composition
	17	12	9-12	English Language Arts, Humanities, Social Studies, AP U.S. History
	18	2	9-12	Humanities
	19	5	9-12	English Language Development
	20	6	10-11	English Language Arts
	21	8	7-12	English Language Arts, ESL, Reading Intervention, Special Education
	22	6	9-12	English Language Development

Table 1: Each interview is listed by teacher ID with years of experience, grade levels, and subjects taught. Phase indicates iteration of the experimentation portion of the protocol, where some interviews engaged teachers in open exploration of ChatGPT and its potential classroom applications while others guided teachers to assess ChatGPT in the context of a specific use case.

from OpenAI was not yet available at the time of interviews (e.g., the Teaching with AI guide [85] and Educator FAQ [83]); however, these resources adapt a subset of technical disclosures described in model documentation and otherwise function as software documentation that defines a few example use cases. We discuss their limitations in Section 5. Teachers were primed to expect the "technical language" and given time to read the document. They were subsequently asked to reflect on the information presented. The interviewer prompted teachers to identify the most meaningful disclosures from the model card for the education domain and explain their reasoning for how the information would affect their classroom practice. Teachers were then asked to identify the disclosures that are not applicable, understandable, or actionable for the education domain and discuss outstanding information needs not addressed by the model card.

3.2.3 Phase 1: Open Exploration. In the first 14 interviews, teachers were asked to engage in open exploration and information-seeking regarding ChatGPT. The interviewer described the motivation for the task: district administrators are seeking teacher perspectives to inform the design of policies overseeing the use of ChatGPT. Teachers were asked to determine whether and how they would integrate the use of ChatGPT in their classroom practice in preparation for responding to the district request with a well-reasoned report. They were asked to share their browser screens and narrate their thought processes as they searched for necessary information

and interacted with the ChatGPT interface. During the process, the interviewer encouraged teachers to lead the exploration and only engaged the teachers in discussion about the quality of ChatGPT responses and answered questions to facilitate teachers' interactions. After observing the challenges teachers faced while independently directing exploration with ChatGPT, in the final eight interviews, we followed an adapted protocol that omitted the open exploration task and instead introduced a guided experimentation task (Phase 2 sessions).

3.2.4 Phase 2: Guided Experimentation. For the final eight interviews, a guided experimentation task followed model card evaluation. Teachers were presented with a list of possible teacher-facing and student-facing applications of ChatGPT that teachers had suggested in previous interviews (e.g., generating essay prompts, writing samples, and feedback on student work). After instructing teachers to select one application as a focal scenario for the experimentation activity, the interviewer led teachers through a series of questions eliciting the student needs and learning goals associated with the scenario. Teachers were asked to describe in detail their approach to the scenario without ChatGPT, including differentiated learning needs, pedagogical considerations, evaluation criteria, and associated classroom management needs. Next, the interviewer instructed teachers to consider how ChatGPT could be used to support the learning scenario. They were asked to share their browser screens and narrate their thought process as they

interacted with the ChatGPT interface, using the needs elicited through the learning context discussion to guide experimentation. During the experimentation process, the interviewer guided teachers to test ChatGPT capabilities related to their stated learning needs.

3.2.5 Debrief. The interviews concluded with a series of debrief questions related to the perceived overuse, misuse, and harms of integrating ChatGPT in the classroom, based on what teachers had gathered from documentation and experimentation. The interviewer prompted teachers to articulate the remaining gaps in information and support required to minimize the risk of harm.

3.3 Analysis

The interview transcripts include question-answering and think-aloud data. We first manually verified the text of the automated transcriptions from Zoom against the video recordings, then conducted inductive qualitative coding in Atlas.ti [37] using a grounded theory approach [105] beginning with in-vivo analysis. During coding, we also had the video synchronized with the transcript and added screenshots of prompts from the video into the transcript. The two authors independently open-coded a common set of three interviews and collaboratively developed an initial codebook, which the authors applied to analyze the remaining transcripts. The coding scheme included references to information seeking (sources, framing, attitudes, mental models, gaps, etc.), experimentation (prompting strategies, response validation, etc.), and integration decision-making (support needs, trade-offs, output characteristics, etc.). Throughout the coding process, the authors wrote reflective memos describing emerging themes and making connections across interviews. The authors then discussed codes and memos, grouping codes to identify higher-level themes and synthesizing findings.

3.4 Positionality

We acknowledge that our research perspectives and approaches are shaped by our own experiences and positionality. Specifically, we are researchers living and working from a school of education in the U.S., with teaching experience and experiences working with school teachers and district personnel on technology integration. We come from a mix of education, computer science, and HCI backgrounds, which we have drawn on to conduct prior research on the fairness of AI in education and sociotechnical approaches to human-centered AI design practices. The pressure to develop new competencies and discover informed, responsible, and pedagogically relevant ways to integrate ChatGPT has been a constant theme of our informal conversations with educators and administrators since its release.

4 FINDINGS

In each interview, we engaged teachers in discussions and activities to assess their information-seeking and experimentation approaches to ChatGPT. In doing so, we identify information gaps, experimentation goals and strategies, and barriers to integration. We summarize our study findings in terms of (1) information seeking and comprehension, (2) capability exploration and experimentation, and (3) classroom integration support needs.

4.1 Information Seeking and Comprehension

We observed teachers' information-seeking processes and examined how teachers made sense of available information regarding ChatGPT. Through our discussions about their prior experiences, we discovered that the adoption process for ChatGPT significantly diverged from typical educational technologies, marked by a scarcity of practical guidance and unconventional avenues of information access. Teachers reported disengagement when sifting through web articles and static documentation, citing frustrations due to time constraints, and the irrelevance of the available information to classroom settings. Here we describe teachers' information sources and document their contribution to information gaps and flawed mental models during comprehension.

4.1.1 Information Sources. Nearly all teachers considered ChatGPT their first exposure to AI tools in education. Several teachers added that they did not know whether their conventional educational technology tools incorporate AI to support software features. In contrast, media narratives about ChatGPT centered on the use of AI. Participants recalled initial discussions prompted by major news outlets at the time of release. Information shared among colleagues in the school setting was dominated by discussion of news articles and perspectives from the media. While most early conversations occurred organically among friends, in a few cases, teachers organized ad-hoc department meetings. Teacher T5 recounted that these sessions remained primarily focused on the media lens. Without coordinated administrative guidance or organized professional development, teachers were encouraged to seek their own resources as needed. For eleven teachers, this need arose from increasing encounters with suspicious student work, later confirmed by students to be generated by ChatGPT. Teachers' information-seeking led them to Twitter posts, podcasts, news sites, education blogs, and a variety of online articles. Six teachers described social media as a primary source of information. Teacher T2 explained:

> T2: *I'm on teacher TikTok and tech TikTok..., so then the convergence of those is like, "Here's AI that teachers are using to make lesson plans or create images to go on their slides". I also got the TikTok videos that are for students that are "Students, if you want to cheat on your essay, use this".*

However, eight teachers reported feeling disengaged from doing personal research regarding ChatGPT, citing time constraints. Indeed, information-seeking practices for ChatGPT are effortful and stand in stark contrast to teachers' information access for other classroom technologies. Though typical technology adoption procedures vary, all teachers reported more district- and company-led support for integrating other tools. Teachers described accessing user-guides on company websites, reviews on Commonsense, or videos of teacher walk-throughs demonstrating the interface. Classroom technologies that are licensed or commonly used across teachers are often accompanied by organized professional development sessions and administrative guidance on requirements, procedures, and learning goals. Teacher T1 described professional development sessions facilitated by a representative from the educational technology company:

> T1: *We also have the option to meet with someone from Writable to like help us roll it out .... It's kind of like two main areas that we talk about. The first is just like the tech support: "How are students being presented this information? How is it integrating with Schoology, how do we need to be setting up our grade books to be compatible with it?" Troubleshooting and problem solving for like, "Okay this doesn't work the way we want so how do we have to adjust?" And then the other aspect, of course, is like, "What do we want students to be doing in this platform?"*

Not all educational technology platforms provide such services. However, across teacher-facing resources, there is an understood standard of both informing teachers about the tool and providing the necessary support and context-specific guidance to integrate it into classroom practice. Despite its presence in classrooms, Chat-GPT is not a dedicated educational technology tool, and teachers lack formal training in its use. As a result, teachers fall back on unconventional sources and abstract information disconnected from their domain.

4.1.2 Perceptions and Negotiated Understandings. Based on available information, teachers arrived at a conflicted perception of ChatGPT. All teachers in our interviews described feeling overwhelmed by its harmful implications for academic integrity. Many discussed their experiences with detection tools, student confrontations, unsuccessful bans, and the loss of traditional methods of assessment. These negative perceptions are compounded by the volume of unknowns:

> T21: *It seems like education as a whole is very hesitant about it, and it's just like people are just going to use it to cheat. Shut it down. ...I think that it's just unknown, like I don't know much about it. I don't know how it could be useful.*

Despite their reservations, teachers felt pressure to embrace ChatGPT to advance their teaching practice for fear of being left behind. These fears extend to ensuring that students have access to tools that may be advantageous for their future opportunities:

> T3: *Especially working in low-income areas where students like aren't having access to the tools that a lot of like their peers in different contexts like across the country or the world do ...No, like we should have access to it too, so like our students can learn to use that too.*

However, with limited information, teachers developed faulty mental models about the functionality of ChatGPT. They perceived ChatGPT as software, defining it through a limited set of use cases. Several teachers framed ChatGPT's capabilities exclusively around information retrieval, likening its advent to that of Sparknotes and Wikipedia. Following this misconception, teacher T17 attributed the information in a ChatGPT-generated study guide for the AP U.S. History exam to a search of online materials published by Barron's and the Princeton Review. The assumption that ChatGPT could perform Google searches and access websites, academic journals, and textbooks led teachers to express overconfidence in the validity of outputs. Additionally, six teachers initially framed ChatGPT as a specialized essay-writing tool, with the misunderstanding that

ChatGPT has a defined and identifiable way of writing. This assumption led teachers to expressed overconfidence in the accuracy of ChatGPT plagiarism detectors and subsequently disengage from exploring other classroom applications.

4.1.3 Information Gaps. Across all interviews, teachers expressed frustration with the applicability of available information in the classroom. In addition to conceptual uncertainties about the underlying technology and its limitations, teachers asked questions about its use and regulation in academic and administrative settings. In Table 2, we present a taxonomy of information needs elicited by teachers in our interviews. Of these, teachers voiced the greatest information concerns about learning-relevant use cases and pedagogical goals of integration:

> T6: *I would be curious what goals the district would have for the use of the technology. Like, what would they want us to do with it? How they would recommend that we use it? More than just samples like, what are the applications that I'm kind of missing?*

Teachers additionally voiced the need for support in managing classroom use, negotiating issues of academic integrity, adapting academic standards, preventing student misuse, and restricting student access to harmful content. In addition to general questions about the privacy and security of student data, teachers expressed concerns about administrative access to conversations to ensure student safety. Teacher T9 explained that these procedures are defined in other student writing contexts:

> T9: *Would there be any sort of like way for the school to know if kids were asking things that were worrisome, whether it's self-harm, weaponry, abuse, etc.? It'd probably be built-in in some way, but what is that way? I know that with like the AP Exams, if you're grading one and you come across something concerning there's a whole process to follow for the school to be alerted so like, what is it here?*

When exposed to the GPT-3 model card, teachers expressed interest in its disclosures but struggled to contextualize the information in their instructional practices. Teachers noted the irrelevance of listed model performance metrics but inquired about ChatGPT's accuracy on classroom tasks, such as summarizing a reading comprehension article. For example, teacher T16 was surprised by the inaccuracies in a ChatGPT-generated summary of an article about Eleanor Roosevelt. Further, though the model card makes transparent that the dataset is composed of text posted to the internet, whether the specific composition of training data included relevant and reputable educational materials remains unclear. Teachers wondered whether ChatGPT has knowledge of specific textbooks, niche works of literature, academic journal papers, historical documents, and other texts commonly used in their practice. Eight teachers expressed that knowing these boundaries would support them in avoiding or assessing the risk of encountering hallucinations and misinformation. Teacher T12 explained why contextualizing training data is essential for avoiding biased perspectives:

> T12: *If we're trying to look at a certain text with a certain like racial consideration in mind, then if I knew like ChatGPT ...has never read books by a Black author*

Needs Category	Information Needs	Deficits in Current Documentation
Data Policies	Safety of student data inputs, privacy and security practices, incorporation into continued training, implications for student ownership of their work, FERPA compliance	Though OpenAI documents their privacy policies [84], disclosures are undiscoverable, and teachers gather false information about data collection by directly asking ChatGPT.
Content Restrictions	Age restrictions, censorship of unsafe content, restriction of unsafe searches	Though the OpenAI privacy policy [84] and FAQ [83] state that use by minors is against the terms of use, there are no documented age restrictions. Documentation does not specify content censorship practices.
Oversight	Teacher, parent, and administrative access to conversations, alerting of concerns for student safety surfaced through chats	Documentation does not specify processes for oversight. The OpenAI privacy policy indicates emailing their legal team [84] in cases where minors share information with the system.
Controlling Misuse	Restricting student misuse, unsafe interactions, guidelines for classroom management	Documentation does not specify guidelines for managing student misuse.
Academic Integrity	Ethical standards for when and how it is appropriate for students to use ChatGPT, how to cite ChatGPT output and discover underlying sources, accuracy of tools to detect ChatGPT content	Documentation discloses the inaccuracy of AI detectors [83] and recommends having students share a record of their interactions with ChatGPT, but provides no further guidance about managing use in support of changing standards for academic integrity.
Educational Alignment	Goals of integration, pedagogical relevance of ChatGPT, more exhaustive illustration of use cases and generative capabilities	Though the OpenAI Teaching with AI guide [85] and various online blogs and articles provide examples, they are few and abstract. Documentation does not discuss the breadth of pedagogical applications or the purposes of integrating ChatGPT in education.
Information Access	Relevance of training data to educational materials, access to textbooks and literary works, availability of historical and current information, internet search capabilities, filtering and preprocessing of training data to mitigate biases	Though the GPT-3 model card discloses training data sources, documentation does not contextualize data disclosures in terms of access to specific content or relevant learning materials. Documentation does not specify data pre-processing or filters upon the training data.
Restrictions	Scope of introduced filters, changing restrictions, filter-restricted content and actions	Documentation does not specify modifications or scope of filter-based restrictions upon model or interface behavior.
Performance	Accuracy of classroom-relevant tasks such as summarizing literary or historical text, evaluation of student engagement and interface usability in educational contexts, disclosures of gaps in testing	Documentation does not provide domain-relevant performance metrics.
Limitations Workarounds	How to assess the risk of encountering biased or inaccurate information, prompts and content to avoid in order to limit the risk of hallucination	Documentation does not provide guidance for avoiding and working around limitations.
Prompting	Prompt engineering practices, chat memory context, formatting inputs such as rubrics and passages of text	Documentation does not provide guidance for navigating the interface or effectively prompting ChatGPT.

Table 2: Information needs expressed by participants in interviews and the degree to which they are addressed in current documentation.

then like that would probably change what I think it's capable of doing, you know, so I might not ask it that question in the first place.

Teachers also pointed to the variety of filters in ChatGPT as an added source of uncertainty. In an effort to limit the generation of harmful content, companies such as OpenAI have increasingly introduced guardrails, placing restrictions on model behavior [70]. However, one history teacher found that though content filters scrub openly hateful content, they may have also censored more controversial items in the Black Panther Party's Ten-Point Program, leading to questions about the degree to which challenging topics in educational materials are sanitized. Two teachers described the experience of discovering that the interface no longer supported using ChatGPT to grade student work, prompting lowered confidence in its changing capabilities.

Finally, nearly all teachers indicated limited knowledge of the ChatGPT interface and required prompting guidance for effectively tailoring outputs to their needs. For example, teacher T20 struggled to include a rubric as an input in the prompt while generating

student writing samples according to rubric specifications. Several teachers were unsure of how to include a piece of student writing in the prompt while denoting it as separate from the instruction. Most teachers expressed dissatisfaction with the academic jargon or overwhelming length of outputs. Teachers noted the lack of such practical information in available documentation and voiced preferences for resources to facilitate interactive exposure to ChatGPT.

4.2 Capability Exploration and Experimentation

To identify teachers' bespoke (learning task-specific) information needs, we engaged teachers in experimentation within the ChatGPT interface to explore its capabilities. In Table 3, we present a taxonomy of teachers' experimentation behaviors. Their tests surfaced additional concerns about the process of prompt engineering and the pedagogical relevance of ChatGPT. Here we present observations about teachers' experimentation goals and prompting strategies, as well as their evaluation of ChatGPT outputs.

4.2.1 Experimentation Goals. Most teachers described their first experimentation goals as obtaining a general understanding of ChatGPT and the implications of its capabilities. They generated writing samples to elicit demonstrations of language understanding and capacity for personal authorship support. Teachers also recounted exploring the limitations of ChatGPT, searching for niche information, and provoking harmful responses.

In the education setting, teachers' early interactions with ChatGPT also often focused on verifying capabilities discussed in media sources:

> T7: *Because I know teachers are curious about it ... I've never done this, but they put things like "Make me a lesson plan on X". Okay. So why don't I try that?*

Teachers predominantly repeated documented use cases involving the creation of either lesson plans or student essays, and simple information searching. Further, due to the common connection between ChatGPT and student cheating, teachers across all interviews reported an experimentation goal of assessing the risk of plagiarism. Teachers took the perspective of a student and entered prompts to generate personal narratives, argumentative essays, comprehension question responses, and citations. In doing so, teachers aimed to identify detectable characteristics of ChatGPT-generated language or evaluate outputs for flaws to share with students as a deterrent for using the tool for ineffective cheating. Across interviews, several teachers repeatedly gravitated back to this limited set of experimentation goals.

When encouraged to think creatively about other applications for ChatGPT, many teachers explored prompts to optimize instructional tasks, preferring to avoid scenarios involving student interaction with the tool. Teachers experimented with compiling lists of resources, creating historical timelines and other reference documents, and generating summaries of textbook chapters. They used ChatGPT to generate lesson plans, writing samples, project ideas, comprehension questions, and parent emails:

> T10: *And then how can we, as teachers, ... use it to make our jobs easier. Like if I need to write a sample response or an essay instead of me sitting down and spending an hour writing this sample response, can I have ChatGPT write me a draft and then like fix it to suit my needs?*

In the second phase of interviews, teachers were guided to adopt the experimentation goal of augmenting student learning. By encouraging teachers to systematically consider the pedagogical motivations and student learning needs associated with potential use cases for ChatGPT, teachers experimented with automating feedback, simplifying the language of documents to student reading levels, and generating differentiated and standards-aligned assessment materials. Several teachers experimented with varying degrees of scaffolding for student writing. For example, teacher T17 shared an idea for transforming students' free writing, brainstorming, and class notes into organized outlines.

Situating generative tasks in student learning also introduced specific requirements for outputs. Consequently, teachers were compelled to more carefully consider the construction of their inputs. Prompt engineering then became an additional experimentation goal as teachers navigated interactions to obtain more useful results.

4.2.2 Prompting Strategies. Across all interviews, teachers expressed reservations about their abilities to effectively interact with the open-ended ChatGPT interface. Teachers faced difficulties phrasing their input prompts. A few teachers worked around this discomfort by copying prompts from samples in online articles. Others initially overcame the cold start by approaching the tool from the perspective of a student to generate pieces of writing, but again struggled with prompting when tasked to consider other applications. A prominent source of frustration stems from teachers' mental models connecting ChatGPT to other known tools. Five teachers compared the prompting process to that of using search engines:

> T20: *So I'm kind of stuck like I don't actually feel like I know really how to use the tool ... it almost feels like I need to know how to engineer the right thing. I guess I'm comparing it a little bit to like a Google search like what's the best keywords to put in to get the right output.*

Without researcher intervention, all teachers initially approached prompting with short inputs no longer than a single sentence. Though a few teachers described prior experiments in which more detailed prompts produced more usable outputs, these details often involved baseline contextual information such as grade level, subject area, and desired length of output text.

In the revised interview protocol, teachers were encouraged to discuss specific student learning needs. Subsequently, the prompting strategy of detailing contextual variables became more involved. When generating lesson materials, teachers specified student characteristics, describing student interests, feelings of belonging in school, and preferred modes of demonstrating understanding. Nearly all teachers incorporated pedagogical requirements, using terms from lesson materials. For example, teacher T19 prompted ChatGPT for feedback on student work, specifically regarding identifying instances of indirect characterization and evaluating the degree to which students' effective use of this writing skill builds suspense. Upon learning that ChatGPT retained the context of previous prompts and outputs, a few teachers began to approach prompting as an iterative process, introducing new information in extended follow-up prompts to refine outputs. For example, one ELA teacher prompted ChatGPT first to generate a set of essay questions about a novel, then used subsequent prompts to revise questions to reflect the language of the Common Core state standards and appeal to specified student interests.

In the second phase of interviews, teachers considered the student experience of prompting, motivating them to rethink prior prompting practices. Though teachers had previously used the strategy of including age and grade level information to manipulate the vocabulary and sentence complexity of outputs, a few teachers expressed concern that the approach may not be appropriate for student-led interactions. When teacher T18 prompted ChatGPT to simplify the language of a reading passage for students of varying language proficiency, their initial approach produced a useful response. However, T18 found the use case unsettling upon reflection:

> T18: *I found this one, "Explain like I am 12" and ... "Please simplify the language you use" ... I feel like students*

Experimentation Goal	Use of ChatGPT	Examples
Obtain General Understanding	Demonstrations of Language Understanding	Generating simple sentences, haiku, style imitation
	Personal Authorship Support	Generating cover letters, emails
	Exploring Limitations	Searching for niche information, provoking harmful responses
Verify Described Capabilities	Instructional materials authorship support	Generating lesson plans, essay prompts
	Student authorship support	Generating student essays, answers to comprehension questions
	Research simplification	Information searching, text summarization
Assess Risk of Plagiarism	Identifying characteristics of ChatGPT-generated language	Generating citations, generating personal narratives and other student essays, assessing capabilities to manipulate vocabulary, tone, and subjectivity
	Plagiarism detection	Generate scores and reports identifying plagiarism
Optimize Instructional Tasks	Academic content authorship support	Learning objectives, lesson plans, project instructions, essay prompts, comprehension questions, writing samples
	Administrative authorship support	Lesson plan documentation, parent emails, recommendation letters
	Research simplification	Information searching, text summarization, resource compilation
	Analyzing student writing	Grading, editing, proof-reading, generating feedback for student work
Augment Student Learning*	Student authorship support	Generating ideas, essay outlines, paragraph completion from outline, diction
	Analyzing student writing	Proof-reading, generating feedback
	Research simplification	Information searching, text summarization, resource compilation
	Content formatting	Generating study materials, simplified explanations
Prompt engineering*	Instructional materials authorship support	Designing appropriate prompting language for sharing with students, specifying or emphasizing output requirements

Table 3: Teachers' goals and behaviors observed in interviews for exploring the capabilities of ChatGPT. Goals marked with an asterisk were observed primarily in the second phase interviews involving guided experimentation.

> *could do that, but I also worry . . . these prompts basically concede a deficit view of their own knowledge.*

In practice, teachers are tasked to design prompts that appropriately specify relevant contextual factors to support pedagogical objectives while accounting for student experiences. The difficulty of engineering such prompts is further compounded by limitations in the outputs of ChatGPT.

4.2.3 Response Validation. Many teachers encountered difficulties with the accuracy of information produced by ChatGPT. Experimentation with the use of ChatGPT for research surfaced gaps in training data, filters and deliberate omissions, and hallucinations. Teachers found factually incorrect summaries of linked articles and lesser known works of literature, and incomplete information about global political issues. A few teachers confronted fabricated citations while searching for resources and expressed frustration with ChatGPT's inability to identify original sources of output content.

Additionally, teachers noted the tendency for ChatGPT to produce generic responses. This behavior occurred inconsistently across similar tasks, exacerbating teachers' uncertainties regarding prompting and expected output. Many teachers found the generated lesson plans to be basic and abstract. Several teachers observed that responses remained mainstream and conservative, limiting resource suggestions to works by well-known authors rather than those less frequently anthologized. For example, when an ELA teacher asked ChatGPT for primary sources for a lesson plan about the Harlem Renaissance, it suggested works by Langston Hughes and Countee

Cullen. Social studies teachers noticed outputs containing hedging and qualifying language to avoid complex stances.

Across all interviews, teachers noted the overwhelming length and complexity of outputs and found them inappropriate for student interactions. Several teachers noticed repetition and circular logic in generated feedback for student writing, noting that such language fails to provide process support for students to implement suggestions. Teacher T9 expressed concern at the litany of criticisms generated in response to a prompt for feedback:

> T9: *This does not help developing writers. What helps them is getting positive feedback on what they're doing right and one or two key points to fix. This would shut somebody down.*

Indeed, ChatGPT lacks pedagogical context. One ELA teacher prompted ChatGPT to design a lesson plan with a constructivist lens but found the pedagogical construct present only in the language of the output rather than in the content or activities described. Several teachers prompted ChatGPT to generate writing prompts, essential questions, or comprehension questions but were disappointed in the outputs that could not distinguish the three categories of materials. For example, teacher T21 prompted ChatGPT for reading comprehension questions for the book Just Mercy, but received discussion questions about how the author approaches issues of discrimination and the importance of empathy and compassion in the pursuit of justice. Further, in several interviews, ChatGPT did not recognize terms critical to instructional goals. For example, while manipulating language to simplify instructions for English

language learners, teacher T20 found that the outputs had changed the original intention behind the assignment:

> T20: *When I'm trying to teach subject-specific vocabulary, I don't want to take out words like "theme". … Making something more simplistic isn't necessarily like teaching what students should do. The first time I put it in there, it said "Give historical context about the poem" versus "Say a little bit about what the poem is about". Those are entirely different things.*

These demonstrated limitations of ChatGPT exacerbate barriers to meaningful adoption. They place further demands on teachers' skillful prompting and ability to translate the educational domain to a technical interface.

4.3 Integration Support Needs

Integrating ChatGPT in classroom practice requires a process of complex decision-making considering technical capabilities, instructional needs and objectives, and administrative procedures. However, teachers faced significant process barriers and persistent challenges with designing classroom processes surrounding ChatGPT and navigating its pedagogical role.

4.3.1 Evaluation Standards. Without standard guidance and evaluation criteria for assessing the behaviors of ChatGPT, teachers struggled to make sense of limitations and how they should affect practice. Given disclosures of model limitations or evidence of low-quality outputs, teachers were unable to assess the severity of risks. Consequently, teachers displayed a highly forgiving attitude towards ChatGPT.

Though outputs often contained factual errors or pedagogically inappropriate content, most teachers were willing to "let it slide." The lack of strategies for evaluating the quality of outputs, combined with the tendency to view ChatGPT as a software product with a limited set of specified functionalities, led many teachers to quickly designate the tool as useful after affirming any level of utility in a single use case. Further, they were subsequently unmotivated to refine the output or find alternative applications. Teacher T22 explained their contentment with personally revising undesired ChatGPT output:

> T22: *There's no rule to say like what ChatGPT spits out is what you have to use. So a lot of times you get mediocre responses or incorrect responses, or just kinda useless responses, but you can always iterate from them.*

The tendency for teachers to compensate for unmet information needs by assuming personal responsibilities is also reflected in their responses to potential harm. Rather than seeking additional information or support, several teachers dismissed disclosures of biases and limitations in the model card as avoidable:

> T12: *Because that's stuff that we're familiar with … so as dangerous as it could be, it seems like the bell curve flattens really quickly. I don't know that I would ever ask it a question that would give it a chance to say something horrible, right?*

The shifted accountability places an additional burden on teachers and creates a false sense of agency. As a result, teachers prematurely abandon their evaluation of model limitations. Model

transparency, especially when set in a process derailed by unmet support needs, fails to meaningfully affect practice.

4.3.2 Classroom Procedural Adaptations. Across interviews, most teachers framed their experiences with ChatGPT around "combating" its involvement in student cheating. For some, this involves added layers of monitoring, over-reliance on AI detectors, and reintroducing hand-written and timed in-class assessments, despite having previously discontinued their use. Other teachers lamented the pressure to adapt all assessments to avoid being what teacher T19 calls "cheatable". Lacking formal guidance, teachers struggled with the design of new assessments. Teacher T16 expressed their frustration with abstract suggestions for teachers to adapt content:

> T16: *It's not useful, because it's like "Turn things project-based so they can't just plug it into ChatGPT", or things like that that aren't necessarily like practical. There's not a lot of stuff available … that is like how to use it effectively.*

Further, teachers struggled to define new classroom procedures to manage and monitor student use of ChatGPT. Though a few teachers shared their solutions, including asking students to print and annotate their chat records to justify their prompting choices, most voiced support needs. Several teachers noted that procedures permitting the use of ChatGPT often require additionally defining adapted evaluation criteria and academic standards. Across all interviews, teachers expressed their discomfort with navigating the added responsibility of significant procedural adaptations on an ad-hoc basis.

4.3.3 Pedagogical Trade-offs. While some teachers lauded the tool as a "game-changer" in providing pre-service or early career teachers with sample lesson plans and administrative writing, others expressed concern about how responsibilities are divided. Teachers navigated the uncertain balance between finding creative applications for the unique technical affordances of ChatGPT and understanding the changing role of their own creativity. As teachers experimented with the ability of ChatGPT to optimize teachers' tasks, they searched for moments of irreplaceable teacher involvement. Teacher T1 explained their hesitation with using ChatGPT to generate writing prompts:

> T1: *I enjoy doing that work, and I think it makes me a better teacher to have done that work and thought about like "how could this be like answered and addressed?".*

Teachers voiced a need for guidance regarding their role in the classroom upon integrating an open-ended interface with the technical capacity to perform numerous tasks that overlap significantly with those traditionally performed by teachers. The instructional role most jeopardized is that of assessing student understanding. Beyond fears of student cheating, student interactions with ChatGPT may reduce teachers' opportunities to observe student learning processes. Further, applications of ChatGPT that aim to provide feedback and process support personalize student learning within the interface and do the teaching work of responding to mistakes and misunderstandings. Such applications have the potential to both augment student learning and erode teacher agency for follow-up:

> T11: *As a teacher, as any type of teacher, not just English, you're trying to see the kids' thinking. And I'm just*

worried that this actually creates a barrier to that, and I can't see if my kid like needs help or doesn't understand the concept, or even like if no one's getting it like. If everybody is kind of like making the same mistake, or pitfall, or whatever, then that's something for me to address …and affects what I'm gonna do with the next day even.

5 DISCUSSION

Our interviews reveal important insights about teachers' experiences with the integration of ChatGPT. We observed their strategies for seeking and making sense of information, exploring technical capabilities, and evaluating applicability in their classroom contexts. Teachers face significant information gaps and support needs, limiting their abilities to interact with and find appropriate applications for ChatGPT. Here, we discuss the ways in which available static documentation inadequately addresses integration barriers and propose a framework for interactive model documentation that empowers teachers to navigate the interplay between pedagogical and technical knowledge. We raise open questions about accountability and implications for AI ethics and the broader education domain.

5.1 Limitations of Model Documentation

Designed to provide readers, including non-technical domain stakeholders, with an understanding of how models work to inform decision-making and enable the pursuit of workarounds for limitations [76], model documentation seems to answer the call for AI literacy in education [97]. However, the practical application of model documentation for end-users, particularly in educational contexts, remained largely untested. In conventional educational technology platforms, layers of software mediate model behaviors such that teachers do not directly interact with–and are often unaware of–AI components. The advent of ChatGPT challenges the conceptualization of model documentation as a technical document reserved for experts and underscores the necessity of adapting it to meet the needs of educators. While current AI documentation practices provide significant disclosures about model limitations, they have limited utility in the education domain. Their formatting and abstract content inadequately support teachers' effective integration of ChatGPT.

5.1.1 Documentation is Static. Unsurprisingly, teachers found model documentation inaccessible and impractical due to their narrow technical focus. However, the negative perceptions of documentation are influenced in part by the same frustrations that teachers expressed about guides and articles in educational publications: the text-based format. As static web documents, OpenAI's Teaching with AI guide [85] and educator FAQ [83] are both subject to this limitation. Static formatting is incompatible with teachers' preferred modes of consuming information about educational technologies, which include short videos of interface walk-throughs, hands-on personal exploration, classroom piloting with student feedback, and troubleshooting sessions with experts. Given well-known time constraints [116], teachers shared their fatigue and disengagement about reading long texts with uncertain utility.

5.1.2 Documentation Restricts Active Discovery. Model documentation does not address teachers' foundational concern of discovering pedagogically relevant applications of ChatGPT. Education blogs and newsletters [24], OpenAI's Teaching with AI guide [85], and even startups [90] have aimed to fill this gap by recommending a set of defined use cases, reducing the complexity of ChatGPT to a sampling of features. While providing high-level advice and sample prompts may inspire teachers, these supports ultimately fail to scaffold teachers' own creativity.

In our interviews, teachers struggled to discover appropriate use cases for ChatGPT and repeatedly returned to the few use cases popularized by media narratives and documented samples, despite finding the resulting outputs generic and incompatible with their own classroom practices. Teachers explored ChatGPT's capabilities through a limited set of generative tasks, primarily composing lesson plans and student essays. Overwhelmed by the open-ended interface, teachers faced a failure of imagination, following documented patterns or otherwise abandoning the integration process. In this way, they face what Ko et al. [57] describe as design and selection barriers, in which users struggle to discover a system's purpose, applications, and behaviors.

Our findings support the work of Spiro et al. [102], which attributes misconceptions and disruptions in knowledge acquisition to the oversimplification of complex systems, over-reliance on pre-compiled knowledge structures, and passive transmission. They underscore the importance of multiple representations, situation-specific knowledge assembly, and active participation in developing flexible knowledge [102]. To conceptualize opportunities for integration, teachers require interaction with a broad range of generative abilities and opportunities to specify the use cases relevant to their pedagogical needs.

Further, documentation providing examples currently fails to scaffold the necessary skills for teachers to modify applications to their unique contexts. We observed teachers' limited prompting strategies and challenges in navigating the ChatGPT interface. Though teachers often identify faults in the resulting outputs, they lack the technical knowledge of prompt engineering to achieve desired behaviors. Without guidance to effectively interact with the ChatGPT interface, teachers lose agency and perceive insurmountable challenges to integration.

5.1.3 Documentation Lacks Domain-Relevance. In the TPACK framework, effective integration of technology requires teachers to integrate technical knowledge with pedagogical and content knowledge [72]. However, while model documentation provides technical information and disclosures about ChatGPT, they critically lack support for teachers to contextualize the information in domain-relevant terms. Though teachers found disclosures about the limitations in dataset composition and the risk of misinformation and representational harms important, they could not imagine how such limitations would manifest in practice. In attempts to understand technical limitations in context, teachers asked questions about whether the training data contained specific textbooks or works of literature and about the comparative degree of risk certain content scenarios may incur for eliciting incomplete or inaccurate information. However, without situating limitations and performance metrics in domain tasks and materials, resources–including the

OpenAI educator FAQ [83]–are too abstract to equip teachers to productively work around limitations in their specific application contexts. As a result, some teachers expressed overconfidence that they could easily identify and avoid disclosed limitations, while others opted to avoid the use case entirely and strengthened their resolution against integrating ChatGPT. These polarized views, in which the challenges of ChatGPT are either uniformly dismissable or insurmountable, illustrate the rigid misconceptions resulting from what Spiro et al. [102] describe as context-independent conceptual representations.

Further, by focusing on technical affordances and limitations, documentation neglects information gaps critical to technology integration in the education domain. Teachers require support navigating their adapted instructional roles and the design of new ChatGPT-integrated classroom procedures in the context of changing educational systems [74]. The use of ChatGPT in classrooms necessitates considerations for administrative oversight, classroom management practices, and assessment methods and standards. Yet, guidance for coordination among classroom roles and processes remains largely missing from documentation. Suggestions that teachers should simply review students' interactions with ChatGPT [83] relegate teachers to the role of reacting to misuse rather than proactively defining practices.

5.2 A Framework for Guided Experimentation

Researchers have called for teachers to experiment and "play" with generative AI to develop new pedagogical techniques and discover system behaviors in the context of instructional goals [72]. In support of this objective, documentation should equip teachers with the appropriate skills to engage in such experimentation. Resources should center the communication of technical information in the context of relevant learning scenarios, bridging technical and pedagogical knowledge [23]. Moreover, in order for documentation to orient and support teachers' use of ChatGPT, it must support interactivity [30]. To address these desired characteristics of documentation, we propose a layered framework for interactive documentation to scaffold teachers' design of classroom applications of ChatGPT that augment student learning. The framework decomposes the open-ended interface of ChatGPT and defines five levels of considerations for developing an integration scenario: learning objectives, student needs, generative artifacts, prompting engineering, and classroom coordination. In ongoing research, we apply the framework with teachers in a series of design interviews and make initial recommendations for its integration in the design of a tool allowing teachers to explore—rather than read about—the technical capabilities of ChatGPT.

5.2.1 Learning Objectives. Because current documentation focuses on the technical affordances of ChatGPT, we observed teachers struggling to fit documented capabilities into their classroom practice. Consequently, their approach to integration relied on a few, often redundant, use cases poorly aligned to their instructional goals. To address the narrow scope of applications, the first dimension of the framework encourages teachers to ground their exploration in the context of learning objectives, without considering ChatGPT. This approach reflects that of the commonly used Understanding by Design curriculum-planning framework [69], which advocates

for crafting learning objectives prior to the development of learning experiences.

We ask teachers to identify the problems of their domain and specify the knowledge and skills, defined by academic standards, that students find most difficult to attain. For example, in a middle-school ELA class, teachers may point to students' abilities to support claims with logical reasoning and relevant evidence or to establish and maintain a formal style [77]. In a high-school social studies class, an objective may be to critically analyze primary sources or to connect historical events to contemporary societal issues. In operationalizing the framework, teachers select a single objective to explore in subsequent stages.

5.2.2 Student Needs. Though our interview participants represented a diversity of school environments, including elite private schools, large underfunded public schools with significant English language learner populations, and remedial institutions, the characteristics of their unique classroom contexts were rarely reflected in their approach to integration. Instead, teachers noted the importance of manually revising the generic outputs of ChatGPT based on their knowledge of their students' abilities and interests. To better reflect these factors, the second dimension of the framework asks teachers to specify the cognitive and contextual needs of students for each learning objective identified in the previous stage, again without considering the technical affordances of ChatGPT. In addition to identifying relevant features of the classroom setting, we ask teachers to consider the reasons for which students may find the objective challenging and the types of knowledge [31] that students may lack. For example, students may struggle to support claims with relevant evidence due to language barriers in reading comprehension, difficulties identifying the main message of passages, lack of interest or engagement in the topic, or insufficient exemplars. In operationalizing the framework, teachers select a single aspect of student needs to explore in subsequent stages.

5.2.3 Situated Generative Capabilities. Discovering applications of ChatGPT requires creativity, but we observed teachers exploring applications of ChatGPT through a limited set of generative tasks. To help teachers imagine additional possibilities, the next dimension presents teachers with a broad range of possible generative artifacts to explore each problem space defined through the previous stage of the framework. We define these options through the instructional roles of technology and the generative roles of ChatGPT.

The use of technology in schools can be expressed through three broad categories: technology for instructional preparation, technology for instructional delivery, and technology as a learning tool [49]. Following these categories, ChatGPT may be used to prepare instructional materials (e.g., generating lesson plans), perform instructional tasks (e.g., providing feedback on student work), or interact with students through defined activities (e.g., acting as a tutoring chatbot). For each category, ChatGPT may function as a tool for authorship support, text analysis, and process prompting. That is, it can generate written work, evaluate content, and act as a conversation partner. For example, to address student difficulties identifying the main message of passages, ChatGPT may be used in instructional preparation by generating passages and comprehension questions as formative assessments, highlighting the sentences within input passages that best represent the main
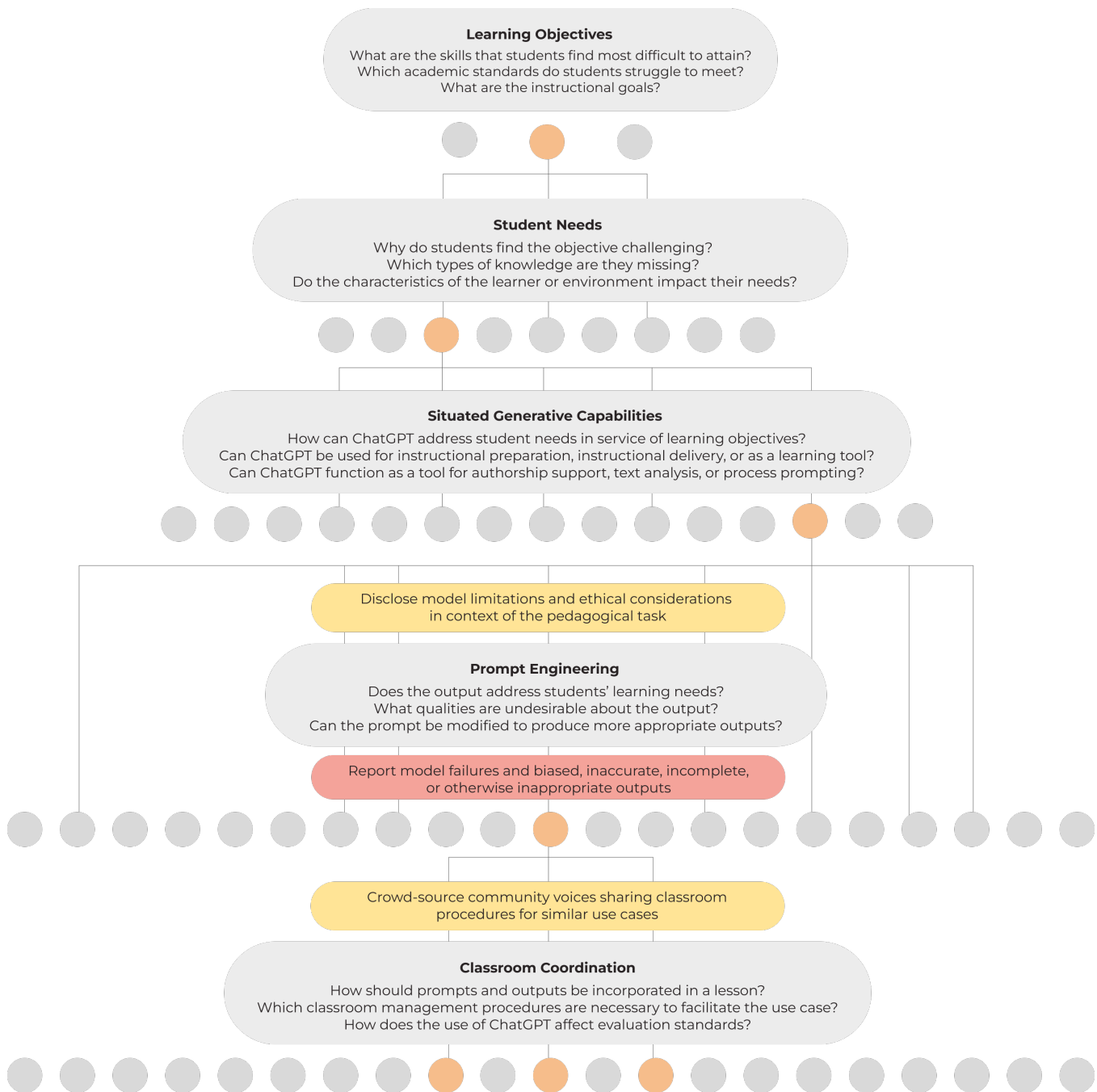
Learning Objectives
What are the skills that students find most difficult to attain?
Which academic standards do students struggle to meet?
What are the instructional goals?

Student Needs
Why do students find the objective challenging?
Which types of knowledge are they missing?
Do the characteristics of the learner or environment impact their needs?

Situated Generative Capabilities
How can ChatGPT address student needs in service of learning objectives?
Can ChatGPT be used for instructional preparation, instructional delivery, or as a learning tool?
Can ChatGPT function as a tool for authorship support, text analysis, or process prompting?

Disclose model limitations and ethical considerations
in context of the pedagogical task

Prompt Engineering
Does the output address students' learning needs?
What qualities are undesirable about the output?
Can the prompt be modified to produce more appropriate outputs?

Report model failures and biased, inaccurate, incomplete,
or otherwise inappropriate outputs

Crowd-source community voices sharing classroom
procedures for similar use cases

Classroom Coordination
How should prompts and outputs be incorporated in a lesson?
Which classroom management procedures are necessary to facilitate the use case?
How does the use of ChatGPT affect evaluation standards?

Figure 2: Stages of considerations to guide teachers' exploration of classroom applications of ChatGPT. Each stage introduces multiple routes to inspire experimentation. In the Situated Generative Capabilities stage, teachers are presented an array of applications matching their contextual specifications in the previous stages. In the Prompt Engineering stage, teachers are iteratively presented with revised prompts in response to their feedback on the quality of outputs while testing.

messages, or suggesting how the teacher might explain and correct a student's misinterpretation of the main message of a passage. To address the same challenge, ChatGPT may be used as a learning tool by students to generate a sentence expressing the main idea of an input passage, generate explanations detailing the contribution of each sentence to the overall message of the passage, or create a conversational tutor asking a series of comprehension questions leading to students identifying the main message of the passage.

5.2.4 Prompt Engineering. Due to the significant information gaps and misconceptions about prompting observed in our interviews, in this stage, our recommended design for the interactive tool involves supplying teachers with an initial prompt based on their specifications in the previous three stages. Beginning with this initial prompt, we scaffold teachers' understanding of prompt engineering by first asking teachers to test the initial prompt in ChatGPT and identify undesired qualities in the model output. For example, the format of the output may be inappropriate in style and length, or the content may inadequately address student needs. We propose suggesting prompt modifications on-demand to address teacher-identified output problems. In doing so, this stage of the framework breaks down the complexity of prompt engineering by framing it as a process of iteratively improving outputs. The scaffolds support teachers in *envisioning* and translating goals into clear intentions provided to the model through prompts [106]. For example, if the output of a query for discussion questions contains vocabulary above the reading level of students, teachers diagnose this flaw in the output and are given a modified prompt including the added specification of language level. If, upon testing the modified prompt, the outputs lack important subject vocabulary, teachers again explain this requirement to receive a modified prompt. The process conveys technical information about ChatGPT's interface features and the structures of well-formed prompts. Further, this stage presents an opportunity to crowd-source model output failures in pedagogical tasks by asking teachers to report false, biased, incomplete, or otherwise unsuitable responses.

5.2.5 Classroom Coordination. The orchestration of roles between teachers and AI systems is critical [93], but complex and unsolved. In our interviews, teachers repeatedly expressed concerns about the impact of ChatGPT integration on classroom management and the standards by which students are assessed. Like the design of use cases and prompts, the design of classroom procedures involving ChatGPT is influenced by context. However, while scholars express optimism for such procedures to evolve from teachers' increasing familiarity with the tool [73], experimentation alone does not sufficiently address the urgent problem of coordinating instructional roles, administrative oversight, classroom management, and lesson procedures in the learning environment. Our recommended design for the final stage of the framework involves crowd-sourced accounts detailing teacher experiences with ChatGPT-integrated classroom procedures, surfaced selectively to teachers based on alignment with specifications in the first three stages of the framework.

In summary, the presented framework is pedagogy-driven, student-focused, and interactive. In ongoing work, we operationalize it through live workshops between teachers and researchers familiar with prompting in ChatGPT and through a scaffolded web-based interface professional development resource for teachers. We contribute this framework to prepare teachers with the skills to leverage ChatGPT for their unique needs. By enabling teachers' discovery of possible applications of ChatGPT in contextually relevant learning scenarios and scaffolding teachers' abilities to tailor interactions to their specifications, the framework addresses key support needs revealed through our interviews. Note that the framework does not address all barriers to integration and is intended to complement—not replace—static documentation and ethical disclosures. However, it provides teachers with a guided approach to the open-ended interface of ChatGPT and an opportunity to infer through experimentation the way such tools function.

5.3 Open Questions

Though we identify teachers' information gaps and support needs about ChatGPT, our findings reveal deeper unknowns about the complexity of its integration into education. To provide teachers with adequate training and support, the roles and responsibilities of various stakeholders require definition, calling into question the degree to which AI developers can be held accountable. We additionally question the effect of model documentation and disclosures on the ethical deployment of AI systems, and we discuss implications for curricular reform, access, and equity.

5.3.1 Accountability. Currently, the burden of adapting to and effectively utilizing ChatGPT falls almost unilaterally on teachers. They are expected to update their instructional competencies and develop new classroom procedures by integrating new technical knowledge in pedagogical skills [73]. These urgent efforts required to address the disruption of ChatGPT compound teachers' demanding workloads [12, 21], contributing to well-documented patterns of burnout and attrition [18].

One way to reduce teachers' burden is to distribute responsibilities to administrators, professional development programs, and teacher training programs. These domain stakeholders have traditionally assumed the responsibility of connecting technical and pedagogical knowledge by training teachers in TPACK competencies for technology integration [19]. Indeed, recent research has discussed the implications for training programs and districts' efforts to better prepare teachers for integrating ChatGPT [117].

However, the responsibilities of OpenAI remain undefined. ChatGPT lacks a critical feedback mechanism commonly found in other educational technology integrations. As teachers described in our interviews, the adoption processes for some educational technology tools might go so far as to involve representatives from companies who are responsible for receiving and responding to bug reports, contributing to the content of professional development, and hosting troubleshooting sessions. More commonly, teachers' discontent about broken functionality or features that fail to address classroom needs is eventually reflected in surveys of tools no longer used and district purchasing decisions. As a result, many conventional educational technology companies have some stake in listening to teachers' feedback and making suitable adjustments to products.

In contrast, the creators of ChatGPT–a free general-purpose tool–are not accountable for responding to the problems raised by educators and parents. Though the U.S. Government's AI Bill of Rights calls for users to "access a person who can quickly consider and remedy problems [they] encounter" [80], such aspirations are not legally binding. There is no mechanism for flagging low-quality and problematic content, demanding transparency about student data, or requesting administrative controls essential to the education domain. Instead, OpenAI states in a recent FAQ [83] that use by minors is against the terms of use, but without changes to the

interface to regulate use, such statements are inconsistent with practitioner realities.

5.3.2 Implications for Ethical AI. Despite industry efforts to require model documentation and ethical statements [9] to address fairness, accountability, and transparency of AI systems, current standards for their contents fail to ensure ethical deployments. Disclosures about dataset composition, performance metrics, and representational harms narrowly construe fairness as a small set of harms that motivate algorithmic fixes [17]. However, these abstract documents are insufficient to address the range of impacts from generative AI systems [101] and critically lack domain-specific considerations. They neglect the full scope of unpredictable context-dependent harms affecting end-users [17, 111]. We observed these harms in teachers' significant information gaps and support needs in response to the deployment of ChatGPT in education. Teachers face practical challenges managing student use, ensuring student safety, navigating curricular disruptions, and urgently seeking new technical literacies. As ChatGPT obscures teachers' abilities to observe and respond to students' understanding, and fears of cheating result in an over-reliance on AI detectors with serious academic consequences, the harms of ChatGPT directly impact student learning.

Further, documentation and ethical disclosures are posthumous, articulating harms as an afterthought to deployment. Such documents deflect and downplay negative societal impacts, shifting blame to improper inputs or malicious use, declaring limitations to caution end-users, and delegating the undefined task of mitigating limitations to practitioners while effectively absolving companies and AI developers of responsibility [65]. In response, researchers have proposed an ethics and societal impact review process as a prerequisite to research funding [15]. Had this process been implemented in the case of ChatGPT, it may have involved a limited and controlled release through risk-assessed sectors, modified interfaces appropriate to partner domains, and early collaborations with domain stakeholders to anticipate usability traps and support needs. However, in the case of ChatGPT, a teaching guide of a limited set of sample prompts and FAQ answers [85] was released in response to months of practitioner outcry.

5.3.3 Implications for Education Systems. The popular narrative comparing ChatGPT to calculators [66, 71] was reflected in our interviews. The debate about calculators in math instruction placed curricular content and the design of educational standards into question. Advocates called for replacing arithmetic-driven lessons that emphasize rote skills and reimagining the content that is relevant to teaching. Though teachers in our interviews expressed similar perspectives about ChatGPT, they were often unable to articulate the qualities of curricular reform. They emphasized the need to eliminate "cheatable" assignments but were waiting for another authority to design the assessments that would be implemented in their place. These unknowns arise in part from a flawed metaphor—ChatGPT is unlike a calculator in that the capabilities of generative AI are far greater than carrying out mechanical operations [66]—and in part from ongoing technical advancements that continue to change the skills students may need in their futures.

One change in response to ChatGPT involves new requirements for teaching technology and addressing the presence of AI in the lives of students. In our interviews, teachers discussed their challenges with finding ways to introduce ChatGPT to students and discuss its limitations. Districts facing similar unknowns have responded by designing AI curriculum [56] and expanding computer science education [63]. However, the curricular implications of preparing students for a future in which every domain of human activity is affected by generative AI [64] remains unknown.

Finally, open questions remain regarding access and equity. Teachers in our interviews often referenced equal access to opportunity as the reason for adopting ChatGPT, motivated by exposing their students to tools and resources available to their peers in other educational contexts. However, our study participants also acknowledged that they themselves were more engaged in learning about ChatGPT and other new technologies than their peers. They expressed worries about their colleagues' resistance to gaining technical experience with these tools. The consequences of varying teacher abilities to limit the disruption of ChatGPT and teach students productive uses of the tool are yet to be seen.

6 CONCLUSION

Effective integration of educational technology relies on teachers' abilities to take pedagogical advantage of technical affordances. Yet, in the case of ChatGPT, teachers are inadequately supported in developing these competencies. Teachers' unguided explorations of ChatGPT's capabilities are limited to a small set of use cases popularized by media narratives and disconnected from their own instructional goals and classroom environments. By perceiving ChatGPT in terms of fixed functionalities, teachers are unequipped to adapt to undesirable model outputs and prematurely abandon their evaluation of its potential classroom applications. Standard static model documentation practices insufficiently address teachers' information gaps and do not prepare teachers to interact directly with models such as ChatGPT. To enable teachers to operationalize model information in their classroom practices, model documentation must center interactivity and guided experimentation, situating information about generative capabilities in learning goals and student needs, and contextualizing disclosures about model limitations in terms of their impact on classroom activities.

ACKNOWLEDGMENTS

REFERENCES

[1] 2024. BERT Base Model (Cased). https://huggingface.co/bert-base-cased Accessed: 2024-01-25.
[2] 2024. Google Cloud Model Cards: Face Detection. https://modelcards. withgoogle.com/face-detection Accessed: 2024-01-25.

[3] Emad Aghajani, Csaba Nagy, Mario Linares-Vásquez, Laura Moreno, Gabriele Bavota, Michele Lanza, and David C Shepherd. 2020. Software documentation: the practitioners' perspective. In *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering*. 590–601.

[4] Emad Aghajani, Csaba Nagy, Olga Lucero Vega-Márquez, Mario Linares-Vásquez, Laura Moreno, Gabriele Bavota, and Michele Lanza. 2019. Software documentation issues unveiled. In *2019 IEEE/ACM 41st International Conference on Software Engineering (ICSE)*. IEEE, 1199–1210.

[5] Emad Aghajani, Csaba Nagy, Olga Lucero Vega-Márquez, Mario Linares-Vásquez, Laura Moreno, Gabriele Bavota, and Michele Lanza. 2019. Software Documentation Issues Unveiled. In *2019 IEEE/ACM 41st International Conference on Software Engineering (ICSE)*. 1199–1210. https://doi.org/10.1109/ICSE.2019.00122

[6] Roberto Aldunate and Miguel Nussbaum. 2013. Teacher adoption of technology. *Computers in Human Behavior* 29, 3 (2013), 519–524.

[7] Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the people: The role of humans in interactive machine learning. *Ai Magazine* 35, 4 (2014), 105–120.

[8] Matthew Arnold, Rachel KE Bellamy, Michael Hind, Stephanie Houde, Sameep Mehta, Aleksandra Mojsilović, Ravi Nair, K Natesan Ramamurthy, Alexandra Olteanu, David Piorkowski, et al. 2019. FactSheets: Increasing trust in AI services through supplier's declarations of conformity. *IBM Journal of Research and Development* 63, 4/5 (2019), 6–1.

[9] Carolyn Ashurst, Emmie Hine, Paul Sedille, and Alexis Carlier. 2022. Ai ethics statements: analysis and lessons learnt from neurips broader impact statements. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 2047–2056.

[10] Evrim Baran, Hsueh-Hua Chuang, and Ann Thompson. 2011. TPACK: An emerging research and development tool for teacher educators. *Turkish Online Journal of Educational Technology-TOJET* 10, 4 (2011), 370–377.

[11] Amy L Baylor and Donn Ritchie. 2002. What factors facilitate teacher skill, teacher morale, and perceived student learning in technology-using classrooms? *Computers & education* 39, 4 (2002), 395–414.

[12] Jaime L Beck. 2017. The weight of a heavy hour: Understanding teacher experiences of work intensification. *McGill Journal of Education* 52, 3 (2017), 617–636.

[13] Sarah W Beck and Sarah R Levine. 2023. Backtalk: ChatGPT: A powerful technology tool for writing instruction. *Phi Delta Kappan* 105, 1 (2023), 66–67.

[14] EM Bender and Batya Friedman. 2019. Data statements for NLP: toward mitigating system bias and enabling better science. In *Preprint at https://openreview.net/forum*.

[15] Michael S Bernstein, Margaret Levi, David Magnus, Betsy A Rajala, Debra Satz, and Quinn Waeiss. 2021. Ethics and society review: Ethics reflection as a precondition to research funding. *Proceedings of the National Academy of Sciences* 118, 52 (2021), e2117261118.

[16] Veronika Bogina, Alan Hartman, Tsvi Kuflik, and Avital Shulner-Tal. 2021. Educating software and AI stakeholders about algorithmic fairness, accountability, transparency and ethics. *International Journal of Artificial Intelligence in Education* (2021), 1–26.

[17] Margarita Boyarskaya, Alexandra Olteanu, and Kate Crawford. 2020. Overcoming failures of imagination in AI infused system development and deployment. *arXiv preprint arXiv:2011.13416* (2020).

[18] Sam Brill and Abby McCartney. 2008. Stopping the revolving door: Increasing teacher retention. *Politics & Policy* 36, 5 (2008), 750–774.

[19] Karen E Brinkley-Etzkorn. 2018. Learning to teach online: Measuring the influence of faculty development training on teaching effectiveness through a TPACK lens. *The Internet and Higher Education* 38 (2018), 28–35.

[20] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[21] Graham Butt and Ann Lance. 2005. Secondary teacher workload and job satisfaction: do successful strategies for change exist? *Educational Management Administration & Leadership* 33, 4 (2005), 401–422.

[22] Davide Castelvecchi. 2016. Can we open the black box of AI? *Nature News* 538, 7623 (2016), 20.

[23] Ismail Celik. 2023. Towards Intelligent-TPACK: An empirical study on teachers' professional knowledge to ethically integrate artificial intelligence (AI)-based tools into education. *Computers in Human Behavior* 138 (2023), 107468.

[24] Betty Chandy. 2023. Five ways teachers can integrate ChatGPT into their classrooms today. https://www.gse.upenn.edu/news/educators-playbook/five-ways-teachers-can-integrate-chatgpt-their-classrooms-today

[25] Kathy Charmaz. 2014. *Constructing grounded theory*. sage.

[26] Kun-Hung Cheng and Chin-Chung Tsai. 2019. A case study of immersive virtual field trips in an elementary classroom: Students' learning experience and teacher-student interaction behaviors. *Computers & Education* 140 (2019), 103600.

[27] Tessa Consoli, Juliette Désiron, and Alberto Cattaneo. 2023. What is "technology integration" and how is it measured in K-12 education? A systematic review of survey instruments from 2010 to 2021. *Computers & Education* (2023), 104742.

[28] Grant Cooper. 2023. Examining science education in chatgpt: An exploratory study of generative artificial intelligence. *Journal of Science Education and Technology* 32, 3 (2023), 444–452.

[29] John W Creswell and J David Creswell. 2017. *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications.

[30] Anamaria Crisan, Margaret Drouhard, Jesse Vig, and Nazneen Rajani. 2022. Interactive model cards: A human-centered approach to model documentation. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 427–439.

[31] Ton De Jong and Monica GM Ferguson-Hessler. 1996. Types and qualities of knowledge. *Educational psychologist* 31, 2 (1996), 105–113.

[32] Ramon Dijkstra, Zülküf Genç, Subhradeep Kayal, Jaap Kamps, et al. 2022. Reading Comprehension Quiz Generation using Generative Pre-trained Transformers.

[33] Renato P. dos Santos. 2023. Enhancing Chemistry Learning with ChatGPT, Bing Chat, Bard, and Claude as Agents-to-Think-With: A Comparative Case Study. arXiv:2311.00709 [cs.HC]

[34] Peggy A Ertmer and Anne T Ottenbreit-Leftwich. 2010. Teacher technology change: How knowledge, confidence, beliefs, and culture intersect. *Journal of research on Technology in Education* 42, 3 (2010), 255–284.

[35] Gregory M Francom. 2020. Barriers to technology integration: A time-series survey study. *Journal of Research on Technology in Education* 52, 1 (2020), 1–16.

[36] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12 (2021), 86–92.

[37] ATLAS.ti Scientific Software Development GmbH. 2023. ATLAS.ti: The Qualitative Data Analysis & Research Software. https://atlasti.com/

[38] Charles R Graham, Jered Borup, and Nicolette Burgoyne Smith. 2012. Using TPACK as a framework to understand teacher candidates' technology integration decisions. *Journal of Computer Assisted Learning* 28, 6 (2012), 530–546.

[39] Andrina Granić. 2022. Educational technology adoption: a systematic review. *Education and Information Technologies* 27, 7 (2022), 9725–9744.

[40] Reza Hadi Mogavi, Chao Deng, Justin Juho Kim, Pengyuan Zhou, Young D. Kwon, Ahmed Hosny Saleh Metwally, Ahmed Tlili, Simone Bassanelli, Antonio Bucchiarone, Sujit Gujar, Lennart E. Nacke, and Pan Hui. 2024. ChatGPT in education: A blessing or a curse? A qualitative study exploring early adopters' utilization and perceptions. *Computers in Human Behavior: Artificial Humans* 2, 1 (2024), 100027. https://doi.org/10.1016/j.chbah.2023.100027

[41] Kunal Handa, Margaret Clapper, Jessica Boyle, Rose E Wang, Diyi Yang, David S Yeager, and Dorottya Demszky. 2023. " Mistakes Help Us Grow": Facilitating and Evaluating Growth Mindset Supportive Language in Classrooms. *arXiv preprint arXiv:2310.10637* (2023).

[42] Will Douglas Heaven. 2023. *ChatGPT is going to change education, not destroy it*. https://www.technologyreview.com/2023/04/06/1071059/chatgpt-change-not-destroy-education-openai/

[43] Khe Foon Hew and Thomas Brush. 2007. Integrating technology into K-12 teaching and learning: Current knowledge gaps and recommendations for future research. *Educational technology research and development* 55 (2007), 223–252.

[44] Khe Foon Hew, Min Lan, Ying Tang, Chengyuan Jia, and Chung Kwan Lo. 2019. Where is the "theory" within the field of educational technology research? *British Journal of Educational Technology* 50, 3 (2019), 956–971.

[45] Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. 2020. The dataset nutrition label. *Data Protection and Privacy* 12, 12 (2020), 1.

[46] Wayne Holmes, Kaska Porayska-Pomsta, Ken Holstein, Emma Sutherland, Toby Baker, Simon Buckingham Shum, Olga C Santos, Mercedes T Rodrigo, Mutlu Cukurova, Ig Ibert Bittencourt, et al. 2021. Ethics of AI in education: Towards a community-wide framework. *International Journal of Artificial Intelligence in Education* (2021), 1–23.

[47] Kenneth Holstein, Bruce M McLaren, and Vincent Aleven. 2019. Designing for complementarity: Teacher and student needs for orchestration support in AI-enhanced classrooms. In *Artificial Intelligence in Education: 20th International Conference, AIED 2019, Chicago, IL, USA, June 25-29, 2019, Proceedings, Part I 20*. Springer, 157–171.

[48] Pi-Sui Hsu. 2016. Examining current beliefs, practices and barriers about technology integration: A case study. *TechTrends* 60 (2016), 30–40.

[49] Fethi A Inan and Deborah L Lowther. 2010. Factors affecting technology integration in K-12 classrooms: A path model. *Educational technology research and development* 58 (2010), 137–154.

[50] Hyangeun Ji, Insook Han, and Yujung Ko. 2023. A systematic review of conversational AI in language education: Focusing on the collaboration with human teachers. *Journal of Research on Technology in Education* 55, 1 (2023), 48–63.

[51] Qinjin Jia, Jialin Cui, Yunkai Xiao, Chengyuan Liu, Parvez Rashid, and Edward F Gehringer. 2021. All-in-one: Multi-task learning bert models for evaluating peer assessments. *arXiv preprint arXiv:2110.03895* (2021).

[52] Rachel Karchmer-Klein and Haruka Konishi. 2023. A mixed-methods study of novice teachers' technology integration: Do they leverage their TPACK

knowledge once entering the profession? *Journal of Research on Technology in Education* 55, 3 (2023), 490–506.

[53] Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences* 103 (2023), 102274.

[54] Rehan Ahmed Khan, Masood Jawaid, Aymen Rehan Khan, and Madiha Sajjad. 2023. ChatGPT-Reshaping medical education and clinical management. *Pakistan Journal of Medical Sciences* 39, 2 (2023), 605.

[55] Seonghun Kim, Yeonju Jang, Seongyune Choi, Woojin Kim, Heeseok Jung, Soohwan Kim, and Hyeoncheol Kim. 2021. Analyzing teacher competency with TPACK for K-12 AI education. *KI-Künstliche Intelligenz* 35, 2 (2021), 139–151.

[56] Alyson Klein. 2022. *Artificial Intelligence Is All Around Us. So This District Designed Its Own AI Curriculum.* https://www.edweek.org/technology/artificial-intelligence-is-all-around-us-so-this-district-designed-its-own-ai-curriculum/2022/08

[57] Amy J Ko, Brad A Myers, and Htet Htet Aung. 2004. Six learning barriers in end-user programming systems. In *2004 IEEE Symposium on Visual Languages-Human Centric Computing*. IEEE, 199–206.

[58] MJ Koehler and P Mishra. 2008. Introducing TPCK. Handbook of technological pedagogical content knowledge (TPCK) for educators, In the AACTE Committee on Innovation and Technology (Eds.), Handbook of technological pedagogical content knowledge (TPCK) for educators (pp. 2-29).

[59] Matthew Koehler and Punya Mishra. 2009. What is technological pedagogical content knowledge (TPACK)? *Contemporary issues in technology and teacher education* 9, 1 (2009), 60–70.

[60] Florian Königstorfer and Stefan Thalmann. 2021. Software documentation is not enough! Requirements for the documentation of AI. *Digital Policy, Regulation and Governance* 23, 5 (2021), 475–488.

[61] Florian Königstorfer and Stefan Thalmann. 2022. AI Documentation: A path to accountability. *Journal of Responsible Technology* 11 (2022), 100043.

[62] Theodore J. Kopcha. 2012. Teachers' perceptions of the barriers to technology integration and practices with technology under situated professional development. *Computers & Education* 59, 4 (2012), 1109–1121. https://doi.org/10.1016/j.compedu.2012.05.014

[63] Lauraine Langreo. 2022. *What Teachers Should Consider Before Using AI-Powered Tools in the Classroom.* https://www.edweek.org/technology/what-teachers-should-consider-before-using-ai-powered-tools-in-the-classroom/2022/12

[64] Robin Li. 2020. *Artificial intelligence revolution: How AI will change our society, economy, and culture.* Simon and Schuster.

[65] David Liu, Priyanka Nanayakkara, Sarah Ariyan Sakha, Grace Abuhamad, Su Lin Blodgett, Nicholas Diakopoulos, Jessica R Hullman, and Tina Eliassi-Rad. 2022. Examining Responsibility and Deliberation in AI Impact Statements and Ethics Reviews. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. 424–435.

[66] Jason M Lodge, Suijing Yang, Leon Furze, and Phillip Dawson. 2023. It's not like a calculator, so what is the relationship between learners and generative artificial intelligence? *Learning: Research and Practice* 9, 2 (2023), 117–124.

[67] Rose Luckin, Karine George, and Mutlu Cukurova. 2022. *AI for school teachers.* CRC Press.

[68] Stephen MacNeil, Andrew Tran, Dan Mogil, Seth Bernstein, Erin Ross, and Ziheng Huang. 2022. Generating diverse code explanations using the gpt-3 large language model. In *Proceedings of the 2022 ACM Conference on International Computing Education Research-Volume 2*. 37–39.

[69] Jay McTighe and Grant Wiggins. 2012. Understanding by design framework. *Alexandria, VA: Association for Supervision and Curriculum Development* (2012).

[70] Cade Metz. 2023. Researchers Poke Holes in Safety Controls of ChatGPT and Other Chatbots. https://www.nytimes.com/2023/07/27/business/ai-chatgpt-safety-research.html

[71] Stephen Mihm. 2023. *What a Calculator Can Tell You About ChatGPT.* https://www.washingtonpost.com/business/2023/09/05/chatgpt-is-2023-s-version-of-the-calculator-and-cliffsnotes/716af25a-4bf0-11ee-bfca-04e0ac43f9e4_story.html

[72] Punya Mishra. 2019. Considering contextual knowledge: The TPACK diagram gets an upgrade. , 76–78 pages.

[73] P Mishra. 2023. ChatGPT is a smart drunk intern: 3 examples.

[74] Punya Mishra and Melissa Warr. 2021. Contextualizing TPACK within systems and cultures of practice. *Computers in Human Behavior* 117 (2021), 106673.

[75] Punya Mishra, Melissa Warr, and Rezwana Islam. 2023. TPACK in the age of ChatGPT and Generative AI. *Journal of Digital Learning in Teacher Education* 39, 4 (2023), 235–251.

[76] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*. 220–229.

[77] Council of Chief State School Officers National Governors Association Center for Best Practices. 2010. *Common Core State Standards: Writing.* https://www.thecorestandards.org/ELA-Literacy/W/7/

[78] Davy Tsz Kit Ng, Jac Ka Lok Leung, Samuel Kai Wah Chu, and Maggie Shen Qiao. 2021. Conceptualizing AI literacy: An exploratory review. *Computers and Education: Artificial Intelligence* 2 (2021), 100041.

[79] Huy A Nguyen, Hayden Stec, Xinying Hou, Sarah Di, and Bruce M McLaren. 2023. Evaluating chatgpt's decimal skills and feedback generation in a digital learning game. In *European Conference on Technology Enhanced Learning*. Springer, 278–293.

[80] White House Office of Science and Technology Policy. 2022. *Blueprint for an AI Bill of Rights.* https://www.whitehouse.gov/ostp/ai-bill-of-rights/

[81] OpenAI. 2020. *GPT-3 Model Card.* https://github.com/openai/gpt-3/blob/master/model-card.md

[82] OpenAI. 2023. ChatGPT. https://chat.openai.com/

[83] OpenAI. 2023. *Educator FAQ.* https://help.openai.com/en/collections/5929286-educator-faq

[84] OpenAI. 2023. *Privacy Policy.* https://openai.com/policies/privacy-policy

[85] OpenAI. 2023. *Teaching with AI.* https://openai.com/blog/teaching-with-ai

[86] Sonmez Pamuk. 2012. Understanding preservice teachers' technology use through TPACK framework. *Journal of computer assisted learning* 28, 5 (2012), 425–439.

[87] Zachary A Pardos and Shreya Bhandari. 2023. Learning gain differences between ChatGPT and human tutor generated algebra hints. *arXiv preprint arXiv:2302.06871* (2023).

[88] David Piorkowski, Daniel González, John Richards, and Stephanie Houde. 2020. Towards evaluating and eliciting high-quality documentation for intelligent systems. *arXiv preprint arXiv:2011.08774* (2020).

[89] Cory Dal Ponte, Sathana Dushyanthen, and Kayley Lyons. 2023. " Close... but not as good as an educator."–Using ChatGPT to provide formative feedback in large-class collaborative learning. *arXiv preprint arXiv:2311.01634* (2023).

[90] Peggy Reimers. 2023. MAGICSCHOOL OFFERS 40+ TIME-SAVING AI TOOLS FOR TEACHERS. https://blog.tcea.org/magicschool/

[91] Margaret Roblyer, Aaron H Doering, et al. 2007. *Integrating educational technology into teaching.* USA: Pearson, 2007.

[92] Kevin Roose. 2023. Don't Ban ChatGPT in Schools. Teach With It. https://www.nytimes.com/2023/01/12/technology/chatgpt-schools-teachers.html

[93] Jeremy Roschelle, James Lester, and Judi Fusco. 2020. *AI and the future of learning: Expert panel report.* Technical Report. Digital Promise.

[94] Kalhan Rosenblatt. 2023. ChatGPT banned from New York City public schools' devices and networks. *NBC News* (2023).

[95] Javeria Salman. 2023. How AI can teach kids to write – not just cheat. https://hechingerreport.org/how-ai-can-teach-kids-to-write-not-just-cheat

[96] Daniel Schiff. 2021. Out of the laboratory and into the classroom: the future of artificial intelligence in education. *AI & society* 36, 1 (2021), 331–348.

[97] Daniel Schiff. 2022. Education for AI, not AI for education: The role of education and ethics in national AI policy strategies. *International Journal of Artificial Intelligence in Education* 32, 3 (2022), 527–563.

[98] Mirjam Schmid, Eliana Brianza, and Dominik Petko. 2020. Developing a short assessment instrument for Technological Pedagogical Content Knowledge (TPACK. xs) and comparing the factor structure of an integrative and a transformative model. *Computers & Education* 157 (2020), 103967.

[99] Sabine Seufert, Josef Guggemos, and Michael Sailer. 2021. Technology-related knowledge, skills, and attitudes of pre-and in-service teachers: The current situation and emerging trends. *Computers in Human Behavior* 115 (2021), 106552.

[100] Natasha Singer. 2023. Despite Cheating Fears, Schools Repeal ChatGPT Bans. *New York Times* (2023).

[101] Irene Solaiman, Zeerak Talat, William Agnew, Lama Ahmad, Dylan Baker, Su Lin Blodgett, Hal Daumé III, Jesse Dodge, Ellie Evans, Sara Hooker, et al. 2023. Evaluating the Social Impact of Generative AI Systems in Systems and Society. *arXiv preprint arXiv:2306.05949* (2023).

[102] Rand J Spiro et al. 1988. Cognitive Flexibility Theory: Advanced Knowledge Acquisition in Ill-Structured Domains. Technical Report No. 441. (1988).

[103] Jill Staake. 2023. *20 Ways Teachers Can Use ChatGPT To Make Their Lives Easier.* https://www.weareteachers.com/chatgpt-for-teachers/

[104] C Stokel-Walker. 2022. AI bot ChatGPT writes smart essays—Should professors worry?[published online ahead of print December 9, 2022]. *Nature News* (2022).

[105] Anselm Strauss and Juliet Corbin. 1990. *Basics of qualitative research.* Sage publications.

[106] Hariharan Subramonyam, Christopher Lawrence Pondoc, Colleen Seifert, Maneesh Agrawala, and Roy Pea. 2023. Bridging the Gulf of Envisioning: Cognitive Design Challenges in LLM Interfaces. *arXiv preprint arXiv:2309.14459* (2023).

[107] Anaïs Tack and Chris Piech. 2022. The AI teacher test: Measuring the pedagogical ability of blender and GPT-3 in educational dialogues. *arXiv preprint arXiv:2205.07540* (2022).

[108] Tamara Tate, Shayan Doroudi, Daniel Ritchie, Ying Xu, et al. 2023. Educational research and AI-generated writing: Confronting the coming tsunami. (2023).

[109] OpenAI Team. 2022. ChatGPT: Optimizing language models for dialogue.

[110] Mark Bedoya Ulla, William F Perales, and Stephenie Ong Busbus. 2023. 'To generate or stop generating response': Exploring EFL teachers' perspectives

on ChatGPT in English language teaching in Thailand. *Learning: Research and Practice* 9, 2 (2023), 168–182.

[111] Michael Veale, Max Van Kleek, and Reuben Binns. 2018. Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. In *Proceedings of the 2018 chi conference on human factors in computing systems*. 1–14.

[112] Rose E Wang and Dorottya Demszky. 2023. Is ChatGPT a Good Teacher Coach? Measuring Zero-Shot Performance For Scoring and Providing Actionable Insights on Classroom Instruction. *arXiv preprint arXiv:2306.03090* (2023).

[113] Rose E Wang, Qingyang Zhang, Carly Robinson, Susanna Loeb, and Dorottya Demszky. 2023. Step-by-Step Remediation of Students' Mathematical Mistakes. *arXiv preprint arXiv:2310.10648* (2023).

[114] Yousef Wardat, Mohammad A Tashtoush, Rommel AlAli, and Adeeb M Jarrah. 2023. ChatGPT: A revolutionary tool for teaching and learning mathematics. *Eurasia Journal of Mathematics, Science and Technology Education* 19, 7 (2023), em2286.

[115] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al.

2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359* (2021).

[116] Shelley B Wepner, Nancy Ziomek, and Liqing Tao. 2003. Three teacher educators' perspectives about the shifting responsibilities of infusing technology into the curriculum. *Action in Teacher Education* 24, 4 (2003), 53–63.

[117] Jeromie Whalen, Chrystalla Mouza, et al. 2023. ChatGPT: Challenges, Opportunities, and Implications for Teacher Education. *Contemporary Issues in Technology and Teacher Education* 23, 1 (2023), 1–23.

[118] Su-Youn Yoon, Eva Miszoglad, and Lisa R. Pierce. 2023. Evaluation of ChatGPT Feedback on ELL Writers' Coherence and Cohesion. arXiv:2310.06505 [cs.CL]

[119] Olaf Zawacki-Richter, Victoria I Marín, Melissa Bond, and Franziska Gouverneur. 2019. Systematic review of research on artificial intelligence applications in higher education–where are the educators? *International Journal of Educational Technology in Higher Education* 16, 1 (2019), 1–27.

[120] Junji Zhi, Vahid Garousi-Yusifoğlu, Bo Sun, Golara Garousi, Shawn Shahnewaz, and Guenther Ruhe. 2015. Cost, benefits and quality of software development documentation: A systematic mapping. *Journal of Systems and Software* 99 (2015), 175–198.