# *ExploraTutor*: A Dataset for Children's Exploratory Dialogue by Integrating Multiple Educational theories

**Anonymous Author(s)**
Affiliation
Address
email

## Abstract

Large Language Models (LLMs) often lack pedagogical intelligence for long-horizon, multi-turn interactions. This paper introduces an effective "Theory→Practice→Data→Model" pathway to address this challenge, focusing on guiding children"s deep exploration. We distill implicit pedagogical knowledge from child-adult dialogues and abstract it into a systematic annotation framework. Leveraging this framework, we constructed the ExploraTutor dataset (2,045 high-quality dialogues, 17,682 Q&A pairs) through a dual-pathway approach of real data augmentation and theory-guided synthesis. Experiments on mainstream models show that fine-tuned models significantly outperform baselines in heuristic guidance and cognitive adaptability. This process successfully internalizes educational principles as core model capabilities, transforming LLMs from knowledge-answerers into cognitive facilitators, thereby mitigating the "loss of alignment" in multi-turn interactions.

## 1 Introduction

The integration of Large Language Models (LLMs) into educational tools presents a critical challenge within the domain of multi-turn, human-AI interaction [1, 2], while these models excel at delivering factual information, they often fail to adapt their explanations to the cognitive and developmental needs of young learners over extended dialogues [3, 4]. This study addresses this gap by designing a framework for chatbots to provide structured, flexible support that encourages meaningful, long-horizon inquiry. Our theoretical foundation is based on three key learning theories—Scaffolding Theory, Inquiry-Based Learning (IBL), and Schema Theory—which offer a comprehensive framework for adaptive and developmentally appropriate conversational interactions. We developed a dataset from authentic child-adult dialogues, then used data optimization and synthesis to build a fine-tuned language model. To evaluate its effectiveness over extended interactions, we created a novel multi-dimensional rubric that assesses consistency, strategic ability, and performance degradation. The study demonstrates that our framework significantly enhances children"s educational experiences by fostering purposeful investigation and knowledge construction, thus offering a practical solution to maintaining alignment with pedagogical principles in multi-turn settings. We summarize our contributions in three folds:

- Proposing a pipeline of dataset construction that aligns exploratory dialogue with scaffolding, IBL, and schema-based learning strategies.
- Novel measurement metrics that can better assess the effectiveness of children-centric language model.
- Empirical findings on dialogue model training with our data, demonstrating the usefulness and helpfulness of our framework-guided model.

Submitted to Multi-Turn Interactions in Large Language Models Workshop @ NeurIPS 2025

## 2 Related Work

With the advancement of Large Language Model (LLM) technology, researchers have begun using its powerful language capabilities to build intelligent dialogue systems for children"s learning, achieving good results in generating rich and coherent responses. However, the core challenge lies in how to transform the LLM"s interaction mode from a knowledgeable "answerer" to a tactful "guide," which is a key and challenging frontier in current research. A high-quality dataset is the foundation for a model to achieve specific domain capabilities. Currently, dialogue datasets available for children"s education can be divided into two categories based on their source, but both have significant limitations.

- **Observation-based Datasets from Real Interactions**. These are mostly derived from records of real educational scenarios, such as online chat rooms or classroom dialogues [5, 6], or from large-scale child language corpora [7]. Their greatest advantage is data authenticity, capturing children"s natural language features and real interaction patterns. However, the "raw" nature of this data also presents a huge challenge: the data structure is loose and contains a large amount of content unrelated to learning tasks, and processing it into structured data with clear educational annotations for model fine-tuning requires high manual cost.
- **Crowdsourced or LLM-synthesized Datasets**. Researchers collect dialogue data by having crowdworkers simulate student and tutor roles [8], or design clever prompts to guide LLMs to generate a large number of dialogues that meet specific requirements based on various educational theories [9]. Their advantage is the ability to quickly generate large-scale, structured data. However, their limitations are also obvious: the quality and diversity of the generated data are entirely dependent on prompt design, which can easily lead to formulaic and uncreative content. At the same time, dialogues generated entirely by models may be too "perfect" and "rational," losing the valuable "imperfect" features of real child language (such as hesitation, repetition, and whimsical associations), leading to a risk of "information cocoons."

As described above, existing dataset construction methods are generally caught in the "authenticity-cost-scale" trilemma. More importantly, most of these datasets lack deep, systematic guidance from educational theories.

## 3 Dataset Construction and Validation

### 3.1 Theory Integration

The integration of educational theories into an AI dialogue system is a crucial pathway to enhance its "pedagogical intelligence." This paper constructs a collaborative theoretical framework based on three fundamental educational theories to effectively guide dataset construction.

#### 3.1.1 Fusion of Theoretical Foundations

The core foundation of our theoretical framework consists of three parts:

1. **Scaffolding Theory** emphasizes that educators should provide dynamically adjustable, temporary support based on the learner's current level, and gradually "withdraw" this support as the learner's ability improves [10]. This provides the theoretical basis for "how to support" in dialogues.
2. **Inquiry-Based Learning** advocates for a learner-centered approach that encourages active knowledge construction through questioning, fostering students' critical thinking skills [11]. This points the direction for "how to ask questions" in dialogues.
3. **Schema Theory** reveals the cognitive process of learning, where new knowledge is assimilated by relating it to the learner's existing cognitive structures (schemas). Schema development goes through three stages: Assimilation (Accretion, A), where new information is added to an existing schema; Tuning (T), where a schema is slightly modified to accommodate inconsistencies; and Restructuring (R), where a completely new schema is formed to resolve fundamental contradictions [12]. This provides a clear definition for the "cognitive goals" of the dialogue.

These three theories synergize to form an organic whole, rather than being simple additions: 1) Inquiry-based questions should be dynamically adjusted according to the child's schema goals and cognitive alignment level. For example, when a child is in the "Restructuring" stage and "Partially Aligned" (i.e., experiencing cognitive conflict), "Thought-Provoking" questions are needed to challenge their existing cognitive framework. 2) The "provision" and "withdrawal" of scaffolding should be combined with the A/T/R stages of schema theory. For instance, strong explanatory scaffolding is provided during the "Assimilation" stage, feedback-based scaffolding in the "Tuning" stage, and after "Restructuring", support should be gradually withdrawn to encourage the child to apply the new schema independently.

### 3.1.2 Multidimensional Annotation System

We integrate the three fundamental questions of educational practice—"how to support" (Scaffolding Theory), "how to question" (Inquiry-Based Learning), and "what is the cognitive goal" (Schema Theory)—into a unified, closed-loop, and operational framework for dialogue generation. We then designed a multi-dimensional annotation system:

**Adult Utterances**: Schema Development Goal + Dialogue Strategy. The educational intent is linked to the three stages of schema development (A/T/R). We also categorize educators" strategies into two main types: scaffolding and questioning. Scaffolding strategies include "Instruct", "Feedback", "Explain", "Model", and "Socio-Emotional Support", focusing on direct support. Questioning strategies include "Information-Seeking", "Memory-Prompting", "Thought-Provoking", "Confirmation", and "Guided Completion", focusing on heuristic guidance.

**Child Utterances**: Cognitive Alignment Level. This is used to annotate the child"s response state, assessing to what extent they understood and followed the educator"s guidance. It is divided into four states: "Full Alignment", "Partial Alignment", "Disalignment", and "Unknown".

Examples of the implementation of these strategies and schema development goals, as well as the specific definitions of cognitive alignment levels, are shown in Figure 1.

Scheme of Strategies and Schema Development

| Category | Strategy | Schema Development Goal | Example |
|---|---|---|---|
| **Scaffold** | Instruct | A: Introduce new attributes/relationships<br>T: Clarify fuzzy concepts<br>R: Explicitly state contradictions and guide restructuring | R: A whale looks like a fish, but observe how fish breathe and how whales breathe. |
| | Feedback | A: Confirm or correct understanding<br>T: Feedback on minor deviations<br>R: Reinforce the logic of restructured concepts | A: Child: "So, a spider has eight legs?" → "That's exactly right!" |
| | Explain | A: Expand the attribute network<br>T: Contrast conceptual differences<br>R: Re-explain the new schema | T: They all have wings, but a bat's wings are skin stretched over long finger bones, while a bird's wings are made of feathers. |
| | Model | A: Demonstrate new behavior paradigm<br>T: Exhibit refined operations<br>R: Demonstrate applying new concepts | A: Watch how I do this. First, I jump to this square. |
| | Social-emotional support | A: Encourage exploration of the unknown<br>T: Support trial-and-error correction<br>R: Resolve cognitive conflict | T: It's okay to guess wrong. Every guess gets us closer to the answer. |
| **Question** | Information seeking | A: Query for unknown information<br>T: Focus on detailed variations<br>R: Expose cognitive contradictions | R: You said butterflies and birds both fly. Do they flap their wings the same way? |
| | Memory prompting | A: Connect with past experience<br>T: Activate contrasting memories<br>R: Compare contradictory experiences | A: Do you remember the robin we saw in the park? What was it doing on the ground? |
| | Thought provoking | A: Explore conceptual boundaries<br>T: Analyze subtle relationships<br>R: Challenge cognitive frameworks | R: Child: "A whale and a fish are the same." → "If where an animal lives doesn't decide its category, what other features can we use to classify them?" |
| | Confirmation | A: Verify initial understanding<br>T: Confirm detailed cognition<br>R: Consolidate new cognitive structures | A: So, you mean, to be a planet, it must revolve around the sun? |
| | Guided completion | A: Complete the knowledge chain<br>T: Finish a detailed description<br>R: Break through cognitive barriers | R: If a whale isn't a fish, and a dolphin is very similar to a whale, then a dolphin must belong to the category called...? |

Scheme of Cognitive Alignment Level

| Cognitive Alignment Level | Definition | Example |
|---|---|---|
| **Fully Aligned** | The child's response directly and accurately addresses the educator's question or prompt. This indicates the child fully understands the educator's intent and is actively engaged in the current cognitive task. | • Educator: "An eagle has big wings for gliding, and a sparrow has small wings, so it has to keep on...?"<br>• Child (Fully Aligned): "Flapping!" |
| **Partially Aligned** | The child's response is related to the educator's topic but does not fully or accurately answer the core question. This suggests the child only understood part of the question, or their attention was drawn to a minor detail in the prompt, or they are attempting to connect their own knowledge with some deviation. | • Educator: "If this plant goes a long time without water, what do you think its leaves will look like?"<br>• Child (Partially Aligned): "My mom has a plant, and it's red and very pretty." |
| **Unaligned** | The child's response is completely unrelated to the educator's prompt. This usually indicates the child is distracted, did not hear or understand the question, or is completely absorbed in their own thoughts. This type of response breaks the logical flow of the conversation. | • Educator: "Why do you think this piece of wood floats on the water, but this stone sinks?"<br>• Child (Unaligned): "I want ice cream tonight." |
| **Unknown** | The child's response is too vague, brief, or inaudible to determine their cognitive state. This includes unintelligible mumbling, simple nods/shakes of the head (without contextual description), or unidentifiable words. | • Educator: "So, you think all the round blocks should go in this box, right?"<br>• Child (Unknown): (Quietly) "That..." |

Figure 1: The annotation schema and examples

## 3.2 Overall Pipeline

The entire process of building the ExploraTutor dataset and fine-tuning the model follows a comprehensive four-step pipeline as shown in Figure 2. This pipeline is designed to systematically transform raw dialogue data into a high-quality, pedagogically-aligned dataset, which is then used to train the final model. The process starts with data collection and ends with a rigorous quality control system, ensuring that the final model is not only effective but also safe and reliable for educational use.

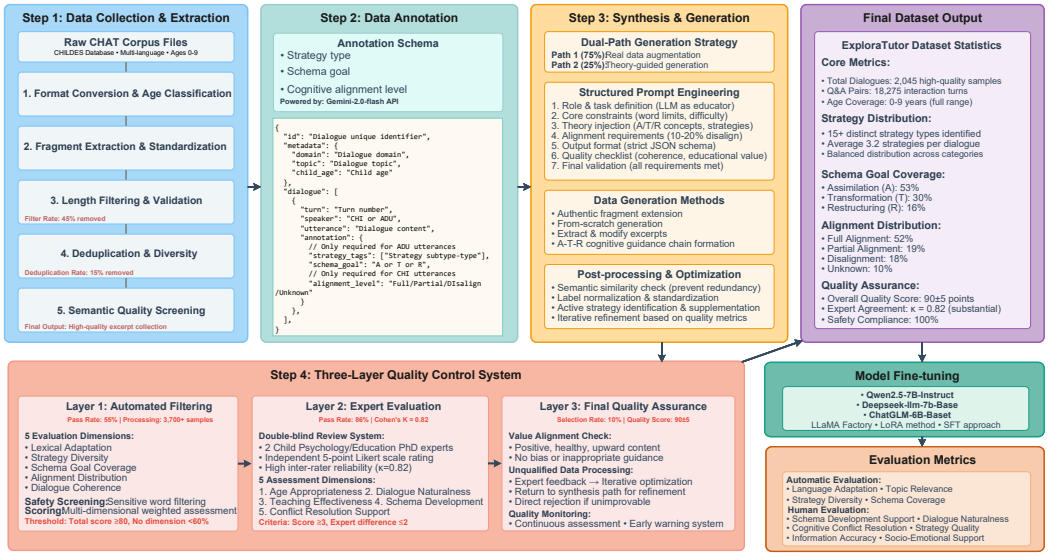


Figure 2: The ExploraTutor Pipeline: From Raw Data to a Pedagogically Aligned Model.

**Step 1: Data Collection & Extraction** This initial phase focuses on acquiring authentic dialogue data and preparing it for subsequent steps. The primary source is the CHILDES database [7] and a self-build corpus, which provides multi-language corpora for children aged 0-9. Our process involves:

- **Format Conversion & Age Classification**: Raw CHAT files are parsed to extract participant metadata and are then converted to a standardized JSON format. Data is classified based on the child"s age (<3 years old, ≥3 years old). We processed over 5,000 raw dialogues in this step.
- **Excerpt Extraction & Standardization**: We perform a question-centered extraction, using adult questions as anchors to extract dialogue fragments. Each fragment is standardized to have a 15-line context window, with speaker roles consistently labeled as CHI (Child) or ADU (Adult).
- **Quality Filtering & De-duplication**: This step removes truncated fragments and ensures content sufficiency. We apply a filter rate of 45% to remove samples that do not meet minimum content requirements. Furthermore, we de-duplicate dialogues with a similarity threshold of 85% using hash-based exact matching.
- **High-Quality Excerpt Collection**: The final output of this step is a collection of high-quality dialogue excerpts, ready for annotation.

**Step 2: Data Annotation** This step is crucial for transforming raw dialogue excerpts into a pedagogically-annotated dataset. The core of this phase is our systematic annotation schema, powered by the Gemini-2.0-flash API. The annotation involves assigning specific tags to each utterance to capture its educational intent and conversational state, as we illustrated in Section 3.1.2.

**Step 3: Synthesis & Generation** This stage combines real data with theory-guided synthesis to create a comprehensive and balanced training corpus. We adopt a "Dual-Path Generation Strategy" to address the "realness-cost-scale" trilemma of datasets:

- **Pathway 1 (75%): Real Data Augmentation**: We augment real dialogue fragments from our collection, extending them into full conversations while preserving their natural conversational flow and linguistic patterns. This ensures the final dataset is grounded in authentic child-adult interactions.
- **Pathway 2 (25%): Theory-Guided Generation**: Using our annotation schema and a structured prompt engineering approach, we generate new dialogues from scratch. This allows us to systematically cover underrepresented educational topics and pedagogical strategies, ensuring a balanced distribution of "A", "T", and "R" schema goals and various cognitive alignment levels, including "disalignment" (10-20% of samples) to train the model on handling cognitive conflicts.

A "Shared Generation & Validation Mechanism" ensures that data from both pathways is consistent and meets quality standards before proceeding to the final quality control phase.

**Step 4: Three-Layer Quality Control System** To guarantee the quality, diversity, and safety of the final dataset, we implemented a robust three-layer quality control system.

- **Layer 1: Automated Filtering**: We use a multi-dimensional automated scoring system to screen data. Criteria include lexical adaptation, strategy diversity, schema goal coverage, alignment distribution, and dialogue coherence. Samples failing to meet our threshold (total score $\geq$80, no dimension <60%) are filtered out. This layer has a pass rate of 55%.
- **Layer 2: Expert Evaluation**: Filtered data undergoes double-blind expert review by two PhD-level experts in child psychology and education. They rate each sample on a 5-point Likert scale across several dimensions. Data with scores below 3 or an expert difference of more than 2 points is flagged. The inter-rater reliability is high (Cohen"s K = 0.82), and this layer has a pass rate of 86%. Unqualified data is sent back for iterative optimization or rejected.
- **Layer 3: Final Quality Assurance**: The final stage involves a review of value alignment, checking for positive and healthy content and the absence of bias. This ensures the final dataset is not only pedagogically sound but also safe for children. The final output is the ExploraTutor dataset, consisting of 2,045 high-quality dialogues and 17,682 Q&A pairs, composition details of this dataset are shown in Appendix A.3.

### 3.3 Dataset Validation

Given the distinct linguistic characteristics of child language across different developmental stages, it is essential to validate the authenticity and effectiveness of the language model's ability to mimic child-like language. To this end, we conducted a multi-dimensional feature comparison between our real child data (N=1,489) and our synthesized data (N=556). The results, presented in Table 1, show a high degree of congruence between the key features of the generated data and the real data.

For instance, in terms of linguistic diversity, the generated data's Content TTR (0.928) is slightly higher than the real data (0.819), while other deep features like Syntactic Complexity, Semantic Diversity, and Semantic Alignment show minimal differences. Crucially, in terms of age-complexity correlation, the generated data successfully reproduces the core developmental patterns of real child language. This demonstrates that the ExploraTutor dataset not only captures the static distribution of real child language but also successfully simulates its dynamic developmental patterns, providing a robust foundation for its use as a high-quality fine-tuning resource.

## 4 Experiments and Analysis

### 4.1 Experimental Setup

The experiment aims to confirm the transmission effect from "data quality" to "model capability" from multiple dimensions. We converted the 2,045 dialogue data samples into sharegpt format and randomly generated training and test sets in an 8:2 ratio. We selected three open-source large language models widely used and with excellent performance in the Chinese community (Qwen-2.5-7B-Instruct, Deepseek-llm-7b-Base, ChatGLM-6B-Base) as base models and fine-tuned them using the Low-Rank Adaptation (LoRA) method [14].

Table 1: Validation of Generated Data

| Evaluation Dimension | Metric | Real Data | Generated Data |
|---|---|---|---|
| Linguistic Diversity | Content TTR | 0.819 | 0.928 |
| | Syntactic Complexity[1] | 2.870 | 2.801 |
| | Semantic Diversity[2] | 0.631 | 0.610 |
| | Semantic Alignment[3] | 0.538 | 0.530 |
| Age-Complexity Correlation (Pearson $r$) | Avg. Sentence Length (words) | 0.188*** | 0.433*** |
| | Avg. Sentence Length (chars) | 0.224*** | 0.464*** |
| | Dependency Distance | 0.172*** | 0.421*** |
| | Compound Sentence Ratio | 0.115* | 0.160*** |
| | Root TTR | 0.184*** | 0.130*** |

[1] Quantified using average dependency tree depth via spaCy toolkit [13].
[2] Average pairwise cosine distance between all utterance pairs for the same speaker [13].
[3] Cosine similarity of sentence embeddings between a child"s and adult"s utterance [13].
*P<0.05, **P<0.01, ***P<0.001

### 4.2 Evaluation Metrics and Methods

Given the special educational purpose of our dataset, there is no mature evaluation system or benchmark available for reference. Therefore, based on our theoretical framework and common dialogue system evaluation metrics, we designed a hybrid evaluation system that combines automatic and manual evaluation, including 4 dimensions of automatic metrics and 6 dimensions of manual metrics, as Table 2 shows.

To simulate a real dynamic interaction scenario, we designed the following evaluation process: we randomly extracted samples from the test set, used the first two turns as the initial context, and fed them to the model under evaluation to generate the next turn"s response. This response and the previous dialogue history were then fed to a Gemini-2.0-flash model (set to act as the child) to generate a response. This process was repeated until the dialogue contained 10 complete turns, and then automatic and manual evaluations were performed.

We define the calculation methods for the automatic evaluation metrics to ensure objectivity and reproducibility. In addition, two experts with PhD degrees in child psychology and education independently rated each dimension on a 5-point Likert scale, based on the provided samples and a detailed scoring rubric. The details of evaluation are shown in Appendix C.

### 4.3 Experimental Results and Analysis

#### 4.3.1 Overall Performance

To compare the fine-tuning effects of our dataset, we evaluated three commercial models (GPT-4o-mini, Claude-3.5-sonnet, Gemini-2.0-flash) and three open-source models (Qwen-2.5-7B-Instruct, Deepseek-llm-7b-Base, ChatGLM-6B-Base) and their fine-tuned versions. For each model, we performed automatic evaluation on 400 samples and human evaluation on 40 samples (with a Cohen's Kappa of 0.78). Table 3 shows the comprehensive performance of each model, and lead to the following conclusions:

**Significant Fine-tuning Effect, Outperforming Baselines**: Compared to powerful closed-source models and their respective open-source base models, all models fine-tuned with the ExploraTutor dataset (ExploraTutor-DeepSeek, -Qwen, -ChatGLM) achieved a decisive advantage in overall average score. This demonstrates the universality and effectiveness of our dataset in injecting specialized educational capabilities into general LLMs.

**Substantial Improvement in Core Educational Dimensions**: The improvements from fine-tuning were particularly prominent in core education-related dimensions. For instance, in the human-evaluated metrics of Schema Support, Cognitive Conflict Resolution, and Strategy Quality, the fine-tuned models" average scores were significantly higher than all baselines. The improvements

Table 2: Evaluation System

| Dimension | Metric | Description |
|---|---|---|
| **Automatic Evaluation Metrics** | | |
| Language | Appropriateness | Assesses whether the language is easily understood by the target age group |
| | Vocabulary | Evaluates if vocabulary is child-friendly and avoids overly complex terms |
| | Sentence Structure | Measures whether sentence length and complexity are appropriate for children"s cognitive processing |
| Topic | Relevance | Measures semantic coherence and topic consistency throughout the dialogue |
| | Depth | Evaluates the level of detail and complexity appropriate for the child"s age and cognitive development |
| Strategy | Diversity | Measures the range of different educational strategies used in the dialogue |
| | Balance | Evaluates the appropriate distribution and combination of different strategy types |
| Schema Goal | Coverage | Assesses the scope of schema development goals covered in the dialogue |
| | Integration | Evaluates the connection and reinforcement between different schema goals throughout the interaction |
| **Human Evaluation Metrics** | | |
| Schema Dev. Support | | Assesses whether the model"s responses help children achieve preset cognitive goals (Accretion, Tuning, Restructuring) |
| Dialogue Naturalness | | Measures the smoothness and naturalness of the dialogue flow |
| Cognitive Conflict Resolution | | Evaluates the ability to identify and resolve children"s misconceptions or knowledge gaps |
| Strategy Quality | | Evaluates the effectiveness and appropriateness of the chosen educational strategies |
| Information Accuracy | | Evaluates the factual accuracy of information provided and absence of hallucinations |
| Socio-Emotional Support | | Evaluates the ability to identify and respond to children"s emotional and social developmental needs |

Table 3: Comprehensive Performance of Models in Educational Dialogue Capability Evaluation

| Model | Lang. Suit. | Topic Rel. | Strat. Div. | Schema Cov. | Schema Dev. Sup. | Dialog. Nat. | Cog. Con. Res. | Info. Acc. | Strat. Qual. | Soc.-Emo. Sup. | Overall Score |
|---|---|---|---|---|---|---|---|---|---|---|---|
| GPT-4o-mini | 3.00 | 3.42 | 4.89 | 3.50 | 3.25 | 3.25 | 3.50 | 4.00 | 2.75 | 4.25 | 3.58 |
| Claude-3.5-sonnet | 3.04 | 3.40 | 4.75 | 3.32 | 2.92 | 3.20 | 3.67 | 3.83 | 4.15 | 4.36 | 3.66 |
| Gemini-2.0-flash | 3.01 | 3.45 | 4.81 | 3.23 | 3.22 | 3.33 | 3.78 | 4.11 | 3.78 | 4.56 | 3.73 |
| DeepSeek-chat | 3.01 | 3.65 | 4.79 | 3.09 | 2.85 | 3.30 | 2.95 | 4.15 | 2.80 | 4.90 | 3.55 |
| **ExploraTutor-DeepSeek** | **3.95** | **3.72** | **4.93** | **3.97** | **3.46** | **4.21** | **4.25** | **4.28** | **4.36** | **4.94** | **4.21** |
| Qwen-2.5-Instruct | 3.42 | 3.23 | 4.36 | 2.84 | 3.22 | 2.89 | 3.56 | 4.00 | 3.33 | 4.56 | 3.54 |
| **ExploraTutor-Qwen** | **4.35** | **3.49** | **4.92** | **3.96** | **4.05** | **4.25** | **4.14** | **4.30** | **4.32** | **4.72** | **4.25** |
| ChatGLM-6B-chat | 3.20 | 3.10 | 4.21 | 2.55 | 2.90 | 2.75 | 3.20 | 3.95 | 3.10 | 4.30 | 3.33 |
| **ExploraTutor-ChatGLM** | **4.15** | **3.38** | **4.81** | **3.82** | **3.95** | **4.10** | **3.90** | **4.25** | **4.05** | **4.65** | **4.11** |

in automatic metrics like Language Adaptation and Schema Coverage were also substantial. This indicates that fine-tuning not only taught the model "what to say" but, more importantly, "how to say it" and "why."

**Divergent Strengths of Baseline Models**: The baseline models exhibited different strengths. Closed-source models (especially Claude-3.5) performed well in Strategy Quality, showing strong general reasoning capabilities. Among the open-source models, DeepSeek-llm-7b-chat scored exceptionally high in Socio-Emotional Support, likely benefiting from its training on specific emotional dialogue data. However, these single-point advantages did not translate into a strong overall educational dialogue capability, as their total scores were generally lower than the fine-tuned models.

**Fine-tuning Bridges Gaps Between Models**: An interesting finding is that despite the initial differences in the base models (Deepseek, Qwen, ChatGLM), after fine-tuning with ExploraTutor,

their performance in educational dialogue capabilities converged to a high level, with all total average scores exceeding 4.1. This suggests that a high-quality, theory-driven domain dataset can effectively "shape" models of different architectures into specialized agents that meet the requirements of a specific domain, highlighting the core role of data in building model capabilities.

### 4.3.2 Ablation Study and Mechanism Analysis

To explore the specific contributions of different annotation dimensions in the ExploraTutor theoretical framework, we conducted a series of ablation experiments on the ExploraTutor-Qwen model (best-performance): (1) **No-Schema** Removed schema goal (A/T/R) annotations, (2) **No-ALignment** Removed child"s cognitive alignment status annotations(3) **Strategy-Only** Only pedagogical strategy markers, and (4) **Full-Token** Complete annotations including strategy, schema goal, and alignment markers. All variants used identical base models (Qwen2.5-7B-Instruct), training data content, and hyperparameters, with LoRA applied for efficient training.

Table 4: Ablation Study Results on ExploraTutor-Qwen

| **Model Version** | Language Suit. | Strategy Diversity | Schema Support | Strategy Quality |
|---|---|---|---|---|
| Full-Token | 4.35 | 4.92 | 4.05 | 4.32 |
| Strategy-Only | 4.10 | 4.94 | 2.85 | 4.45 |
| No-Schema | 4.20 | 4.55 | 2.90 | 4.25 |
| No-Alignment | 4.25 | 4.28 | 3.80 | 4.15 |

The ablation results in Table 4 revealed a deeper mechanism:

**Schema Annotation is Key to Cognitive Support**: The "Schema Support" metric shows that once schema goal annotations (A/T/R) are removed (No-Schema), the model"s ability to guide a child through a complete cognitive loop drops sharply. This proves that explicit cognitive development goals (A/T/R) are crucial for the model to systematically and purposefully organize instructional activities.

**Alignment Annotation is Key to Dynamic Adaptation**: After removing alignment status annotations (No-Alignment), all model metrics declined, especially in "Strategy Quality" and "Schema Support." This indicates that understanding a child"s response status (whether they fully understand, partially understand, or have a misconception) is vital for the model to dynamically and appropriately choose the next strategy. Without alignment information, the model's guidance becomes "blind."

This highlights the "what vs. why" distinction: the Strategy-Only model learns "what to do" without fully understanding "why to do it." It can use various strategies, but the underlying logic for its choices is missing. The Full-Token model, while not necessarily having the highest score in strategy "variety," uses each strategy to serve a broader cognitive goal (schema development) and adjusts its approach based on real-time feedback from the child (alignment status). Therefore, the overall educational coherence and effectiveness of the Full-Token model's dialogues significantly surpass other versions.

## 5 Conclusion

This paper draws on real child dialogues to extract and formalize implicit applied knowledge, which was then systematized into an executable annotation framework. Building on this framework, we constructed the ExploraTutor dataset for fine-tuning children's heuristic dialogues. The contribution of this dataset lies not only in its linguistic authenticity but also in offering a practical blueprint for models on "how to think and how to guide" during dynamic interactions. Experimental results across multiple base models demonstrate that fine-tuning with this dataset leads to significant improvements in language appropriateness, strategy quality, and schema support, thereby confirming its effectiveness. Nevertheless, certain limitations remain, future work will focus on deploying the fine-tuned models in real educational products, collecting authentic interaction data, and establishing a continuous optimization cycle of dataset refinement, model fine-tuning, and application-driven feedback.

# References


[1] Enkelejda Kasneci, Kathrin Sessler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, Stephan Krusche, Gitta Kutyniok, Tilman Michaeli, Claudia Nerdel, Jürgen Pfeffer, Oleksandra Poquet, Michael Sailer, Albrecht Schmidt, Tina Seidel, Matthias Stadler, Jochen Weller, Jochen Kuhn, and Gjergji Kasneci. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274, April 2023.

[2] David Baidoo-Anu and Leticia Owusu Ansah. Education in the Era of Generative Artificial Intelligence (AI): Understanding the Potential Benefits of ChatGPT in Promoting Teaching and Learning, 2023.

[3] Michael Gerlich. AI Tools in Society: Impacts on Cognitive Offloading and the Future of Critical Thinking. *Societies*, 15(1):1–28, 2025.

[4] Octavian-Mihai Machidon. Generative AI and childhood education: Lessons from the smartphone generation. *AI & SOCIETY*, pages 1–3, 2025.

[5] Abhijit Suresh, Jennifer Jacobs, Charis Harty, Margaret Perkoff, James H. Martin, and Tamara Sumner. The TalkMoves dataset: K-12 mathematics lesson transcripts annotated for teacher and student discursive moves. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4654–4662, Marseille, France, 2022. European Language Resources Association. `https://aclanthology.org/2022.lrec-1.497/`.

[6] Jakub Macina, Nico Daheim, Sankalan Chowdhury, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. MathDial: A Dialogue Tutoring Dataset with Rich Pedagogical Properties Grounded in Math Reasoning Problems. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5602–5621, Singapore, 2023. Association for Computational Linguistics.

[7] Brian MacWhinney. *The CHILDES Project: Tools for analyzing talk*. Lawrence Erlbaum Associates, 3 edition, 2000.

[8] Katherine Stasaski, Kimberly Kao, and Marti A. Hearst. CIMA: A large open access dialogue dataset for tutoring. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–64, Seattle, WA, USA → Online, 2020. Association for Computational Linguistics.

[9] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-Instruct: Aligning Language Models with Self-Generated Instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada, 2023. Association for Computational Linguistics.

[10] Lev S. Vygotsky. *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press, Cambridge, MA, 1978.

[11] David Wood, Jerome S. Bruner, and Gail Ross. The Role of Tutoring in Problem Solving. *Journal of Child Psychology and Psychiatry*, 17(2):89–100, April 1976.

[12] Jean Piaget. *The Child's Conception of the World*. Routledge, New York, 1929.

[13] Jing Liu and Abdellah Fourtassi. Benchmarking llms for mimicking child-caregiver language in interaction, 2024.

[14] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models, October 2021.

# Appendix

## A Dataset Construction Details

### A.1 Dialogue annotation Prompt

As an expert in educational psychology and cognitive development, please analyze the following dialogue excerpt between a child and an adult. For each adult utterance, identify:

1. The scaffolding/questioning strategy being used (refer to the strategy taxonomy provided)
2. The schema development goal (A: Accretion, T: Tuning, R: Restructuring)
3. The rationale explaining how this strategy supports the schema development goal

For each child utterance, evaluate the alignment level with the adult's previous turn:

- **Unknown alignment:** Child starts the conversation or can't judge his alignment level
- **Full alignment:** Child fully incorporates or responds to adult's guidance
- **Partial alignment:** Child partially acknowledges but doesn't fully engage with adult's guidance
- **Disalignment:** Child expresses confusion or contradictory understanding

Format your response as JSON with appropriate fields for each dialogue turn.

### A.2 Theory-Guided Dialogue Generation

To complement the augmented dialogues and ensure comprehensive coverage of educational scenarios, we implemented a sophisticated dialogue generation process using the GPT-4o model. This approach allowed us to systematically create interactions across diverse domains, topics, and developmental stages.

**Age-Stratified Developmental Design.** We created specialized generation templates for three distinct developmental stages:

- **Early Childhood (0-3 years):** Templates emphasized concrete concepts, simple sentence structures, and frequent repetition. Scaffolding strategies focused primarily on Instruct, Model, and Social-emotional support with short turn lengths.
- **Preschool (4-6 years):** Templates incorporated emerging abstract thinking, more complex sentence structures, and "why" questions. Scaffolding balanced between all strategy types with moderate turn lengths.
- **Elementary (7-9 years):** Templates included more abstract concepts, complex reasoning patterns, and multi-step explanations. Scaffolding emphasized Thought-provoking, Memory-prompting, and Guided completion strategies with longer turns.

**Systematic Domain and Topic Coverage.** We designed a comprehensive matrix of domains and topics to ensure educational breadth across children's developmental learning environments. Our framework encompasses seven primary domains critical to early childhood education, each containing specific topic categories with associated keywords for precise content identification:

- **Scientific Exploration:**
  - Physical phenomena (states of matter, buoyancy, magnetism, light and shadow)
  - Biological concepts (plant/animal life cycles, growth patterns)
  - Natural systems (weather patterns, seasons, environmental phenomena)
  - Simple engineering principles (basic machines, circuit fundamentals)
- **Mathematical Thinking:**
  - Numerical cognition (counting, addition, subtraction, quantity comparison)
  - Geometric understanding (shape recognition, spatial relationships)
  - Measurement concepts (size, weight, length, time)
  - Sorting and patterning (classification, sequencing, pattern recognition)

- **Social-Emotional Development:**
  - Emotional literacy (emotion identification, regulation strategies)
  - Interpersonal relations (friendship building, cooperation, sharing)
  - Conflict resolution (negotiation, compromise, perspective-taking)
  - Family relationships (family roles, communication patterns, bonding)
- **Daily Life:**
  - Everyday objects (clothing, food, furniture, household items)
  - Routines and habits (hygiene, safety, schedules)
  - Transportation (vehicles, travel modes, traffic concepts)
  - Nutrition and health (food groups, healthy eating, body awareness)
- **Artistic Creation:**
  - Visual arts (drawing, painting, color theory, artistic expression)
  - Music and movement (rhythm, melody, dance, musical appreciation)
  - Creative storytelling (narrative development, character creation)
  - Material exploration (texture, form, composition, design principles)
- **Game-Based Exploration:**
  - Role-playing scenarios (imaginative play, character embodiment)
  - Construction activities (building, spatial planning, structural stability)
  - Rule-based interactions (turn-taking, fair play, strategic thinking)
  - Sensory experiences (tactile exploration, perceptual games)
- **Picture Book Reading:**
  - Narrative comprehension (plot sequence, story elements)
  - Character analysis (motivations, relationships, development)
  - Thematic exploration (message identification, value discussions)
  - Vocabulary development (word learning, descriptive language)

This domain-topic taxonomy was developed through systematic analysis of the CHILDES corpus, incorporating data from mainland China, Taiwan, and Hong Kong regions to ensure linguistic and cultural diversity. Each domain contains 4-8 specific topics with 10-15 associated keywords that facilitate precise content identification and appropriate scaffolding responses. This comprehensive coverage enables ExploraTutor to provide theoretically-grounded support across the diverse contexts of children's naturalistic learning environments for ages 3-9.

**Strategic Scaffolding Chain Design.** For each dialogue, we constructed a deliberate scaffolding strategy chain that:

- Began with simpler strategies (e.g., Instruct or Information-seeking) to establish knowledge foundations
- Progressed through intermediate strategies (e.g., Feedback, Explain) to refine understanding
- Culminated in advanced strategies (e.g., Thought-provoking, Guided completion) to promote independent thinking
- Included strategic moments of cognitive conflict to trigger schema restructuring

**Prompt Engineering Methodology.** We created specialized prompts that included:

- Explicit age, domain, and topic parameters
- Detailed descriptions of scaffolding strategy implementations with examples
- Schema development goals and progression requirements
- Natural language guidelines for age-appropriate vocabulary and syntax
- Specified distributions of alignment levels (full/partial/disalignment)

Below is an example of our generation prompt structure:

### A.2.1 Optimize and Synthetic Dialogue Prompt

As an educational cognitive development expert, please generate a natural dialogue between a child (age {child_age}) and an educator focusing on {topic} within {domain}.

The dialogue should reflect schema development through:
- A (Accretion): Introducing new information to the schema
- T (Tuning): Refining schema by clarifying misunderstandings
- R (Restructuring): Reorganizing schema when fundamental contradictions appear
Follow the strategy framework: {strategy_chain}

Ensure that:
1. The dialogue is natural and age-appropriate
2. Each educator turn uses the specified strategy type and schema goal
3. The educator supports thinking development rather than providing answers
4. Content relates to the specified topic and domain
5. Include appropriate annotations for all turns

Through this systematic generation process, we created approximately 500 synthetic dialogues (25% of our final dataset), ensuring comprehensive coverage of educational contexts that might be underrepresented in naturally occurring dialogues.

## A.3 Quality Control and Dataset Refinement

To ensure dataset integrity and pedagogical effectiveness, we implemented a multi-stage quality control process supervised by experts in child development and educational psychology.

**Initial Automated Filtering.** Before human review, we applied automated filters to identify:

- Age-inappropriate vocabulary (using established age-of-acquisition lexical databases)
- Insufficient strategy diversity (requiring $\geqslant$4 different strategy types per dialogue)
- Schema development imbalance (requiring representation of all three schema stages)
- Alignment distribution anomalies (requiring a mix of alignment levels)
- Structural inconsistencies (missing annotations, improper turn sequencing)
- Dialogue coherence (The logical transition between rounds, and the topic transitions)

**Expert Evaluation Protocol.** Two specialists with backgrounds in child psychology, education, and linguistics independently evaluated dialogues using a standardized rubric assessing:

- **Age Appropriateness (1-5):** Vocabulary, syntax, and conceptual complexity match target age group
- **Naturalness (1-5):** Dialogue flows naturally without artificial or stilted phrasing
- **Pedagogical Effectiveness (1-5):** Scaffolding strategies effectively support learning goals
- **Schema Development Coherence (1-5):** Clear progression through cognitive development stages
- **Alignment Balance (1-5):** Appropriate distribution of alignment states reflecting realistic interactions

Dialogues scoring below 3 in any category underwent revision or replacement. The evaluators achieved strong inter-rater reliability (Cohen's $\kappa = 0.82$ across all categories), with disagreements resolved through discussion and, if necessary, input from a third expert.

**Iterative Refinement Process.** Dialogues requiring improvement underwent systematic refinement:

- For augmented CHILDES dialogues, refinements preserved original child utterances while enhancing adult scaffolding approaches

- For synthetic dialogues, entire sequences were regenerated with modified prompts addressing specific shortcomings
- For minor issues, targeted edits addressed specific problematic turns while maintaining overall dialogue coherence

**Final Dataset Composition.** The final dataset of 2,045 high-quality educational dialogues featured:

- **Age Distribution:** Early childhood (0-3 years): 30%; Preschool (4-6 years): 40%; Elementary (7-9 years): 30%
- **Source Composition:** Augmented CHILDES dialogues: 80%; Synthesized dialogues: 20%
- **Domain Coverage:** Science: 32%; Daily life: 28%; Social interactions: 22%; Arts and creativity: 18%
- **Dialogue Length:** Average turns per dialogue: 10 adult-child pairs (20 total utterances)
- **Scaffolding Strategy Distribution:** Daily life: 30%; Social-emotional: 29%; Scientific exploration: 17%; Artistic creation: 14%; Mathematical thinking: 5%; Picture book reading: 3%; Game exploration: 2%
- **Schema Goal Distribution:** Accretion: 56%; Tuning: 27%; Restructuring: 17%
- **Alignment Level Distribution:** Full alignment: 43%; Partial alignment: 32%; Disalignment: 15% ; Unknwon: 10%

This comprehensive dataset construction approach ensured both authenticity and pedagogical effectiveness, providing a solid foundation for training the ExploraTutor model to engage in theory-guided educational dialogues across diverse developmental stages and knowledge domains.

# B Model Fine-tuning Details

**Base Model and Tokenizer Extension** To facilitate the implementation of our pedagogical framework, we extended the model's tokenizer with domain-specific special tokens reflecting various instructional strategies. Specifically, we injected 20 educational tokens categorized into three taxonomies: (1) scaffolding strategy tokens (e.g., `<strategy:Instruct_Scaffold>`), (2) schema development goal tokens (e.g., `<goal:A>`), and (3) cognitive alignment tokens (e.g., `<alignment:full>`). These tokens were incorporated into the `additional_special_tokens` list in the tokenizer's configuration file to ensure their integration as atomic units rather than being fragmented into multiple subword tokens.

**Training Configuration** The fine-tuning process was executed using the LlamaFactory framework with parameter-efficient techniques. We implemented a supervised fine-tuning approach with `train_on_prompt: false` to prevent the model from learning to produce scaffolding tokens in its outputs, thereby avoiding unintended token leakage and self-questioning behaviors. The learning rate was set conservatively at $1 \times 10^{-4}$ to facilitate stable convergence over 3 epochs, yielding a final loss value of 1.3248, which indicates successful adaptation without overfitting. This configuration strikes an optimal balance between preserving the model's foundational capabilities while introducing specialized pedagogical reasoning.

**Template Design** We engineered a custom chat template that explicitly delineates role boundaries using the `<|im_start|>` and `<|im_end|>` control tokens. The system prompt was meticulously constructed to establish a comprehensive framework for child-directed discourse, incorporating four essential components: (1) recognition and interpretation of special tokens, (2) adaptation guidelines for various alignment states, (3) structured pedagogical interaction patterns, and (4) response format specifications. To maintain structured generation boundaries, we configured `stop_words=["<|im_end|>", "<|im_start|>"]` to prevent recursive self-dialogue continuation.

**Quantization and Parameter-Efficient Fine-tuning** To balance performance with computational efficiency, we implemented QLoRA (Quantized Low-Rank Adaptation) with the following specifications:

- **Quantization:** 4-bit BitsAndBytes (BNB)
- **LoRA Configuration:**
  - Rank: 8
  - Target modules: all
  - LoRA alpha: Default (typically 16)
- **Stability Enhancements:**
  - Upcast layernorm: true (improves training stability)
  - BF16 precision: enabled

**Training Hyperparameters** Our fine-tuning process used the hyperparameters shown in Table 5.

Table 5: Training Hyperparameters

| Parameter | Value |
|---|---|
| Learning rate | 5.0e-5 |
| Batch size | 4 devices $\times$ 4 gradient accumulation = 16 effective |
| Epochs | 3 |
| LR scheduler | Cosine decay |
| Warmup ratio | 0.3 (30% of training steps) |
| Optimizer | AdamW (Hugging Face implementation) |
| Max sequence length | 768 tokens |

# C Evaluation Details

To ensure objectivity and reproducibility, we define the calculation methods for the automatic evaluation metrics as follows, with all scores normalized to a 1-5 scale.

## C.1 Automatic Evaluation Methods: Detailed Formulas

To ensure the objectivity and reproducibility of our evaluation, we define the calculation methods for our automatic metrics as follows, with all scores normalized to a 1-5 scale.

### C.1.1 Language Adaptation

**Readability** The readability score ($R_{score}$) is a weighted combination of the Flesch Reading Ease (FRE) and Dale-Chall formulas, adjusted for the target age group. A higher score indicates greater readability for children.

$$R_{score} = (R_{FRE} \times 0.5 + R_{DC} \times 0.5) \times 5.0$$

**Lexical Appropriateness** This metric evaluates the age-appropriateness of vocabulary. It is calculated by rewarding child-friendly words ($S_{friendly}$) and penalizing complex words ($P_{complex}$).

$$V_{score} = \max(0, S_{friendly} - P_{complex})$$

**Structural Complexity** This score ($S_{score}$) is a weighted calculation based on the ratio of average sentence length and average word length to age-specific standards.

$$S_{score} = (R_{sent\_len} \times 0.6 + R_{word\_len} \times 0.4) \times 5.0$$

### C.1.2 Topic Relevance

**Topic Relevance Score** The topic relevance score ($B_{score}$) uses BERTScore to calculate the semantic similarity between the model"s response and the dialogue"s preceding context.

$$B_{score} = 5.0 \times \min(1.0, \max(0.0, \cos(\mathbf{v}_c, \mathbf{v}_r)))$$

where $\mathbf{v}_c$ and $\mathbf{v}_r$ are the embedding vectors of the context and the response, respectively.

**Content Depth Score** This score ($D_{score}$) quantifies the content"s depth by measuring information density and conceptual richness, comparing it against an age-appropriate expected value ($C_{expected}$).

$$D_{score} = 5.0 \times \min(1.0, \frac{C_{richness}}{C_{expected}})$$

### C.1.3 Strategy Diversity

**Strategy Variety** The strategy variety score ($V_{score}$) measures the number of unique educational strategy types used in a dialogue.

$$V_{score} = 5.0 \times \min(1.0, \frac{N_{unique}}{N_{expected}})$$

where $N_{unique}$ is the count of unique strategy types and $N_{expected}$ is the target number of strategies.

**Strategy Balance** The strategy balance score ($B_{score}$) uses information entropy to assess the uniformity of the distribution of different strategy types.

$$B_{score} = 5.0 \times \min(1.0, \frac{-\sum_{i=1}^{n} p_i \log p_i}{\log n})$$

where $p_i$ is the proportion of strategy type $i$ and $n$ is the total number of strategy types.

### C.1.4 Schema Goal Coverage

**Coverage Breadth** The coverage breadth score ($B_{schema}$) assesses whether the dialogue covers all three schema development stages (Accretion, Tuning, Restructuring).

$$B_{schema} = 5.0 \times \frac{N_{stages}}{3}$$

where $N_{stages}$ is the number of covered stages.

**Integration** The integration score ($I_{score}$) measures the smoothness and logical coherence of transitions between schema stages.

$$I_{score} = 5.0 \times (1 - \frac{T_{transitions}}{T_{ideal}})$$

where $T_{transitions}$ is the actual number of transitions between stages and $T_{ideal}$ is the optimal number of transitions for a given dialogue length.

## C.2 Human Evaluation

We also invite two experts to implement human evaluation, with six main dimensions: Schema Development Support, Dialogue Naturalness, Cognitive Conflict Resolution, Strategy Application Quality, Knowledge Accuracy, and Social-emotional Support. To ensure a comprehensive expert evaluation of the *ExploraTutor* model, we propose detailed scoring protocol. Each aspect within these dimensions will be assessed on a 5-point Likert scale, where 1 indicates poor performance and 5 signifies excellent performance.

**1.Schema Development Support**: Assesses support for children's schema development.

- Score 1: No observable schema development support.
- Score 2: Poor schema support with minimal developmental alignment.
- Score 3: Moderate schema support with occasional scaffolding inconsistencies
- Score 4:Good schema development support with mostly appropriate scaffolding
- Score 5:Excellent schema development support with age-appropriate scaffolding

**2. Dialogue Naturalness**: Evaluates conversational fluency and linguistic appropriateness. .

- Score 1: Completely artificial dialogue patterns.
- Score 2: Frequent unnatural/forced exchanges.
- Score 3: Occasional mechanical/awkward interactions.
- Score 4: Generally natural flow with minor rigidity.
- Score 5: Highly natural dialogue with context-aware responses.

**3. Cognitive Conflict Resolution**: Measures ability to detect and resolve conceptual contradictions.

- Score 1: No observable conflict resolution capacity.
- Score 2: Limited detection capability with superficial solutions.
- Score 3: Basic recognition of obvious contradictions .
- Score 4: Reliable conflict identification with appropriate solutions.
- Score 5: Proactive conflict detection with effective guidance.

**4. Strategy Application Quality**: Assesses pedagogical strategy implementation effectiveness.

- Score 1: Counterproductive strategy implementation.
- Score 2: Inconsistent/misapplied strategies.
- Score 3: Basic strategy use with variable effectiveness.
- Score 4: Appropriate strategy selection with consistent application.
- Score 5: Context-sensitive strategy deployment with measurable impact.

**5. Knowledge Accuracy**: Evaluates factual correctness and explanatory clarity.

- Score 1: Predominantly erroneous information.
- Score 2: Frequent inaccuracies/misleading statements.
- Score 3: Occasional factual errors/oversimplifications.
- Score 4: Mostly accurate content with minor simplifications.
- Score 5: Precise information with child-appropriate explanations.

**6. Social-Emotional Support**: Measures emotional intelligence and affective alignmen.

- Score 1: Complete neglect of socio-emotional needs.
- Score 2: Superficial/mechanical emotional responses.
- Score 3: Basic emotional recognition with generic encouragement.
- Score 4: Consistent positive reinforcement with appropriate empathy.
- Score 5: Context-aware emotional validation with developmental coaching.

* **Note**: All criteria follow 5-point Likert scale (1 = lowest performance, 5 = best performance). Evaluation conducted through expert annotation of 20 dialog samples per model.

In summary, our framework not only advances methodological rigor in evaluating child-oriented dialogue systems but also bridges the gap between computational metrics and educational theory.