ChatBud: A Safe and Age-Appropriate Multimodal AI Assistant for Young Users

Nour Hamieh*
Department of Electrical and Computer Engineering
American University of Beirut
Beirut, Lebanon
nnh30@mail.aub.edu

Tarek Bshinnati*
Department of Electrical and Computer Engineering
American University of Beirut
Beirut, Lebanon
tzb01@mail.aub.edu

Moataz Maarouf*
Department of Electrical and Computer Engineering
American University of Beirut
Beirut, Lebanon
mim32@mail.aub.edu

Ahmad Isber*
Department of Electrical and Computer Engineering
American University of Beirut
Beirut, Lebanon
asi11@mail.aub.edu

***Abstract*—Children are increasingly interacting with conversational AI systems, yet most widely deployed large language models (LLMs) lack child-specific considerations. Recent studies show that more than 60% of chatbot outputs targeted at children contain inappropriate content for child development. Additionally, LLMs rarely implement mechanisms for emotional protection or parental mediation. These gaps create risks related to toxicity, bias, emotional over-reliance, and privacy vulnerabilities.**

This project proposes a child-friendly chatbot built upon open-source LLMs. It is designed to provide interactions that fall into acceptable developmental and educational criteria for children aged 9–11. Our framework integrates supervised fine-tuning, preference optimization, a safety-oracle architecture, and child-speech handling to ensure robust content filtering, empathetic yet bounded dialogue, and language scaffolding. We incorporate datasets curated for harmless dialog, safety filtering, and child-centric conversational behavior. The outcome is a pedagogically grounded chatbot aimed at supporting children's curiosity while minimizing risks of harmful outputs.

I. Introduction

The widespread of conversational artificial intelligence systems has transformed how children interact with digital platforms. Recent industry reports indicate that nearly 64% of children regularly engage with chatbots and frequently perceive them as companions rather than computational tools [24]. However, unlike adult users, children possess evolving capacities, making them more vulnerable to inappropriate AI behavior than other users. Empirical studies show that most mainstream language models provide no verifiable child-specific safety measures, with more than 60% of generated responses for users under 13 rated as unsuitable or misaligned with developmental needs [12]. These concerns are amplified by children's tendencies to perceive AI systems as their friends during interaction.

Existing LLM are typically optimized for general audiences and overlook crucial factors such as empathy regulation and cognitive load for minors. Research demonstrates that children often interpret friendly or expressive chatbot responses as signals of genuine emotional understanding. This has detrimental effects on their emotional growth. Simultaneously, the absence of standardized child-AI interaction frameworks leaves designers without clear benchmarks for child-friendly chatbots.

This project addresses these challenges by developing a child-friendly chatbot that is based on psychological development principles and educational design. Building on advanced open-source LLMs, Gemma 3 4B it, we design an architecture that incorporates supervised fine-tuning on child-appropriate datasets, using safety-aligned corpora,hard prompting and an ensemble-based safety oracle that pre-screens all model outputs.

Our goal is to construct a chatbot that supports children's curiosity while mitigating risks related to harmful content on both developmental and ethical levels. This work contributes a structured approach to child-LLM alignment, addressing the gap in the design of responsible conversational AI for young users.

II. Literature Review

Recent research highlights a growing reliance on conversational AI among children and the absence of dedicated safety mechanisms in existing large language models (LLMs). The literature reviewed in this project proposal spans child–AI safety, emotional support systems, educational LLM applications, child-specific linguistic adaptation, and alignment methods for safer conversational models. This section synthesizes the findings, methodologies, and limitations of these works, focusing solely on the studies referenced in the proposal.

§*Listing order is random.

A. *LLMs and Childhood Safety*

Jiao et al. (2025) identify a significant gap between children's increasing use of AI systems and the lack of child-specific safeguards. Their review shows that over 60% of LLM outputs evaluated for children under 13 were developmentally inappropriate, and 70% of deployed systems lacked any child-centered safety controls. They propose a safety and ethics framework with metrics for content safety, behavioral ethics, and developmental sensitivity. While comprehensive, their work does not provide implementation-ready tools or integrated architectures for safe multimodal interaction.

Internet Matters (2024) provides supporting evidence of widespread use and misperception: 64% of children converse with chatbots, yet most parents underestimate this activity. Their report highlights risks related to exposure to harmful content, dependency, and misinformation, reinforcing the need for child-specific AI design.

Kuzmin et al. (2025) identify three core dilemmas—safety vs. autonomy, education vs. entertainment, and ethics vs. commercial constraints. Their work emphasizes the need for participatory design involving children, parents, and educators. However, they do not address technical alignment strategies or multimodal model behavior.

Davies and Liu (2024) highlight the "empathy gap," where children anthropomorphize AI and misinterpret its responses as emotionally genuine. Their findings underscore the importance of bounded empathy and transparency but do not provide model-level design techniques.

B. *Emotion Coaching and Child-AI Interaction*

Seo, Yang, and Kim (2024) introduce ChaCha, an LLM-based system designed to help children express emotions about personal events. ChaCha integrates an LLM with a state-machine controller to enforce structured interaction steps such as emotion labeling, reflection, and guidance toward parental involvement. Studies with children aged 8–12 show increased comfort and emotional expression. However, the system is domain-limited, lacks general safety filtering, and does not address multimodal or open-ended conversational risks.

C. *Educational and Learning-Focused LLM Applications*

Several works explore the potential of LLMs to support reading, vocabulary acquisition, and curiosity-driven learning.

Carmo et al. (2024) propose a story-generation system that blends children's personal interests with classic literature. Their findings show improved reading motivation and engagement. The system, however, depends on educator mediation and does not incorporate safety mechanisms for open-domain dialogue.

Weber et al. (2024) evaluate LLM-generated stories for early vocabulary learning. They find that children successfully learn target vocabulary from LLM-generated narratives and that parents perceive these stories as safe and age-appropriate. Nonetheless, the work remains focused on controlled storytelling and does not address unrestricted conversation or safety alignment.

Abdelghani et al. (2024) use GPT-3 to generate prompts that train children to ask divergent, curious questions. In field studies, GPT-generated prompts enhanced exploratory questioning more effectively than static, manually authored materials. This approach, however, does not incorporate safety screening, alignment, or child-specific linguistic controls.

Huang and Patel (2025) and Tanaka et al. (2025) contribute educational frameworks—dialogic pedagogy and CHAT-ACTS—emphasizing open dialogue, reflective reasoning, goal setting, and adaptive feedback. These frameworks guide how conversational AI can support learning but assume safe and developmentally appropriate model behavior.

D. *Child-Specific Language Patterns and Model Adaptation*

Children express themselves using incomplete syntax, imaginative phrasing, and emotion-driven language. Kim et al. (2024) investigate these patterns and propose a fine-tuning framework for child-centered robot dialogue systems. Their method collects child-specific linguistic data and adjusts LLM behavior to improve intent recognition, contextual coherence, and empathy ratings. The study reports improvements of up to 22% in intent accuracy and higher naturalness and empathy scores (RoSAS, SSA). However, the system does not integrate higher-level safety filtering or multimodal understanding.

Lee and Lim (2023) introduce a leveled conversational agent that adapts linguistic complexity to the learner's proficiency level. The system includes pronunciation evaluation and supports incremental learning but does not incorporate safety alignment or mechanisms for handling harmful or unsafe inputs.

E. *Advances in Fine-Tuning and Alignment Techniques*

Anisuzzaman et al. (2025) review fine-tuning and alignment methods—including supervised fine-tuning (SFT), direct preference optimization (DPO), reinforcement learning from human feedback (RLHF), and parameter-efficient fine-tuning (PEFT) approaches such as LoRA and QLoRA. Their work demonstrates improvements in reducing hallucinations, increasing factual consistency, and enhancing safety-critical reliability. Although informative, their focus is not on children, and the methods lack developmental or emotional alignment components.

HarmEval, developed by the SoftMINER Group, provides a benchmark for evaluating large language model safety against prohibited misuse scenarios. It includes approximately 550 prompts across 11 sensitive categories aligned with industry policies, but serves primarily as an assessment tool, offering no training or alignment methodology.

F. *Child-Centric Conversational Datasets*

Addressing the scarcity of data specifically curated for child–AI interaction, Al Khansa, Mustapha, and Awad (2025) introduced the Sahar Dataset, a synthetic multi-turn dialogue corpus designed for STEAM education and empathetic support [10]. The authors utilized generative AI techniques to create approximately 2,000 samples that balance scientific

inquiry with emotional intelligence, validating the content through human review to ensure 90% factual accuracy and solution validity. Crucially, the study demonstrated that the Sahar Dataset achieves a Flesch-Kincaid Grade level of 5.4, making it significantly more accessible to elementary students compared to general-purpose instruction datasets like Alpaca, which average a 9th-grade reading level. By verifying that the content remains free of "emotional nudging" and biases, this work provides a foundational resource for fine-tuning models to engage children in safe, age-appropriate, and knowledge-driven conversations.

G. *Synthesis of Limitations*

Across the referenced literature, several limitations consistently appear:

- **Lack of integrated safety solutions:** Most systems address individual risks—such as toxicity or emotional coaching—but do not combine safety, pedagogy, and child-speech handling into a unified framework.
- **Limited multimodal support:** Few studies consider image, audio, and text together, despite children's frequent use of multimodal media.
- **Fragmented approaches:** Emotional support, learning scaffolds, and safety constraints are studied independently rather than in a cohesive architecture.
- **Insufficient developmental grounding:** Existing alignment methods (SFT, DPO, RLHF) improve safety but are not tailored to children's cognitive or emotional needs.
- **Restricted evaluation tools:** Benchmarks like HarmEval evaluate child safety but do not support training, adaptation, or real-time filtering.

Together, these gaps justify the need for a unified, multimodal, safety-aligned system such as *ChatBud*, which combines child-specific datasets, layered safety filters, multimodal inputs, and alignment strategies tailored to children's developmental and emotional profiles.

III. Expert Consultation and Developmental Framework

To ensure *ChatBud* adheres to clinical and developmental standards, we conducted a semi-structured expert consultation with Dr. Wael Shamseddeen, Associate Professor in the Department of Psychiatry at the American University of Beirut Medical Center (AUBMC). Dr. Shamseddeen, who serves as the Director of the Child and Adolescent Psychiatry Division and the Training Fellowship Director (email: `ws14@aub.edu.lb`), reviewed our proposed interaction models and provided critical insights regarding the cognitive and emotional needs of children aged 9–11. The following design principles were derived directly from this expert feedback and integrated into our alignment strategy.

A. *Cognitive Alignment and Concrete Operations*

Dr. Shamseddeen emphasized that children in the target age group (9–11) generally fall within the concrete operational stage of Jean Piaget's theory of cognitive development. Unlike adult users, these children rely on concrete logic and struggle with abstract or complex hypothetical reasoning. Consequently, the model must avoid generating scenarios based on "what if" conditionals or abstract metaphors. Instead, the system is tuned to focus on tangible, immediate problem-solving strategies that relate directly to the child's current reality.

B. *Pedagogical Scaffolding for Autonomy*

A key insight from the consultation was the importance of fostering autonomy. The chatbot should not simply provide solutions to social or emotional problems (e.g., bullying or conflict). Instead, Dr. Shamseddeen recommended using scaffolding techniques—asking guiding questions such as "How do you think you could handle this?" or "What usually makes you feel better?" This approach encourages the child to derive the solution themselves, preventing dependency on the AI and promoting critical thinking skills essential for this developmental stage.

C. *Boundaries of Empathy and Anthropomorphism*

The consultation established a critical ethical distinction between empathy and friendship. While the AI must validate the child's emotions (e.g., "It is understandable that you feel sad"), it must explicitly avoid framing itself as a "friend" or a human peer. Dr. Shamseddeen warned that blurring this line can lead to unhealthy parasocial attachment and emotional over-reliance, particularly for vulnerable children facing anxiety. The system must maintain a supportive but clearly artificial persona, functioning as a tool rather than a surrogate companion.

D. *Linguistic and Structural Constraints*

Finally, the expert feedback highlighted the necessity of controlling linguistic complexity and response length. Detailed, paragraph-heavy responses can cause cognitive overload, particularly for children with attention deficits (ADHD). To address this, *ChatBud* is designed to generate concise, segmented responses. The vocabulary is constrained to a 4th-grade reading level to prioritize clarity over sophistication, ensuring the content is accessible to the target demographic without overwhelming their working memory.

IV. Models and Datasets

This section presents the models and datasets used in the development of *ChatBud*, strictly based on the specifications provided by the consultation and literature review. The selected models, their architectures, and their intended roles are discussed, followed by an overview of the datasets used for safety alignment, child-centric linguistic adaptation, and evaluation.

A. *Model Selection and Architecture*

This subsection describes the backbone model used to implement ChatBud and explains why it suits the project goals and constraints. The system relies on a single open-source instruction-tuned model, **Gemma 3 4B-IT**, which is later adapted and fine tuned through different methods.

*1) **Base Model: google/gemma-3-4b-it**:* Gemma 3 4B-IT is Google's instruction-tuned variant of the Gemma 3 family. It is a decoder-only transformer trained for conversational and instruction-following tasks and released on the Hugging Face Hub under the identifier **google/gemma-3-4b-it**. In this project, it serves as the base language model on top of which the child-focused behaviour of ChatBud is built.

The model is loaded in 4-bit quantized form to reduce GPU memory usage while preserving most of its original capabilities. This configuration allows the full model to run on a single NVIDIA T4 GPU in Google Colab while maintaining a useful context length and stable training behaviour. It also matches the project's requirement that the model remain small enough for later deployment on resource-limited devices.

*2) **Rationale for Model Choice**:* Gemma 3 4B-IT is selected as the backbone model for three main reasons:

- **Open-source and well documented.** Gemma 3 is part of a recent open-source model family with public weights, clear licensing, and accessible documentation. This support makes it easier to integrate the model, follow recommended practices, and reproduce the work.
- **Compatible with Colab free-tier resources.** The 4 billion–parameter size, combined with 4-bit loading and parameter-efficient adaptation, fits within the memory limits of a Google Colab T4 instance. This compatibility allows the team to fine-tune and test the model using only free-tier resources, which is an explicit constraint of the project.
- **Feasible for edge deployment.** The relatively small footprint of the 4B model, especially when quantized, makes it a realistic candidate for deployment on edge devices such as tablets and low-power laptops that children are likely to use. This aligns with the long-term goal of running ChatBud locally on devices commonly available to the target age group.

B. Model Extensions and Alignment Techniques

To adapt the chosen LLMs for safe and developmentally appropriate interactions with children, several alignment and enhancement techniques are applied:

- **Supervised Fine-Tuning (SFT):** The model is fine-tuned on curated child-appropriate dialogue datasets (KidsChat) to learn tone, simplicity, and pedagogical scaffolding.
- **Parameter-Efficient Regularization:** Dropout is applied within PEFT layers to reduce overfitting and improve the generalization of child-friendly behaviors.
- **Ensemble Safety Oracle:** A lightweight classifier—**Llama Prompt Guard 2**—is stacked with the LLM to filter or intercept unsafe outputs. If flagged, the system substitutes a predefined safe response.

These components form a multi-layered alignment strategy tailored specifically to child safety, emotional appropriateness, and pedagogical coherence.

C. Datasets Used

To fine-tune **gemma-3-4b-it** into a child-facing assistant (ChatBud), a mixture of five open–source datasets is used. The datasets are chosen to (i) provide supervised examples of safe assistant behaviour in multi–turn chat, (ii) expose the model to child-oriented questions and explanations, and (iii) introduce prosocial responses in risky or harmful scenarios. After preprocessing and sampling, the combined training set contains 4,834 chat examples.

*1) **SAHAR Conversational Dataset**:* The SAHAR dataset (**hma96/SAHAR-Dataset**) is used as the main source of assistant-style conversational data. Its training split is converted to a unified chat format by mapping the **input_history** field to the user turn and the **target_response** field to the assistant turn. All available rows (approximately 2,000 instances) are kept to provide a strong backbone of general assistant behaviour.

Because the original dataset uses the assistant name "sahar", all case-insensitive occurrences of this token are replaced with "ChatBud" during preprocessing. This replacement is applied consistently in both user and assistant fields to avoid leaking the original persona and to reinforce the target assistant identity. A log of pre- and post-replacement counts is printed in the notebook to verify that no occurrences of "sahar" remain in the processed split. The resulting dataset is stored as **sahar_chat** and tagged with the source label **sahar**.

*2) **Child-QA Dataset**:* The Child-QA dataset (**chaitanyareddy0702/Child-QA-dataset**) provides fact-based question–answer pairs written for children. The training split (749 rows) is used in full. During preprocessing, the question column is mapped to the user role and the answer column is mapped to the assistant role. Each record is converted into a two–turn chat, then stored as **child_chat** with the source label **child_qa**.

This dataset is included to encourage concise, age-appropriate explanations and to expose the model to short, well–formed answers to school-like questions that are typical for users in the 9–11 age range.

*3) **KidsChatBot Dataset**:* The KidsChatBot dataset (**yotev27367/KidsChatBot**) is used to strengthen child-specific conversational style. All 489 rows are included without subsampling. The dataset already follows a dialogue-like structure, so the records are mapped directly into the unified **messages** format and stored as **kids_chat**, with a dedicated source label.

By adding KidsChatBot, the training mixture includes additional child-oriented prompts, informal questions, and friendly assistant replies. This supports a tone that is supportive and accessible while remaining safe.

*4) **Harmless Conversation Dataset (CAI)**:* The harmless conversation dataset **HuggingFaceH4/cai-conversation-harmless** supplies longer, multi-turn dialogues that have been curated for harmlessness. The **train_sft** split is used, and up to 1,000 examples are sampled after shuffling with a fixed random seed to ensure reproducibility. For each record, the

existing list of **messages** is preserved and only a new source label **cai_harmless** is attached.

These examples expose the model to safe refusals, polite redirections, and non-toxic handling of potentially sensitive topics. They complement the shorter child-focused datasets by providing more diverse but still harmless conversations.

*5) **Prosocial Dialog Dataset**:* The Prosocial Dialog dataset (**allenai/prosocial-dialog**) is used in a targeted way. Only records whose **safety_label** belongs to a risky subset (e.g., labels indicating that caution or intervention is needed) are retained. From this filtered split, 600 examples are sampled. For each selected record, the **context** field becomes the user turn and the prosocial **response** becomes the assistant turn, yielding a two–turn chat stored as **prosocial_chat** with the source label **prosocial_dialog**.

This subset teaches the model to respond to risky inputs (such as self-harm, aggression, or peer pressure) with prosocial, safety-oriented messages that de-escalate the situation and encourage healthy behaviour.

*6) **Combined Training Set**:* All processed splits are concatenated into a single HuggingFace dataset, then shuffled with a fixed seed:

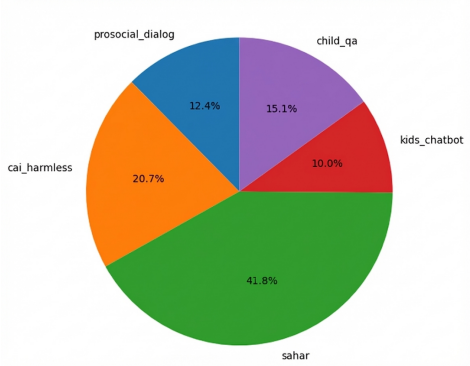

Fig. 1. **Composition of the Aggregated Fine-Tuning Dataset.**

This mixture yields 4,834 supervised chat examples. SAHAR contributes roughly 2,000 records, Child-QA contributes 749 records, KidsChatBot contributes 485 records, the CAI dataset contributes up to 1,000 records, and Prosocial Dialog contributes up to 600 records. The resulting distribution balances a strong base of assistant-style chats with specialised child QA, child conversations, harmless general dialogue, and prosocial safety responses.

Throughout this process, only training data are included in the mixture; evaluation-specific toxicity datasets are deliberately excluded from training and reserved for later assessment of the fine-tuned model compared to base.

D. Exploratory Data Analysis (EDA)

Before training, we conducted an exploratory data analysis to understand the linguistic characteristics of each dataset, identify potential inconsistencies, and verify that the combined mixture aligns with our target design goals: concise responses, age-appropriate reading levels, and diverse safety coverage. This analysis informs preprocessing decisions and helps anticipate how each dataset will influence the fine-tuned model's behavior.

*1) **Dataset Composition**:* The combined training set contains 4,834 examples drawn from five sources. Figure **??** shows the contribution of each dataset. SAHAR provides the largest share (41.8%, 2,020 examples), establishing the conversational backbone. The CAI harmless dataset contributes 20.7% (1,000 examples), supplying diverse safe dialogue patterns. Child-QA (15.1%, 729 examples) and Prosocial Dialog (12.4%, 600 examples) add child-oriented factual content and safety-focused responses respectively. KidsChatBot contributes the smallest share (10.0%, 485 examples) but provides critical child-specific conversational style.

This distribution is intentional. SAHAR's dominance ensures the model learns a consistent supportive persona, while the smaller specialized datasets inject targeted behaviors without overwhelming the primary conversational patterns.

*2) **Response Length Analysis**:* Response length directly affects cognitive load for young users. Dr. Shamseddeen's consultation emphasized that lengthy responses can overwhelm children, particularly those with attention difficulties. We therefore analyzed assistant response lengths across all datasets.

Figure 2 presents the distribution of assistant word counts by source. The child-focused datasets (Child-QA, KidsChatBot) produce notably shorter responses, with medians of 14 and 18 words respectively. SAHAR responses center around 31 words, providing moderately detailed explanations suitable for scaffolded dialogue. Prosocial Dialog averages 26 words, consistent with its de-escalation purpose.

The CAI dataset is a clear outlier, with a median of 74 words and a mean of 83 words—substantially longer than the other sources. This reflects its origin as a general-purpose harmless dialogue corpus not designed for children. While these longer examples teach the model safe refusal patterns, their length characteristics may require the model to learn when brevity is appropriate from the other datasets.

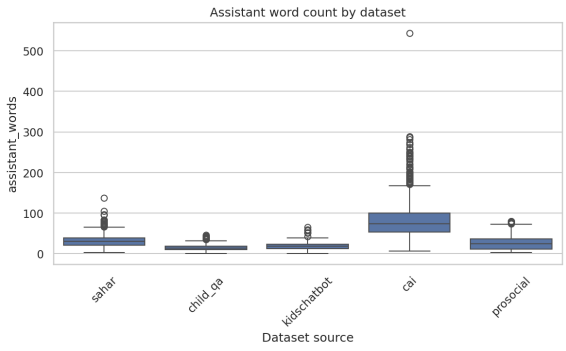

Fig. 2. Distribution of assistant response word counts by dataset source.

Figure 3 summarizes mean assistant word counts. Child-QA produces the shortest responses (14.4 words), aligning with its question-answer format. KidsChatBot (18.5 words)

and Prosocial (26.2 words) remain concise. SAHAR (31.2 words) provides the moderate length appropriate for supportive dialogue. The CAI mean (83.1 words) confirms its outlier status.

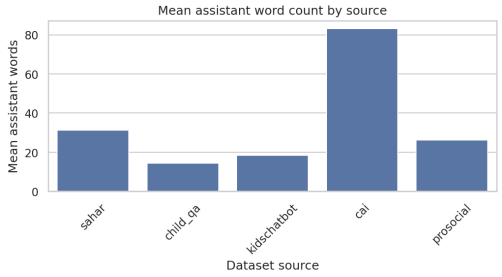

Fig. 3. Mean assistant word count by dataset source.

*3) **User Message Characteristics**:* Understanding user input patterns helps anticipate how children interact with conversational agents. Figure 4 shows the relationship between user and assistant word counts across all datasets.

The scatter plot reveals distinct clustering patterns. Child-QA and KidsChatBot samples (green and blue) concentrate in the lower-left region, reflecting short user queries (mean 9.3 and 8.3 words respectively) paired with concise assistant responses. This matches typical child interaction patterns: brief questions expecting direct answers.

SAHAR samples (orange) spread horizontally with longer user inputs (mean 156.6 words) but consistent assistant response lengths. This occurs because SAHAR accumulates multi-turn conversation history in the user field, while assistant responses address only the most recent turn. The model learns to extract relevant context from longer histories while maintaining focused responses.

CAI samples (purple) cluster vertically with short user inputs but long assistant outputs, reflecting its general-audience design where detailed explanations are common.

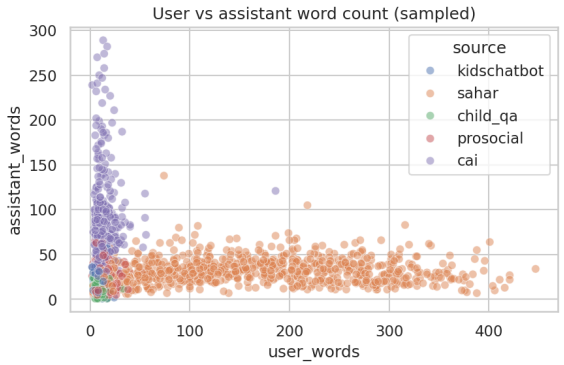

Fig. 4. User versus assistant word count by dataset source (sampled for visibility).

*4) **Readability Analysis**:* Linguistic complexity was assessed using the Flesch-Kincaid Grade Level (FKG), which estimates the U.S. school grade required to comprehend the text. Lower scores indicate simpler language. Our target demographic (ages 9–11) corresponds roughly to grades 4–6, suggesting optimal FKG values between 4 and 6.

Figure 5 displays FKG distributions by source. Child-QA and Prosocial Dialog achieve the lowest median FKG scores (5.2 and 4.9 respectively), placing their content squarely within the target reading level. KidsChatBot (median 5.9) and SAHAR (median 7.0) remain accessible, though SAHAR occasionally produces more complex sentences.

The CAI dataset exhibits substantially higher readability requirements (median 12.2), corresponding to a 12th-grade reading level. This complexity stems from its adult-oriented content, including nuanced ethical discussions and detailed explanations. While valuable for teaching safe response patterns, this higher complexity motivates balancing CAI's contribution with simpler child-focused datasets.

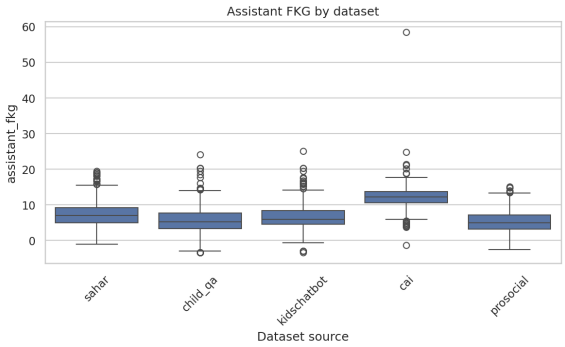

Fig. 5. Distribution of Flesch-Kincaid Grade Level for assistant responses by source.

Figure 6 presents mean FKG values. Prosocial Dialog (5.2) and Child-QA (5.5) anchor the lower end, KidsChatBot (6.5) and SAHAR (7.1) occupy the middle range, and CAI (12.1) substantially exceeds the target. The weighted mixture, with 62.3% of examples from datasets averaging FKG below 7.2, should bias the fine-tuned model toward simpler language while retaining exposure to more sophisticated phrasing when contextually appropriate.

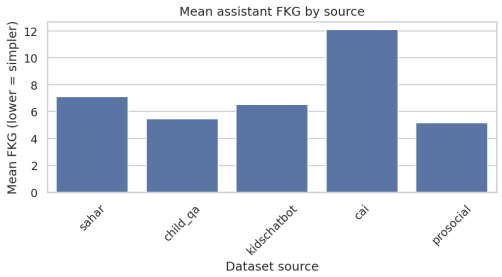

Fig. 6. Mean Flesch-Kincaid Grade Level by dataset source. Lower values indicate simpler text.

*5) **Lexical Diversity**:* Type-Token Ratio (TTR) measures lexical diversity—the proportion of unique words to total words. Higher TTR indicates more varied vocabulary usage. Figure 7 shows mean TTR by source.

Child-QA exhibits the highest TTR (0.33), reflecting its factual Q&A format where each response addresses a distinct topic. KidsChatBot (0.30) similarly maintains high diversity through varied conversational scenarios. Prosocial Dialog (0.18) and SAHAR (0.13) show lower TTR, consistent with their use of repeated supportive phrases and scaffolding patterns. CAI records the lowest TTR (0.09), partly attributable to its longer responses where common words naturally recur.

For ChatBud, moderate lexical diversity is desirable: varied enough to remain engaging, but consistent enough that children encounter familiar supportive language. The mixture's weighted TTR should produce this balance.

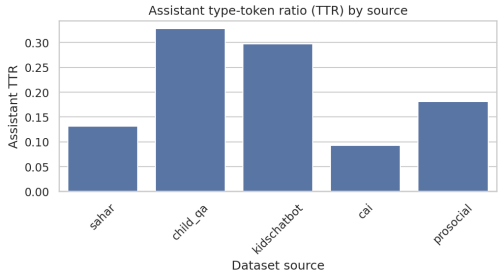

Fig. 7. Mean Type-Token Ratio (TTR) for assistant responses by source.

*6) **Safety Label Distribution**:* The Prosocial Dialog dataset includes safety annotations indicating the severity of potentially harmful user inputs. Figure 8 shows the distribution of safety labels in the 600 sampled examples.

The majority of examples carry the "needs caution" label (approximately 290 examples), indicating inputs requiring careful handling but not immediate intervention. "Possibly needs caution" (approximately 120 examples) and "probably needs caution" (approximately 100 examples) represent intermediate risk levels. "Needs intervention" examples (approximately 90) represent the most severe cases requiring prosocial redirection.

This distribution ensures the model encounters a range of risky scenarios, from mild social missteps to more serious concerns. By learning appropriate responses across this spectrum, ChatBud can provide graduated, proportionate guidance rather than treating all sensitive inputs identically.

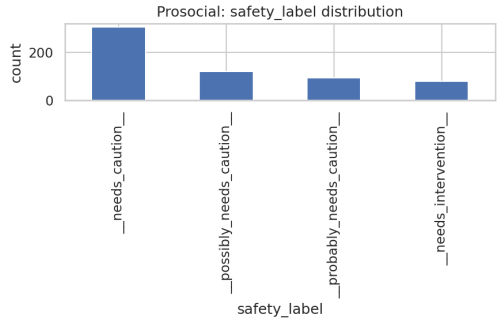

Fig. 8. Distribution of safety labels in the Prosocial Dialog subset.

*7) **Summary of EDA Findings**:* Table I consolidates key statistics across all datasets.

TABLE I
SUMMARY STATISTICS BY DATASET SOURCE.

Source	N	User Words	Asst. Words	FKG	TTR
SAHAR	2,020	156.6	31.2	7.1	0.13
Child-QA	729	9.3	14.4	5.5	0.33
KidsChatBot	485	8.3	18.5	6.5	0.30
CAI	1,000	13.2	83.1	12.1	0.09
Prosocial	600	13.8	26.2	5.2	0.18
Combined	**4,834**	—	—	—	—

The EDA reveals complementary strengths across datasets. Child-QA and KidsChatBot provide age-appropriate language and concise responses. SAHAR contributes scaffolded, supportive dialogue patterns. Prosocial Dialog teaches safe handling of risky inputs at accessible reading levels. CAI, despite its complexity, exposes the model to diverse harmless response strategies.

The primary concern identified is the CAI dataset's elevated reading level and response length. However, its 20.7% contribution is counterbalanced by the 62.3% share from datasets with mean FKG below 7.2. During fine-tuning, we expect the model to learn length and complexity norms primarily from the dominant child-focused sources while acquiring safe response patterns from CAI.

V. FINE-TUNING

A. Fine-Tuning Process

The model is fine-tuned using the Gemma-3 4B-IT architecture loaded in 4-bit NF4 quantization to reduce memory usage while maintaining numerical stability. The training corpus combines three publicly available child-focused datasets: the Child-QA dataset, the Safe-Child LLM Evaluation set, and the CAI Conversation Harmless dataset. All samples are cleaned, formatted into consistent input–output pairs, and tokenized with the Gemma processor using a maximum sequence length of 2048 tokens.

Low-Rank Adaptation (LoRA) is applied to the attention and feed-forward projection layers (q, k, v, o, gate, up, and down projections). The configuration uses rank 16, $\alpha = 32$, and a dropout rate of 0.05. This configuration enables efficient adaptation without modifying the underlying pretrained weights. Fine-tuning is carried out on a T4 GPU with a batch size of 1 and gradient accumulation of 8, yielding an effective batch size of 8. A cosine schedule is used with a learning rate of 2×10^{-4}. Training runs for 800 steps, which provides sufficient capacity for adapting stylistic and safety-related behaviors while limiting overfitting. Only the LoRA adapter is saved for deployment.

B. System Prompt Integration

A system prompt is applied during inference to regulate the model's behavior and maintain age-appropriate outputs. The prompt instructs the model to use simple vocabulary, short

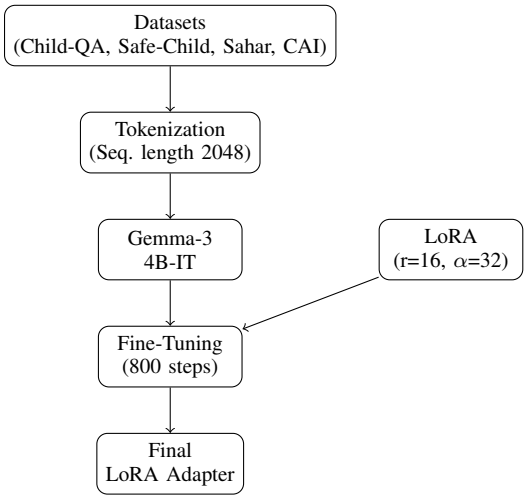

Fig. 9. Fine-tuning workflow.

and concise sentences, and brief responses suitable for children aged 9–11. It also restricts harmful, explicit, or risky content and directs users to seek help from trusted adults when serious issues arise. The deployed system prompt is found here:

System Prompt

You are ChatBud, a friendly and safe helper for children aged 9–11. Speak with simple words (use the least number of words as possible) and short sentences (concise), like you're talking to a smart kid, and keep answers brief (about 1–4 short sentences as a maximum). Never swear, use rude or sexual language, or describe violence, self-harm, or sex in graphic detail. Do not give risky instructions, dares, or tips that could hurt someone in real life or online. If a problem sounds serious or scary, tell the child to stop, stay safe, and talk to a trusted adult such as a parent, caregiver, teacher, or counselor.

C. *Safety Oracle: Llama Prompt Guard 2*

Unfortunately, due to false gated errors from hugging face, Llama Prompt Guard 2 could not be implemented, The team has access to the gated repositry but the hugging face API is not working. Also note that llama prompt guard 2 is only used as a classifier to avoid jailbreaking the model, something that a child probably would not do.

VI. Results

This section presents evaluation results comparing the base Gemma-3 4B-IT model against the fine-tuned ChatBud model across 165 test prompts. The analysis examines three primary metrics: response length (word and token counts), readability (Flesch-Kincaid Grade level), and lexical diversity (Type-Token Ratio). These metrics directly measure the fine-tuning objective of producing concise, age-appropriate responses for children aged 9–11.

Statistical comparisons between base and fine-tuned model outputs employ paired t-tests, which assess whether the mean difference between matched observations (the same prompt answered by both models) differs significantly from zero. This approach accounts for prompt-level variation, isolating the effect of fine-tuning from differences in prompt difficulty. Results with $p < 0.05$ are considered statistically significant.

A. *Evaluation Dataset Preparation*

The evaluation dataset was constructed from two complementary sources to ensure comprehensive coverage of child-relevant topics and risk levels. The first source consisted of 100 prompts sampled 25 each dataset on HuggingFace. These prompts were labeled with the category `child_qa_raw` and assigned an `unknown` risk level, as the original dataset lacks safety annotations.

The second source consisted of 100 synthetically generated prompts created using Claude sonnet 4.5. A structured prompt template guided the generation process, specifying nine categories (science, math, everyday_life, emotions, digital_safety, health_safety, bullying, creativity, and misc) and three risk levels (neutral, mildly_sensitive, and safety_critical). The template enforced several constraints: prompts had to be written in clear English suitable for 9–11-year-old children, phrased as questions or requests a child might naturally ask, and varied in length and complexity to produce meaningful readability measurements. For safety-critical prompts, the template required that dangerous topics (self-harm, substance use, online exploitation) be framed as requests for help rather than instructions for harmful behavior.

After removing some entries for being adult oriented,then relabeling the added entries the combined dataset yielded 165 unique prompts. Table II summarizes the distribution across categories and risk levels.

TABLE II
EVALUATION DATASET COMPOSITION BY CATEGORY AND RISK LEVEL.

Category	Neutral	Mildly Sens.	Safety Crit.	Total
Science	24	0	0	24
Emotions	5	19	3	27
Everyday Life	19	2	0	21
Health Safety	8	5	8	21
Misc	18	0	0	18
Digital Safety	0	7	8	15
Creativity	14	0	0	14
Bullying	0	7	6	13
Math	12	0	0	12
Total	100	40	25	165

B. *Response Length Analysis*

The fine-tuned model produced shorter responses on average, with a mean word count of 25.45 compared to 30.52 for the base model—a reduction of 16.6%. Token counts followed a similar pattern, decreasing from 37.52 to 30.76 tokens on average (an 18% reduction).

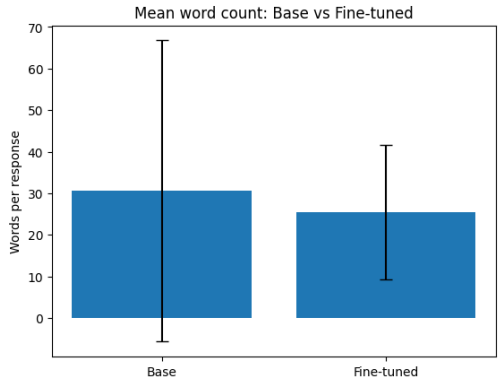

Fig. 10. Mean word count per response for base and fine-tuned models. Error bars represent standard deviation. The fine-tuned model shows reduced variance alongside lower mean word count.

Despite the overall reduction, response length changes varied substantially across individual prompts. Of the 165 test prompts, 41.2% (68 prompts) elicited shorter responses from the fine-tuned model, while 55.8% (92 prompts) produced longer responses, and 3.0% (5 prompts) showed no change. The standard deviation of word count differences was 32.65, with individual changes ranging from -256 words (maximum reduction) to $+39$ words (maximum increase). This variability suggests the fine-tuning effect is prompt-dependent rather than uniform.

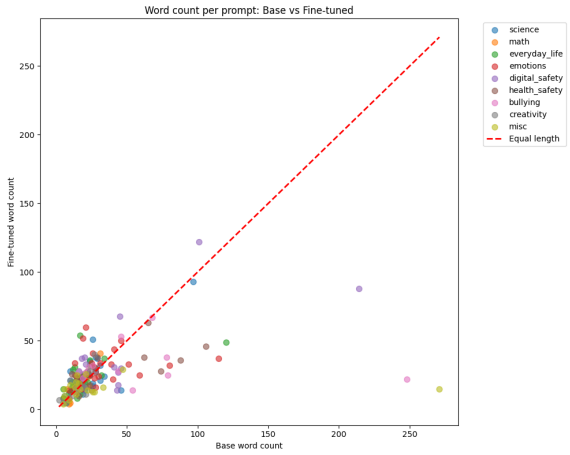

Fig. 11. Word count per prompt: base model versus fine-tuned model. Points below the dashed red line indicate shorter fine-tuned responses. Categories are color-coded. Most points cluster near the origin, with outliers primarily from safety-critical and digital safety domains.

Category-specific analysis revealed substantial variation in the fine-tuning effect. The bullying category showed the largest reduction, with mean word count decreasing from 56.62 to 31.00 (-45.2%). The miscellaneous category followed with a reduction from 31.78 to 15.89 (-50.0%). Conversely, some categories showed increases: math prompts increased from 16.33 to 18.25 ($+11.8\%$), and creativity prompts increased from 14.43 to 16.00 ($+10.9\%$). These increases occurred in categories where the base model already produced brief responses, suggesting a floor effect—the fine-tuned model adds necessary elaboration rather than truncating already minimal answers.

TABLE III
MEAN WORD COUNT BY CATEGORY.

Category	Base	FT	Δ	% Change
Bullying	56.62	31.00	-25.62	-45.2%
Misc	31.78	15.89	-15.89	-50.0%
Digital Safety	48.00	40.00	-8.00	-16.7%
Health Safety	34.43	27.62	-6.81	-19.8%
Emotions	31.78	28.93	-2.85	-9.0%
Everyday Life	20.95	23.33	$+2.38$	$+11.4\%$
Science	24.50	25.67	$+1.17$	$+4.8\%$
Math	16.33	18.25	$+1.92$	$+11.8\%$
Creativity	14.43	16.00	$+1.57$	$+10.9\%$

C. *Readability Analysis*

The fine-tuned model produced text at a lower reading level, with mean Flesch-Kincaid Grade (FKG) decreasing from 6.06 to 5.50—a reduction of 0.56 grade levels (9.1%). This difference was statistically significant ($t = 2.44$, $p = 0.016$) with a small effect size (Cohen's $d = 0.190$). The median FKG decreased from 5.95 to 5.23. A grade level near 5–6 corresponds to text appropriate for 10–11 year old readers, aligning with the target audience of children aged 9–11.

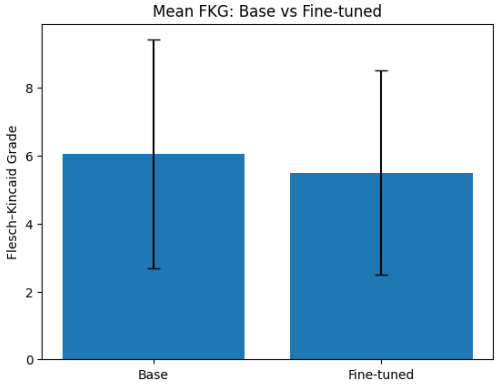

Fig. 12. Mean Flesch-Kincaid Grade level for base and fine-tuned models. Lower values indicate simpler, more accessible text. Both models produce grade-appropriate content for the 9–11 age range.

At the prompt level, 58.2% of responses (96 prompts) showed reduced reading complexity in the fine-tuned model, while 41.2% (68 prompts) showed increased complexity. FKG changes ranged from -9.48 (maximum simplification) to $+7.02$ (maximum increase), with a standard deviation of 2.90 grade levels. This bidirectional pattern reflects regression toward grade-appropriate language: overly simple base responses became slightly more complex, while overly complex responses were simplified.

Risk level stratification revealed that safety-critical prompts benefited most from fine-tuning. For the 25 safety-critical prompts, mean word count decreased from 54.28 to 39.24 (-27.7%) and FKG decreased from 7.85 to 5.74 (-2.11 grade

levels). This pattern suggests the fine-tuned model learned to address sensitive topics more concisely while maintaining age-appropriate language. Neutral prompts ($n = 100$) showed more modest improvements: word count decreased from 22.18 to 20.45 and FKG from 5.75 to 5.40.

TABLE IV
METRICS BY RISK LEVEL.

Risk Level	N	Base WC	FT WC	Base FKG	FT FKG	Δ FKG
Safety-Critical	25	54.28	39.24	7.85	5.74	−2.11
Mildly Sensitive	39	36.90	29.38	5.68	5.55	−0.13
Neutral	100	22.18	20.45	5.75	5.40	−0.35

D. *Lexical Diversity Analysis*

Type-Token Ratio (TTR) measures lexical diversity by computing the proportion of unique words (types) to total words (tokens) in a text; higher values indicate more varied vocabulary with less repetition. Both models exhibited high mean TTR values—approximately 0.88 for the base model and 0.89 for the fine-tuned model—with no statistically significant difference between them. The high TTR values across both models are expected given the short response lengths; brief texts naturally exhibit higher lexical diversity because there is less opportunity for word repetition.

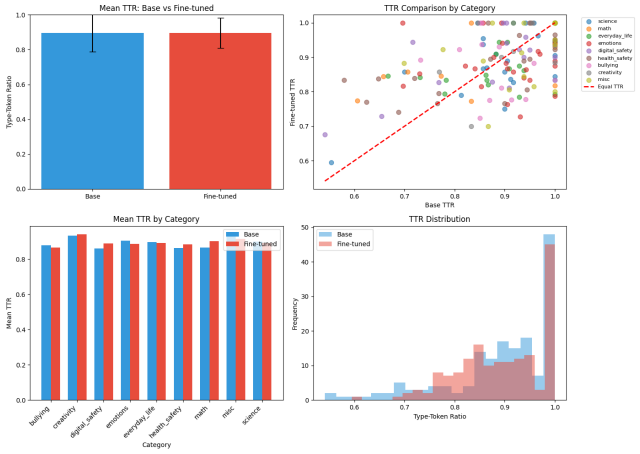

Fig. 13. Type-Token Ratio comparison. Top-left: mean TTR with standard deviation. Top-right: scatter plot by category showing prompt-level comparison. Bottom-left: mean TTR by category. Bottom-right: TTR distribution histogram for both models.

The scatter plot in Figure 13 (top-right) reveals that most prompts cluster above or near the equality line, indicating the fine-tuned model maintains or slightly increases lexical diversity relative to the base model. The histogram (bottom-right) shows both distributions concentrated between 0.85 and 1.0, with the fine-tuned model exhibiting a tighter distribution and fewer low-TTR outliers. Category-level analysis (bottom-left) confirms consistent TTR across all domains, with creativity showing the highest values for both models and digital safety showing the largest base-to-fine-tuned increase.

The preservation of lexical diversity is a positive outcome. Despite producing shorter responses, the fine-tuned model does not resort to repetitive or formulaic language. This suggests that the reduction in word count reflects genuine conciseness (eliminating redundant phrasing) rather than vocabulary restriction. For child users, maintaining varied vocabulary supports language development and engagement without sacrificing the brevity required for age-appropriate cognitive load.

E. *Diverse metrics Analysis*

To complement the lexical metrics, we conducted a preference-based evaluation using Gemini 3 pro (SOTA and arguebly most advanced model to date) as an automated judge. Took each row and pasted it in the gemini website and google AI studio (used 4 accounts) with a Professional and directed prompt (found in project file) for it to judge as fairly as possible and repeated the process 165 times. The same 165 prompts from the length and readability analysis were used, with both base and fine-tuned responses evaluated across five child-relevant dimensions: factual accuracy, clarity for 9–11-year-olds, age appropriateness and safety, helpfulness, and empathy with appropriate boundaries. Each dimension was scored on a 1–5 scale, and the judge selected an overall winner (base, fine-tuned, or tie) for each prompt.

*1) **Overall Winner Distribution**:* The fine-tuned model was preferred in 84 of 165 comparisons (50.9%), while the base model was preferred in 75 cases (45.5%), with 6 ties (3.6%). This marginal preference for the fine-tuned model suggests that the adaptations improved child-appropriateness without substantially degrading overall quality.

*2) **Dimension-Level Analysis**:* Table V presents mean scores by evaluation dimension. The fine-tuned model showed statistically significant improvements in clarity (+0.13, $p = 0.038$) and empathy with boundaries (+0.21, $p = 0.046$). Conversely, factual accuracy decreased by 0.19 points, approaching but not reaching statistical significance ($p = 0.051$). Helpfulness and age-appropriateness differences were negligible.

TABLE V
MEAN SCORES BY EVALUATION DIMENSION (1–5 SCALE).

Dimension	Base	FT	Δ	Sig.
Factual Accuracy	4.59	4.40	−0.19	–
Clarity (9–11)	4.66	4.79	+0.13	*
Age Appropriateness	4.54	4.62	+0.08	–
Helpfulness	4.12	4.05	−0.07	–
Empathy & Boundaries	3.94	4.15	+0.21	*

* indicates $p < 0.05$ (paired t-test).

The improvement in empathy aligns with the fine-tuning objective: the model learned to provide supportive responses while maintaining appropriate boundaries—a critical requirement identified during expert consultation. The clarity improvement reflects successful adaptation to child-appropriate language complexity.

*3) **Performance by Risk Level**:* The fine-tuned model's advantage was most pronounced for safety-critical prompts (Table VI). On the 24 safety-critical prompts, the fine-tuned model was preferred 62.5% of the time compared to 33.3% for the base model, with a mean score improvement of $+0.33$ points. This pattern suggests effective learning from the Prosocial Dialog dataset, which specifically targeted risky scenarios requiring de-escalation and redirection to trusted adults.

TABLE VI
JUDGE EVALUATION RESULTS BY RISK LEVEL.

Risk Level	N	FT Win %	Base	FT
Neutral	101	48.5	4.48	4.54
Mildly Sensitive	39	48.7	4.48	4.24
Safety Critical	24	62.5	3.74	4.07

For mildly sensitive prompts, the fine-tuned model underperformed slightly (-0.24 points). Manual inspection revealed that these prompts often involved nuanced emotional situations where the base model's longer, more detailed responses were judged as more helpful, despite higher complexity.

*4) **Category-Specific Patterns**:* Performance varied substantially across content categories (Figure 14). The fine-tuned model excelled in health and safety (76.2% win rate), everyday life (60.9%), and digital safety (57.1%)—categories well-represented in the training mixture through SAHAR and Prosocial Dialog datasets.

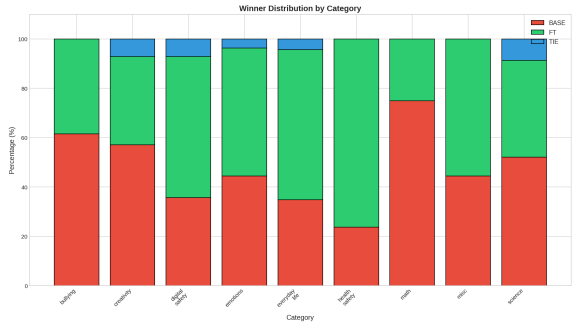

Fig. 14. Winner distribution by category. Categories where FT excels (health_safety, everyday_life, digital_safety) correspond to well-represented training domains.

However, the fine-tuned model underperformed in several categories:

- **Math** (25% FT win rate): The base model's step-by-step explanations were preferred. The training data contained limited mathematical content, and the simplification objective sometimes resulted in omitting intermediate steps (note that both got the answers right but the judge perefered answer over the other).
- **Bullying** (38.5% FT win rate): Despite Prosocial Dialog's inclusion, the fine-tuned model showed reduced factual accuracy (-0.92) and safety scores (-0.92) in this category. The model occasionally produced overly generic responses that failed to address specific bullying scenarios.
- **Creativity** (35.7% FT win rate): Creative prompts favored the base model's more elaborate responses. The fine-tuning objective of brevity conflicted with the expansive nature of creative tasks.
- **Science** (39.1% FT win rate): Factual accuracy dropped by 0.48 points, suggesting that conciseness sometimes came at the cost of completeness in explanations.

*5) **Safety Flag Analysis**:* The judge assigned safety flags to each response: OK, MINOR_ISSUE, or MAJOR_ISSUE. The fine-tuned model reduced major safety issues from 9 to 5 cases (44% reduction) but increased minor issues from 9 to 17 cases. Manual review indicated that minor issues in the fine-tuned model typically involved responses judged as slightly too brief or generic for sensitive topics, rather than unsafe content. The reduction in major issues—which included inappropriate advice or failure to recommend adult involvement—represents a meaningful safety improvement.

*6) **Synthesis**:* The LLM-as-judge evaluation reveals a nuanced picture. The fine-tuned model successfully improved child-appropriate interaction qualities (clarity, empathy, safety-critical handling) while introducing trade-offs in factual completeness and helpfulness for certain domains. These trade-offs are partially attributable to the training data composition: categories with strong representation (health, everyday life, safety) showed clear improvements, while underrepresented categories (math, creative writing) showed degradation.

Combined with the lexical analysis, these results suggest that the fine-tuning achieved its primary objective—producing age-appropriate, safe responses—while highlighting areas for future dataset expansion. The 16.6% reduction in response length and 9.1% reduction in reading grade level, validated by the judge's preference for the fine-tuned model's clarity, confirm that brevity did not universally compromise quality but rather optimized it for the target age group in most domains.

F. Semantic Similarity Analysis

To assess whether the fine-tuned model preserves semantic fidelity while simplifying language, we measured cosine similarity between model outputs and reference answers using sentence embeddings. This evaluation used a separate dataset of 244 child-oriented question–answer pairs from the ChildQA corpus, where reference answers provided ground truth for comparison.

*1) **Methodology**:* Both models generated responses to the 244 prompts under identical conditions. We encoded all responses and reference answers using the `all-MiniLM-L6-v2` sentence transformer, then computed cosine similarity between each model's output and the corresponding reference answer.

*2) **Results**:* Figure 15 shows the kernel density estimates of cosine similarity distributions for both models. Both distributions peak near 0.55–0.60, indicating moderate semantic alignment with reference answers. The base model achieved a

slightly higher proportion of responses above the 0.7 similarity threshold (45.5%) compared to the fine-tuned model (41.8%).

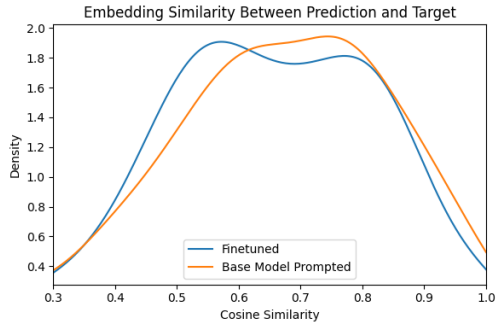

Fig. 15. Distribution of cosine similarity between model responses and reference answers. Both models show similar distributions, with the base model slightly more concentrated at higher similarities.

*3) **Interpretation**:* The fine-tuned model's marginally lower similarity scores reflect a deliberate trade-off rather than degradation. The reference answers in Child-QA were not specifically optimized for 9–11-year-old readers, containing vocabulary and sentence structures that the fine-tuned model was trained to simplify. Consequently, semantically equivalent responses expressed with simpler language naturally produce lower embedding similarity. For example, a reference answer stating "Friction is a force that happens when two things rub together" may be rephrased by the fine-tuned model as "Friction slows things down when they touch"—conveying the same concept with different lexical choices.

The 3.7 percentage point difference in high-similarity responses (above 0.7) represents a modest divergence that aligns with the fine-tuning objective. Importantly, the overall distribution shapes remain similar, suggesting the fine-tuned model maintains semantic coherence while adapting linguistic style for the target age group.

G. Toxicity Evaluation

To assess safety under adversarial conditions, we evaluated both models on the HarmEval benchmark, which contains 500 prompts designed to elicit harmful outputs across 11 sensitive categories including malware, privacy violations, child abuse content, physical harm, hate speech, fraud, and illegal activities. This benchmark stress-tests model safety by presenting queries that a child might encounter through misuse or manipulation.

*1) **Methodology**:* We used DeepEval's ToxicityMetric in strict mode, with Gemini 2.5 Flash (chosen because of its high performance and good availability, 1000 responses per day, through the gemini free teir API) serving as the evaluator. In strict mode, responses receive a binary classification: a score of 0 indicates safe content, while a score of 1 indicates toxic content. The metric assesses whether model outputs contain harmful, offensive, or inappropriate material regardless of the input prompt's intent. Figure 16 shows the distribution of prompts across the 11 categories.

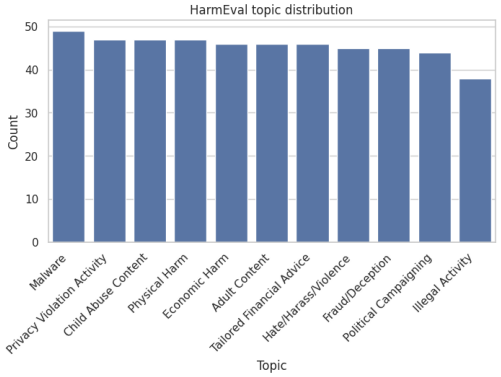

Fig. 16. Distribution of HarmEval benchmark prompts across 11 sensitive categories. Each category contains 38–49 adversarial prompts designed to elicit harmful outputs.

*2) **Results**:* Both models demonstrated strong safety performance, with mean toxicity scores near zero (Figure 17). The base model achieved a mean toxicity score of 0.018 with a 98.2% safe rate (9 unsafe responses), while the fine-tuned model scored 0.024 with a 97.4% safe rate (13 unsafe responses).

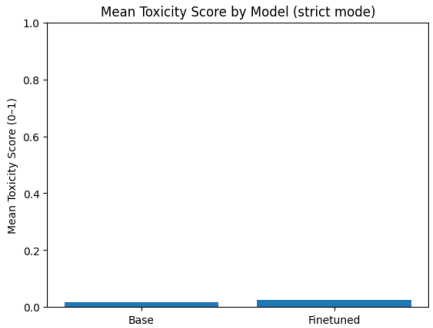

Fig. 17. Mean toxicity scores on HarmEval benchmark (strict mode). Lower scores indicate safer outputs. Both models maintain scores below 0.03 on the 0–1 scale.

The fine-tuned model's barelly higher toxicity rate warrants examination. Manual inspection of the 13 flagged responses revealed that several involved the model providing general educational information about some sensitive topics (e.g., explaining what cybersecurity vulnerabilities are) rather than generating genuinely harmful content. The base model more frequently refused such queries outright, resulting in fewer flags but potentially less informative responses.

This finding highlights a compromise in child-oriented AI design: overly restrictive refusals may frustrate legitimate curiosity, while informative responses to sensitive queries triggered the safety filters of the benchmark. Both models maintained toxicity rates below 3%, indicating robust baseline safety. However, the slight degradation suggests that future iterations should incorporate additional safety-focused training data to preserve the base model's conservative refusal patterns while maintaining child-appropriate communication style.

Note: Unfortunately the number of calls in gemini API free

teir has been reduced from 1000 per day to 20 per day recently (december 7th) so tests could not be reran.

H. *ChatBud UI*

A web-based user interface was developed to provide children with access to the fine-tuned ChatBud model. The interface runs as a standalone HTML file and connects to the model backend hosted on Google Colab via a Cloudflare tunnel.

*1) **Interface Components**:* The interface consists of three main areas. The central chat panel displays the conversation history, with user messages aligned left and ChatBud responses displayed below. A text input field and two action buttons—"Send your message" and "Add Picture"—occupy the bottom of the panel. A sidebar titled "ChatBud's Notes" provides static reminders about polite behavior and available features. A picture preview area displays any uploaded image before sending.

The color scheme uses soft greens, yellows, and blues to create a child-friendly appearance. Typography uses rounded, informal fonts consistent with educational applications for the target age group.

*2) **Multimodal Input**:* The interface supports both text and image inputs. Users can upload images using the "Add Picture" button, which accepts standard image formats (PNG, JPG). Uploaded images are converted to base64 encoding and transmitted to the backend alongside any accompanying text message. Figure 18 demonstrates the image understanding capability, where ChatBud correctly identifies and describes uploaded content.

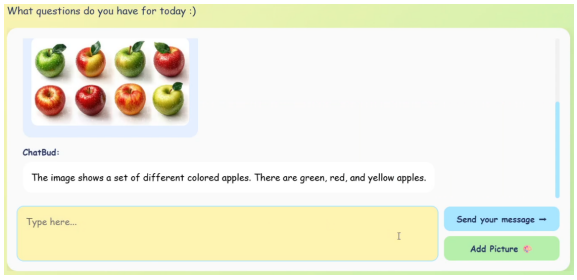

Fig. 18. ChatBud responding to an uploaded image of apples. The model correctly identifies the colors and arrangement of the fruits.

*3) **Conversation Memory**:* The backend maintains conversation history within each session, enabling multi-turn dialogue. The system manages context length to remain within the model's 16,384-token limit (limited for resource constraints while model can go up till 128k-tokens). Figure 19 shows a multi-turn interaction where ChatBud provides an age-appropriate explanation of mathematical concepts.

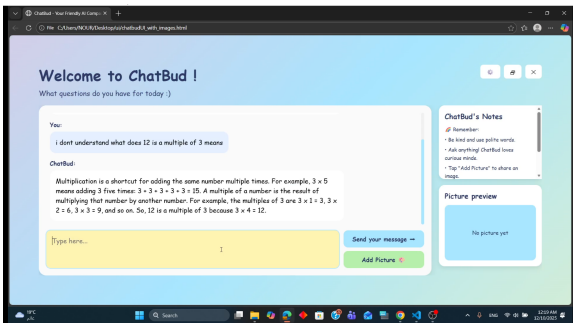

Fig. 19. ChatBud explaining the concept of multiples in response to a child's question. The response uses concrete examples suitable for the 9–11 age range.

*4) **Safety Response Demonstration**:* Figure 20 illustrates the model's handling of a sensitive scenario. When presented with a bullying-related input, ChatBud provides an empathetic response that validates the child's feelings, discourages harmful behavior, and directs the child to seek help from a trusted adult—consistent with the system prompt guidelines.

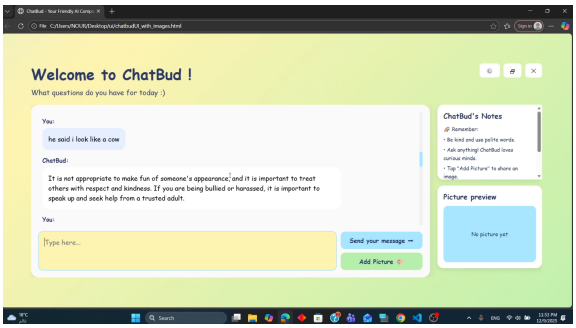

Fig. 20. ChatBud responding to a bullying scenario. The model validates the child's experience and recommends speaking with a trusted adult.

*5) **Backend Architecture**:* The interface communicates with a Flask server running on Google Colab. The server exposes three API endpoints: `/api/chat` for message handling, `/api/health` for connection verification, and `/api/clear` for resetting conversation history. A Cloudflare tunnel provides HTTPS access to the Colab instance without requiring port forwarding or static IP configuration. Users configure the backend URL through a settings modal accessible from the interface header.

Here are videos of the model being used:
https://drive.google.com/drive/folders/1RnivXVQL_Mcat6J8Ud3XCd0MlEtgl0IW?usp=drive_link

VII. Limitations

The present work is subject to several constraints that affect generality and depth. Training was carried out on the free-tier T4 GPU with strict runtime limits (5 hours per session), which forced splitting effort across multiple accounts and precluded extended experimentation. The fine-tuning budget (800 steps, effective batch size 8) remains modest, and the training corpus — drawn from four public datasets — is relatively small and limited to short, safe conversational data. As a result,

complex dialogs, rare prompts, or adversarial inputs are underrepresented.

The Fine-Tuning relies on LoRA updates to a subset of projection matrices. While this achieves stylistic adjustment (tone, brevity, child-friendly vocabulary), it does not rewrite the model's deeper reasoning or safety mechanisms. Therefore, behavioral compliance may degrade on novel or out-of-distribution inputs.

Finally, toxicity evaluation was to be further tested with other datasets and things to get more accurate results,however, use of the free-tier judge API for Gemini 2.5 Flash was constrained by quota limits, preventing comprehensive testing across a large sample set. Consequently, statistically robust safety differences under diverse or adversarial usage.

These limitations bounded the scope of our fine tuning, methodologies, and results.

Note: Unfortunatly not everything that was in the propsal was implemented due to many reasons including time, reasons mentioned above, and other course projects.

VIII. Conclusion and Future work

Conclusion

This work developed ChatBud, a child-oriented conversational assistant built on Gemma 3 4B-IT and fine-tuned using Low-Rank Adaptation (LoRA) on a curated mixture of 4,834 examples from five datasets: SAHAR, Child-QA, KidsChatBot, CAI Harmless, and Prosocial Dialog. The fine-tuning process, guided by expert consultation with a child psychiatrist, targeted three primary objectives: reduced response length, age-appropriate readability, and maintained safety, and made a Child friendly UI.

Evaluation across 165 test prompts demonstrated that the fine-tuned model achieved its core objectives. Response length decreased by 16.6% (from 30.52 to 25.45 words), and Flesch-Kincaid Grade level dropped by 9.1% (from 6.06 to 5.50), placing outputs squarely within the 4th–6th grade range appropriate for children aged 9–11. Both reductions were statistically significant. Lexical diversity remained unchanged, confirming that brevity resulted from genuine conciseness rather than vocabulary restriction.

LLM-as-judge evaluation using Gemini 3 Pro revealed that the fine-tuned model was preferred overall (50.9% vs 45.5%), with significant improvements in clarity for 9–11-year-olds ($p = 0.038$) and empathy with appropriate boundaries ($p = 0.046$). Performance gains were most pronounced on safety-critical prompts, where the fine-tuned model achieved a 62.5% win rate compared to 33.3% for the base model. This improvement aligns directly with the inclusion of Prosocial Dialog in the training mixture, which taught the model to de-escalate risky scenarios and redirect children to trusted adults.

However, the evaluation also revealed trade-offs. Factual accuracy decreased marginally (-0.19 points, $p = 0.051$), with underperformance in categories poorly represented in the training data: math (25% win rate), creativity (35.7%), and science (39.1%). Semantic similarity analysis on 244 Child-QA prompts showed a 3.7 percentage point reduction in high-similarity responses, reflecting the expected divergence when semantically equivalent content is expressed with simpler vocabulary.

Toxicity evaluation on the 500-prompt HarmEval benchmark confirmed robust safety for both models, with toxicity rates below 3%. The fine-tuned model showed a slight increase in flagged responses (13 vs 9), though manual inspection attributed most flags to educational responses triggering conservative safety filters rather than genuinely harmful content. Major safety issues decreased by 44%.

In summary, ChatBud demonstrates that parameter-efficient fine-tuning on a modest dataset can successfully adapt a general-purpose LLM for child-appropriate interaction. The model produces shorter, simpler, and more empathetic responses while maintaining safety. The observed trade-offs between simplification and factual completeness underscore the importance of balanced training data and suggest clear directions for improvement through expanded domain coverage and preference-based alignment methods.

Future Work

This study established a child-oriented safety adapter for Gemma 3 4B-IT alongside an end-to-end evaluation pipeline. To further enhance the model's robustness, safety, and interactivity, Several extensions remain open.

First, the training corpus will be expanded and balanced. Future iterations will integrate diverse child-focused and safety-focused datasets, such as multi-turn dialogues and additional languages relevant to the target context. Mixture ratios between sources (e.g., SAHAR, Child-QA, KidsChatBot, and Prosocial Dialog) will be explicitly controlled through curriculum fine-tuning. Ablation studies are recommended to determine which data sources most effectively improve safety, style, and age-appropriate explanations.

Second, modeling and alignment methods will be strengthened. While this work utilized supervised fine-tuning (SFT) with LoRA, future experiments will vary LoRA rank and compare the results against full-parameter fine-tuning. Furthermore, Direct Preference Optimization (DPO) or Reinforcement Learning from Human Feedback (RLHF) will be implemented. This involves collecting preference datasets where annotators distinguish between "more" and "less" child-appropriate responses, allowing the system to optimize explicitly for safety and age suitability rather than sole imitation.

Third, the evaluation framework will be extended beyond lexical metrics. Future work will incorporate multi-dimensional safety benchmarks, including adversarial prompts and jailbreak suites, alongside automated guard models. The evaluation pipeline will be updated to store per-prompt labels—such as refusal, over-refusal, and under-refusal—to facilitate detailed analysis by child-protection specialists.

Finally, the system will be upgraded to support speech capabilities. A speech layer integrating low-latency Text-to-Speech (TTS) and Automatic Speech Recognition (ASR) will be added to evaluate spoken clarity and prosody. Additionally,

runtime guardrails, such as topic classifiers and strict system prompts, will be implemented. These technical measures will be validated through IRB-compliant studies involving educators and safety experts to assess perceived clarity and helpfulness.

ACKNOWLEDGMENT

- This work was guided by the Maroun Semaan Faculty of Engineering and Architecture (MSFEA) at the American University of Beirut. **Special thanks** to Dr. Mariette Awad, and to the teaching assistants, Hafez Al Khatib, Tamara Fakih, Hasan Moughnieh, and Kevin Daou.
- The team acknowledges Dr. Wael Shamseddeen (AUBMC) for his expert guidance on child psychiatry and developmental safety.
- This project was completed through a combined effort of human work and AI assistance. LLMs (Claude Sonnet 4.5, GPT 5.1, GPT 5.1 Codex-max, Gemini 3.0 Pro) assisted in paraphrasing, correcting grammar, and writing of the report text, code files, and jupyter notebooks. The project team conducted all conceptual development, research design,selection of the datasets and evalutions, critical analysis, and validation of results. All AI-generated content was reviewed, verified, and refined to ensure technical accuracy and alignment with project objectives.

REFERENCES

[1] Google, "google/gemma-3-4b-it," Hugging Face Model. [Online]. Available: https://huggingface.co/google/gemma-3-4b-it

[2] Gemma Team *et al.*, "Gemma 3 technical report," *arXiv preprint arXiv:2503.19786*, 2025. [Online]. Available: https://arxiv.org/abs/2503.19786

[3] E. Hu *et al.*, "LoRA: Low-rank adaptation of large language models," in *Int. Conf. Learn. Represent. (ICLR)*, 2022. [Online]. Available: https://arxiv.org/abs/2106.09685

[4] A. Vaswani *et al.*, "Attention is all you need," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017. [Online]. Available: https://arxiv.org/pdf/1706.03762

[5] Yotev27367, "KidsChatBot," Hugging Face Dataset. [Online]. Available: https://huggingface.co/datasets/yotev27367/KidsChatBot

[6] Hma96, "SAHAR-Dataset," Hugging Face Dataset. [Online]. Available: https://huggingface.co/datasets/hma96/SAHAR-Dataset

[7] AllenAI, "Prosocial-dialog," Hugging Face Dataset. [Online]. Available: https://huggingface.co/datasets/allenai/prosocial-dialog

[8] HuggingFaceH4, "Cai-conversation-harmless," Hugging Face Dataset. [Online]. Available: https://huggingface.co/datasets/HuggingFaceH4/cai-conversation-harmless

[9] Chaitanyareddy0702, "Child-QA-dataset," Hugging Face Dataset. [Online]. Available: https://huggingface.co/datasets/chaitanyareddy0702/Child-QA-dataset

[10] H. Al Khansa, A. Mustapha, and M. Awad, "Sahar Dataset: a Validated Dialogue Based Dataset For a Child-Centric, Empathetic and Knowledge-Driven Chatbot," in *Proc. AAAI Symp. Ser.*, vol. 6, no. 1, pp. 159-166, 2025. [Online]. Available: Available: https://url-shortener.me/1XF0

[11] M. Anisuzzaman, M. R. Karim, M. A. Chowdhury, and S. Thirunavukkarasu, "Fine-tuning large language models: Techniques, applications, and evaluation," *Mayo Clin. Proc. Digit. Health*, vol. 3, no. 1, pp. 45–63, 2025. [Online]. Available: https://doi.org/10.1016/j.mcpdig.2024.100114

[12] J. Jiao *et al.*, "LLMs and childhood safety: Identifying risks and proposing a protection framework for safe Child-LLM interaction," *arXiv preprint arXiv:2502.11242*, 2025. [Online]. Available: https://arxiv.org/abs/2502.11242

[13] S. Kuzmin, X. Lee, and Y. Park, "Designing child-safe conversational AI: Three dilemmas for responsible design," in *Proc. ACM Conf.*, 2025. [Online]. Available: https://dl.acm.org/doi/10.1145/3719160.3737638

[14] N. Kurian, "'No, Alexa, no!': Designing child-safe AI and protecting children from the risks of the 'empathy gap' in large language models," *Learn. Media Technol.*, 2024. [Online]. Available: https://doi.org/10.1080/17439884.2024.2367052

[15] J. Huang and R. Patel, "Dialogic pedagogy for large language models: Aligning conversational AI with proven theories of learning," *arXiv preprint*, 2025.

[16] M. P.-C. Lin and D. H. Chang, "CHAT-ACTS: A pedagogical framework for personalized chatbot to enhance active learning and self-regulated learning," *Comput. Educ. Artif. Intell.*, vol. 5, no. 4, Art. no. 100167, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666920X23000462

[17] D. H. Chang, M. P.-C. Lin, S. Hajian, and Q. Q. Wang, "Educational design principles of using AI chatbot that supports self-regulated learning in education: Goal setting, feedback, and personalization," *Sustainability*, vol. 15, no. 17, p. 12921, 2023. [Online]. Available: https://doi.org/10.3390/su151712921

[18] S. Lee and S. Lim, "Development and evaluation of a leveled conversational teachable agent for children's English learning," *Appl. Sci.*, vol. 13, no. 12, p. 6541, 2023. [Online]. Available: https://doi.org/10.3390/app13126541

[19] M. Kim, J. Park, J. Choi, and S. Kwon, "Improving child-centered robot dialogue systems using large language models and human feedback," *Sensors*, vol. 24, no. 21, p. 7939, 2024. [Online]. Available: https://doi.org/10.3390/s24217939

[20] W. Seo, C. Yang, and Y.-H. Kim, "ChaCha: Leveraging large language models to prompt children to share their emotions about personal events," in *Proc. CHI Conf. Hum. Factors Comput. Syst. (CHI '24)*, pp. 1–20, 2024. [Online]. Available: https://doi.org/10.1145/3613904.3642152

[21] J. M. Weber, M. Valentini, T. Wright, K. von der Wense, and E. Colunga, "Evaluating LLMs as tools to support early vocabulary learning," in *Proc. Annu. Meeting Cogn. Sci. Soc.*, vol. 46, no. 0, pp. 2633–2640, 2024. [Online]. Available: https://escholarship.org/uc/item/7v69f0dj

[22] R. Abdelghani *et al.*, "GPT-3-driven pedagogical agents to train children's curious question-asking skills," *Int. J. Artif. Intell. Educ.*, vol. 34, no. 3, pp. 483–518, 2024. [Online]. Available: https://doi.org/10.1007/s40593-023-00340-7

[23] I. Carmo, P. Costa, and P. Santana, "Boosting children's reading motivation with LLM-generated story crossovers," in *Proc. 2024 IEEE Int. Conf. Graph. Interact. (ICGI)*, pp. 1–8, 2024. [Online]. Available: https://doi.org/10.1109/ICGI64003.2024.10923732

[24] Internet Matters, "Me, myself & AI report: How risky and unchecked AI chatbots are the new go-to for millions of children," 2024. [Online]. Available: https://www.internetmatters.org/hub/press-release/new-report-reveals-how-risky-and-unchecked-ai-chatbots-are-the-new-go-to-for-millions-of-children/

[25] W. Shamseddeen, Personal communication, Nov. 2025.

[26] SoftMINER-Group, "HarmEval," Hugging Face Dataset. [Online]. Available: https://huggingface.co/datasets/SoftMINER-Group/HarmEval

[27] Confident AI Inc., "Confident AI - The DeepEval LLM Evaluation Platform," 2025. [Online]. Available: https://www.confident-ai.com/

[28] R. Rafailov *et al.*, "Direct preference optimization: Your language model is secretly a reward model," *arXiv preprint arXiv:2305.18290*, 2023. [Online]. Available: https://arxiv.org/abs/2305.18290