ContextQ: Generated Questions to Support Meaningful Parent-Child Dialogue While Co-Reading

Griffin Dietz Smith Apple Seattle, WA, USA griffind@apple.com

Siddhartha Prasad **Brown University** Providence, RI, USA siddhartha_prasad@brown.edu

Matt J. Davidson Apple Austin, TX, USA matt_davidson@apple.com

Leah Findlater Apple Seattle, WA, USA lfindlater@apple.com R. Benjamin Shapiro Apple Seattle, WA, USA nerd@apple.com

ABSTRACT

Much of early literacy education happens at home with caretakers reading books to young children. Prior research demonstrates how having dialogue with children during co-reading can develop critical reading readiness skills, but most adult readers are unsure if and how to lead effective conversations. We present ContextQ, a tabletbased reading application to unobtrusively present auto-generated dialogic questions to caretakers to support this dialogic reading practice. An ablation study demonstrates how our method of encoding educator expertise into the question generation pipeline can produce high-quality output; and through a user study with 12 parent-child dyads (child age: 4-6), we demonstrate that this system can serve as a guide for parents in leading contextually meaningful dialogue, leading to significantly more conversational turns from both the parent and the child and deeper conversations with connections to the child's everyday life.

CCS CONCEPTS

• Social and professional topics → Children; • Applied com $puting \rightarrow \mathit{Interactive learning environments}$; • Human-centered **computing** → *Human computer interaction (HCI).*

KEYWORDS

dialogic reading; literacy; co-reading; parent-child interaction; large language models; question generation

ACM Reference Format:

Griffin Dietz Smith, Siddhartha Prasad, Matt J. Davidson, Leah Findlater, and R. Benjamin Shapiro. 2024. ContextQ: Generated Questions to Support Meaningful Parent-Child Dialogue While Co-Reading. In Interaction Design and Children (IDC '24), June 17-20, 2024, Delft, Netherlands. ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3628516.3655809

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored For all other uses, contact the owner/author(s).

IDC '24, June 17-20, 2024, Delft, Netherlands

© 2024 Copyright held by the owner/author(s). ACM ISBN 979-8-4007-0442-0/24/06. https://doi.org/10.1145/3628516.3655809

1 INTRODUCTION

Early childhood reading level is one of the most important predictors of lifetime success [22]. Yet, over a third of children entering kindergarten lack reading readiness skills (e.g., phonemic awareness and vocabulary) and start formal schooling already below grade level [6]. These children who start behind rarely catch up [19, 24, 47]. Much of the necessary early literacy education happens at home, through caretakers reading books with young children, but most adult readers do not realize that children learn so much more when parents and children talk about the stories they read [28]. Specifically, through a style of co-reading called "dialogic reading" [52, 57], adults lead contextually meaningful dialogue with children by asking specific types of story-related questions that develop vocabulary, phonemic awareness, recall, expressive fluency, and the building of connections between stories and a child's life. This practice can move children ahead of peers in tests of language ability [52], build children's expressive vocabulary [48], improve comprehension [15], and increase their scores on measures of linguistic complexity [57].

Enacting dialogic reading requires awareness of the importance of book-based conversation, knowledge of what questions are good [57], and cognitive attention to generate such questions while simultaneously reading aloud, which is a difficult task for adult readers to perform. Prior work has shown that eBooks that present adults with appropriate questions for their child increase on-topic dialogue during co-reading [44]. However, existing systems are hard-coded with educator-written prompts, making them inflexible and unscalable [48], or they target comprehension rather than language development, and so are focused on asking specific kinds of contentbased questions that are less conducive to open-ended dialogue [58]. Generating pedagogically-sound dialogic reading prompts at scale remains an open challenge.

Large language models (LLMs) present a promising new avenue for such dialogic question generation because they can use context to create questions that go beyond the immediate page (e.g., asking readers to make predictions or relate narrative events to their own experiences). However, while LLMs like GPT3 have helped make generative AI more accessible to the general public, operationalizing these models for specific applications can be challenging. Developers of LLM applications typically rely on adaptations like fine-tuning in order to overcome output quality concerns for specific tasks [4, 27], but this process is computationally expensive, requires high-quality task-specific data, and demands machine learning (ML) domain expertise [4]. For a task like dialogic question creation, determining quality in outputs also requires educational domain knowledge, so curation and annotation of new training data can necessitate dozens to hundreds of hours of expert time [55]. In this paper we demonstrate a new technique that is effective and orders of magnitude cheaper to instantiate.

As an alternative to fine-tuning, we draw inspiration from instructional rubrics and computational formal methods to create a self-assessing and self-correcting system. Instructional rubrics are a classroom tool in which teachers provide scoring criteria for students to self-assess their own work [1, 45, 50]. By creating a rubric to identify a good dialogic question, we encode educator expertise to evaluate data quality with comparatively little expert time and without curating a dataset. Then, to mimic self-evaluative practices for applying these rubrics within our system, we turn to counter-example guided inductive synthesis (CEGIS), a formal methods technique [8, 42]. Traditionally, this method uses test cases to self-supervise and self-correct outputs in program synthesis. By building a similar layer into a question generation system that uses the rubric scores to give targeted feedback for improvement, we can identify and generate high quality questions without fine-tuning.

In this work, we describe the design, development, and user evaluation of ContextQ, an automated LLM-enabled system for supporting dialogic questioning during parent-child co-reading. ContextQ has two components, a question generation module and a tablet-based reading interface. The question generation module uses an LLM to generate dialogic questions and applies a rubric to identify suitable outputs, providing feedback to the LLM as needed. The tablet-based reading interface then presents the selected questions unobtrusively to parents during co-reading to spur contextually meaningful dialogue. A qualitative evaluation of ContextQ investigates its impact on reading-time conversation and how parents use it as a tool in practice.

Specifically, our research contributes:

- an approach for improving LLM output quality by applying expert-informed rubrics in a self-supervising system
- ContextQ, an LLM-supported tablet application for dialogic questioning that leverages the rubric-based approach for question generation
- a qualitative evaluation with 12 parent-child dyads that demonstrates how ContextQ can be a supportive tool to guide parents in leading deeper conversations that fit learning objectives for their child (aged 4-6)

2 RELATED WORK

To situate these contributions, we review literature on dialogic reading as an educational practice, systems to support dialogic reading, and methods for question generation.

2.1 Dialogic Reading as an Educational Practice

Children's early experiences with books play an important role in reading readiness. Picture book reading can help children develop skills related to vocabulary, phonemic awareness, print meaning, narrative structure, and more [17]. But *how* we read to children matters just as much as how frequently we read to them [3, 52].

Dialogic reading is a style of co-reading in which the adult reader encourages the child to talk about the picture book while reading and in doing so models progressively more sophisticated language [52]. Children who have been read to following this dialogic technique can move significantly ahead of their peers on tests of expressive language ability even after just a few weeks [52]. They will also talk more often while reading and for longer than children who have been read to in a traditional manner [52]. Critically, these findings have held time and again across children from different countries, learning in different settings (e.g., home or daycare), and coming from a wide range of economic backgrounds [2, 32, 49, 52].

The most popular strategy for facilitating dialogic reading follows the CROWD acronym [51, 57], which outlines the types of prompts a parent might use to foster meaningful conversation. These question types are completion, recall, open-ended, wh-, and distancing (see Table 1). Each of these types of prompts has a specific educational objective, and successful dialogic reading practice will include questions of different types to advance each of these pedagogical goals. Notably, other recent strategies to facilitate parent-child conversation focus more heavily on abstract, openended discussion, but these methods were not developed for reading contexts [30, 31].

While CROWD describes the types of prompts that could foster meaningful co-reading conversation, additional characteristics of the question itself can influence its suitability in a given context. In pedagogical practice, an authentic question is loosely defined as one without a prescribed answer, motivated by a genuine desire to learn something from the respondent [36, 40]. More authentic questions also typically map to higher levels in taxonomies that categorize questions by the complexity of thought and expressive language needed to respond [10, 11, 18]. By nature of their uncertainty and complexity, these questions are more germane to dialogue, and the number of authentic questions asked by educators has functioned as a heuristic by which to quantify the quality of classroom discussion [36, 37].

It is not enough just to ask a question. Asking the right question means asking a question that fosters contextually meaningful conversation, aligns with pedagogical goals, and has the correct degree of complexity and authenticity.

2.2 Systems to Support Dialogic Reading

Given the nuances of asking "good" questions, leading dialogic questioning can be a rather demanding task for the adult reader. To do so, they must know a) that asking questions is a valuable practice and b) what makes a good question. To address these challenges, researchers developed a variety of strategies aimed at supporting family dialogic reading practices.

Early intervention studies brought parents into lab settings for in-person training led by highly-skilled teachers [35, 51]. In practice, these trainings pose access challenges to parents with limited time and resources [2, 23]. Researchers have developed video-taped trainings for caretakers to watch at home, but doing so still requires awareness and time on the part of the parent [2, 43].

Question Type	Description	Educational Objective	Example Question
Completion	Ask children to complete a sentence or phrase from the book being read.	Build phonemic awareness and introduce the structure of language.	I'll huff, and I'll puff, and I'll blow the house
Recall	Ask children about story plot, typically spanning multiple pages.	Build story comprehension, typically around sequences of events or overall themes.	Which house couldn't the Big Bad Wolf blow down?
Open ended	Encourage children to express their own ideas and opinions about the story.	Discuss the story beyond the content of the text (i.e., to make predictions or inferences).	How do you think the pigs felt when the wolf tried to get them?
Wh-	Solicit descriptive details, and begin with what, when, who, where or why.	Develop vocabulary or evaluate comprehension of a page.	What did the first pig make his house out of?
Distancing	Relate the story to children's lives and lived experiences.	Build connections between the story and the child's own life.	What's a time that someone broke something of yours? How did you feel?

Table 1: CROWD identifies five types of questions that could initiate dialogue. Each question type has a specific educational objective [52, 57]. The last column includes examples of each prompt type for the story, "The Three Little Pigs."

eBooks present a promising opportunity to encourage parentchild engagement through appropriate interaction design [26, 48]. Prior research has explored the use of an on-screen agent to drive dialogic questioning [43, 48, 54]. Strouse et al. developed an eBook for 3-year-old children with pre-recorded videos of a preschool teacher that would periodically appear in the top corner of the screen to ask a question [43], and Troseth et al. developed an enhanced eBook fro children aged 3-5 in which a character from the book appeared on-screen to ask the child questions [48]. More recently, Xu et al. created a conversational agent to ask questions during reading with bilingual families with children aged 3-6 [54]. These studies revealed that agent-driven dialogue in eBooks can lead to increased conversational turns and lexical diversity (i.e., number of unique words) between a parent and child [48], increases in parent's dialogic question-asking on pages without character-driven prompts [48], and improvements in children's story comprehension and story vocabulary [43].

Rather than relying on an agent to ask questions, StoryVisit instead presented questions to a remote adult co-reader [38]. This interface supported children under age 6 and long-distance adults in reading children's books together through a videochat co-reading interface, with the intention of increasing the duration of video calls [38]. However, the authors found the presented questions in this app were only used about 5% of the time, which they argued was due to the presence of too many other features [38].

While the above interfaces all utilized hard-coded questions for a handful of stories, StoryBuddy is an interface for presenting ML-generated questions to children aged 3–8 while reading [58]. It incorporates an AI agent that can read to the child, ask questions generated from a question answering model, and assess the correctness of responses. StoryBuddy also has a corresponding parent co-reading mode which allows the parent to take over any or all of these activities. However, the emphasis of StoryBuddy is in developing reading comprehension through the asking and evaluation of questions with a clear correct answer and on building a system that

can minimize parent involvement as needed [58], as opposed to focusing on developing narrative sense through rich parent-child dialogue, often spurred by open-ended and distancing questions that do not necessarily have a distinctly correct response.

This prior research demonstrates the promise of utilizing eBooks to drive parent-child dialogue [38, 48] and the potential for machine learning to scale this approach through question generation [58]. However, there are no systems to date aimed at automatically generating and presenting dialogic questions to adult readers in co-reading situations.

2.3 Methods for Question Generation

Question answer generation (QAG) is a field in natural language processing aimed at automatically generating both questions and answers from a body of text. Work in the QAG space primarily focuses on improving accuracy of generated questions and answers using rule-based (e.g., [12, 29]) or neural-network-based (e.g., [14, 16, 46]) models. This accuracy measures similarity to humangenerated questions, but does not generally emphasize pedagogical value.

The model underlying StoryBuddy, on the other hand, focuses on reading comprehension and was trained on a FairytaleQA, a dataset of over 10,000 questions and answers hand-written by educational experts [55, 56, 58]. While more recent QAG datasets have begun exploring open-ended and "unanswerable" questions [9], we could not find evidence of work that seeks to generate authentic questions for the purpose of driving dialogue between co-readers.

Given their broad applicability, large language models (LLMs) present a promising approach to such dialogic question generation. However, these foundation models have known output quality issues that developers must address [4, 27], and there is a further set of knowledge, expertise, and experience from educators in the context of dialogic question generation that we must consider. For the most part, LLMs (like neural-network-based QAG models) are adapted for specific use-cases through fine-tuning, but fine-tuning is

(often prohibitively) computationally expensive and requires a highquality dataset [4]. While one option for question generation using LLMs would be to hire educators with the background and skillset to generate pedagogically-sound ground-truth data and then to finetune using that dataset, this approach would necessitate dozens to hundreds of hours of expert time, require significant computational resources, and be inflexible to changes as base models improve.

To address the limitations around data annotation, we have developed an approach to encode educator expertise into the system directly and with orders of magnitude less expert time by using a rubric. Instructional rubrics are short documents to describe the quality of a written assignment to students [1]. They encode an educator's expertise and expectations to support students' self-evaluative practices when learning to write [1, 21, 45, 50].

For an LLM to leverage a rubric without computationally intensive fine-tuning, it must be able to self-supervise its own output. Prior work in machine learning has explored self-supervision for model training (e.g., GANs [20]). Within the space of computational formal methods and program synthesis (e.g., CEGIS [8, 42]), selfsupervision results in additional specifications to input into the synthesizer. In the context of LLMs, self-supervision often takes the form of safeguards and guardrails that accompany deployed models to detect and block potentially harmful inputs or outputs [34, 39]. Prior work has shown how fine-tuned LLMs can successfully apply rubrics to open-ended data [25] and how off-the-shelf models might be applied to automate deductive coding of such data [41, 53], lending promise to the approach of leveraging rubrics for self-supervision. In this work, we explore how LLMs can evaluate and self-correct their own output for a task-specific application dialogic question generation.

3 CONTEXTO

ContextQ comprises two components: a question generation module and a co-reading interface.

3.1 Question Generation Module

To generate high-quality dialogic questions, we borrow the notion of instructional rubrics from writing education. Inspired by this practice, we created a rubric that encodes educator expertise for identifying a good dialogic question. To implement the self-evaluative process of applying rubrics, we took inspiration from computational formal methods to build a layer into our question generation architecture that provides self-correction via counter-examples. We demonstrate through ablation (i.e., comparison via removal of specific parts of the system architecture) how this approach yields better questions than directly generated outputs.

3.1.1 A Rubric for Dialogic Questions. In the broadest sense, a question is "suitable" to present to a family while co-reading if it could result in pedagogically meaningful dialogue between coreaders. While this suitability may be influenced by extrinsic factors (e.g., a child's age), there are aspects of prompts that intrinsically make them more likely to elicit conversation. We looked to identify these implicit characteristics and codify them through a rubric that could classify and explain suitability decisions.

Following Boyatzis' procedures for creating codebooks from theory and data [5, 13], we began this process by constructing codes

reflecting relevant research (see Section 2.1). In determining potential codes, three researchers had a series of discussions about the frameworks that guided dialogic reading and its desired outcomes. The initial rubric was separated by each CROWD prompt type and considered question wording, authenticity, and complexity. We then reviewed and revised these codes in the context of a supervised co-reading session between one researcher and their child. Finally, two of these researchers independently applied the candidate rubric to a set of LLM-synthesized questions. These researchers had an agreement rate (Cohen's κ) of 0.79 and met to discuss and resolve all disagreements. These resolutions were used to produce a candidate theory-driven rubric.

We then iterated upon this theory-based rubric through four talk-aloud interviews with a literacy expert ¹ totaling less than 3 hours. In each session, we presented this expert with text from a children's story and dialogic questions deemed suitable by the rubric. We asked them to rate each question's likelihood of provoking meaningful contextual dialogue and to explain these decisions. We collected video and audio recordings of these interview sessions, enabling researchers to analyze sessions in detail.

We then applied iterative coding to these session recordings, looking for concepts and dimensions not captured by the extant rubric. This analysis informed new rubric iterations, which were used in subsequent sessions with the expert. Dimensions that surfaced and informed rubric revisions included the lexical structure of questions, their thematic relevance to the story, and their relevance to children's lived experiences. We achieved saturation after four sessions.

Two researchers then engaged in iterative coding, applying the rubric on a small set of questions generated from a range of children's stories and meeting to discuss and resolve identified ambiguity. These discussions led to rubric clarifications to ensure alignment with both theory and data described above.

Finally, these researchers independently coded 116 questions synthesized from four children's stories and had an inter-rater reliability (Cohen's Kappa) of 0.87. The final rubric considers level, authenticity, grammatical structure, relevance, and other question type-specific criteria to determine suitability (see Appendix A for complete rubric).

3.1.2 Question Generation Architecture. While generative models can be relied upon to quickly synthesize a huge number of candidate questions, this approach granted us few guarantees about the quality or correctness of these outputs. ContextQ addresses this problem by applying the rubric in a two-part architecture, whose interactions are described in Figure 1.

The **question synthesizer** uses the source book text and the CROWD template to engineer LLM prompts to generate candidate dialogic questions. If provided feedback on unsuitable questions from the suitability recognizer, the synthesizer augments the prompt in-context [33].

The **suitability recognizer** takes on the self-assessment and self-correction jobs of rubric application. It first operationalizes the dialogic questioning rubric through a series of prompts; these are

¹The literacy expert is an experienced researcher and former professor of literacy focused on critical literacy and classroom curriculum, multilingual classroom communities, teacher inquiry, and collaborative teacher education.

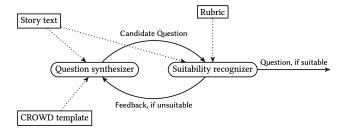


Figure 1: The question generation module utilizes a question synthesizer and suitability recognizer to produce high-quality dialogic questioning prompts. Dotted lines denote fixed inputs.

fed into an LLM alongside candidate questions in order to assess output quality. When a question is recognized as unsuitable, the recognizer prompts the LLM to identify *why* this decision was made. The candidate question and natural-language explanation of its unsuitability are then provided as a counter-example [8, 42] back to the question synthesizer.

While traditional fine-tuning methods would require dozens of hours of expert time for data annotation, we encode domain expertise in the rubric with just three hours of expert time, and feedback is incorporated without computationally expensive model training. All feedback produced by the suitability recognizer is acted upon by the question synthesizer by means of prompt adjustment, rather than any model updates. These prompts are not only specific to the task, context, and domain, but also robust to model updates and behavior changes (i.e., as base LLMs improve, we do not need to fine-tune updated models). Additionally, because the pool of candidate questions can be cheaply increased, this synthesis-recognition loop can run for several iterations until the desired number of suitable outputs is met. We used OpenAI GPT-3.5-Turbo as the LLM in our implementation because of its combined accuracy and speed. We include example prompts in Appendix B.

3.1.3 Question Quality Evaluation. To understand the impact of the suitability recognizer's rubric-based self-assessment approach and to evaluate the quality of questions produced by ContextQ's question generation module, we compared it to an ablated architecture consisting of only the question synthesizer. In total, we generated 330 questions across five stories with half produced by each system. We presented these questions to four primary-school educators who are employees at our company. They were asked to rate the likelihood of each question to foster contextually meaningful dialogue between a parent and a child on a scale from 1 (very unlikely) to 5 (very likely). Each rater scored an equal number of questions from each system while blinded to which system produced each question. Although 3 is typically considered a neutral choice on a 5-point scale, raters articulated that they used it to mean that the question could lead to contextually meaningful dialogue for certain families (e.g., depending on child age). In other words, the question was suitable, but depending on the family may be more or less fruitful.

Reasoning that a question's score depends on both the generation system and the biases of the rater themself, we ran an ordinal logistic regression with assigned ratings as a dependent variable and system and rater as the independent variables. Controlling for rater, we found that questions generated by ContextQ were 1.64 times more likely to have a higher rating than those generated by the ablated architecture lacking the suitability recognizer (95% CI: [1.11, 2.42]).

Of the 165 questions produced by ContextQ's question generation module, educators gave 131 of them a score of 3 or higher, representing an overall suitable question rate of 79%. By contrast, the ablated architecture had a suitability rate of 69%.

3.2 Co-Reading Interface

Having created a question generation module that could reliably output suitable dialogic questions, we then turned to developing a user-friendly interface to display these questions in an unobtrusive way to caretakers. Critically, for this prototype, the type of CROWD question generated and presented was randomly selected; because not every question type can be generated on each page (e.g., recall questions cannot appear on the first page and completion prompts are only suitable in instances with rhymes or repetition), if the system tries and fails to generate a question of a certain type three times, it will move onto a different question type.

We designed the tablet-based interface such that the question would appear in a colored tooltip at the bottom of the screen, without covering any of the content on the page (see Figure 2). This tooltip has a small info button in the top right corner that presents a three-sentence pop-up describing the importance of dialogue during co-reading.

3.2.1 Iterative Design Testing. To correct bugs and ensure usability, we conducted iterative testing with 8 parent-child groups (seven dyads and one two-child triad) remotely over video conferencing software. The parents were employees of a large technology company with a mix of technical and non-technical roles and their children ranged in age from 4 to 7 years old (M=6.1, SD=1.2). We selected this age range as one where children would be comprehensible and able to engage in dialogue but not yet reading independently. (In practice, many 7 year olds preferred reading themselves over being read to, hence the narrowed age range in our later evaluation study.)

After giving informed consent, families either installed our prototype application onto their iPad or we screenshared an iPad simulator in cases when installation was not possible (e.g., incompatible software version). We then observed as families read two children's stories—one without on-screen prompts and one with them—before collecting feedback from them on how to improve their experience. For this iterative design testing, we used hard-coded dialogic questions rather than generated ones in order to get feedback on specific questions and question types.

Based on these observations and families' feedback we made several design decisions. First, parents expressed confusion on when the question should be read, so we chose to move the prompt from the top of the screen to the bottom to suggest reading it after finishing the page. Second, because younger children were distracted by the presence of touch points on the screen, we also



Figure 2: ContextQ presents dialogic questions to parents via a tablet-based reading application.

removed an audio button from the tooltip which played synthesized speech of the prompt out loud. Finally, we implemented a page-turn button in place of page turn swipes to allow families to point at and touch the page during dialogue; we later increased the tap target dimensions of this button based on iterative testing observations of mistaps.

Once we ran several consecutive sessions without further design updates, we proceeded to a system evaluation of the connected interface and question generation model, described in the next section.

4 EVALUATION STUDY

With ContextQ, we aimed to design a dialogic reading interface to support parents in leading contextual dialogue. We conducted a qualitative study to understand the impact of this system on reading-time conversation and how parents use this system in practice. Through this study we aimed to answer the following research questions:

- (1) How does a dialogic reading interface impact parent-child reading-time conversation as compared to a traditional eBook interface?
- (2) How do parents use the dialogic reading interface as a tool in practice?

4.1 Participants

As we aimed to study the impact of the system on parents as they read aloud to their child, we recruited 12 parent-child dyads to participate in a one hour online study session. To qualify for the study, families had to speak fluent English and all child participants needed to be 4–6 years old (see Table 2 for participant details). To include a range of backgrounds, families were recruited through public schools serving communities with low and mid socioeconomic status (based on median household income; N=5 dyads) and via word of mouth (N=7). Families came from 4 different US states and participated using either family-owned iPads or iPads that we lent to participants for this study. They received a \$50 gift card in exchange for their time.

4.2 Procedure

Sessions were one hour long and were conducted remotely via videoconferencing software. Families were instructed to join the call from a device other than the iPad they would be using for the study.

During the session, we first obtained informed consent before installing the application onto the family's iPad. As the app installed, we asked the family for background information on their typical co-reading habits, including when, how often, and why they choose to read together.

We then asked families to use the application to read two short stories (each 300 words split across 6 pages), which were included in the app for the study. We selected and modified two of Aesop's fables, *The Lion and the Mouse* and *The Fox and the Stork*, to match in style, complexity level, page count, and word count; the first two families read different stories of similar complexity. While reading

Family	Child	Parent	Typical Co-Reading	Typical	Recruitment
ID	Age	Relationship	Minutes per Week	Book Type	Avenue
1	6	Mother	20-30	Print	School
2	5	Mother	60-75	Print	School
3	5	Mother	140	Print	School
4	4	Mother	30-50	Print	School
5	6	Mother	60-75	Print	Word of Mouth
6	5	Father	100	Print	Word of Mouth
7	4	Mother	120-150	Print	Word of Mouth
8	6	Mother	0-60	Print	School
9	6	Mother	140	Print and eBook	Word of Mouth
10	6	Mother	105-140	Print	Word of Mouth
11	5	Father	105	Print and eBook	Word of Mouth
12	6	Mother	110	Print	Word of Mouth

Table 2: Participant demographic information

the first story, the app displayed a standard eBook interface (No Questions) and during the second story the app also presented a generated question on each page (With Questions). We did not indicate that questions would appear in the second book or provide any instruction about how families should use the presented questions.

We included the No Questions interface to help us understand each family's typical reading behaviors and to serve as a point of comparison for families when reflecting on the presented prompts. Therefore to minimize potential learning effects that could impact the types of questions parents asked or the conversations families had, we chose to always present this No Questions interface first. Critically, we told families directly at the start of the reading phase that we were "trying to learn about the conversations families have while reading" so the changing interface between books (No Questions to With Questions) did not clue the parent into the research objectives halfway through the session. We alternated the order in which the stories themselves were read.

After reading both books, we conducted a semi-structured interview to get feedback from families about their experience using the app, their prior knowledge about dialogic reading, and how they might put dialogic reading into practice in their homes. While we focused on the parent perspective during this interview, children were present and permitted to respond as well. We include a full list of interview questions as well as the text of both stories in our Supplementary Materials.

4.3 Data Analysis

While we carefully considered the order in which families used the two versions of the interface and read the two stories, the primary goal of the study was to gauge qualitative reactions to using the dialogic question system rather to experimentally compare system versions. We thus focus our analysis and findings on the qualitative data while presenting quantitative comparisons more briefly.

We collected video and audio recordings of participants from all sessions. These sessions were automatically transcribed via the video conferencing software and those transcriptions were manually corrected by the research team. We applied a semantic and realist thematic analysis to this data, taking a combined deductive and inductive approach [7]. Our final codebook is included in the Supplementary Materials.

We coded interview data at a per interview question unit, focusing on the goals parents have for reading, the challenges they face when engaging in dialogue, the perceived benefits of the app, and any concerns with the app.

For the reading phase, we identified segments of contextually meaningful dialogue to analyze how conversation and behaviors change when supported by a dialogic reading interface. In this case, contextual meant the dialogue specifically related to the book reading activity and meaningful meant it further connected to the story's content (i.e., plot, themes, lessons, or pictures). This definition therefore excludes off-topic conversation (e.g., the child asking for a snack) and dialogue about the activity, but not the content (e.g., who presses the page turn button). Given the focus of our research, our unit of analysis for this book-reading phase of the study was a dialogic interaction, which we identified as a conversation during a pause in reading separated from the prior or next dialogic interaction via 1) a page turn, 2) an off-topic conversation, or 3) a clear transition statement (e.g., "Something else I was going to ask...").

While we focus our analysis on qualitative findings that convey how ContextO influenced reading behaviors, we report a more quantitative perspective of the collected data to explore how much of an impact the presented questions had. To quantify dialogue, we consider the number of contextually meaningful conversational turns from both the parent and the child, where a contextually meaningful conversation turn is an utterance by the parent or child related to the book's contents. We then look to individual dialogic interactions to lend additional perspective on the contents of conversations by reporting the prevalence of each CROWD prompt type (i.e., the first question asked in a dialogic interaction) and the number of dialogic interactions covering each identified conversation topic (e.g., plot or morals). Finally we consider how parents used ContextQ as a tool in practice by reporting on the number of times parents used, modified, or ignored presented dialogic questions.

5 RESULTS

From our thematic analysis, we identified three key themes in this data that describe how family conversations differ when supported by ML-generated dialogic questions, and how such a reading interface might impact family reading practices.

5.1 Theme 1: Generated Questions as a Parent Guide for Conversation

Parent participants described ContextQ as a "guide" to help them better engage with their child. Parents by and large reported that they do typically try to talk to their child about books while reading them, but they also described the challenges they regularly faced around leading conversation including forgetfulness, exhaustion, or difficulties thinking of questions to ask. By contrast, when reflecting on their experiences using the app, they described how the presented dialogic questions could help them overcome these common challenges.

When discussing their typical home reading behaviors parents identified challenges that often prevent them from starting conversations. Some, like P10, reflected on "questions that sometimes we forget to ask because we're just reading," suggesting that a focus on getting through the story can distract from the goal of dialogue. Other parents reported that even when they remember to ask questions, "it's kind of hard to think of new ways to make it kind of a conversation," especially when "kids love to reread books over and over again" (P11).

Not all parents start conversations while reading though, as some described that when reading at home they often consciously choose not to ask questions because they lack the energy or presence to think of them. P11 typically reads to his children right after work, and can find it hard to completely pull his mind away from other tasks: "If I am worried about something or I've got something brewing in the back of my head it's harder to think or want to make conversation." P8, on the other hand, reads to his child at bed time, but says that one downside of this is that "the questions are less interesting because there's the—there's just like a tiredness, both of child and adult." P3 similarly points to exhaustion as a reason why she may choose to skip conversation while reading, especially when she's had a long day. "Sometimes it's—it's hard...when you're at the end of your day, if you've been working all day, to have the questions to say: How are you feeling? What did you notice?"

However, parents believed that ContextQ would serve as a reminder to talk about stories, an inspiration when thinking of questions, and a tool for tired parents. For the forgetful, P8 appreciates the presence of questions "that are there to remind me to ask her," and P3 explains that the on-screen "prompts can also have you pause and remember what the point of reading it to them is. It's not just to—to read for them to—to listen. It's for them to actually respond as well." Parents like P4 who typically have a hard time thinking of questions to ask appreciated that "it's not a question that I have to come up with myself to try to spark conversation." As P10 reported: "I like the fact that...they gave you the questions to ask them. That way, you know, you don't really have to think about it. You just, you know, go along with it." Finally, when reflecting on exhaustion, P3 also believes that ContextQ can provide support for parents "so that

sometimes if you're tired, you don't have to come up with the questions yourself." Whether forgetful, distracted, tired, or struggling to think of a prompt, parents appreciated that ContextQ was "kind of a guide [that] could be helpful for parents reading to kids" (P9).

With participants describing ContextQ as an effective parent guide, we would expect to see more meaningful contextual dialogue when presented with dialogic questions than when reading without them. Indeed, we find that parents and children alike had significantly more meaningful conversational turns when reading with presented generated questions (Parents: Range: 7-74, M=30.8, SD=23.2; Children: Range: 3-67, M=25.1, SD=20.2) than without (Parents: Range: 0-28, M=13.4, SD=9.3; Children: Range: 0-25, M=10.6, SD=8.8). A paired-samples Wilcoxon test found this difference was significant for both parents (z=-3.06, p=0.002, r=0.88) and children (z=-2.94, z=0.004, z=0.88).

5.2 Theme 2: Depth Beyond Comprehension

ContextQ deliberately presented all five types of CROWD prompts to parents, including the comprehension-based wh- and recall prompt types more commonly found in prior work, as well as the more authentic open-ended and distancing types. We posited that the more authentic questions would lead to richer dialogue, but recognized the need for attention and comprehension checks as well.

Parents reflecting on their reading experience generally reported that they disliked the plot-based comprehension questions so central to prior work because they "just ask you to just repeat what you just read" (P6). When discussing which prompts were least useful, P7 specifically called out a wh- prompt that appeared for her while reading: "What promise did the mouse make to the lion?" This parent explains that this sort of question doesn't achieve the objectives she has when having dialogue with her child because "it doesn't really trigger anything except like, you know, rote memorization" and that asking strictly comprehension-based questions causes her child to tune out: "Do you know what this means? Do you remember what happens on this page? I frankly think I lose his attention if I do it that way." P4 further described how these comprehension questions are less helpful to her when determining how to start a conversation because they're comparatively easy for her to come up with on her own: "I'm not a very creative person...but, like, asking, like, reading comprehension questions doesn't really require any creativity of me."

On the other hand, parents expressed an appreciation for ContextQ's authentic questions that allowed them to achieve a level of depth with their discussions beyond strict comprehension. For example, these questions asked children to consider the character's emotions (e.g., "How do you think the mouse felt after she helped the lion get out of the net?") or to relate events in the story to their own life (e.g., "Have you ever been in a situation where you needed someone's help? How did they assist you?"). When reflecting on how the presented prompts could be most useful, P5 explained: "I thought the questions easily kind of let you get both characters' perspectives, and also prompted me to go a little bit deeper in what I would have thought to connect the story to...I just feel like these are open, kind of, you know, those open-ended questions that you should be asking." More specifically, parents, like P12, wanted to use questions as a way to connect stories to long-running and ongoing

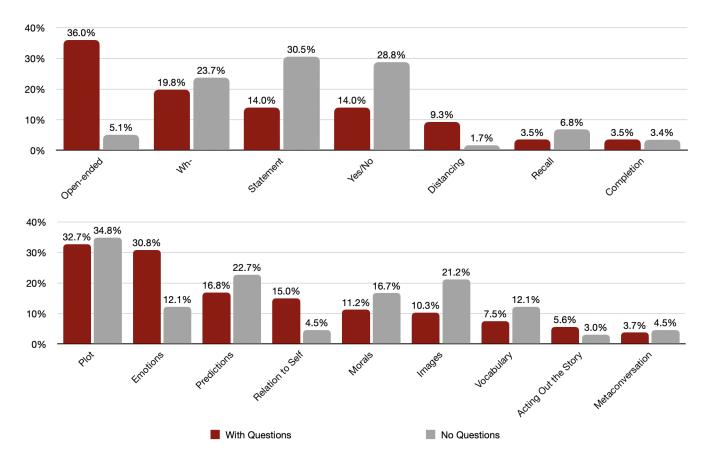


Figure 3: Top: For each interface, the proportion of parent-led dialogic interactions that began with each question type. Bottom: Proportion of dialogic interactions in each interface that touched on a particular topic. Meta conversations were discussions about the reading experience (e.g., do you like the story). Note: conversations could have more than one topic so percentages for each interface do not sum to 100%.

conversations with her child: "It creates a bigger discussion...because I think you can use the books and the stories as, like, life metaphors to just kind of help with, like, life situations. So to really apply it." As P2 explains, these questions drive the conversations deeper "because sometimes I'm not thinking about, like, what the potential lessons are in the book."

In particular, parents appreciated being presented with openended or distancing questions that tied to lessons they were trying to teach in day-to-day life around morals and emotional awareness. P12 felt the presented questions were "designed to help pull out like the messages of the story, create a conversation around the meaning, and help kind of think through the lessons", and she felt that this lesson-based questioning "creates a bigger discussion." Notably, it was often the questions related to emotions, in particular, that resonated with parents, because they mirrored ongoing socialemotional lessons the parent was trying to teach at home. P3 has been teaching her child how to express herself and her emotions and thinks guided discussions about books and characters could help with this effort: "We've been actually working a lot on feelings and describing feelings. So I think a lot of the questions were actually

helping to discuss feelings." Similarly, P7 felt that questions for children related to emotions (e.g., "What do you think the mouse felt when she saw the lion trapped in the net?") were "helpful" and "age appropriate" because "at this age...especially is when they're trying to understand feelings and things like that."

Parents articulated a clear preference for open-ended and distancing questions over comprehension questions and for discussions about life lessons over discussions about plots. We examined the types of questions parents asked to start conversations and the topics of those conversation (identified through the qualitative coding process) across the No Question and With Question interfaces to study the extent to which ContextQ supported this preference. When considering the type of question that parents asked to start a conversation, we find that 45.3% of parent-led conversations started with open-ended or distancing questions when using the With Questions interface; by contrast only 6.8% of conversations with the No Questions interface began with one of those question types (see Figure 3). Similarly, when looking at topics discussed during these dialogic interactions identified via our inductive coding process, 30.8% of conversations when reading with presented dialogic questions related to emotions as opposed to 12.1% when reading without

them. In contrast, other common conversation topics (e.g., plot) were much more balanced between the first and second story (see Figure 3).

One critical caveat to this finding is that the system itself did not display these question types in equal distribution. For example, each story only had a single page for which a suitable completion prompt could be created, but suitable open-ended prompts are easily created for any page of the story. As a result of this difference and the question type randomization built into the system, a completion prompt appeared on screen only once across the full study, whereas open-ended prompts were displayed 36 times (Recall: 9, Wh-: 12, Distancing: 14). This bias in presentation certainly affected the question type counts above, but simultaneously aligned with parents' articulated preferences.

5.3 Theme 3: Parents as a Mediator Between the App and the Child

Finally, we looked at how parents used the presented questions when reading to support conversation with their child and found that parents read the question as written on the screen only about a third of the time (i.e., for 23 of the 72 presented questions). In the remaining instances we observed ways a parent served as a mediator between the app and the child by rewording questions or selectively choosing to disregard them.

A large subset of those rewordings consisted of minor changes where the parent swapped a word or reordered the question, but stayed close to the presented prompt. In 7 cases, a parent split apart a two part question, asking those parts separately, without otherwise changing the wording of the prompt. In an additional 12 cases, the parent slightly reworded the question. Most of these rewordings had minimal affect on the meaning of the prompt (e.g., "How do you think the lion felt when the mouse ran across his nose?" became "How do we think the lion felt after the mouse ran across his nose?"), but some of the changes added clarity to the prompt itself. For example, P2 prompted their child with the on-screen prompt, "Can you think of a time when you really wanted to do something but you had to stop yourself? How did you feel?" However, their child misunderstood the prompt and instead discussed how the character felt during this part of the story. On the following page, the app presented P2 with a similar question ("Can you think of a time when you really wanted to do something, but you knew shouldn't? What did you do?"), but this time P2 chose to reword the question to clarify C2's prior misunderstanding: "Can you think of a time when you really wanted to do something-you, [C2's name]-but you shouldn't do it? Like you knew you weren't supposed to do it."

Aligning with this observed behavior, in the post-session interviews, P3 spoke directly about how she reworded a question for clarity when she felt her child wouldn't understand: "I think that actually the one about repaying...I don't know that she would actually understand, like, what it means to repay. Um, so maybe to give back."

On 7 occasions, we observed instances that we classified as major rewording of the on-screen question. These rewordings appeared to serve one of two purposes. In some cases, the parent drastically simplified the question for the child. For example, P12 took the question "How do you think the stork felt when they realized they had been tricked by the fox?" and reworded it to say "Do you think"

the stork's going to like that?" In doing so, she converted an openended question into a yes/no prompt. While this change may not lead to as rich of a discussion, a parent may opt to make such a change when the child is getting restless or is unable to articulate an answer to the more complex question.

The other reason we observed question rewording was for instances of malformed questions, which appeared three times throughout the study. We know from our ablation study on the question generation module (see Section 3.1.3) that not every questions the model outputs is perfectly suitable, so we were interested in how parents would deal with flawed outputs. Twice the system asked how a character was feeling about certain events that would be more appropriately asked about a different character. In both of these cases, the parent mediated between the app and the child to appropriately reword the prompt. For example, P10 corrected a prompt in one such instance while reading it aloud: "How do you think the stork felt when he saw the fish dinner served in a jar with a—I think it's how do you think the fox felt?"

The other case of a malformed question appeared during the second session before we identified a rare edge-case bug that caused a question to show on the incorrect page. In this instance, P2 instructed C2 to disregard the question and turn the page because it could not be answered until later in the story.

Finally, in addition to this one instance of skipping due to a malformed prompt, parents chose to ignore the question on the page 22 other times during the study (i.e., 23 total ignored questions out of 72 presented). On five of those occasions, they chose to ask a different question, whereas on the remaining occasions, they turned the page without asking a question. In some cases, the app prompted parents with two very similar questions in a row; the parent asked the question when it first appeared and then chose not to ask it again on the next page. In other cases, the parents simply reported that they did not initially notice the (deliberately unobtrusive) questions at first, and only began to read them once they had. Finally, parents may choose to ignore questions when they feel the setting and moment is not conducive to meaningful dialogue; as P7 explained, "depending on the mood we can use it or skip through." Ultimately, the on-screen questions are there to support the parent without distracting the child, and it is up to the parent to decide how and when to mediate between their child and the app itself.

6 DISCUSSION

Through our system and evaluation, we demonstrate that we can generate questions that support meaningful contextual dialogue during parent-child co-reading scenarios.

In this section, we discuss how generating questions without answers can lead to richer discussion and better support parent preferences, how unobtrusive design supports parents without distracting a child, how the suitability recognizer approach might generalize to other applications, and ways to extend this system in future work.

6.1 Questions Without Answers for Rich Discussion and Parent Ease

Through our analysis, we found that parents viewed ContextQ as a guide to help them start conversations, leading to significantly more meaningful conversational turns from both parents and children. Parents utilized the suggested questions the majority of the time, but mediated between the app and the child to modify the presented question to fit the context. When studying the contents of those conversations, we learned that parents appreciated open-ended and distancing questions that tied to real-life situations over comprehension questions, because these authentic questions were more difficult to come up with, led to deeper discussion, and provided more opportunities for bonding.

Returning to related work on systems for dialogic questioning, we found a number of prior systems that successfully used onscreen agents to drive dialogue about books [43, 48, 54], although
systems that relied on an adult co-reader for this job previously
found that extraneous features distracted from this task and lead
to low question utilization [38]. By contrast, with ContextQ's simple interface, we found high question utilization and significant
increases in the amount of dialogue.

Prior models and systems that generated questions focused on comprehension questions with predefined answers [56, 58]. In educational contexts, those questions support students in developing reading comprehension by allowing them to respond and then, critically, check the correctness of the response [58]. When we began this project we established that solely assessing comprehension was not the goal we were striving for. Instead we followed the CROWD method to generate questions that are germane to dialogue in order to encourage parents toward reading practices that have been shown to advance children's early literacy skills. Parents found suggested comprehension questions like those so dominant in prior work comparatively unhelpful. We initially included more authentic open-ended and distancing questions because they support rich discussion, but we found that these were the questions parent preferred the most but felt least equipped to come up with on their own.

Interestingly, this finding reflects more recent research in supporting parent-child conversation outside of reading contexts (i.e., at mealtimes) [30, 31]. This body of work trains parents in conversations about recalling past events, explaining concepts, or discussing the future, and shows such talk led to more conversation turns [31] and more abstract references in children's everyday language [30]. The work we present in this paper contributes to this growing evidence on the importance of and preference for open-ended conversation strategies between parents and children.

6.2 Generalizing the Suitability Recognizer to Other Applications

Traditional fine-tuning methods are computationally expensive and require high-quality data that can demand dozens to hundreds of hours of expert time to curate or annotate [4, 55]. We developed the rubric as a way to improve model output quality by encoding expertise in the absence of an appropriate dataset and the suitability recognizer in order to apply that expertise within a system. In this paper, we show that this technique is highly effective while also

requiring orders of magnitude less expert time to implement (just 3 hours), no additional computation for training, and no dataset.

One limitation of the specific instantiation of this approach presented in this paper is that our rubric is based on the input of a single expert. While we show that just one expert's input can still yield higher quality generated questions, it is possible that incorporating input from multiple experts in rubric development could yield even better results, albeit at the cost of additional expert time.

While dialogic question generation is one application of this approach, we imagine it could be generalized to other use-cases when expertise is needed to evaluate output quality or when finding large-scale datasets is challenging. For example, considerable work has explored the space of using LLMs for short story generation. Fine-tuning could leverage datasets of stories, but most high-quality stories fall under copyright. Alternatively, developers could consult narratologists to identify what makes a compelling story. Encoding that expertise into a suitability recognizer could provide a filter to identify and correct for high-quality stories.

6.3 Future Work

Looking ahead, we see several opportunities to extend the work presented in this paper.

6.3.1 Generating Prompts from Images. ContextQ used only the text of a story as input when generating questions. However, during the evaluation study, we observed how families occasionally chose to discuss the images in the story. Furthermore, while we selected stories that were entirely comprehensible without images, some children's books (e.g., Milo's Hat Trick by Jon Agee) include pages where the story is conveyed with images alone. Future work might consider leveraging multimodal LLMs to consider images as input for question generation alongside story text.

6.3.2 Considering Additional Context when Assessing Suitability. Observations of parents opting to ignore questions due to their similarity to prior questions suggests an opportunity to consider additional context beyond the text on the page. At the simplest, future iterations of this system could include in suitability recognition some check that the generated question is sufficiently semantically different from previously presented questions or allow the family to select which question type should be presented.

Taking this notion of context a step further though, future work may consider the individual child when determining suitability or what type of question to ask. For example, it may make sense for a child reading a book for the fifteenth time to be asked different (e.g., more cognitively complex) questions than a child hearing it for the first time. Similarly, a parent may want to ask different questions to their four year old than to their six year old, even when reading the same story, due to the older child's more advanced vocabulary, narrative understanding, or theory of mind. We witnessed how parents engaged in a mediation process to adjust questions to suit their child, but given the right contextual information, it is possible for a system to take on some of this work as well.

6.3.3 Longitudinal Evaluation. Finally, dialogic questioning as a practice has proven so successful across socio-economic groups, cultures, and languages because researchers have demonstrated

that caretakers and educators trained in this practice can help children jump months ahead in early literacy skills [2, 32, 49, 52]. We developed ContextQ, though, with the understanding that such training does not reach most parents; indeed when asked at the end of the study, only three of the twelve participating parents reported prior formal knowledge of the importance of asking questions to children while reading. Our work shows that machine-generated prompts support untrained parents in leading contextual dialogue while co-reading, but future work should explore if this approach can 1) show quantifiable differences in a large-scale experimental controlled setting, 2) teach parents to lead more consistent or effective dialogue over time, even in the absence of technological support, and 3) lead to childhood literacy gains in a manner that mirrors prior studies on the effectiveness of dialogic reading when introduced with formal training.

7 CONCLUSION

In this work, we presented ContextQ, a tablet-based reading application to unobtrusively present dialogic questions to parents during co-reading. We generate these questions by leveraging an LLM in conjunction with a suitability recognizer that encodes educational expertise in a rubric to identify high-quality outputs. This system produces higher-quality questions with orders of magnitude less expert time and no additional training. Through a qualitative evaluation, we find that ContextQ serves as a guide for parents leading to more contextually meaningful conversation turns from both the parent and the child, supports parents in guiding deeper conversations that tie to real-life learning objectives through more open-ended and distancing questions, and enables parents to mediate question content and wording between the app and their child. Looking ahead, this work presents exciting opportunities for technology to close gaps in early literacy education by supporting caretakers in making the most of reading time with their children.

8 SELECTION AND PARTICIPATION OF CHILDREN

Families were recruited from an early childhood center in Connecticut, an elementary school in New Jersey, and an elementary school in Texas; additional families were recruited via word of mouth in Texas and California. All contacted families received an email with study information that included a link to contact the research team to express interest in participating. The research team then confirmed eligibility (i.e., within age-range and English speaking), scheduled sessions, and shipped loaner iPads to families as needed for this study. All children provided verbal assent to participate in the study and to be video recorded, and parents signed consent forms prior to video and audio data collection. Families received a \$50 gift card in exchange for participating.

ACKNOWLEDGMENTS

We would like to thank Katie Van Sluys, Katie Morrow, Gabriella Meyers, and Megan Kinsey for contributing their expertise to rubric development and evaluation.

REFERENCES

- Heidi Goodrich Andrade. 2000. Using rubrics to promote thinking and learning. Educational leadership 57, 5 (2000), 13–19.
- [2] David H Arnold, Christopher J Lonigan, Grover J Whitehurst, and Jeffery N Epstein. 1994. Accelerating language development through picture book reading: replication and extension to a videotape training format. *Journal of educational* psychology 86, 2 (1994), 235.
- [3] Isabel L Beck and Margaret G McKeown. 2001. Text talk: Capturing the benefits of read-aloud experiences for young children. The reading teacher 55, 1 (2001), 10-20.
- [4] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258 (2021).
- [5] Richard E Boyatzis. 1998. Transforming Qualitative Information: Thematic Analysis and Code Development. Sage Publications.
- [6] Ernest L Boyer and Carnegie Foundation for the Advancement of Teaching (Princeton NJ). 1991. Ready to learn: A mandate for the nation. Carnegie Foundation for the Advancement of Teaching Princeton, NJ.
- [7] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. Qualitative research in psychology 3, 2 (2006), 77–101.
- [8] Milan Češka, Christian Hensel, Sebastian Junges, and Joost-Pieter Katoen. 2019. Counterexample-Driven Synthesis for Probabilistic Program Sketches. In Formal Methods – The Next 30 Years, Maurice H. ter Beek, Annabelle McIver, and José N. Oliveira (Eds.). Springer International Publishing. Cham. 101–120.
- [9] Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. arXiv preprint arXiv:1808.07036 (2018).
- [10] Arthur Costa and Lawrence F Lowery. 1989. Techniques for teaching thinking. Routledge.
- [11] Arthur L. Costa. 2001. Toward a Model of Human Intellectual Functioning. Association for Supervision and Curriculum Development.
- [12] Sérgio Curto, Ana Cristina Mendes, and Luísa Coheur. 2011. Exploring linguistically-rich patterns for question generation. In Proceedings of the UC-NLG+ eval: Language Generation and Evaluation Workshop. 33–38.
- [13] Jessica T. DeCuir-Gunby, Patricia L. Marshall, and Allison W. McCulloch. 2011. Developing and Using a Codebook for the Analysis of Interview Data: An Example from a Professional Development Research Project. Field Methods 23, 2 (2011), 136–155. https://doi.org/10.1177/1525822X10388468 arXiv:https://doi.org/10.1177/1525822X10388468
- [14] Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pretraining for natural language understanding and generation. Advances in neural information processing systems 32 (2019).
- [15] Brooke Graham Doyle and Wendie Bramwell. 2006. Promoting emergent literacy and social-emotional learning through dialogic reading. The Reading Teacher 59, 6 (2006), 554–564.
- [16] Xinya Du, Junru Shao, and Claire Cardie. 2017. Learning to ask: Neural question generation for reading comprehension. arXiv preprint arXiv:1705.00106 (2017).
- [17] Elisabeth Duursma, Marilyn Augustyn, and Barry Zuckerman. 2008. Reading aloud to children: the evidence. Archives of disease in childhood 93, 7 (2008), 554-557
- [18] Kylie S. Flynn. 2011. Developing Children's Oral Language Skills through Dialogic Reading: Guidelines for Implementation. TEACHING Exceptional Children 44, 2 (2011), 8–16. https://doi.org/10.1177/004005991104400201 arXiv:https://doi.org/10.1177/004005991104400201
- [19] David J Francis, Sally E Shaywitz, Karla K Stuebing, Bennett A Shaywitz, and Jack M Fletcher. 1996. Developmental lag versus deficit models of reading disability: A longitudinal, individual growth curves analysis. *Journal of Educational* psychology 88, 1 (1996), 3.
- [20] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. Commun. ACM 63, 11 (2020), 139–144.
- [21] Heidi Watts Goodrich. 1996. Student self-assessment: At the intersection of metacognition and authentic assessment. Harvard University.
- [22] Donald J Hernandez. 2011. Double jeopardy: How third-grade reading skills and poverty influence high school graduation. Annie E. Casey Foundation (2011).
- [23] Annemarie H Hindman, Barbara A Wasik, and Emily K Snell. 2016. Closing the 30 million word gap: Next steps in designing research to inform practice. *Child Development Perspectives* 10, 2 (2016), 134–139.
- [24] Connie Juel. 1988. Learning to read and write: A longitudinal study of 54 children from first through fourth grades. *Journal of educational Psychology* 80, 4 (1988), 437
- [25] Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, et al. 2023. Prometheus: Inducing Fine-grained Evaluation Capability in Language Models. arXiv preprint arXiv:2310.08491 (2023).

- [26] Ofra Korat and Tal Or. 2010. How new technology influences parent—child interaction: The case of e-book reading. First Language 30, 2 (2010), 139–154.
- [27] Diane M Korngiebel and Sean D Mooney. 2021. Considering the possibilities and pitfalls of Generative Pre-trained Transformer 3 (GPT-3) in healthcare delivery. NPJ Digital Medicine 4, 1 (2021), 93.
- [28] Marina Krcmar and Drew P Cingel. 2014. Parent-child joint reading in traditional and electronic formats. Media Psychology 17, 3 (2014), 262–281.
- [29] Igor Labutov, Sumit Basu, and Lucy Vanderwende. 2015. Deep questions without deep understanding. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 889–898.
- [30] Kathryn Leech, Ran Wei, Jeffrey R Harring, and Meredith L Rowe. 2018. A brief parent-focused intervention to improve preschoolers' conversational skills and school readiness. *Developmental psychology* 54, 1 (2018), 15.
- [31] Kathryn A Leech et al. 2021. An intervention to increase conversational turns between parents and young children. *Journal of Child Language* 48, 2 (2021), 399–412
- [32] Young Sook Lim and Kevin N Cole. 2002. Facilitating first language development in young Korean children through parent training in picture book interactions. Bilingual Research Journal 26, 2 (2002), 367–381.
- [33] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. arXiv:2107.13586 [cs.CL]
- [34] Todor Markov, Chong Zhang, Sandhini Agarwal, Florentine Eloundou Nekoul, Theodore Lee, Steven Adler, Angela Jiang, and Lilian Weng. 2023. A holistic approach to undesired content detection in the real world. In Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 37. 15009–15018.
- [35] Suzanne E Mol, Adriana G Bus, Maria T De Jong, and Daisy JH Smeets. 2008. Added value of dialogic parent-child book readings: A meta-analysis. Early education and development 19, 1 (2008), 7–26.
- [36] Martin Nystrand. 1997. Opening dialogue: Understanding the dynamics of language and learning in the English classroom. Teachers College Press.
- [37] Martin Nystrand, Lawrence L Wu, Adam Gamoran, Susie Zeiser, and Daniel A Long. 2003. Questions in time: Investigating the structure and dynamics of unfolding classroom discourse. *Discourse processes* 35, 2 (2003), 135–198.
- [38] Hayes Raffle, Glenda Revelle, Koichi Mori, Rafael Ballagas, Kyle Buza, Hiroshi Horii, Joseph Kaye, Kristin Cook, Natalie Freed, Janet Go, et al. 2011. Hello, is grandma there? let's read! StoryVisit: family video chat and connected e-books. In Proceedings of the SIGCHI conference on human factors in computing systems. 1195–1204.
- [39] Traian Rebedea, Razvan Dinu, Makesh Sreedhar, Christopher Parisien, and Jonathan Cohen. 2023. Nemo guardrails: A toolkit for controllable and safe Ilm applications with programmable rails. arXiv preprint arXiv:2310.10501 (2023).
- [40] Caroline Schaffalitzky. 2022. What makes authentic questions authentic? Dialogic Pedagogy: An International Online Journal (2022).
- [41] Bruce Sherin. 2013. A computational study of commonsense science: An exploration in the automated analysis of clinical interview data. Journal of the Learning Sciences 22, 4 (2013), 600–638.
- [42] Armando Solar-Lezama, Liviu Tancau, Rastislav Bodik, Sanjit Seshia, and Vijay Saraswat. 2006. Combinatorial Sketching for Finite Programs. SIGARCH Comput. Archit. News 34, 5 (oct 2006), 404–415. https://doi.org/10.1145/1168919.1168907
- [43] Gabrielle A Strouse, Katherine O'Doherty, and Georgene L Troseth. 2013. Effective coviewing: Preschoolers' learning from video after a dialogic questioning intervention. *Developmental psychology* 49, 12 (2013), 2368.
- [44] Zachary D Stuckelman, Gabrielle A Strouse, and Georgene L Troseth. 2021. Value added: Digital modeling of dialogic questioning promotes positive parenting during shared reading. *Journal of Family Psychology* (2021).
- [45] Todd H Sundeen. 2014. Instructional rubrics: Effects of presentation options on writing quality. Assessing writing 21 (2014), 74–88.
- [46] Duyu Tang, Nan Duan, Tao Qin, Zhao Yan, and Ming Zhou. 2017. Question answering and question generation as dual tasks. arXiv preprint arXiv:1706.02027 (2017).
- [47] Joseph K Torgesen and Stephen R Burgess. 1998. Consistency of reading-related phonological processes throughout early childhood: Evidence from longitudinalcorrelational and instructional studies. Word recognition in beginning literacy 161 (1998), 188.
- [48] Georgene L Troseth, Gabrielle A Strouse, Israel Flores, Zachary D Stuckelman, and Colleen Russo Johnson. 2020. An enhanced eBook facilitates parent–child talk during shared reading by families of low socioeconomic status. Early childhood research quarterly 50 (2020), 45–58.
- [49] Marta C Valdez-Menchaca and Grover J Whitehurst. 1992. Accelerating language development through picture book reading: A systematic extension to Mexican day care. Developmental psychology 28, 6 (1992), 1106.
- [50] Despina Vasileiadou and Konstantinos Karadimitriou. 2021. Examining the impact of self-assessment with the use of rubrics on primary school students' performance. *International Journal of Educational Research Open* 2 (2021), 100031.

- [51] Grover J Whitehurst, Jeffrey N Epstein, Andrea L Angell, Adam C Payne, Deanne A Crone, and Janet E Fischel. 1994. Outcomes of an emergent literacy intervention in Head Start. *Journal of Educational Psychology* 86, 4 (1994), 542.
- [52] Grover J Whitehurst, Francine L Falco, Christopher J Lonigan, Janet E Fischel, Barbara D DeBaryshe, Marta C Valdez-Menchaca, and Marie Caulfield. 1988. Accelerating language development through picture book reading. 24, 4 (1988), 552.
- [53] Ziang Xiao, Xingdi Yuan, Q. Vera Liao, Rania Abdelghani, and Pierre-Yves Oudeyer. 2023. Supporting Qualitative Analysis with Large Language Models: Combining Codebook with GPT-3 for Deductive Coding. In 28th International Conference on Intelligent User Interfaces. ACM. https://doi.org/10.1145/3581754. 3584136
- [54] Ying Xu, Kunlei He, Valery Vigil, Santiago Ojeda-Ramirez, Xuechen Liu, Julian Levine, Kelsyann Cervera, and Mark Warschauer. 2023. "Rosita Reads With My Family": Developing A Bilingual Conversational Agent to Support Parent-Child Shared Reading. In Proceedings of the 22nd Annual ACM Interaction Design and Children Conference. 160–172.
- [55] Ying Xu, Dakuo Wang, Mo Yu, Daniel Ritchie, Bingsheng Yao, Tongshuang Wu, Zheng Zhang, Toby Jia-Jun Li, Nora Bradford, Branda Sun, et al. 2022. Fantastic Questions and Where to Find Them: FairytaleQA–An Authentic Dataset for Narrative Comprehension. arXiv preprint arXiv:2203.13947 (2022).
- [56] Bingsheng Yao, Dakuo Wang, Tongshuang Wu, Zheng Zhang, Toby Jia-Jun Li, Mo Yu, and Ying Xu. 2021. It is Al's Turn to Ask Humans a Question: Question-Answer Pair Generation for Children's Story Books. arXiv preprint arXiv:2109.03423 (2021).
- [57] Andrea A Zevenbergen and Grover J Whitehurst. 2003. Dialogic reading: A shared picture book reading intervention for preschoolers. On reading books to children: Parents and teachers 177 (2003), 200.
- 58] Zheng Zhang, Ying Xu, Yanhao Wang, Bingsheng Yao, Daniel Ritchie, Tong-shuang Wu, Mo Yu, Dakuo Wang, and Toby Jia-Jun Li. 2022. StoryBuddy: A Human-AI Collaborative Chatbot for Parent-Child Interactive Storytelling with Flexible Parental Involvement. In CHI Conference on Human Factors in Computing Systems. 1–21.

A RUBRICS

A.1 Suitability Rubric

Type	Level	Authenticity	Grammatical Structure	Relevance to Book	Other Suitability Criteria
Completion	1	Inauthentic	 Completion phrase is at most one sentence long Blank should be at the end of the phrase 	Completion phrase is on current page	Deals with rhyming or repeated phrases
Recall	2	Inauthentic	 The question is not a composite of multiple questions Starts with an interrogative adverb/pronoun 	• Asks child to summarize thematically important events	 Asks child to summarize elements of plot or describe sequences of events Answer cannot be determined solely from the current page
Open- Ended	3	Authentic	Starts with an inter- rogative adverb/pro- noun	 Questions should relate to story themes, soliciting speculation about or foreshadowing upcoming story events Relates to the current page of the story 	 Solicits ideas or opinions about story elements or asks child to speculate about something related to the story (e.g., plot, characters, setting) Does not directly ask about child's personal experiences, but child may need to draw on personal experiences to answer Children should not easily be able to opt-out of answering the question Question discourages one word answers
Wh-	1	Inauthentic	 Start with an interrogative pronoun Not a composite of multiple questions 	 Details should pertain to objects or characters that are thematically important to the story plot Answer is in the text or pictures on the current page 	• Focuses on story details
Distancing	3	Authentic	• Start with an inter- rogative adverb/pro- noun or a verb	• Relates to the current page of the story	 Explicitly asks the child about their experiences Relates to the current page Cannot be answered in one word

Table 3: Question suitability rubric used to determine if a question of a given type generated from the system is suitable for presentation. Level is further described in Appendix ??.

A.2 Question Levels

Question Level	Description	
Level 1	Information recall questions focused on what can immediately be seen (or read) in the text. Questions ask studen	
	define, describe, list, or name attributes or utility of objects or characters in the text.	
Level 2	Open ended questions used to solicit the child's feedback. Questions involve information processing, asking students to analyze, compare, contrast, group, infer, sequence, or synthesize information gathered from the text.	
Level 3	Questions are related to the story plot, but may also relate to the child's personal experiences or remote events. Questions ask the child to apply, evaluate, hypothesize, imagine, judge, predict, or speculate about the story and their own experiences.	

Table 4: Levels for rubric suitability as adapted from Costa & Lowery [10] and Flynn [18].

B PROMPTS

This appendix section gives example prompts for generating and suitability checking an open-ended question. Words in all caps and square brackets were included verbatim as prompt variables. Words in parentheses were replaced with the relevant piece of text.

B.1 Generation Prompt

The initial prompt is used to generate a candidate question.

Act as a early childhood reading instructor, producing 'dialogic reading' prompts that encourage conversation and engagement with the text.

Generate an 'openEnded' prompt, that encourages the child to express their own ideas and opinions about the story.

This prompt should allow for creativity and imagination. Avoid questions that can be answered with a simple yes or no.

Remember that the language you use to create [PROMPT] and the themes you pull from the text must be age appropriate for 4-6 year olds.

Make sure [PROMPT] is to the point, and is not verbose.

Read the following text and use it to better understand the characters and events of the main text block. DO NOT use any of the text for prompting.

```
(previous page text)
"With that context, generate a prompt of type 'openEnded' for this main text:
(current page text)
Format your response in JSON using exactly the template below:
{
    "prompt": PROMPT
}
```

B.2 Suitability Check Prompt

Once a candidate prompt is generated, suitability is determined through a series of prompts that capture items in the suitability rubric (see Appendix A). If a prompt is deemed unsuitable, a feedback line is included within the generation prompt text for the next query that includes the unsuitable prompt and the explanation. Here is an example suitability prompt that determines authenticity.

```
Act as a early childhood reading instructor. You will be judging if [PROMPT] is [AUTHENTIC], given the [CURRENT_PAGE] and [PREVIOUS_PAGES] of the story book as context.

Remember that [AUTHENTIC] is defined as follows:

True if [PROMPT] does NOT have a prescribed answer on [CURRENT PAGE] or [PREVIOUS PAGES].

False if [PROMPT] has a prescribed answer, which can be determined from [CURRENT PAGE] or [PREVIOUS PAGES].

With all of this in mind, please help define the [AUTHENTICITY] of [PROMPT]:

Format responses in JSON using exactly the template below:

{
    "Authentic": {[AUTHENTIC]'s value, one of True or False},
    "Explanation": {Explanation of why [AUTHENTIC] was chosen},
}

[PREVIOUS PAGES]: (text of previous 5 pages)
[CURRENT PAGE]: (current page text)
[PROMPT]: (prompt)
```