DAPIE: Interactive Step-by-Step Explanatory Dialogues to Answer Children's Why and How Questions

Yoonjoo Lee
School of Computing, KAIST
Daejeon, Republic of Korea
yoonjoo.lee@kaist.ac.kr

Tae Soo Kim
School of Computing, KAIST
Daejeon, Republic of Korea
taesoo.kim@kaist.ac.kr

Sungdong Kim
NAVER AI Lab
Seongnam, Republic of Korea
sungdong.kim@navercorp.com

Yohan Yun
School of Computing, KAIST
Daejeon, Republic of Korea
yohanme@kaist.ac.kr

Juho Kim
School of Computing, KAIST
Daejeon, Republic of Korea
juhokim@kaist.ac.kr

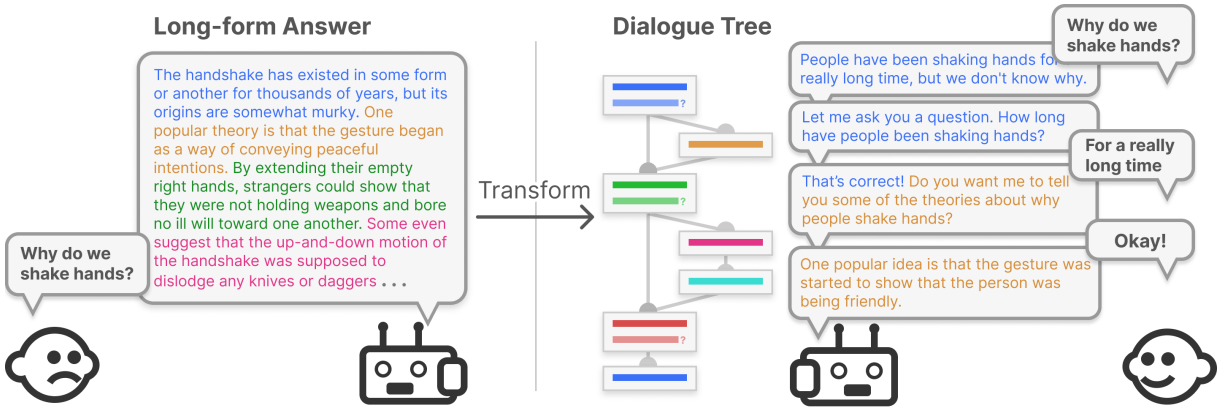

Figure 1: While existing conversational agents (CAs) are able to answer children's diverse "why" and "how" questions by retrieving explanations from the internet, these explanations are lengthy and require the child to perform complex reasoning. In this work, we present DAPIE, a novel system that transforms existing long-form answers into interactive dialogues. Through an AI-based pipeline, the system generates dialogue trees that present explanations step-by-step while prompting the child to engage with them and check their understanding.

ABSTRACT

Children acquire an understanding of the world by asking "why" and "how" questions. Conversational agents (CAs) like smart speakers or voice assistants can be promising respondents to children's questions as they are more readily available than parents or teachers. However, CAs' answers to "why" and "how" questions are not designed for children, as they can be difficult to understand and provide little interactivity to engage the child. In this work, we propose design guidelines for creating interactive dialogues that promote children's engagement and help them understand explanations. Applying these guidelines, we propose DAPIE, a system that answers children's questions through interactive dialogue by employing an AI-based pipeline that automatically transforms existing long-form answers from online sources into such dialogues. A user study (N=16) showed that, with DAPIE, children performed better in an immediate understanding assessment while also reporting higher enjoyment than when explanations were presented sentence-by-sentence.

CCS CONCEPTS

• **Human-centered computing** → **Interactive systems and tools**; **Empirical studies in HCI**; **Natural language interfaces**.

KEYWORDS

Children, Conversational Agents, Dialogue, Question Answering, Natural Language

ACM Reference Format:
Yoonjoo Lee, Tae Soo Kim, Sungdong Kim, Yohan Yun, and Juho Kim. 2023. DAPIE: Interactive Step-by-Step Explanatory Dialogues to Answer Children's Why and How Questions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), April 23–28, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 22 pages. https://doi.org/10.1145/3544548.3581369

CHI '23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9421-5/23/04…$15.00
https://doi.org/10.1145/3544548.3581369

1 INTRODUCTION

Asking "why" and "how" questions is an important theory-building mechanism for children, as answers to those questions can help the child gain a causal understanding of the world [14]. Children as young as three years old can formulate sophisticated questions to resolve gaps in knowledge or perceived inconsistencies [13] on a variety of phenomena: natural, biological, physical, cultural, and social [14]. As children can frequently perceive such gaps or inconsistencies, they require responders who could provide quality answers to their questions in time [79]. Thus, the availability of responders is important for young children. Moreover, knowing that a responder is available [79] can encourage children to ask more questions and improve learning. In terms of availability, conversational agents (CAs), like Alexa or Google Assistant, can offer great value in acting as responders to children's questions. Through CAs, children gain access to a huge amount of information available on the internet without being fluent in reading or writing [28]. Moreover, unlike adult responders, CAs are always available to answer children's questions [78] and have become increasingly common in home settings [64].

However, beyond challenges caused by CA's inaccurate speech-to-text translation, prior work has demonstrated CA's answers are frequently inadequate for children [28, 58]. For example, when asked *"Why do polar bears have white fur?"*, Google Assistant responds with the following: *"Polar bears have white fur so that they can camouflage into their environment. Their coat is so well camouflaged in Arctic environments that it can sometimes pass as a snow drift. Interestingly, the polar bear's coat has no white pigment; in fact, a polar bear's skin is black and its hairs are hollow."*

These types of long responses are challenging for children to understand because they often require the child to possess the prior knowledge needed and to interpret possibly complex reasoning chains [58]. Furthermore, existing CAs provide long responses at once without prompting, which leads to them not being able to identify what a child did not understand or to engage them in a conversation. These challenges stem from the fact that most CAs are powered by computational pipelines designed for adults. For example, information retrieval models, which are commonly used models in pipelines to make CAs, identify relevant passages from the internet and bring these raw long answers to be presented as a response [36, 45, 85]. On the other hand, generative long-form question answering (LFQA) models have also been designed to generate, long-form answers to given questions [25, 52, 53, 63]. However, these models typically provide the answers only focusing on answers' factuality or accuracy with no careful considerations for children. To answer children's "why" and "how" questions by leveraging CA's availability and their ability to connect children to vast amounts of information, we aim to transform long answers into interactive conversations that can enhance children's understanding and engagement.

To identify effective techniques for answering children's questions and for presenting explanations in a form that is comprehensible for children, we propose design guidelines for step-by-step interactive dialogues that scaffold children's understanding on their own "why" and "how" questions. The guidelines present common conversational turns and strategies that can be employed to engage children in conversations, diagnose their understanding, and provide adequate interventions to help them overcome difficulties. To construct these guidelines, we first conducted an iterative inductive analysis on challenges and lessons from prior literature in child development, and conducted consultancy sessions with four child education experts to refine and validate our guidelines.

Applying these guidelines, we propose DAPIE (**D**ialogic **A**nswering via **P**iecemeal **I**nteractive **E**xplanations), a novel system that answers children's questions through step-by-step interactive dialogues that adapt explanations contained in existing long answers. To power this system, we propose an AI-based pipeline that automatically transforms existing long answers into dialogue trees that CAs can perform. The pipeline consists of two main steps that adhere to our design guidelines: (1) decompose and structure the long-form answer into chains of sub-explanations (e.g., tree structure) that provide the information step-by-step; and (2) augment the tree with additional dialogue turns to diagnose the children's understanding and provide adaptive interventions. Through a technical evaluation of the modules in our pipeline, we found that our pipeline outperformed baseline techniques according to measures that correspond to our design guidelines.

To understand whether our interactive explanations improve children's understanding of the information and engagement, we conducted a within-subjects study with 16 participants aged five through seven. They experienced DAPIE and the baseline which provide a sentence at a time using the same source as DAPIE. Our study revealed that children when using our system got a significantly higher score in an immediate assessment and showed a significantly higher level of engagement than when using the baseline system. Children reported that DAPIE was a better teacher, and provided more comprehensible and enjoyable dialogue. We believe that DAPIE is a first step at extending CAs and smart speakers to interactively and adaptively answer children's questions to foster children's curiosity and enhance their understanding about the world.

The contributions of this work are as follows:

- Design guidelines for supporting explanations that answer children's "why" and "how" questions through step-by-step and interactive dialogues.
- DAPIE, a system that serves interactive dialogues through an AI-based pipeline that transforms existing long-form answers into dialogues that follow our guidelines.
- Findings from a user study demonstrating how these generated interactive dialogues can help children's understanding and promote engagement.

2 RELATED WORK

We review research on (1) children's question answering behavior, (2) child-CA conversations, and (3) existing long-form question answering techniques.

2.1 Children's Question Answering Behavior

Children ask many questions to acquire information and develop knowledge about the world [13, 32]. Beyond fact-based questions, children also ask"why" and "how" questions that require explanations about causal relationships or mechanisms [13, 44]. According

to developmental psychologists, questions and answers (QAs) help children construct complex causal knowledge [13], so QA-based conversations with more knowledgeable others (e.g., parents, instructors, or CAs) are important for children's development [82]. Studies observed that children prefer answers with satisfactory amounts of information [14, 27] and will ask follow-up questions if not satisfied [26, 62], indicating that question asking is not simply to seek attention. Due to children's well-known need for information, substantial research (covered by our guidelines in Section 3) investigated how to effectively answer children's questions. However, these factors are rarely considered in the design of existing CAs. To address this gap, we qualitatively analyzed literature related to answering children's question and consulted experts in children development to propose design guidelines for creating dialogues that effectively answer children's "why" and "how" questions.

2.2 Child-CA Conversations

Recently, several studies investigated the experiences of children with conversational agents (CAs) in the form of smart speakers or voice assistants [22, 58, 74]. By observing interactions in natural settings, these studies demonstrated that children interact with CAs on diverse topics [74] and ask questions on various domains (e.g., science, culture, language) [58]. Further, studies showed that children view CAs as friendly, trustworthy, safe, and always available for them [22, 58, 89]—providing evidence on the potential of CAs as conversational partners for children. Despite their potential, research on CAs has revealed several challenges regarding child-CA interactions [38]. For example, CAs' speech recognition frequently misinterprets or fails to understand children's speech [38] which leads to breakdowns in conversation [74]. To resolve these breakdowns, Cheng et al. [9] observed that more capable adults can provide scaffolding strategies, and Xu et al. [89] designed conversational patterns to guide children's responses and prevent breakdowns. However, beyond such breakdowns, children can struggle to parse and understand CAs' responses as they can be long and complex [28, 58] and little research has investigated how to scaffold children's understanding of CAs' responses. In this work, we aim to maximize the benefits of CAs as conversational partners for children by introducing a novel approach that automatically creates interactive dialogues that answer children's questions on-the-fly, but with the adequate scaffolding that children necessitate.

2.3 Existing Question Answering Datasets and Applications

In natural language processing (NLP), QA tasks involve (1) extracting unambiguous short answers from text or (2) generating free-form answers for given questions and passages. Long-form Question Answering (LFQA) , a subset of the latter, aims to answer more open-ended questions that require explanations. To drive research in this area, researchers have constructed various datasets (e.g. SQuAD [66], Natural Questions [45], ELI5 [25]) and, based on these datasets, have investigated the structure of long-form answers [86] and methods to improve the faithfulness of answers [76]. On the other hand, conversational question answering (CQA) aims to generate multi-turn QAs where a questioner and an answerer converse with each other (e.g., CoQA [67], QuAC [12], QReCC [1]). Although CQA models offer more interactivity than conventional single-turn QA, communication is one-way and lacks considerations on how the answerer can help or engage with the questioner. In HCI, researchers have devised QA systems that consider users' understanding or engagement to help them learn programming [84], math [8], and factual knowledge [72]. Although these systems provide educational benefits, significant manual effort is needed to create diverse QA dialogues. To decrease effort and increase diversity, Promptiverse [46] introduces a human-AI approach to annotate knowledge graphs which are then automatically traversed to generate QAs. However, this work focused on general learners and lacks consideration for children, who require different support.

3 GUIDELINES FOR DESIGNING EXPLANATORY DIALOGUES

In this work, we propose a set of guidelines that outline how to construct explanatory dialogues to answer children's "why" and "how" questions. While research in child development, cognitive psychology, and education has investigated how children ask questions and understand explanations, little work has compiled and organized the findings into design lessons. To address this gap, we present guidelines that describe how to deliver explanations through dialogues that are catered to children's understanding and engagement.

3.1 Method

To identify relevant literature, we conducted a keyword-based search on Google Scholar and the ACM Digital Library using the terms: *"question-asking behavior of children"*, *"answering children's questions"*, and *"explanations for children"*. Through several cycles, we expanded our set of search terms by collecting keywords mentioned in sampled literature, and sampling more literature by combining the terms. Details in the Supplementary Materials.

Based on the collected papers, three of the authors conducted iterative coding through inductive analysis to organize the findings and lessons in the papers, and discover recommendations from the data. Any discrepancies in coding were negotiated until mutual agreement was achieved. Based on the analysis, we categorize these recommendations into design guidelines.

To verify and revise our guidelines, we then conducted a design consultancy with four experts in child education. The experts all had majored in child education (one M.S., three Ph.D.) and two also had more than five years of experience teaching children. During the consultancy, the experts were asked to evaluate our framework by revising dialogues that the authors made by applying the framework, and to apply the framework themselves by designing dialogues for given pairs of questions and long-form answers.

By qualitatively analyzing the experts' feedback, we found the following general guides which served to support and extend our guidelines. First, all experts mentioned that it is essential to *"decompose information into smaller steps"* when explaining verbally due to children's limited attention span and working memory. To decide on what information to omit when simplifying explanations, experts suggested considering importance (i.e., what the child needs

to know) and acceptance (i.e., what the child can understand). After providing a small chunk of information, three experts recommended asking the child if they understood as each child might understand differently and might need further explanations based on their prior knowledge. They suggested that an explainer can then provide *"further information if the child understands, or provide adjusted explanations if they do not"*. When checking children's understanding, two experts mentioned that true/false or multiple choice questions are used often in practice. More detailed feedback is reflected in our guidelines below.

3.2 Guidelines

Based on our analysis of literature and design consultancy with experts, we present guidelines for effectively constructing interactive dialogues that answer "why" and "how" questions from children. The first section in our guideline describes how answers can be decomposed into a step-by-step explanation to scaffold children's reasoning. Beyond reasoning, children can struggle to understand answers as they may lack prior knowledge and can have difficulties in staying engaged. Therefore, the second section describes how, during an explanation, the explainer can interact with the child to promote engagement and check their understanding to provide suitable interventions.

3.2.1 Constructing a Chain of Explanatory Units. To present explanations step-by-step, we suggest that explainers construct chains of explanatory units. Specifically, explainers should decompose the explanation into sub-units, identify relevant sub-units and their relationships, and present these units based on their identified relationships. By building relationships between concepts as they learn new concepts, explainers can help children achieve meaningful learning [2].

Decomposing Complex Explanations into Simpler Units When providing complex explanations, we suggest that explainers decompose the explanation into simpler units to help children understand complex concepts while avoiding significant cognitive load [23, 58]. By decomposing, explainers can lower complexity by unpacking the varying factors, entities, and relations contained in an explanation so that the child can process each one independently [31].

Identifying Relevant Units and Relations To aid and guide children's reasoning, explainers should identify and denote the factors in an explanation that are necessary to achieve an understanding. By highlighting relevant factors for a child, the child can identify them as well, focus on them, and reason about the relationships between them [41].

Connecting the Units With the relevant units and relationships identified, explanation should be presented in a cumulative and causal manner. Specifically, an explanation should indicate how one unit in the explanation leads to the subsequent unit, since this guides children to carry out sophisticated reasoning on the relationships between the presented information [73]. Moreover, prior work demonstrated that children show greater curiosity and learning when explanations elaborate on causal connections [37].

3.2.2 Designing Interactive and Understandable Dialogues. Beyond constructing chains of explanations, we recommend that explainers expand on these chains to carry out interactive dialogues that support understanding and promote engagement. For this purpose, we suggest a systematic structure for guided dialogues between an explainer and a child, and recommend multiple strategies for the explainer to engage with the child and to adapt explanations to the child's level of understanding. Specifically, we adapted the three key components of effective dialogues with children (i.e., questions, feedback, and scaffolding), which were designed for contexts where adults ask questions to children [75], to our context where a child asks questions to an explainer. Several studies have applied these components to create CAs for preschool-aged children [87, 89].

We propose a dialogue structure where each turn of dialogue is composed of three sub-turns: **Feedback**, **Explanation**, and **Question**. The explainer first provides feedback by building upon the child's utterance. Then, the explainer provides an explanatory unit relevant to the child's answers. The explainer then ends the turn with a question that invites the child to engage in the dialogue. For example, when a child cannot understand an explanation about how bees make honey, the explainer can respond with *"That's alright. Nectar is a sugary liquid that flowers produce. Do you get that?"* where the sentences represent feedback, explanation, and question, respectively. Below, we describe each sub-turn, the various roles that each sub-turns can perform to support children, and strategies through which these roles can be performed.

Feedback involves the explainer verbally commenting on the child's response to the explainer's prompt. For example, the explainer corrects the child's answer (contingency) or praises their attempt to answer through contingency feedback (encouragement). Direct and specific feedback helps children clarify their confusion and increase their engagement [4, 48, 80].

Explanation delivers information after the explainer has provided feedback on the child's response. Explanations can perform two roles: *extension* and *adjustment*.

For *extension*, the explainer elaborates on the topic by providing a new explanatory unit like further details or new pieces of information to deepen the dialogue [7, 13, 15, 40, 62]. *Adjustment* is adapting the explanation to the child's developmental levels, and the cognitive and linguistic demands the child faces during the dialogue with the aim of facilitating child's understanding. For adjustment, explainer gives an explanatory unit that was already provided once, but adapts its content or language.

We propose several strategies for both extension and adjustment sub-turns (Table 1). Dialogues can be designed with these strategies to help children understand new information provided in an extension, or, in case the child was unable to understand, the strategies can be applied to create additional alternative explanations to use as adjustments. Local strategies (i.e., Simplifying, Providing examples, Summarizing, Providing analogies, Providing personifications, Representing or demonstrating) are applied individually on explanation sub-turns according to the intervention that the child requires. Global strategies (i.e., Textual simplification, Explicitly mentioning coherence, Global adjustment, Highlighting relevancy) apply to all explanation sub-turns in a dialogue, and act as general support that can benefit all children. Table 1 presents local and global strategies and guidelines for each strategy.

	Strategies	Guidelines
Local	**Simplifying**	**G1. Use language that matches child's level of comprehension [24].** **G1-1. Change scientific, technical, or formal terminology into simpler language.**
	Providing examples	**G2. Provide various examples that represent new or unfamiliar concepts [71].** **G3. Clearly explain the relationship between the original concept and examples to help generalization [10, 42].** **G4. Consider child's prior knowledge when choosing examples.** **G5. Provide examples with high similarity to the original concept [29].**
	Summarizing	G6. Clearly indicate the core principles of a concept. G7. First provide an immediate and summarized answer to a question before diving into the details [50].
	Providing analogies	G8. Consider the child's unique interest and experiences when choosing a comparison target [30, 56, 80]. G9. Explicitly guide the child to recognize the similarities between the source and target [80, 81]. G10. Choose a target that presents similar entities and relations to those in the source [30].
	Providing personifications	G11. Explain unfamiliar or complex entities and concepts by personifying them or granting them human attributes [33]. G12. Personification is more effective when the entity or concept shares similarities with humans [30, 34].
	Representing or demonstrating	G13. Use representations and demonstrations to illustrate or visualize concepts.
Global	**Textual simplification**	**G14. Apply simplification to all information [58] by considering the average child (local strategy G1-1 simplifies for a specific child).** **G15. (Lexical) Replace difficult terms with simpler ones [39].** **G16. (Syntactical) Simplify the syntactic construction of sentences [19].** **G17. (Length) Use intermediate-length sentence [27].**
	Explicitly mentioning coherence	**G21. Explicitly mention the coherency between turns [42] or use explicit linking language [23, 49] (e.g., "before that", "then").** **G22. For cause-and-effect relationships, explicitly mention how an event leads to the response (e.g., "When all the pieces touch, energy can travel from the battery to the light") [43].**
	Global adjustment	G18. Adapt explanations according to a child's level of prior knowledge [42, 50]. G19. Adapt explanations according to a child's personal experiences. G20. Propagate adjustments made in prior turns through the whole dialogue.
	Highlighting relevancy	G23. Redirect children's attention to the crucial content to help them engage in deeper processing [23, 42].

Table 1: Guidelines for local and global strategies that can be used in adjustment and extension sub-turns. Strategies in bold were used in devising our system, DAPIE, presented in Section 4.

Question After providing information to a child, the explainer can ask the question the child to invite them to participate and engage in the dialogue by answering the question. Our guidelines suggest three roles for questions: *guiding*, *diagnosing understanding*, and *eliciting prior knowledge*. Table 2 presents description of the roles and strategies for asking questions.

4 DAPIE: CONVERSATIONAL AGENT TO SUPPORT INTERACTIVE DIALOGUES

Based on our guidelines, we propose DAPIE, a novel system that automatically transforms existing long-form answers into interactive explanatory dialogues for children. Particularly, our computational pipeline applies our guidelines through state-of-the-art NLP techniques (e.g., large language model (LLM)-based few-shot generation) to structure and augment long-form answers into comprehensible and interactive dialogue trees. Though this, DAPIE can leverage and adapt existing long-form answers on the internet, which are inaccessible to children, to answer children's various

Role	Description	
Guiding	Explainers can use *guiding questions* to scaffold children's understanding by helping them narrow down their focus [17] or by leading them to consider other information [26, 61, 88].	**Let children know what information is missing or what information they can ask about [17, 26, 47, 61, 88].**
		Guide children to understand detailed information from a prior turn [17].
Diagnosis	Explainers should *diagnose* children's understanding to provide interventions if they failed to understand [42]. Diagnosis is more effective and reliable if the child is prompted to apply information from the explanation.	**Providing all but one piece of information and asking children to fill-in-the-blank.**
		Ask children to give predictions.
		Ask children to self-explain.
		Ask children to decide whether a fact generalizes across items of varying similarity.
Eliciting	Explainers should *check or ask about children's prior knowledge* to adjust explanations with knowledge that is more familiar to the child. [7].	**Ask children about their knowledge.**
		Ask experience-based questions.

Table 2: Guidelines for the different roles that questions can take and specific strategies that explainers can apply to enable these roles. Strategies in bold were used in devising our system, DAPIE, presented in Section 4.

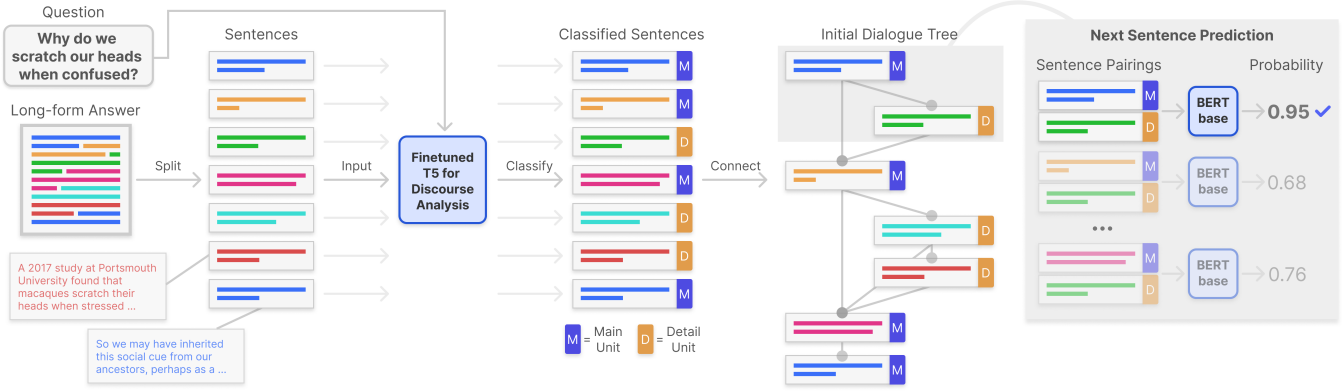

Figure 2: Overview of the first step (Section 4.1) in our computational pipeline for DAPIE. Starting from an existing long-form answer, it constructs a chain of explanatory units (i.e., an initial dialogue tree). First, it splits a long-form answer into sentences and judges their relevance to a given question, whether it is a main or detail unit, with a T5-based analyzer [86]. Then, the sentences are connected to each other according to their pair-wise relevance score (i.e., next sentence prediction by BERT model [20]) to form a tree structure.

'why" and "how" questions. Our pipeline follows the two main processes from our guidelines: (1) *constructing* chains of explanatory units from the long-form answer; and (2) *designing* an interactive and understandable dialogue by augmenting the chains.

For examples of the final outputs generated by our pipeline, see Figure 5 for a dialogue tree, and Appendix A for a thread from a dialogue tree.

4.1 Constructing Chains of Explanatory Units

In the first phase, our pipeline constructs chains of explanatory units by structuring the explanation in the long-form answer (Fig. 2). This phase involves (1) decomposing answers into units, (2) identifying relevant units, and (3) connecting the units into step-by-step explanatory chains.

4.1.1 Decompose. Our pipeline decomposes a long-form answer by splitting it into its constituent sentences: each sentence is an explanatory unit. We assume that each sentence represents one explanatory unit as writers are frequently encouraged to encapsulate one point or thought per sentence. By qualitatively analyzing a sample of 10 QA pairs from the "BBC Science Focus Magazine", two of the authors verified that this assumption generally held for professionally written explanations.

4.1.2 Identify. Our guidelines suggest that explainers should identify factors that are relevant to guide children's focus. As existing answers can include auxiliary information that is less relevant to a question [86], we employ the T5-based discourse analyzed by Xu et al. [86] to distinguish between relevant and auxiliary information. By employing this model, our pipeline first classifies the sentences in a long-form answer into their functional roles: "summary", "answer", "example", or "auxiliary information". Then, it assigns those classified as "summary" or "answer" as *main* units (i.e., directly

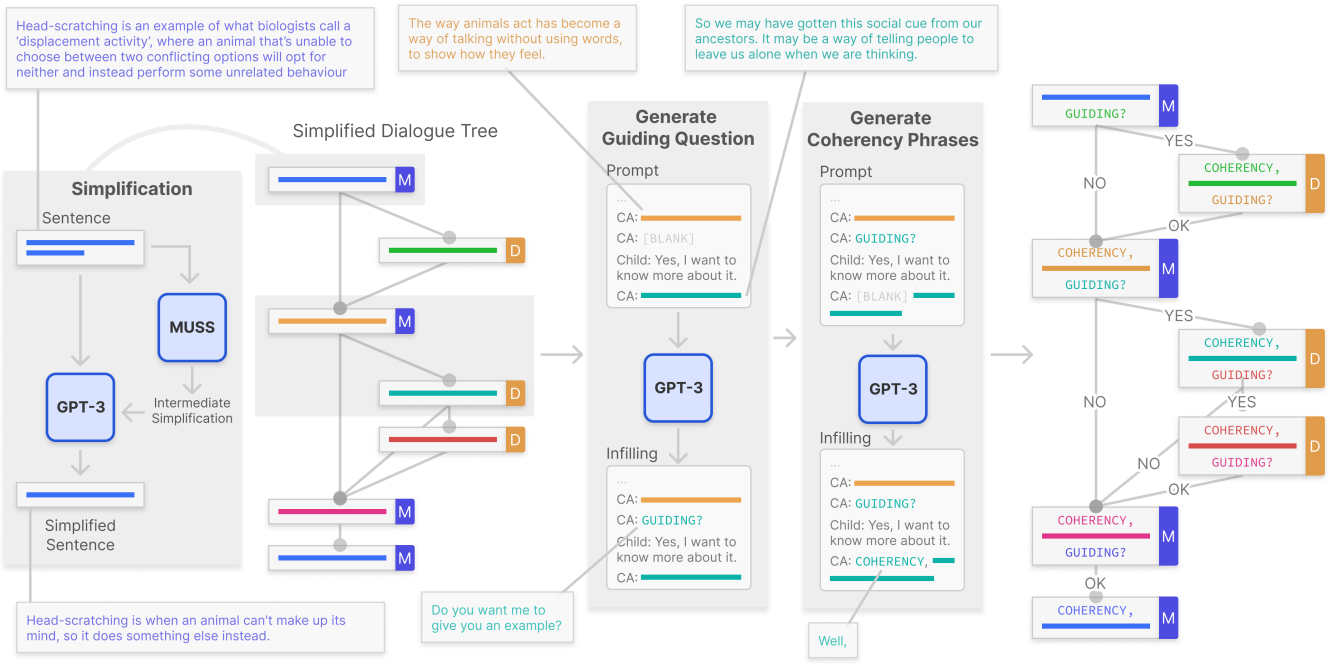

Figure 3: In the second step, the pipeline simplifies all the explanations in the dialogue tree, and then integrates guiding prompts and coherency phrases to bridge children's understanding across consecutive explanations.

relevant), and those classified to the other two roles as *detail* units (i.e., less relevant).

4.1.3 Connect. With the units classified, the pipeline connects the *main* units in the order that they appear in the answer. This chain serves as the main thread of the dialogue tree. As *detail* units are less relevant, we incorporate these as optional extensions for when a child desires more information. As our guideline suggests that explanations should be cumulative, we connect each *detail* unit to the *main* unit that it is most likely to build on. For this, the pipeline uses BERT base model [20] to performs next sentence prediction (NSP) between all *detail* and *main* units. We used an additional model to connect units, instead of relying on their ordering in the answer, as our qualitative analysis of 10 QA samples showed that detail units were not consistently adjacent to their relevant main units. For details, see *"Connecting Explanatory Units"* in the Supplementary Materials.

With these steps, the pipeline produces an initial dialogue tree where each turn consists of an extension explanation sub-turn. Following this tree provides all of the most relevant information to the child's question, with optional branches that provide additional details.

4.2 Designing the Interactive and Understandable Dialogue

In the second phase (Fig. 3 and 4), the pipeline augments the initial dialogue tree by incorporating feedback and questions to interact with a child, and adjustment explanations to scaffold their understanding. We devised the augmentations based on our guidelines. For questions (Table 2), the pipeline incorporates (1) guiding questions to lead children to further information (i.e., other main or detail units), (2) diagnosis questions to check children's understanding of main units through fill-in-the-blank questions, and (3) questions that elicit prior knowledge to identify what causes difficulties in understanding. For adjustment explanations, we adopted simplification and provided examples as local strategies (Table 1). Analogies and personifications were not used as they are similar to exemplification, but only apply in narrower situations. Additionally, the pipeline applies global strategies (Table 1) to simplify and mention coherency in all units.

Our pipeline's goal is to maintain the core information in an explanation, but to incorporate additional turns that are coherent and follow our guidelines. With this goal, we employ an LLM as these models can produce text that coheres with the given context and follows given examples (i.e., few-shot learning). Specifically, we use GPT-3 [5] to extend the *dialog inpainting* technique [16] that generates simulated dialogues where an LLM fills in questions from a "reader" and an "author" answers with sentences from a document. We extend this technique to simulate dialogues where a CA interacts with a child through feedback-explanation-question sub-turns. To employ our extended technique, *turn inpainting*, we designed dialogue templates by imagining dialogues where a CA follows our guidelines to provide questions and adjustment explanations to a child. With the same 10 QA samples analyzed in Section 4.1.1, we performed prompt engineering to iterate on the templates until they produced satisfactory results. As LLMs can generate harmful words (e.g., swear words, vulgar words), our pipeline checks each

(A) Generate Guiding Question

CA: [Explanation sub-turn in t_n]
CA: [BLANK]
Child: Yes, I want to know more about it.
CA: [Explanation sub-turn in t_n + 1]

(B) Generate Coherency Phrase

[Turns t_1 to t_{n-1}]
CA: [Explanation sub-turn in t_n]
CA: [Guiding question from t_n to t_n + 1]
Child: Yes, I want to know more about it.
CA: [BLANK] [Explanation sub-turn in t_n + 1]

(C) Generate Diagnosis Question

CA: [Explanation sub-turn in t_n]
CA: Let me ask you a question. [BLANK]
Child: The answer is [Answer for t_n].

(D) Generate Elicit Question

CA: [Explanation sub-turn in t_n]
CA: [BLANK]
Child: Hmm. I don't know.
CA: It's okay. [Definition for term in t_n]

(E) Generate Term-based Example

CA: [Definition for term in t_n] Did you get it?
Child: No, I couldn't understand it.
CA: Don't worry. Let me give you examples. As you know well, [BLANK]. They are all [Term in t_n].

(F) Generate Clause-based Example

CA: [Explanation sub-turn in t_n] Did you get it?
CA: No, I couldn't understand it.
CA: Don't worry. Let me give you an example. [Clause in t_n] is like [BLANK]

Table 3: Prompt templates used as input for GPT-3 to produce the functionalities in our pipeline. The few-shot examples that are prepended to each template are available in the Supplementary Materials.

generation output for such words and re-generates if found—our pipeline never had to re-generate during this work.

4.2.1 Simplify. Due to children's developing language skills, it can be beneficial to simplify all of the explanation sub-turns in the dialogue tree (G14 in Table 1). For simplification, our pipeline first uses MUSS [59], a sentence simplification model with controllable attributes for the degree of lexical, syntactic, and length simplification (G15, G16, G17 in Table 1). While adequate for syntactic and length simplification, we observed that this model's lack of knowledge led to limited or incorrect lexical simplifications. Thus, our pipeline uses GPT-3, which contains vast language knowledge, to simplify sentences one more time by combining the original sentence, the MUSS simplification, and few-shot examples into an input prompt (T1 in Supplementary Materials).

4.2.2 Integrating Guiding Questions. According to our guidelines, guiding questions can help children to engage further with a topic by previewing information to come ("Guiding" in Table 2). To generate these questions, our pipeline uses *turn inpainting* by constructing a template (Table 3A) with two consecutive explanation turns, t_n and t_{n+1}. With this template and few-shot examples as input, the model fills in a guiding question, in place of `[BLANK]`, that asks the child if they want to learn about the second turn. In the dialogue tree, the CA asks the question and moves to the next turn when the child accepts. If the next turn can be a main or detail unit, the CA asks the guiding question to the detail unit. With the guiding questions, the pipeline also uses *turn inpainting* to generate phrases that explicitly describe the coherency between turns and how an event leads to the response (global strategy G21, G22 in Table 1). For details, see *"Creating Coherency Phrases"* in Supplementary Materials.

4.2.3 Designing Diagnosis Questions. As our guideline suggests, it is crucial that an explainer checks whether a child understood an explanation and, if they did not, to provide suitable adjustments ("Diagnosis" in Table 2). We chose to generate fill-in-the-blank questions as the other strategies require free-form responses that are difficult to verify with existing techniques. To generate these questions, the pipeline identifies two potential difficulties in the explanations to use as the "blank": unfamiliar terms and complex cause-effect relationships. According to surveyed literature, these two were common challenges in children's understanding [24, 30], and are core factors for answering "why" and "how" questions (i.e., prior knowledge and mechanistic reasoning). With GPT-3, our pipeline identifies these difficulties, and then extracts a word/phrase from the difficulty to use as the answer for the diagnosis question For details, see *"Identifying Difficulties and Correct Answers"* in Supplementary Materials.

With the difficulties and answers extracted, the pipeline generates a diagnosis question using GPT-3 with a template (Table 3C) where the CA asks a `[BLANK]` question and the child gives the answer extracted by the pipeline (prompt T6 in Supplementary Materials). Finally, to narrow down children's possible answers, we use GPT-3 to create alternative but wrong answers (prompt T7 in Supplementary Materials).

In the dialogue tree, diagnosis questions are asked after the main turns. For correct answers, the dialogue provides contingency feedback ("Feedback" in Sec. 3.2.2) like "That's correct!"and asks the guiding question. For incorrect answers, the dialogue provides feedback (i.e., "Hmm, I don't think so") and moves to adjustment turns according to the difficulty (i.e., term or cause-effect). We provide adjustment turns after the main turns as our consulted experts suggested that children should first be provided with information relevant to their question and then, if needed, provided with support. They explained that this retains the engagement of children who can understand, while guaranteeing fallback support for those who cannot.

4.2.4 Adjustment Turns for Difficult Terms. To verify whether the child failed to understand because they did not know the difficult term, the pipeline generates a question to elicit prior knowledge ("Eliciting prior knowledge" in Table 2). The pipeline constructs a template (Table 3D) where the CA asks a `[BLANK]` question, the

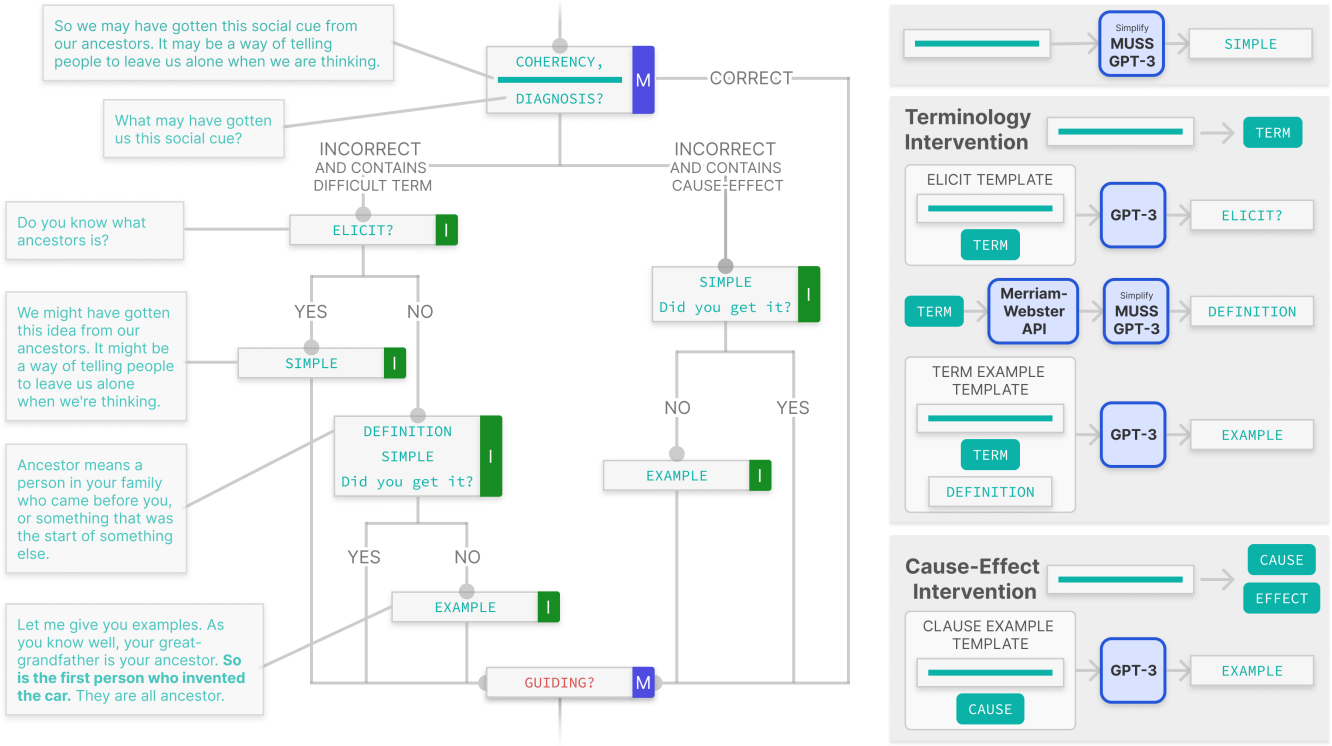

Figure 4: In the final step, the pipeline constructs adjustment sub-trees that scaffold children's understanding of (1) difficult terms, or (2) cause-effect relationships. The types and number of adjustments provided depend on children's answers to intermediate questions. After receiving the corresponding adjustments, the dialogue proceeds to the guiding questions for the next turn in the dialogue tree.

child responds that they do not know, and so the CA provides a definition for the difficult term. To visualize how the adjustment turns are connected in the dialogue tree, refer to Figure 4.

Simplification Turn. If the child answers that they know the term, the pipeline offers a more simplified explanation (G1 in Table 1). For this, the explanation in the main turn is simplified again using the same simplification method as in Section 4.2.1.

Definition Turn. If the child does not know the term, the pipeline provides an extension explanation for the term. As LLMs can hallucinate [60] (i.e., generate false information), we retrieve definitions from a verified source, Merriam-Webster API[1], instead of generating them. After retrieving, our pipeline simplifies the definitions using our simplification method. The definition turn provides this definition and a simple diagnosis question asking the child if they understood or not. A simpler diagnosis is used to not exhaust children with frequent quizzing.

Term Exemplification Turn. If the child could not understand the definition, the CA should provide an additional adjustment to help them understand. Based on our guidelines, the pipeline generates various examples to illustrate the unfamiliar term by creating a template (Table 3E) where the CA provides `[BLANK]` examples of the unfamiliar term and explicitly indicates that they are all examples of the term (G2 and G3 in Table 1). Also, the prompt includes few-shot examples illustrating effective exemplification—i.e., familiar to children and have high similarity to the original concept (G4 and G5 in Table 1) We identified that, beyond questions and coherency phrases, *turn inpainting* could also produce context-relevant adjustment explanations (e.g., examples) based on a simulated dialogue.

4.2.5 Adjustment Turns for Cause-Effect Relationships. Causal reasoning can be challenging for children due to limited prior knowledge on various cause-effect relationships [30]. To help children understand cause-effect relationships, the pipeline constructs two consecutive adjustment turns: a simplification turn, and a clause exemplification turn. The simplification turn is the same as that in the adjustment turns for difficult terms. To view how the adjustment turns are connected, refer to Figure 4.

Clause Exemplification Turn. When simplification is insufficient, the pipeline creates an example of another similar cause-effect relationship. We generated examples based on the causes as we observed that generating from the effects lead to broader and more unrelated examples (G5 in Table 1). The pipeline constructs a dialogue template (Table 3F) where the CA compares the cause in the explanation to another `[BLANK]` phenomena. While this dialogue template was designed to create examples, the pipeline occasionally creates analogies—possibly due to the presence of commonly used analogies in GPT-3's training data.

[1]https://dictionaryapi.com/

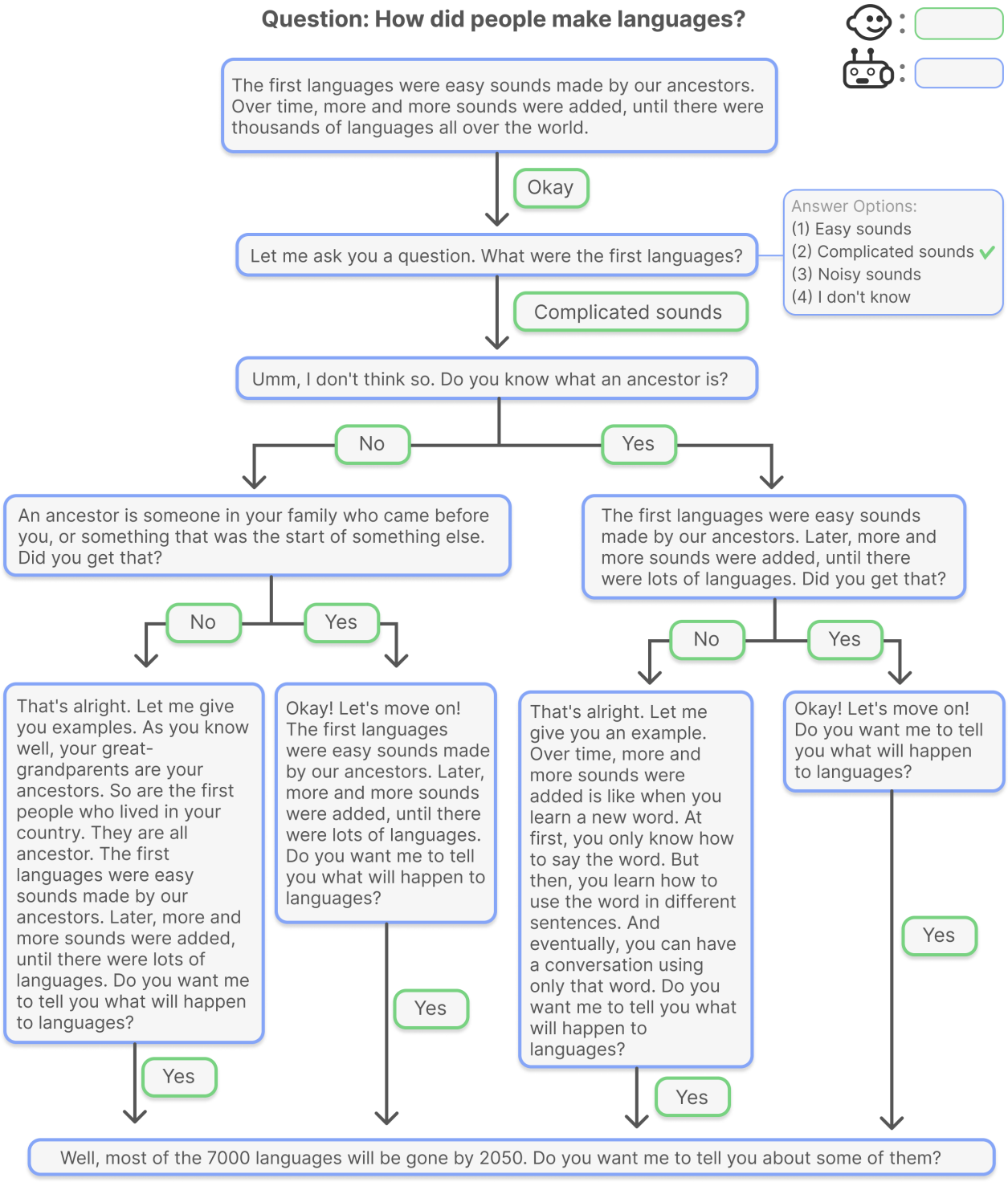

Figure 5: Subtree taken from the dialogue tree generated by our computational pipeline from the answer to the question *"How did people make languages?"* The dialogue tree shows how the CA can provide an explanation, diagnose the child's understanding, and the provide interventions (e.g., definition, example) to help them if they have difficulties understanding.

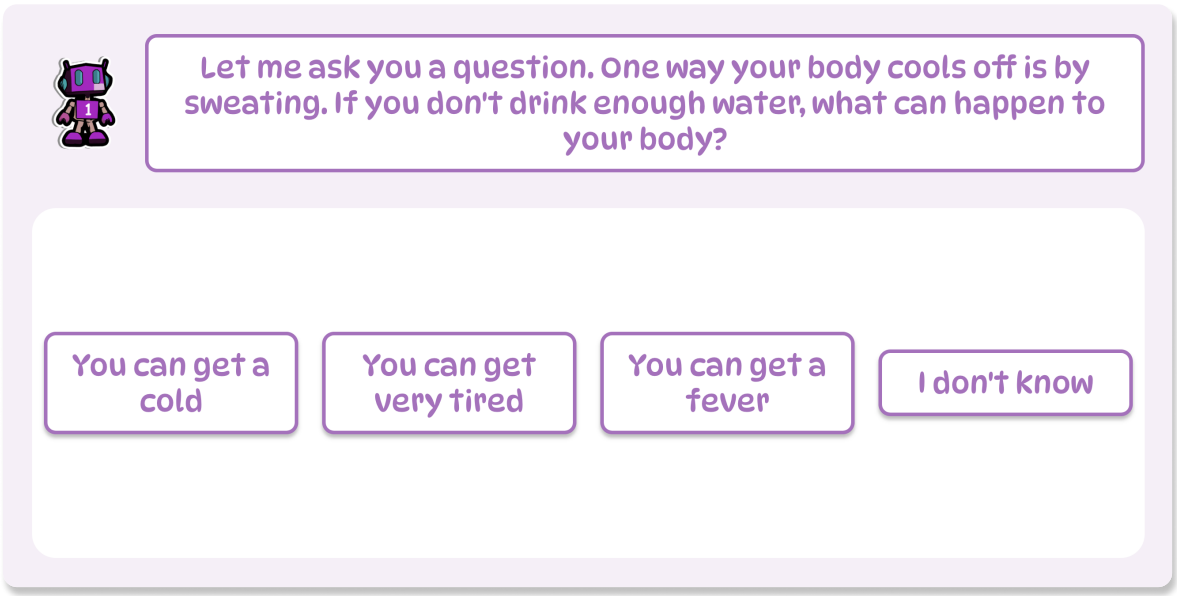

Figure 6: In the interface for DAPIE, the CA's explanation is shown on the panel above and spoken through text-to-speech. After speaking the explanation, the interface presents the child's response options in the panel below.

4.3 Interface

We developed DAPIE (Fig. 6), a CA that serves the interactive dialogues generated by our pipeline. DAPIE interacts with the child by first saying an utterance, and then reading options to which the child can respond. The CA, implemented as a web-based interface, was designed based on prior work and insights from the expert consultancy.

For each turn of a dialogue, DAPIE first says their utterance and also displays the text. The speaking rate was set lower than default to prevent overloading the child, and we show text as an additional modality to benefit children who are able to read. After the agent finishes its utterance, it reads out the options the child can respond with one-by-one while revealing them below the utterance. DAPIE pauses briefly after each time it speaks to let the child think and process what was just said.

After all the options have been read, the child can respond by clicking on one of the options. Although we could have designed DAPIE as a fully voice-based interface, automatic speech recognition is still limited and error-prone—specifically with children's speech [38]. Since our goal is to help children understand and engage with explanations, we designed our interface so that the child can accurately express their intent through clicks—protecting their understanding from being hindered by speech recognition errors. Furthermore, experts suggested limiting the child's response options, instead of allowing free-form responses, as children might already be cognitively burdened from understanding the explanations.

When the child clicks on an option, the text is selected and it is read to them again. This allows the children to re-listen to any option (or even the utterance) in case they may have forgotten or failed to hear what was first said. While an option is selected, the child can click on the option again to choose it and the dialog will then proceed to the next turn corresponding to that choice. This interaction repeats until the child reaches the end of the dialogue, after which they return to the main menu.

5 TECHNICAL EVALUATION

We evaluated DAPIE through a pipeline evaluation and a user study (Section 6). The pipeline evaluation was conducted to verify the performance of each step in the pipeline, and the user study was to evaluate whether DAPIE's interactive dialogues, as a whole, help children's understanding and engagement. To validate our pipeline, we conducted human evaluations on the five main sub-modules used to generate the dialogues: simplification, exemplification, guiding question, diagnosis question, and elicit question. We focused on evaluating the text generation modules for questioning and adjusting (Section 4.2), since the modules for chain construction (Section 4.1) use off-the-shelf models. We provide a comprehensive analysis of each sub-module's performance compared to their baselines, and an in-depth post-analysis according to characteristics of the input question-answer pairs (e.g., source type, domain, and question type) to understand whether our pipeline is generalizable.

5.1 Method

We compared the quality of generated sentences from the five sub-modules (i.e., simplification, exemplification, guiding question, diagnosis question, and elicit question) to those from corresponding baseline models. For the evaluation metrics, we saw that existing metrics for evaluating generated text depend on adult-centric

Sub-module	Measuring questions
Simplification	QS1. Which one uses more common and easier words while preserving the core information of the original text? (G15) QS2. Which one has a simpler sentence structure while preserving the core information in the original text? (G16) QS3. Which one excluded more unimportant information while preserving the core information of the original text? (G17)
Exemplification	QC1. Which example is more helpful for understanding of the context? (G2) QC2. Which example would be more familiar to a child? (G4) QC3. Which one follows the following rule better? (Rule: The context and the (example/analogy) have both similarities and differences.) (G5) QC4. Which one is more relevant to the context?
Guiding Question	QG1. Which one would be more understandable to a child? QG2. Which one is closer to what a teacher or tutor would ask? QG3. Which one is more grammatically correct? QG4. Which question is more proper in the [BLANK] to connect the previous and following context?
Diagnosis Question	QD1. Which one would be more understandable to a child? QD2. Which one is closer to what a teacher or tutor would ask? QD3. Which one is more grammatically correct? QD4. Which question checks the understanding of a child about the context more properly?
Elicit Question	QE1. Which one would be more understandable to a child? QE2. Which one is closer to what a teacher or tutor would ask? QE3. Which one is more grammatically correct? QE4. Which question checks the prior knowledge of a child more effectively? QE5. Which question is more relevant or adequate with respect to the given following part (answer) of the [BLANK]?

Table 4: Measuring questions used in our human evaluation of the pipeline sub-modules to measure the performance of the sub-modules compared to the baselines. The table also notes relevant guideline strategies (Table 1) that motivated certain question.

datasets so they cannot adequately evaluate text generated for children. Also, several of our modules perform novel tasks for which evaluation criteria do not exist. Therefore, by referring to our guidelines and commonly used measures in NLP, we defined different evaluation questions for each sub-module (Table 4).

Test Data Collection: To evaluate whether our pipeline can be applied to various types of explanations, we collected test data (N=32) from multiple sources and domains. For sources, we chose an expert-generated source, BBC Science Focus Magazine, and a user-generated source, ELI5 dataset [25]. From each source, we collected data that corresponds to four domains: natural phenomena, biology, physics, and cultural and social conventions. Then, for each source and domain combination (total of 8), we collected two QA pairs for "why" questions and two for "how" questions. This totaled to 32 QA pairs (and more details are available in the Supplementary Materials).

Baselines: For the baselines, we selected state-of-the-art models for existing tasks and, for our new tasks, we used an LLM as it could perform the task from given instructions. For exemplification, we adopted GPT-3 with a zero-shot prompt (T11 in the Supplementary Materials) as a baseline since there is no appropriate specialized model for this task—Wang et al.'s [83] model is not open-sourced and is retrieval-based while our task is generative. For simplification, we adopted the state-of-the-art model MUSS [59]. For questioning, we trained a dialogue inpainting model following Dai et al. [16] using four datasets: QuAC [12], QReCC [1], DailyDialog [55], and Taskmaster [6]. Specifically, we finetuned the T5-large model [65] for three epochs with a learning rate of $3e^{-4}$.

Procedure: Inspired by the ACUTE-EVAL method [54] which is widely used for comparing generated dialogues, we showed a human evaluator two *responses*, one from our sub-module and one from the baseline. Both responses were generated from the same input *context*. For simplification and exemplification, we provided the original sentence and a context sentence as input context. For question generation, we provided a multi-turn dialogue including the `[BLANK]` that the model filled in. The human evaluators looked at the input context and the two generated *responses* side-by-side. The evaluator was also shown the measuring questions that corresponded to the sub-module and, for each question, were asked to make a choice between the two responses or to choose "tie". For each data point, we assigned three evaluators to collect three trials of such pairwise judgments, and used majority voting to designate whether our sub-module performed better, the baseline did, or whether it was a tie (e.g., the majority chose "tie"). We hired crowd workers as evaluators as the experts from our consultancy

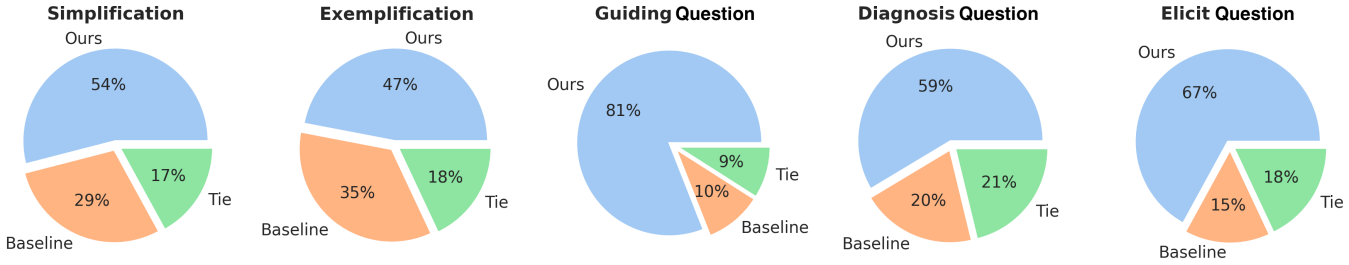

Figure 7: Overall human evaluation results for the five sub-modules. Outputs from our pipeline were assessed favorably compared to those from the baselines for all of the sub-modules.

	Simplification			Exemplification				Guiding Question			
Measuring Questions	**QS1**	**QS2**	**QS3**	**QC1**	**QC2**	**QC3**	**QC4**	**QG1**	**QG2**	**QG3**	**QG4**
Ours	**52%**	**54%**	**57%**	**49%**	**46%**	**42%**	**49%**	**81%**	**80%**	**88%**	**76%**
Baseline	34%	32%	21%	40%	38%	32%	31%	9%	11%	6%	16%
Tie	12%	14%	22%	11%	16%	26%	20%	10%	9%	6%	8%

	Diagnosis Question				Elicit Question				
Measuring Questions	**QD1**	**QD2**	**QD3**	**QD4**	**QE1**	**QE2**	**QE3**	**QE4**	**QE5**
Ours	**55%**	**70%**	**38%**	**69%**	**64%**	**83%**	**72%**	**53%**	**64%**
Baseline	27%	21%	17%	17%	14%	9%	21%	17%	14%
Tie	18%	9%	45%	14%	22%	8%	6%	30%	22%

Table 5: Human evaluation results on five sub-modules shows, for each measurement question, the percentage of workers that preferred our pipeline's outputs, the baseline's outputs, or chose that it was a tie. For all five sub-modules, our pipeline outperformed the baselines.

mentioned that even the general public can evaluate how helpful and easy it would be for a child to understand given content. We recruited crowd workers from Amazon Mechanical Turk who were in the US and had task approval rates higher than 98%. Each worker evaluated all five sub-modules (five pairs per sub-module) and answered one gold standard question. Our task took around 30 minutes and we paid workers $6 for their time. The final inter-annotator agreement was rated as fair (Fleiss's kappa=0.338).

5.2 Results

In short, the results generated by our pipeline were assessed favorably when compared to their corresponding baselines across all the criteria (Fig. 7). As seen in Table 5, the biggest difference between our sub-modules and the baselines was for the **guiding questions** (QG, 81% vs 10%), while the smallest difference was for **example** generation (QC, 47% vs. 35%). Evaluators judged that our **simplification** module generated text with easier words (QS1, 52% vs. 34%) and simpler sentence structures (QS2, 54% vs. 32%) while preserving the core information in the original text. The differences between our **examples** and the baseline's were relatively small in terms of helpfulness (QC1, 49% vs. 40%) and familiarity (QC2, 46% vs. 38%) since both generally composed examples with easier words. Regarding the **guiding (QG), diagnosis (QD), and elicit (QE) questions**, there were apparent differences between *turn inpainting* and the baseline across all criteria. We presume that the substantial differences are derived from whether the models explicitly considered that the recipient of the question is a child or not. The baseline failed to generate adequate outputs, since the model had likely never seen dialogues with children at training time. Specifically, evaluators rated all the questions from our *turn inpainting approach* to be more understandable for a child (QG1, QD1, QE1) and closer to what a teacher would ask (QG2, QD2, QE2). We found that the difference for grammatical errors is significant for guiding questions (QG3) while this gap is reduced for diagnosis questions (QD3) since these are usually simpler, e.g., "Did you get that?" or "What's the problem?"

For the QA pairs from both expert-generated and user-generated sources, our sub-modules produced better generations than the baseline did (Fig. 8 and 9). In terms of questioning, the differences between our sub-modules and the baselines were irrespective of source and domain. For simplification and exemplification, the difference between the models was greater in the expert-generated sources, which are written more formally. When analyzing by domain, the differences in simplification and exemplification seem to be greater in the Physics domain, which contains the most scientific content. Furthermore, we observed that the relatively poor exemplification performance in the Biology domain (39% vs. 38%) was due to our sub-module significantly underperforming QC3 (27% vs. 42%). This implies that our exemplification outputs often included text duplicated from the given context, but the outputs still contained valid examples as they were assessed better for the

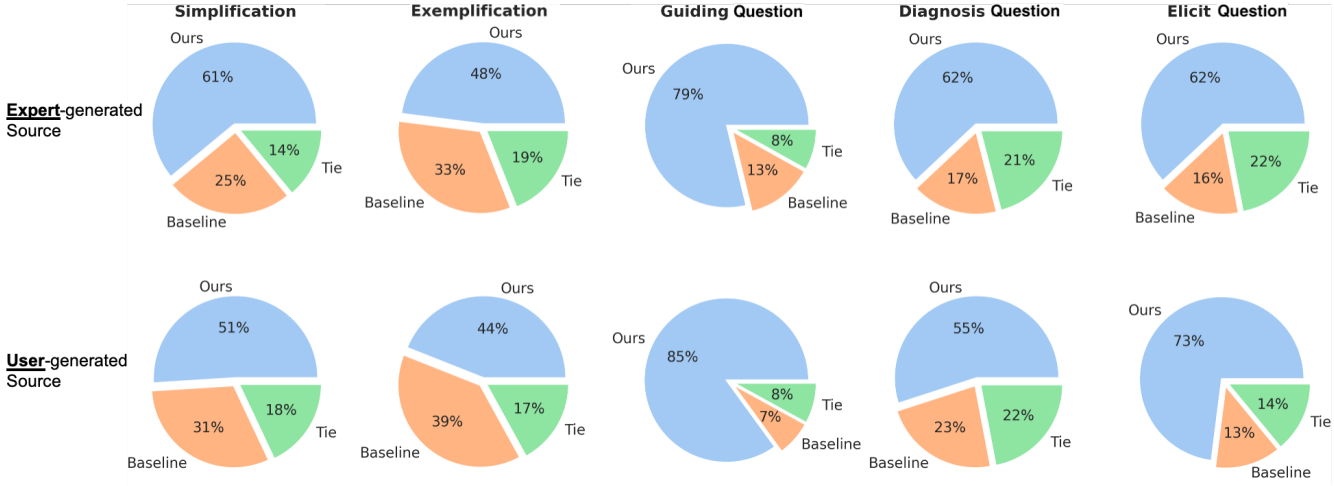

Figure 8: Human evaluation results on the five sub-modules according to the type of the source (i.e., expert-generated or user-generated). For all five sub-modules, our pipeline outperformed the baselines in the both type of sources.

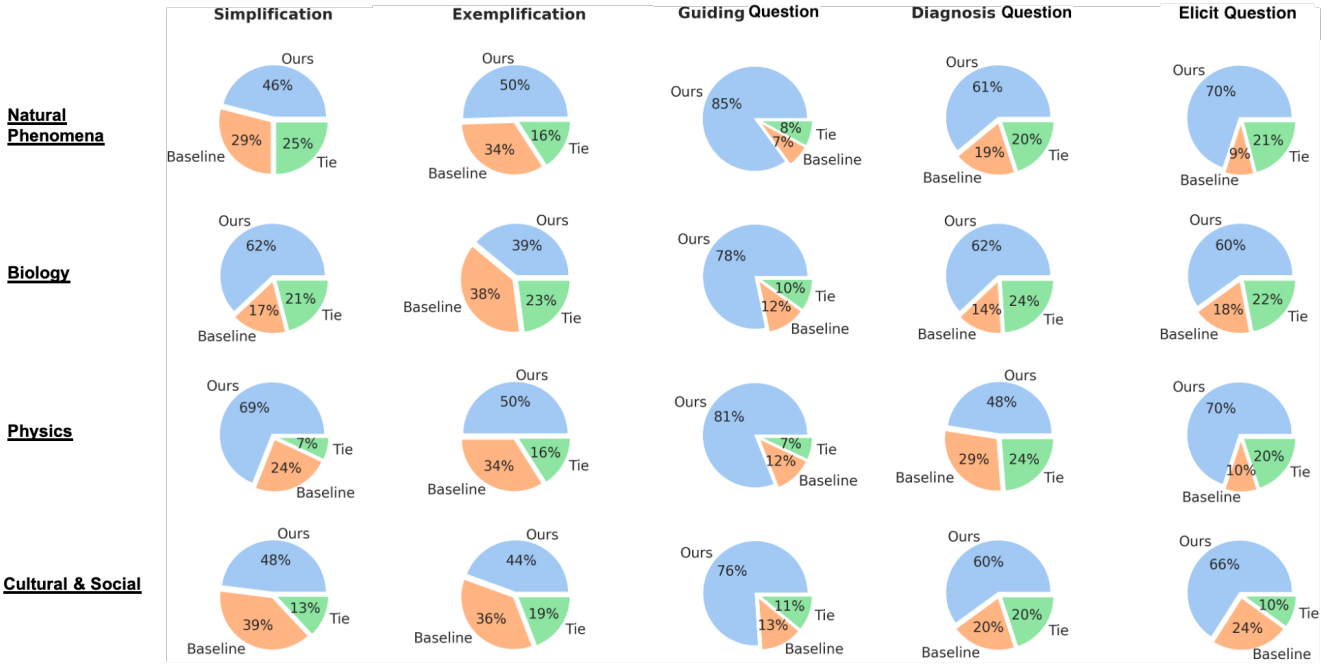

Figure 9: Human evaluation results on five sub-modules according to the domain. For all five sub-modules, ours outperformed the baselines in all the domains.

other criteria (i.e., QC1, QC2, and QC4). Nevertheless, we conclude that our sub-modules overall perform better than the baselines regardless of source and domain, but show greater differences in formal or scientific text for simplification and exemplification.

Additionally, as LLMs can generate hallucinations (i.e., non-factual or nonsensical information) [60] or errors (e.g., grammatically incorrect or incoherent text), we qualitatively analysed 20 sample generations for each of our sub-modules or baselines. For simplification, our sub-module and baseline produced hallucinations for the same 10% of the inputs and, for exemplification, our sub-module produced fewer hallucinations (20%) than the baseline (30%). Overall, both our sub-modules and baselines tended to hallucinate minor superfluous or incorrect details for the same inputs, but our sub-modules were generally more robust to hallucinations.

Also, while we recognize that exemplification had a relatively higher chance of hallucinations, the potential negative impact of these is the least significant as examples do not modify the original explanation and are the last resort support in our dialogues—only provided when children fail to understand after several adjustments. For question generation, all of our sub-modules produced few errors (<10%) while the baselines frequently produced questions that were incoherent or grammatically incorrect. For more details on this analysis, see the Supplementary Materials.

6 USER STUDY

We conducted a controlled study to investigate whether interactive conversations from DAPIE improve children's understanding of concepts (RQ1) and increases engagement (RQ2) compared to a simpler CA that provides the same information sentence-by-sentence.

	Sample Ratio
Female	31.25%
Age	
5-year-old	12.50% ($N = 2$)
6-year-old	31.25% ($N = 5$)
7-year-old	56.25% ($N = 9$)
Predominant Home Language	
English	62.50% ($N = 10$)
Other (Korean)	37.50% ($N = 6$)
Race/Ethnicity	
Asian	50.00% ($N = 8$)
White	18.75% ($N = 3$)
Black	12.50% ($N = 2$)
Other	18.75% ($N = 3$)
Parents' Education	
Bachelor's degree or higher	81.25% ($N = 13$)
Other	18.75% ($N = 3$)
Usage of CA	
Daily or Weekly	25.00% ($N = 4$)
Monthly	25.00% ($N = 4$)
Rarely	50.00% ($N = 8$)
N	16

Table 6: Demographics of the participants in our study.

6.1 Participants and Apparatus

We recruited 16 participants (5 female, 11 male) aged five through seven through snowball sampling and by posting advertisements on online forums (e.g., Twitter, Reddit, and the online communities of several colleges). Before the study, we assessed children's English language proficiency using a computer-based assessment (i.e., Quick Interactive Language Screener [51]) to ensure that participants can understand the questions in the conversations, assessment, and usability survey. Table 6 summarizes the participants' demographic information.

For explanatory material, we selected question and answer pairs for four domains: Natural Phenomena, Biology, Physics, and Cultural and Social Science. We chose these domains as they are commonly asked by children [58] and are topically diverse helps us test generalizability. For each domain, we selected two QA pairs, one for each question type (i.e., "why" and "how"). Participants were assigned a total of four questions where they saw one question per domain and two questions per question type.

We compared DAPIE's interface to a baseline interface with the same UI. However, instead of providing interactive conversations, the baseline provided the information in the original answer by presenting one sentence at a time. After each sentence, the interface showed "Okay" as the only option the user could click to respond with. Unlike existing real-world voice-based CAs which provide lengthy explanations in a single turn, this baseline provides greater interactivity and allows the user to consume the information step-by-step. Thus, we believe this is fair baseline since it provides a higher level of interactivity than what is supported in existing voice-based CAs. Each session lasted about 60 minutes and participants were compensated with $50.

6.2 Study Procedure

The study was conducted remotely. Children participated in the study from their homes and communicated with the researcher via a video conferencing tool[2]. Children first followed a simple tutorial dialogue where they were introduced to DAPIE and the baseline, and learned to use the systems by answering a few simple questions like "Are you ready?" Then, the children went through the explanations for four questions. They interacted with each condition for half of the questions—the order was counterbalanced. Each question was from a different domain (i.e., Natural Science, Biology, Physics, and Social Science). As learning Natural Science can affect understanding of Biology and vice versa, we grouped the domains such that participants saw questions from Natural Science and Biology in one condition, and Physics and Social Science in the other—limiting learning effects across conditions. Additionally, participants saw one "why" question and one "how" question in each condition—the order was counterbalanced. After each question, we conducted an assessment that asked about specific knowledge in the explanation. After each condition, we conducted usability surveys and semi-structured interviews. After the children completed the study, we conducted semi-structured interviews with their parents regarding their child's experience with the CAs. The child's screen and their camera video were recorded. The procedure was pre-approved by the IRB of the authors' institution.

6.3 Measures

For measures, we evaluated the participants' understanding of the information in the dialogues, their engagement with the dialogue, and their perceived usability of the interface. We also qualitatively analyzed the interview data.

6.3.1 Immediate Assessment. To assess children's understanding of concepts from the dialogue, we developed three questions for each dialogue. The questions assessed children's recall and understanding of facts introduced in the dialogue. These questions were different from the explanations embedded in the dialogue and did not overlap with the diagnosis questions provided in the dialogues. We designed these questions by consulting experts on children's

[2]https://zoom.us/

learning and language development. All the assessment items are included in the Supplementary Materials.

For all of the questions, we first asked children open-ended questions and allowed them to freely formulate their answer. If they were unable to provide the correct answer, we provided them with two answer options to choose from. Children received a score of 2 if they answered correctly without options, a score of 1 if they required options to answer correctly, and a score of 0 if they could not answer correctly.

6.3.2 Engagement. The evaluation of engagement was based on coders' assessments of each child's engagement in terms of three behaviors: eye gaze, verbal comments, and nonverbal comments. Eye gaze considered instances when participants stared at places other than the screen where the explanation was presented, which has been used as a negative indicator of engagement in children's book reading [35] and video watching [87]. On the other hand, verbal and nonverbal comments were considered as positive indicators. Verbal comments considered when a participant would verbally answer the CA's question, ask a question about on-topic information, or react to agent's response. These comments could be either to their parent or to the agent, although the agent could not understand these. Nonverbal comments included pointing at the screen, moving the cursor around, or clicking the interface to re-play the CA's explanation. Two coders observed the study sessions and, for each dialogue turn, recorded whether the turn included these engagement behaviors. Then, for each of the behaviors, we calculated $(number\_of\_turns\_with\_the\_behavior)/(total\_number\_of\_turns)\times 100$. The IRR calculated by Intraclass Correlation for the two coders was 0.73, which is considered as substantial [68].

6.3.3 Usability. For usability, we used a survey to elicit children's enjoyment on the usage experience, and their perceived trust towards DAPIE and the baseline. We adapted the four questions from the Giggle gauge [21] to measure enjoyment, and adapted two questions from Richards and Calvert's survey [69] for measuring children's perceived trust. For all items, children were first asked to indicate whether they agree with a statement (i.e., "yes" or "no") and then asked to clarify the magnitude to which they agree or disagree (i.e., "a bit" or "definitely"), leading to four possible ordinal response: "definitely no", "a bit no", "a bit yes", and "definitely yes" [87]. Finally, we asked them to compare both CAs for each dimension.

6.3.4 Interview Data. We qualitatively analyzed the video recordings of participants's usage of DAPIE, and the interview data from the children and parents. One of the authors iteratively coded the data through inductive analysis, and the other authors reviewed and verified the coding results.

6.4 Results

Children performed better on the immediate understanding assessment and were more engaged in the dialogues when using DAPIE compared to the baseline. To statistically analyze each measure under different conditions, we first conducted a Shapiro-Wilk test to determine if the data was parametric (P) or non-parametric (NP). Then, to compare between conditions, we used a paired t-test (if parametric) and a Wilcoxon signed-rank test (if non-parametric).

6.4.1 Immediate Assessment Score Analysis. Out of a maximum of 12 points for the understanding assessment, children's score when they used DAPIE ($M = 7.43, SD = 2.57$) significantly outperformed scores with the baseline ($M = 5.13, SD = 2.52$). The difference equates to correctly answering one more question out of six questions ($p < 0.05$, NP).

We observed that this greater understanding with DAPIE was possibly due to how it provided adaptive explanations (i.e., simplifying, defining difficult words, or providing examples). Through the system, fifteen children received adaptive explanations more than once. We did not ask children whether they understood an explanation after they received each adjustment to not influence their understanding. However, we observed that, on average, the children went through a whole thread of adjusted explanations at least once per dialogue, which might show that children needed all the adjustments once per dialogue. We consider that children who received adaptive explanations were able to digest the information more easily, leading to a better understanding. P14 mentioned, *"DAPIE is more like my dad or kind teachers who explain again more easily when I couldn't understand".*

Additionally, we observed that DAPIE questions could encourage more active learning. When DAPIE provided diagnosis or elicit questions with multiple choices, some of the children kept moving their cursors while they thought of an answer (C1, C2, C3, C8), and others expressed excitement after getting the correct answer (C1, C6, C12). C7 said, *"I can focus more to get correct answers. I'm happy when I get answers".* This finding is aligned with prior work that argues that meaningful learning occurs when learners think about the information presented rather than just passively receiving it [11].

Furthermore, children generally felt that the language from DAPIE was easier to understand than the language from the baseline. C4, C5, C9, and C14 mentioned that they liked DAPIE more as its explanations were "easier". For example, C9 said, "*DAPIE is like my friend because words are easier than the other one so it makes me more comfortable*".

6.4.2 Engagement Analysis. By analyzing the participants' behaviors, we observed that DAPIE could promote engagement. Children's gaze distraction when using DAPIE ($M = 11.9\%, SD = 0.16$) was statistically lower than when they used the baseline ($M = 29\%, SD = 0.28, p < 0.01$). Also, participants made verbal comments more frequently when using DAPIE ($M = 18.7\%, SD = 0.11$) than when they used the baseline ($M = 11.5\%, SD = 0.13$), but there was no statistical difference ($p > 0.05$, NP). However, children's non-verbal comments when using DAPIE ($M = 11\%, SD = 0.12$) were statistically lower than with the baseline ($M = 18.36\%, SD = 0.19, p < 0.05$). In addition to these behaviors, we also analyzed overall interaction time and observed that children used DAPIE for longer ($M = 7.03\,minutes$, $SD = 3.03$) than the baseline ($M = 2.80\,minutes$, $SD = 1.33$).

Fifteen children chose to listen to the explanation on detailed units when guiding questions were given. Although this could make the dialogues longer, the children focused on the explanations and did not blindly skip or ignore these dialogue turns. Thus, with DAPIE, participants saw an average of three times more turns (20) than with the baseline (7) while also being more engaged.

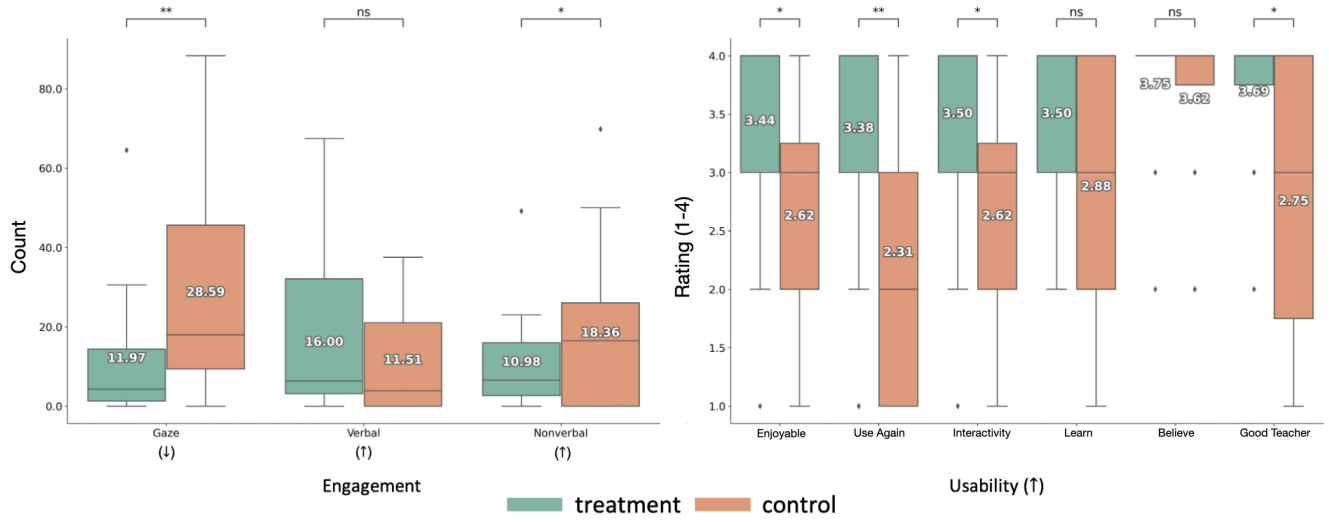

Figure 10: Results of engagement and usability analysis. **, *, and ns indicate significance of $p < 0.01$, $p < 0.05$, and $p > 0.05$, respectively.

Moreover, when DAPIE provided diagnosis and guiding questions, some participants reacted with a smile (C3, C6, C12, C13) or said their answer out loud while also providing rationales (C1, C6, C14), showing their engagement and curiosity. Thus, children tended to be more engaged in the interactive dialogue, which allowed the child to decide whether they wanted more details or not. This finding aligns with prior work that showed granting agency to a child can increase engagement in an activity [57].

For gaze distraction, we observed that it usually occurred when the child wanted to say something to their parents or when they could not concentrate on the dialogue (e.g., distracted by surrounding noises or lost interest). With DAPIE, we observed that the former happened more frequently. Several of the children looked at their parents to explain their rationale, boast about being correct, and talk about other related topics. However, with the baseline, we observed more of the latter behavior where some children looked at other things while the system was talking and clicked the "Okay" button as soon as it was displayed.

6.4.3 Usability Analysis. Overall, the children felt that the dialogues from DAPIE were more enjoyable and trustworthy than the baseline's (Fig. 10). In terms of enjoyment, children felt that DAPIE was significantly more enjoyable and interactive. They expressed that they would like to use DAPIE again ($M = 3.38, SD = 1.05$) significantly more than the baseline ($M = 2.31, SD = 1.10, p < 0.05$). In terms of trust, children felt DAPIE was significantly a better teacher ($M = 3.69, SD = 0.58$) than the baseline ($M = 2.75, SD = 1.15, p < 0.05$). Children also felt that they learned new things with DAPIE ($p = 0.07$) and that the system was more trustworthy ($p = 0.60$), but these differences were not significant.

Children noted DAPIE's easier language, questions and corrections, and various explanations as reasons for their positive reactions. For example, C5 said *"DAPIE is a good teacher. It's easier, and it makes me understand new information"*. C6 mentioned that they wanted to use DAPIE again because it provided "a lot of stories", and C9 felt he learned new things from the system because it provided more detailed information, which he wanted. While most children liked the interactive experience, C12 noted that the extra effort involved in answering and clicking could be burdensome.

Regarding trustworthiness, which showed the smallest difference between the two conditions, children generally believed that both systems were trustworthy since they both provided new information, which made the systems appear smart. However, several children commented that DAPIE seemed smarter because it acted like a teacher—e.g., helping them understand, correcting their answers, providing questions. On the other hand, some children also perceived the baseline as more trustworthy and intelligent since it talked in longer sentences.

6.4.4 Parents' Perception. In their interviews, parents mentioned that DAPIE was more interactive and provided explanations that were easier to digest, which aligned with their children's comments. Also, parents were surprised that their children could focus on longer dialogues and expressed that this was due to the interactivity. Others mentioned that, by watching how their children enjoyed interacting with DAPIE, they also learned about how they should interact with their child.

Some parents emphasized the benefits of the adaptive explanation provided by DAPIE. They explained how the system acted similarly to how parents or teachers might adapt explanations for their children. For example, parents mentioned that the questions provided by DAPIE were similar to the questions they wanted to give their child when they interacted with the baseline. P14 said, "*When DAPIE explained, I wanted to ask whether my child knows the meaning of these words like "experts" and "product," but [the system] asked these questions to my daughter, so I like it.*"

Parents emphasized the benefits of DAPIE's more accessible explanations and its correction feedback, and related this to their own challenges in delivering scientific information. They mentioned that explaining science to their children is challenging as it is difficult to understand the information and to transform it into simpler expressions (P9, P11, P12, P13). For example, P9 said, *"My son asks me these questions quite frequently, but I couldn't always know the answers, so I googled the question to get information and change it in a way that my child could understand. This process is challenging, so sometimes I can't care about whether my son understands or not."* Moreover, P8, P14, and P16 said that they were able to learn about new information from DAPIE, and that interacting with DAPIE could serve as a bonding activity for parents and children. P8 mentioned, *"It reminded me of what I learned in school. Interestingly, we can learn together"*. However, some parents also mentioned that it was challenging to detect AI-related errors in the turns generated by DAPIE. In fact, several parents were unable to recognize that the dialogues were generated by an AI model at first, and thus failed to recognize any significant errors in the explanations.

7 DISCUSSION

In this section, we discuss the potential of AI-based interactive dialogues for children and parents, how our approach could enrich information understanding for user groups beyond children, implications for dialogue design, and limitations and future work.

7.1 AI-based Interactive Dialogues for Children and Parents

In a real-life setting, DAPIE can be beneficial for answering children's questions when parents are unable or unavailable. Parents may struggle to understand the information needed to answer children's diverse and unpredictable questions [61] and, as expressed by parents in our study, they must dedicate significant effort to make the information comprehensible for their children. Through our system, children can satisfy their curiosity by accessing and consuming explanations on the internet, whenever they need them, without parents' effort.

Furthermore, we believe that parents and DAPIE can collaborate and combine their expertise to produce more meaningful experiences for their children: parents as experts in interacting with their children, and DAPIE as an expert on the domain. DAPIE can easily retrieve an explanation, simplify it, and design initial interventions—time-consuming and burdensome tasks for parents. By handing off this burden, parents can then focus on their child by observing how they interact with the system and noticing any difficulties. Future extensions of DAPIE could allow the parent to mediate in the dialogue and offer this knowledge to DAPIE, which it can then use to generate more personalized interventions. This presents an effective case of human-AI collaboration that leverages the respective strengths of the human and the AI.

7.2 AI Errors: Propagate Outputs but Intercept Errors

While children were generally positive about generated turns, our AI-based pipeline could occasionally produce unsatisfactory or nonsensical turns due to its inability to propagate outputs across turns. As turns are considered separately when generating questions, the pipeline could lose context about the dialogue and generate confusing questions. In one dialogue, a sentence used the pronoun "these" to refer to an entity in the previous turn, and the pipeline generated a diagnosis question with "these", "that" and "those" as answer options. This question frustrated child participants as the options were nonsensical but, if they did not choose "these", they were told that they were wrong. Further, as the pipeline does not consider what was provided in prior turns, it could generate similar questions across multiple turns. For example, one dialogue explained aspects of gravity in each turn and provided several diagnosis questions where the answer was "gravity"—causing children to feel tired and bored. While these failure cases indicate that the pipeline should propagate information across turns, we observed that propagating AI's outputs could lead to error propagation. Thus, to prevent errors from propagation, the pipeline must also incorporate modules to detect, filter and/or correct failed generations before they are propagated. Additionally, the system can include simple feedback buttons (e.g., "bored", "bot was not smart") for children to explicitly express intent, and for the system to mitigate the cost of errors by providing other explanations.

7.3 Beyond Young Children and Beyond Question Answering

Although our system targets young children (ages five to seven), we believe that our guidelines and the general structure of our pipeline can generalize to support older children and even adults. As learning theory [70, 82] suggests reducing support (i.e., fading) according to children's ability, the pipeline can be adjusted to make more challenging and cognitively engaging dialogues for older children. For example, the pipeline could provide longer explanations per turn or simplify less to help children learn new terms. Also, instead of the recall-based diagnosis questions, which bored some children in our study, the pipeline could generate self-explanation questions which challenge the child to explain what they have just learned.

A significant merit of our computational pipeline is that it maintains the core information in the original answer, but *wraps* it with interactivity. Beyond supporting children and question answering, this functionality can be extended to enable a new form of reading support. In learning contexts, our pipeline could be extended to assist in the active reading of documents (e.g., textbooks) by creating document-based dialogue turns that guide readers to other relevant parts of the document, diagnose their understanding, and provide adaptive interventions. Whereas current document-based QA models (e.g., Qasper [18]) focus on facilitating information-seeking, this approach could generate dialogues that focus on cognitively engaging users in the reading activity.

7.4 Limitations and Future Work

As our pipeline generates dialogues while retaining core information, parents in our study said that they rarely noticed any harms or dangers in the dialogues. However, LLMs can exhibit biases [3] and our pipeline could propagate or even amplify various biases (e.g., gender, race, and culture). For these reasons, safeguards are needed to prevent negative impact on children. For example, our system could apply NLP techniques to recognize and mitigate biases [77] or,

instead of generating on-the-fly, allow parents to verify dialogues before their children can access them.

Our study had several limitations. First, the study compared our full system to a baseline that provides sentences one-by-one. We adopted this design to evaluate the comprehensive experience supported by our system, but this makes it difficult to discern the effect of each component (e.g., simplification, questions). Second, as we assessed children's understanding immediately after each dialogue, the effect of the system on children's long-term retention is unclear. Third, although our target scenario is for when children ask "why" and "how" questions, we did not allow children in our study to ask the questions to strictly control the experiment design. Future work could explore how children ask questions with our system through a deployment study. Finally, our participant pool was skewed regarding race, age, and usage of voice assistants. Future work could conduct studies with younger children and more diverse demographics.

8 CONCLUSION

This work proposes design guidelines for creating interactive dialogues that help children understand answers to their "why" and "how" questions. Applying these guidelines, we developed DAPIE, a novel AI-based system that automatically transforms existing long-form answers from online sources into interactive dialogues. Through a technical evaluation and a user study, we found that DAPIE shows reliable performance in generating interactive dialogues, and these dialogues are effective in promoting children's understanding and enjoyment. With recent advances in generative AI models, we hope that DAPIE paves the way for providing children with more accessible and engaging forms to learn and consume information.

ACKNOWLEDGMENTS

This work was supported by KAIST-NAVER Hypercreative AI Center. The authors would like to thank our participants for their positive engagement during the studies and the reviewers as their feedback helped us improve our paper. Finally, we also thank Dongsun Yim for the valuable feedback and advice.

REFERENCES

[1] Raviteja Anantha, Svitlana Vakulenko, Zhucheng Tu, Shayne Longpre, Stephen Pulman, and Srinivas Chappidi. 2021. Open-Domain Question Answering Goes Conversational via Question Rewriting. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 520–534.

[2] David Paul Ausubel. 2012. *The acquisition and retention of knowledge: A cognitive view.* Springer Science & Business Media.

[3] 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).

[4] Elizabeth Baraff Bonawitz and Tania Lombrozo. 2012. Occam's rattle: children's use of simplicity and probability to constrain inference. *Developmental psychology* 48, 4 (2012), 1156.

[5] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. https://doi.org/10.48550/ARXIV.2005.14165

[6] Bill Byrne, Karthik Krishnamoorthi, Chinnadhurai Sankar, Arvind Neelakantan, Daniel Duckworth, Semih Yavuz, Ben Goodrich, Amit Dubey, Andy Cedilnik, and Kyu-Young Kim. 2019. Taskmaster-1: Toward a realistic and diverse dialog dataset. *arXiv preprint arXiv:1909.05358* (2019).

[7] Sonia Q. Cabell, Laura M. Justice, Anita S. McGinty, Jamie DeCoster, and Lindsay D. Forston. 2015. Teacher–child conversations in preschool classrooms: Contributions to children's vocabulary development. *Early Childhood Research Quarterly* 30 (2015), 80–92. https://doi.org/10.1016/j.ecresq.2014.09.004

[8] William Cai, Hao Sheng, and Sharad Goel. 2020. MathBot: A Personalized Conversational Agent for Learning Math.

[9] Yi Cheng, Kate Yen, Yeqi Chen, Sijin Chen, and Alexis Hiniker. 2018. Why Doesn't It Work? Voice-Driven Interfaces and Young Children's Communication Repair Strategies. In *Proceedings of the 17th ACM Conference on Interaction Design and Children* (Trondheim, Norway) *(IDC '18)*. Association for Computing Machinery, New York, NY, USA, 337–348. https://doi.org/10.1145/3202185.3202749

[10] Michelene T.H. Chi, Miriam Bassok, Matthew W. Lewis, Peter Reimann, and Robert Glaser. 1989. Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science* 13, 2 (1989), 145–182. https://doi.org/10.1016/0364-0213(89)90002-5

[11] Michelene T. H. Chi and Ruth Wylie. 2014. The ICAP Framework: Linking Cognitive Engagement to Active Learning Outcomes. *Educational Psychologist* 49, 4 (2014), 219–243. https://doi.org/10.1080/00461520.2014.965823 arXiv:https://doi.org/10.1080/00461520.2014.965823

[12] Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question Answering in Context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 2174–2184. https://doi.org/10.18653/v1/D18-1241

[13] Michelle M Chouinard, Paul L Harris, and Michael P Maratsos. 2007. Children's questions: A mechanism for cognitive development. *Monographs of the society for research in child development* (2007), i–129.

[14] Kathleen H. Corriveau and Katelyn E. Kurkul. 2014. "Why Does Rain Fall?": Children Prefer to Learn From an Informant Who Uses Noncircular Explanations. *Child Development* 85, 5 (2014), 1827–1835. http://www.jstor.org/stable/24033022

[15] Catherine Crain-Thoreson, Michael P Dahlin, and Terris A Powell. 2001. Parent-child interaction in three conversational contexts: Variations in style and strategy. *New directions for child and adolescent development* 2001, 92 (2001), 23–38. https://doi.org/10.1002/cd.13

[16] Zhuyun Dai, Arun Tejasvi Chaganty, Vincent Zhao, Aida Amini, Mike Green, Qazi Rashid, and Kelvin Guu. 2022. Dialog Inpainting: Turning Documents to Dialogs. In *International Conference on Machine Learning (ICML)*. PMLR.

[17] Judith H. Danovitch, Candice M. Mills, Kaitlin R. Sands, and Allison J. Williams. 2021. Mind the gap: How incomplete explanations influence children's interest and learning behaviors. *Cognitive Psychology* 130 (2021), 101421. https://doi.org/10.1016/j.cogpsych.2021.101421

[18] Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. 2021. A Dataset of Information-Seeking Questions and Answers Anchored in Research Papers. In *NAACL*.

[19] Jan De Belder and Marie-Francine Moens. 2010. Text simplification for children. (01 2010).

[20] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[21] Griffin Dietz, Zachary Pease, Brenna McNally, and Elizabeth Foss. 2020. Giggle Gauge: A Self-Report Instrument for Evaluating Children's Engagement with Technology. In *Proceedings of the Interaction Design and Children Conference* (London, United Kingdom) *(IDC '20)*. Association for Computing Machinery, New York, NY, USA, 614–623. https://doi.org/10.1145/3392063.3394393

[22] Stefania Druga, Randi Williams, Cynthia Breazeal, and Mitchel Resnick. 2017. "Hey Google is It OK If I Eat You?": Initial Explorations in Child-Agent Interaction. In *Proceedings of the 2017 Conference on Interaction Design and Children* (Stanford, California, USA) *(IDC '17)*. Association for Computing Machinery, New York, NY, USA, 595–600. https://doi.org/10.1145/3078072.3084330

[23] Gerald G. Duffy, Laura R. Roehler, Michael S. Meloth, and Linda G. Vavrus. 1986. Conceptualizing instructional explanation. *Teaching and Teacher Education* 2, 3 (1986), 197–214. https://doi.org/10.1016/S0742-051X(86)80002-6

[24] Mary Evans, Shelley Moretti, Deborah Shaw, and Maureen Fox. 2003. Parent Scaffolding in Children's Oral Reading. *Early Education and Development - EARLY EDUC DEV* 14 (07 2003), 363–388. https://doi.org/10.1207/s15566935eed1403_5

[25] Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long Form Question Answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 3558–3567. https://doi.org/10.18653/v1/P19-1346

[26] Brandy N. Frazier, Susan A. Gelman, and Henry M. Wellman. 2009. Preschoolers' Search for Explanatory Information within Adult: Child Conversation. *Child Development* 80, 6 (2009), 1592–1611. http://www.jstor.org/stable/25592097

[27] Brandy N Frazier, Susan A Gelman, and Henry M Wellman. 2016. Young children prefer and remember satisfying explanations. *Journal of Cognition and Development* 17, 5 (2016), 718–736.

[28] Radhika Garg, Hua Cui, Spencer Seligson, Bo Zhang, Martin Porcheron, Leigh Clark, Benjamin R. Cowan, and Erin Beneteau. 2022. The Last Decade of HCI Research on Children and Voice-Based Conversational Agents. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 149, 19 pages. https://doi.org/10.1145/3491102.3502016

[29] Susan A Gelman. 1988. The development of induction within natural kind and artifact categories. *Cognitive Psychology* 20, 1 (1988), 65–95. https://doi.org/10.1016/0010-0285(88)90025-4

[30] Usha Goswami. 2001. *Analogical Reasoning in Children.* 437 – 470.

[31] Graeme S. Halford. 2009. *Children's understanding: The development of mental models.* Erlbaum.

[32] Paul L Harris. 2012. *Trusting what you're told: How children learn from others.* Harvard University Press.

[33] Jiangbo Hu, Camilla Gordon, Ning Yang, and Yonggang Ren. 2020. "Once Upon A Styoar": A Science Education Program Based on Personification Storytelling in Promoting Preschool Children's Understanding of Astronomy Concepts. *Early Education and Development* 32 (05 2020), 1–19. https://doi.org/10.1080/10409289.2020.1759011

[34] Kayoko Inagaki and Giyoo Hatano. 1987. Young Children's Spontaneous Personification as Analogy. *Child Development* 58, 4 (1987), 1013–1020. http://www.jstor.org/stable/1130542

[35] Joan N Kaderavek, Ying Guo, and Laura M Justice. 2014. Validity of the children's orientation to book reading rating scale. *Journal of Research in Reading* 37, 2 (2014), 159–178.

[36] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP).* 6769–6781.

[37] Deborah Kelemen. 2019. The Magic of Mechanism: Explanation-Based Instruction on Counterintuitive Concepts in Early Childhood. *Perspectives on Psychological Science* 14 (04 2019), 174569161982701. https://doi.org/10.1177/1745691619827011

[38] James Kennedy, Séverin Lemaignan, Caroline Montassier, Pauline Lavalade, Bahar Irfan, Fotios Papadopoulos, Emmanuel Senft, and Tony Belpaeme. 2017. Child Speech Recognition in Human-Robot Interaction: Evaluations and Recommendations. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction* (Vienna, Austria) *(HRI '17)*. Association for Computing Machinery, New York, NY, USA, 82–90. https://doi.org/10.1145/2909824.3020229

[39] Yea-Seul Kim, Jessica Hullman, Matthew Burgess, and Eytan Adar. 2016. SimpleScience: Lexical Simplification of Scientific Terminology. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing.* Association for Computational Linguistics, Austin, Texas, 1066–1071. https://doi.org/10.18653/v1/D16-1114

[40] Alison King. 1994. Guiding knowledge construction in the classroom: Effects of teaching children how to question and how to explain. *American educational research journal* 31, 2 (1994), 338–368.

[41] Christina Krist, Christina V. Schwarz, and Brian J. Reiser. 2019. Identifying Essential Epistemic Heuristics for Guiding Mechanistic Reasoning in Science Learning. *Journal of the Learning Sciences* 28, 2 (2019), 160–205. https://doi.org/10.1080/10508406.2018.1510404 arXiv:https://doi.org/10.1080/10508406.2018.1510404

[42] Christoph Kulgemeyer. 2018. Towards a framework for effective instructional explanations in science teaching. *Studies in Science Education* 54, 2 (2018), 109–139. https://doi.org/10.1080/03057267.2018.1598054 arXiv:https://doi.org/10.1080/03057267.2018.1598054

[43] Katelyn E Kurkul, Eleanor Castine, Kathryn Leech, and Kathleen H Corriveau. 2021. How does a switch work? The relation between adult mechanistic language and children's learning. *Journal of Applied Developmental Psychology* 72 (2021), 101221. https://doi.org/10.1016/j.appdev.2020.101221

[44] Katelyn E Kurkul and Kathleen H Corriveau. 2018. Question, explanation, follow-up: A mechanism for learning from others? *Child Development* 89, 1 (2018), 280–294.

[45] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association for Computational Linguistics* 7 (2019), 452–466. https://doi.org/10.1162/tacl_a_00276

[46] Yoonjoo Lee, John Joon Young Chung, Tae Soo Kim, Jean Y Song, and Juho Kim. 2022. Promptiverse: Scalable Generation of Scaffolding Prompts Through Human-AI Hybrid Knowledge Graph Annotation. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 96, 18 pages. https://doi.org/10.1145/3491102.3502087

[47] Cristine H Legare. 2014. The contributions of explanation and exploration to children's scientific reasoning. *Child Development Perspectives* 8, 2 (2014), 101–106. https://doi.org/10.1111/cdep.12070

[48] Cristine H Legare and Tania Lombrozo. 2014. Selective effects of explanation on learning during early childhood. *Journal of experimental child psychology* 126 (2014), 198–212.

[49] Gaea Leinhardt, Kevin Crowley, and Karen Knutson. 2015. *Building islands of expertise in everyday family activity.* Routledge.

[50] Gaea Leinhardt and Michael Steele. 2005. Seeing the Complexity of Standing to the Side: Instructional Dialogues. *Cognition and Instruction - COGNITION INSTRUCT* 23 (03 2005), 87–163. https://doi.org/10.1207/s1532690xci2301_4

[51] Dani Levine, Amy Pace, Rufan Luo, Kathy Hirsh-Pasek, Roberta Michnick Golinkoff, Jill de Villiers, Aquiles Iglesias, and Mary Sweig Wilson. 2020. Evaluating socioeconomic gaps in preschoolers' vocabulary, syntax and language process skills with the Quick Interactive Language Screener (QUILS). *Early Childhood Research Quarterly* 50 (2020), 114–128.

[52] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.* 7871–7880.

[53] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems* 33 (2020), 9459–9474.

[54] Margaret Li, Jason Weston, and Stephen Roller. 2019. ACUTE-EVAL: Improved Dialogue Evaluation with Optimized Questions and Multi-turn Comparisons. https://doi.org/10.48550/ARXIV.1909.03087

[55] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957* (2017).

[56] Paul Light and George Butterworth. 2017. *Chapter 7. Desituating cognition through the construction of conceptual knowledge.* Routledge.

[57] Mike E.U. Ligthart, Mark A. Neerincx, and Koen V. Hindriks. 2020. Design Patterns for an Interactive Storytelling Robot to Support Children's Engagement and Agency. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction* (Cambridge, United Kingdom) *(HRI '20)*. Association for Computing Machinery, New York, NY, USA, 409–418. https://doi.org/10.1145/3319502.3374826

[58] Silvia B. Lovato, Anne Marie Piper, and Ellen A. Wartella. 2019. Hey Google, Do Unicorns Exist? Conversational Agents as a Path to Answers to Children's Questions. In *Proceedings of the 18th ACM International Conference on Interaction Design and Children* (Boise, ID, USA) *(IDC '19)*. Association for Computing Machinery, New York, NY, USA, 301–313. https://doi.org/10.1145/3311927.3323150

[59] Louis Martin, Angela Fan, Éric de la Clergerie, Antoine Bordes, and Benoît Sagot. 2021. MUSS: Multilingual Unsupervised Sentence Simplification by Mining Paraphrases. *arXiv preprint arXiv:2005.00352* (2021).

[60] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661* (2020).

[61] Candice M. Mills, Judith. H. Danovitch, Victoria N. Mugambi, Kaitlin R. Sands, and Candice Pattisapu Fox. 2022. "Why do dogs pant?": Characteristics of parental explanations about science predict children's knowledge. *Child Development* 93, 2 (2022), 326–340. https://doi.org/10.1111/cdev.13681 arXiv:https://srcd.onlinelibrary.wiley.com/doi/pdf/10.1111/cdev.13681

[62] Candice M Mills, Kaitlin R Sands, Sydney P Rowles, and Ian L Campbell. 2019. "I want to know more!": Children are sensitive to explanation quality when exploring new information. *Cognitive Science* 43, 1 (2019), e12706.

[63] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2021. WebGPT: Browser-assisted question-answering with human feedback. *CoRR* abs/2112.09332 (2021). arXiv:2112.09332 https://arxiv.org/abs/2112.09332

[64] Nielsen. 2018. (Smart) speaking my language: Despite their vast capabilities, smart speakers are all about the music.

[65] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research* 21, 140 (2020), 1–67. http://jmlr.org/papers/v21/20-074.html

[66] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing.* 2383–2392.

[67] Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A Conversational Question Answering Challenge. *Transactions of the Association for Computational Linguistics* 7 (2019), 249–266. https://doi.org/10.1162/tacl_a_00266

[68] Darrel A Regier, William E Narrow, Diana E Clarke, Helena C Kraemer, S Janet Kuramoto, Emily A Kuhl, and David J Kupfer. 2013. DSM-5 field trials in the United States and Canada, Part II: test-retest reliability of selected categorical

diagnoses. *American journal of psychiatry* 170, 1 (2013), 59–70.

[69] Melissa N. Richards and Sandra L. Calvert. 2017. Measuring young U.S. children's parasocial relationships: toward the creation of a child self-report survey. *Journal of Children and Media* 11, 2 (2017), 229–240. https://doi.org/10.1080/17482798.2017.1304969 arXiv:https://doi.org/10.1080/17482798.2017.1304969

[70] Laura R Roehler and Danise J Cantlon. 1997. Scaffolding: A powerful tool in social constructivist classrooms. (1997).

[71] Rod D. Roscoe and Michelene T. H. Chi. 2007. Understanding Tutor Learning: Knowledge-Building and Knowledge-Telling in Peer Tutors' Explanations and Questions. *Review of Educational Research* 77, 4 (2007), 534–574. https://doi.org/10.3102/0034654307309920 arXiv:https://doi.org/10.3102/0034654307309920

[72] Sherry Ruan, Liwei Jiang, Justin Xu, Bryce Joe-Kun Tham, Zhengneng Qiu, Yeshuang Zhu, Elizabeth L. Murnane, Emma Brunskill, and James A. Landay. 2019. *QuizBot: A Dialogue-Based Adaptive Learning System for Factual Knowledge*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300587

[73] Rosemary Russ, Rachel Scherr, David Hammer, and Jamie Mikeska. 2008. Recognizing mechanistic reasoning in student scientific inquiry: A framework for discourse analysis developed from philosophy of science. *Science Education* 92 (05 2008), 499 – 525. https://doi.org/10.1002/sce.20264

[74] Alex Sciuto, Arnita Saini, Jodi Forlizzi, and Jason I. Hong. 2018. "Hey Alexa, What's Up?": A Mixed-Methods Studies of In-Home Conversational Agent Usage. In *Proceedings of the 2018 Designing Interactive Systems Conference* (Hong Kong, China) *(DIS '18)*. Association for Computing Machinery, New York, NY, USA, 857–868. https://doi.org/10.1145/3196709.3196772

[75] Catherine Snow. 1983. Literacy and language: Relationships during the preschool years. *Harvard educational review* 53, 2 (1983), 165–189.

[76] Dan Su, Xiaoguang Li, Jindi Zhang, Lifeng Shang, Xin Jiang, Qun Liu, and Pascale Fung. 2022. Read before Generate! Faithful Long Form Question Answering with Machine Reading. https://doi.org/10.48550/ARXIV.2203.00343

[77] Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating Gender Bias in Natural Language Processing: Literature Review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 1630–1640. https://doi.org/10.18653/v1/P19-1159

[78] Anuj Tewari and John Canny. 2014. What Did Spot Hide? A Question-Answering Game for Preschool Children. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 1807–1816. https://doi.org/10.1145/2556288.2557205

[79] Barbara Tizard and Martin Hughes. 2008. *Young children learning*. John Wiley & Sons.

[80] Araceli Valle and Maureen Callanan. 2006. Similarity Comparisons and Relational Analogies in Parent-Child Conversations About Science Topics. *Merrill-Palmer Quarterly* 52 (01 2006), 96–124. https://doi.org/10.1353/mpq.2006.0009

[81] Stella Vosniadou and Marlene Schommer. 1988. Explanatory Analogies Can Help Children Acquire Information from Expository Text. Technical Report No. 460.

[82] Lev Semenovich Vygotsky and Michael Cole. 1978. *Mind in society: Development of higher psychological processes*. Harvard university press.

[83] Shufan Wang, Fangyuan Xu, Laure Thompson, Eunsol Choi, and Mohit Iyyer. 2022. Modeling Exemplification in Long-form Question Answering via Retrieval. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Seattle, United States, 2079–2092. https://doi.org/10.18653/v1/2022.naacl-main.151

[84] Rainer Winkler, Sebastian Hobert, Antti Salovaara, Matthias Söllner, and Jan Marco Leimeister. 2020. *Sara, the Lecturer: Improving Learning in Online Education with a Scaffolding-Based Conversational Agent*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376781

[85] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *arXiv preprint arXiv:2007.00808* (2020).

[86] Fangyuan Xu, Junyi Jessy Li, and Eunsol Choi. 2022. How Do We Answer Complex Questions: Discourse Structure of Long-form Answers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Dublin, Ireland, 3556–3572. https://doi.org/10.18653/v1/2022.acl-long.249

[87] Ying Xu, Valery Vigil, Andres S. Bustamante, and Mark Warschauer. 2022. "Elinor's Talking to Me!":Integrating Conversational AI into Children's Narrative Science Programming. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 166, 16 pages. https://doi.org/10.1145/3491102.3502050

[88] Ying Xu and Mark Warschauer. 2020. A Content Analysis of Voice-Based Apps on the Market for Early Literacy Development. In *Proceedings of the Interaction Design and Children Conference* (London, United Kingdom) *(IDC '20)*. Association for Computing Machinery, New York, NY, USA, 361–371. https://doi.org/10.1145/3392063.3394418

[89] Ying Xu and Mark Warschauer. 2020. Exploring Young Children's Engagement in Joint Reading with a Conversational Agent. In *Proceedings of the Interaction Design and Children Conference* (London, United Kingdom) *(IDC '20)*. Association for Computing Machinery, New York, NY, USA, 216–228. https://doi.org/10.1145/3392063.3394417

A EXAMPLE DIALOGUE GENERATED BY OUR PIPELINE

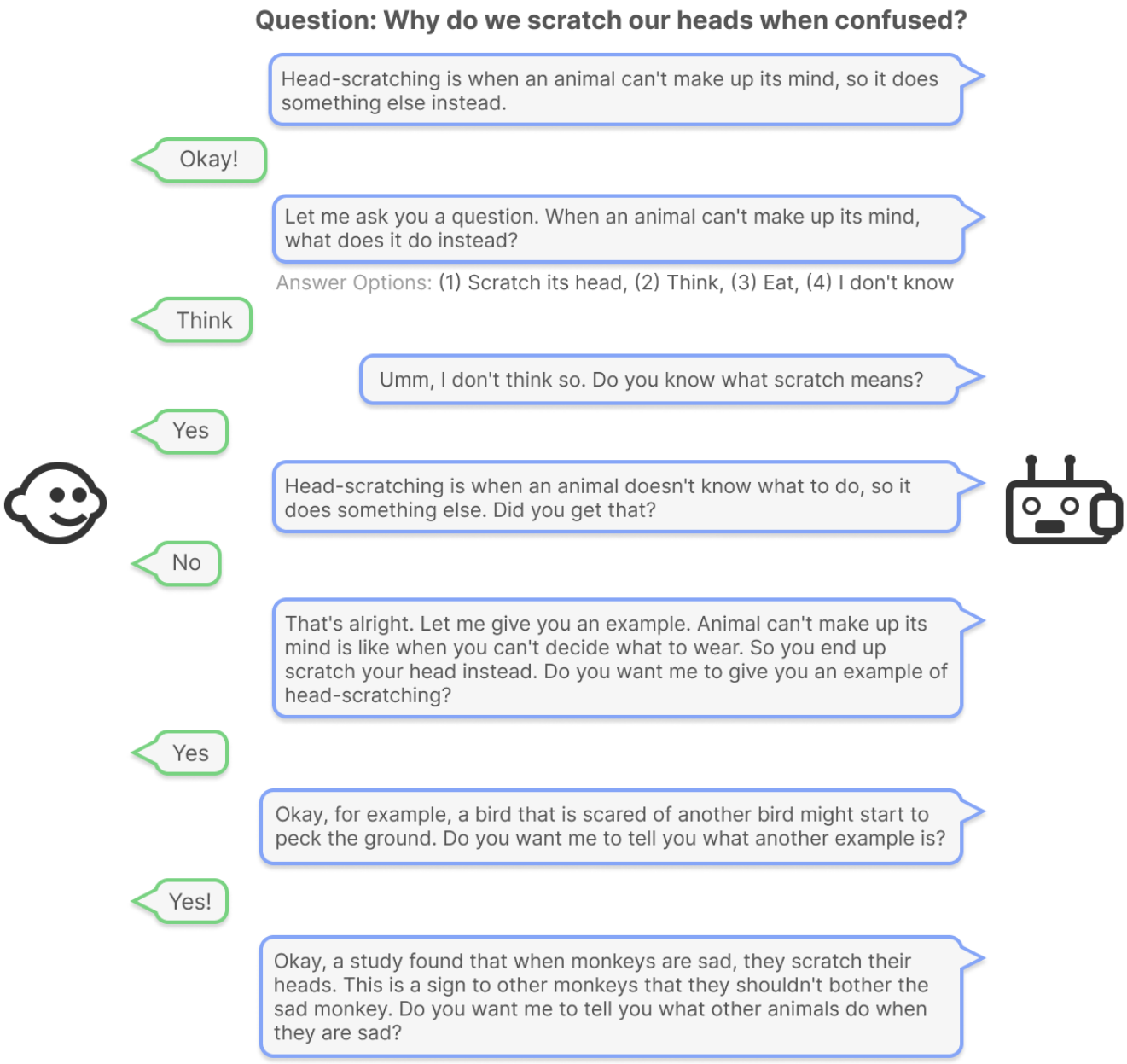

Figure 11: Example thread from the dialogue tree generated for the answer to the question "Why do we scratch our heads when confused?"