Sahar Dataset: a Validated Dialogue Based Dataset For a Child-Centric, Empathetic and Knowledge-Driven Chatbot

Hadi Al Khansa[1], **Ahmad Mustapha**[1], **Mariette Awad**[1]

[1]The Department of Computer and Electrical Engineering, American University of Beirut, Beirut, Lebanon
hma96@main.aub.edu, amm90@main.aub.edu, ma162@aub.edu.lb

Abstract

Artificial intelligence, particularly large language models (LLMs), has had a significant impact in many fields, including chatbots and virtual assistants. With the popularity of ChatGPT, the trend of human-AI collaboration through LLM based chatbots is growing, reaching an ever-expanding audience. A key group that requires special attention is children. A chatbot designed for children should be both knowledgeable and empathetic. While chatbots are essentially fine-tuned versions of LLMs, fine-tuning these models for this specific purpose presents a challenge due to the lack of readily available datasets that address both scientific queries and empathetic situations. This data shortage can be addressed by using generative AI techniques to create synthetic dataset samples. As such, we propose in this paper the use of ChatGPT prompting to generate the Sahar Dataset, a multi-turn student-centric chatbot interaction dataset that supports both STEAM and empathetic related dialogues. Our results show that the Sahar dataset is readable by 5th grade students according to the Flesch-Kincaid Grade score, while other popular datasets like Alpaca require a 9th grade reading level. Moreover, we obtained an IRB for human evaluations, and the results show that 90% of the dataset's STEAM is factual, and the empathetic dialogues lead to valid solutions to the child's problem 90% of the time.

1 Introduction

In a STEAM-based educational environment, students are encouraged to apply critical thinking to develop innovative solutions to real-world problems (Yakman 2008). As AI continues to advance, its role in education has expanded, particularly in online learning environments where it automates instructional tasks and facilitates assessment generation (Seo et al. 2021). However, current AI-driven educational tools often lack contextual understanding tailored to children's cognitive development, learning styles, and interdisciplinary STEAM principles.

Moreover, AI-driven child-friendly chatbots, which have been recognized as valuable resources when parental guidance is unavailable, could play a significant role in STEAM education by fostering inquiry-based learning and interdisciplinary engagement. Recent studies suggest that well-designed child-friendly chatbots can provide opportunities for self-expression, potentially supporting emotional well-being. However, their impact depends on design factors, ethical considerations, and responsible implementation (Aarts et al. 2022; Cooper and Ireland 2018; Santos, Ong, and Resurreccion 2020; Zhang et al. 2022). As a result, integrating AI-powered chatbots into STEAM education presents a unique opportunity to enhance human-AI collaboration.

The most recent advancements in AI-powered Chatbots have leaned heavily on large language models (LLM), as they have proven their ability to sustain longer and more coherent dialogues compared to the conventional natural language processing techniques utilized in chatbot development (Lee 2023). LLMs are redefining Human-AI Collaboration by enabling adaptive, context-aware interactions that enhance human cognition and decision-making. Unlike traditional AI systems, LLMs engage in fluid, iterative exchanges, allowing users across disciplines to refine ideas, explore complex problems, and generate creative solutions.

However, the effectiveness of these AI-driven collaborations depends on whether the models have been trained on diverse, representative, and ethically minded datasets. This is particularly critical in STEAM education and empathy-driven AI for children, where biased, incomplete, or poorly curated datasets risk reinforcing stereotypes and failing to capture the depth of human emotions and cognitive diversity. Therefore, to maximize the benefits of LLMs for human-AI collaboration in an educational setting, these models should be fine-tuned to create a STEAM-oriented, empathetic multi-turn chatbot. These will be finely attuned to the academic and emotional needs of elementary school children. However, the primary challenge impeding the development of such LLM-powered STEAM and empathetic chatbots for children is the lack of datasets that can be employed to adapt these models to become helpful assistants for children.

In STEAM learning, AI-powered tutors must navigate challenges like explaining abstract scientific principles in engaging ways, fostering creativity in the arts, and adapting to diverse cognitive styles. High-quality datasets that integrate real-world problem-solving, historical context, and interdisciplinary connections ensure that AI can support critical thinking and curiosity, rather than merely providing correct answers. Beyond academic learning, empathy in AI-driven education is essential for nurturing social-

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

emotional intelligence in children. AI learning assistants must be trained on datasets enriched with child-centric narratives, inclusive language, and psychologically sound reinforcement strategies, ensuring that interactions promote emotional growth and avoid reinforcing harmful biases or misinformation. Poorly curated datasets can lead to AI-generated responses that lack cultural sensitivity, misinterpret emotional cues, or perpetuate outdated stereotypes, limiting AI's potential as a truly adaptive and compassionate learning companion.

To address this dataset sparsity issue, we propose utilizing synthetic data generation techniques. LLMs like ChatGPT have already been successfully employed to generate data for fine-tuning smaller LLMs on specific tasks, such as following instructions and coding (Ding et al. 2023; Köksal et al. 2023; Rohan et al. 2023; Ubani, Polat, and Nielsen 2023; Zhou et al. 2023). We decided to prompt ChatGPT to simulate dialogues between a child and their caretaker chatbot on various STEAM and empathetic topics. All content generated by ChatGPT underwent a thorough human review process that is IRB (AUB 2025) approved, and this, to ensure factual accuracy and appropriateness.

Because the content in the Sahar dataset should be comprehensible for elementary school children, we evaluated the readability of our dataset using the Flesch-Kincaid Grade (Kincaid et al. 1975) US Grade readability score, and we compared the proposed dataset's scores to a popular fine-tuning instruction dataset, and to the vanilla responses of popular LLMs. Our results show that the Flesch-Kincaid Grade scores are 4.18 for the empathy content and 6.94 for the STEAM content, which are significantly lower than the average scores of unprompted LLMs (around 8 and 11, respectively). Moreover, the IRB (AUB 2025) approved 14 human evaluations indicate that the dataset is suitable for 5th grade students, its STEAM content is 90% factual, and its empathetic dialogues provide viable solutions to the child's problem 92% of the time.

Moreover, while LLMs can show human-like empathy, they suffer from the problem of 'nudging' (Kurian 2024) . Nudging occurs when an LLM pushes a user from a factual exchange to more personal discussions. Nudging has been flagged as a coercion tool and a danger of LLMs that pushes users towards certain behaviors, and it becomes risky when children cannot provide informed consent on how the conversation is evolving. This becomes even more serious when combined with the risk of having the child's personal information gathered by the LLM monitored or leaked by the LLM provider. Therefore, in this research, we also conducted an experiment to validate that the STEAM content of the proposed dataset is free of emotional nudging.

This highly curated and highly empathetic knowledge-driven dataset is the first of its kind, and it can be used in multiple ways. First, it is useful on its own to fine-tune a model specialized on empathetic and STEAM dialogue with children. Second, the it can be incorporated into larger datasets as a distillation-based method to prevent catastrophic forgetting, where a model overfits other tasks it is finetuned on, and loses the ability over time to conduct STEAM and empathetic dialogue with children (Song et al. 2025).

Unlike previous research, this paper contributes to the literature by:

- Presenting a ChatGPT-based approach for generating a dataset designed for elementary school children, with a focus on STEAM education and empathetic dialogues.
- Conducting human validation of the Sahar dataset to ensure its quality and relevance.
- Demonstrating the dataset's readability for elementary school children.
- Validating that the STEAM portion of the Sahar dataset remains nudge free including biases or leading prompts.

The rest of the paper is divided such that Section 2 presents related work on augmenting datasets with synthetic data produced by LLMs while Section 3 briefly describes the curated Sahar dataset. Section 4 demonstrates how the dataset was collected and processed. Section 5 demonstrates the evaluations of the dataset. Section 7 presents limitations. Section 8 concludes the work.

2 Related Work

Table 1 provides a concise overview of existing children chatbots from the literature. These chatbots primarily rely on conventional NLP techniques and utilize pre-defined sets of questions, answers, and dialogue flows tailored for specific tasks. However, it's important to note that none of the chatbots listed in Table 1 offer comprehensive coverage of general STEAM topics and demonstrate empathy towards children's broader concerns.

To bridge this gap, there is a need for the development of children chatbots by harnessing the capabilities of LLMs. LLMs can be effectively trained to create chatbots capable of engaging in more extended, diverse, and coherent dialogues while specializing in STEAM subjects and empathetic dialogues, as highlighted in (Lee 2023).

While STEAM-related content is abundantly available in sources such as Wikipedia scraped datasets and select children's websites like kids.kiddle.co, it's important to recognize that this data remains unstructured. Consequently, it may not serve as a suitable foundation for training an LLM to effectively engage in STEAM discussions with children.

Furthermore, datasets like "daily_dialog", "prosocial-dialog" and "Customer Support on Twitter" are indeed multi-turn in nature. However, they predominantly cover typical adult dialogues, which do not equip an LLM with the necessary insights to address a child's emotional needs, as noted in references (Kim et al. 2022; Li et al. 2017; Axelbrooke 2017)

In light of these considerations, the primary challenge in developing such a chatbot lies in the absence of datasets that showcase chatbot-child interactions related to STEAM subjects and empathetic dialogues. Addressing this data gap will be crucial to the successful creation of chatbots tailored to engage with children effectively in these domains.

One solution to overcome the lack of a dataset is to use AI assisted synthetic data generation techniques. Ubani et al. (Ubani, Polat, and Nielsen 2023) augmented training sets in

Table 1: A summary of the literature around child centered chatbots.

Authors	Contribution	Methods	Citations
Cooper et al. (2018) (Cooper and Ireland 2018)	Chatbot that can aid children on the autism spectrum.	AIML programming language to define chat patterns.	44
Liu et al. (2022) (Liu et al. 2022)	Chatbots as book talk companions for children.	Google Actions Console used to build a Q&A chatbot about books.	170
Aarts et al. (2022) (Aarts et al. 2022)	Help children with sleeping disorders.	Predefined a set of dialogues a child can have with Snoozy based on interviews.	22
Mageira et al. (2022) (Mageira et al. 2022)	Teach high school students cultural content in the English or French languages.	Snatchbot website and predefined knowledgebase.	275
Rajwal (2023) (Rajwal 2023)	A proposed framework for a Chatbot that can help improve the emotional intelligence of children.	Google's DialogFlow and Assistant and a knowledgebase of dialogues.	2
Xu et al (2024)(Xu et al. 2024)	Propose ”Sound Guardian” an AI Chatbot augmented framework for teaching children about sound systems	Build an educational game and use an AI Chatbot as an assistant	1

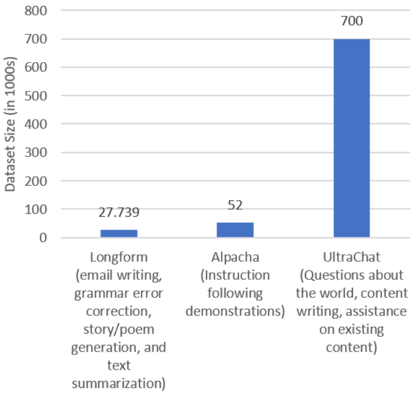
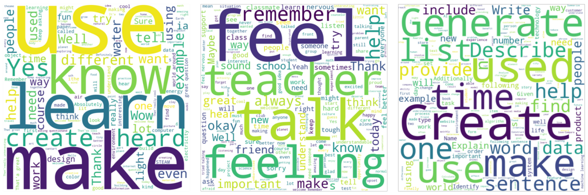


Figure 1: The content type and size of the three most common datasets generated or augmented using LLMs.

low resource scenarios using data obtained from prompting LLMs such as ChatGPT, and showed that with task specific ChatGPT prompts it is possible to outperform other state of the art approaches for data augmentation. Furthermore, such large LLMs have been prompted in the literature to produce synthetic instruction datasets that were used to successfully improve the performance of smaller LLMs. Figure 1 summarizes the most prominent LLM generated and augmented instruction datasets found in the literature (Ding et al. 2023; Köksal et al. 2023; Rohan et al. 2023).

While it may be tempting to use these datasets to finetune a STEAM specialized and empathetic chatbot for children, such datasets follow an instruction-answer paradigm, and manually inspecting such datasets shows that many of the target responses are challenging for children to read. In fact, running the Flesch-Kincaid Grade (Kincaid et al. 1975) on such datasets shows that they require a 9th US grade reading level to easily understand their content, which is not suitable for elementary school children.

For these reasons, this paper uses ChatGPT to produce the Sahar dataset, a child-chatbot multi-turn dataset covering both STEAM and empathetic dialogues for elementary school children.



(a) STEAM content (b) Empathetic content (c) Alpaca dataset

Figure 2: Word clouds from the Sahar and Alpaca datasets

3 Sahar The Dataset

The curated Sahar dataset consists of 281 simulated dialogues between a child and a caretaker named Sahar. 210 of these dialogues are STEAM related, and 71 dialogues are about empathetic situations where the student may need guidance.

Each dialogue contains around 7 turns on average so around 2000 samples were prepared by formatting the dataset such that the chat history is the input and the chatbot's response given the chat history is the output. Given the advancements in parameter efficient finetuning techniques, the Sahar dataset could be sufficient to orient an LLM to a child's STEAM and empathetic chatbot. (Xu et al. 2023).

To validate the need of the Sahar dataset, we compared it to the Alpaca dataset - which is one of the most popular instruction datasets available on Huggingface (0Hu 2016), on several metrics. Finally, we classified our data to understand its distribution across topics and sentiments. We classified the data by first finding the contextualized embeddings for our class words and the dialogues, and then assigning each

Metric	Sahar Dataset	Alpaca
Flesch-Kincaid Grade	5.4	9.3
Target Response Length (words)	33.37	44.18
Entropy	4.64	4.6

Table 2: Readability and entropy on the Sahar and Alpaca datasets.

dialogue to the class word it is closest to in the embeddings space. The classes and the distribution of the chats among them can are shown in the Appendix.

Furthermore, we had 14 reviewers with different backgrounds and high English proficiencies inspect the generated data to ensure the information quality and multi-turn style. The study was approved by the IRB (AUB 2025), and the reviewers read and approved the digital consent form before starting the survey. Section 5 details the reviewers' evaluation process and results.

Figures 2a and Figure 2b show the word maps for both STEAM and the empathetic content of the Sahar dataset. The most prominent words in STEAM include make, use, create, learn, heard, and know all of which are associated with the STEAM pedagogy that encourages critical thinking. Secondary words include wow, yes, thank, fun, great question, try, tell, sure, and keep asking, which are positive reinforcement words used in the context of engaging multi-turn dialogues. On the other hand, the empathetic content is dominated by phrases such as talk, feel, feeling, teacher, help, feel better, and feel nervous, because this part of the Sahar dataset focuses on dialogues where the child is expressing his/her feelings to a teacher regarding a specific situation and then seeking the teacher's help and advice.

For the sake of comparison, Figure 2c shows the word map for the Alpaca dataset. The most prominent words include explain, create, make, generate, provide, describe, list, and find, which are used in imperative sentences to order the model to follow an instruction and produce a result. Furthermore, unlike our STEAM content, the Alpaca dataset lacks the conversational words discussed previously.

Moreover, Table 2 compares our dataset to the Alpaca dataset on the average Flesch-Kincaid US Grade readability, the average target response length, and entropy. Our dataset achieves a 5.4 Flesch-Kincaid score meaning that it is readable by 5th grade students, while the Alpaca dataset requires 9th grade reading level. We also compared the entropy of our dataset to Alpaca's. Since our dataset has a high entropy level comparable with Alpaca's, we have a preliminary indication that the quality of the English text and redundancy levels in both datasets are similar. This, combined with the significantly shorter target responses of our dataset makes it more appealing for finetuning an LLM Chatbot to converse with lower elementary school children. Furthermore, both STEAM and empathetic content were classified as shown in Table 3 and Table 4. Clusters are shown in Figure 8 and Figure 9 of Appendix A.

A third of the STEAM content revolves around science, another third is about engineering and technology, and the last third is about art and mathematics. While art is often neglected in STEAM education (Liu and Wu 2022), we ensured that more than 20% of the topics revolve around art. Only 13.3% of the chosen STEAM topics revolve around mathematics because of the difficulty we faced getting ChatGPT to produce reliable mathematics outside basic theories. Therefore, we decided that mathematics shouldn't be a primary goal for an LLM finetuned on this dataset as shown in Table 3.

Topic	Count	Percentage (%)
Science	69	32.8
Technology	39	18.6
Engineering	29	13.8
Art	45	21.4
Mathematics	28	13.3

Table 3: The distribution of the STEAM topics of the Sahar dataset

Sentiment	Count	Percentage (%)
Excited	6	8.5
Upset	8	11.3
Proud	6	8.5
Frustrated	17	23.9
Anxious	25	35.2
Indifferent	9	12.7

Table 4: Sentiment distribution of empathetic situations in the Sahar dataset.

As for the gathered empathetic situations, about 70% of them result in a child having a negative sentiment. We choose such a distribution because we believe that finetuning a LLM capable of correctly guiding a child through a difficult situation is a more critical and challenging task than finetuning a LLM to encourage the child in positive events.

4 Building The Sahar Dataset

4.1 Curating a Dataset Using ChatGPT

In this section we explain how Sahar dataset was prepared using ChatGPT. While instruction generation prompting asks ChatGPT to produce an instruction, a response to an instruction, or both, our methodology relies on prompting ChatGPT to generate multi-turn dialogues by alternating between the roles of a curious elementary school child needing assistance and an empathetic, kind, and encouraging chatbot. To automate the generation process, a python script that uses ChatGPT API (gpt-3.5-turbo API) was used to send the prompts to ChatGPT, collect and store the responses.

4.2 The STEAM Content

The first step to curating the STEAM content was collecting a list of STEAM topics that would spark the curiosity of a child. We collected a list of 210 topics by first querying children websites (https://www.kiddle.co/, https://www.sciencekids.co.nz/, and https://academickids.com/,), and then prompting ChatGPT to generate related topics. The topics include but

Prompt

Generate a conversation between a teacher chatbot named SAHAR and an elementary school student. SAHAR, the chatbot is a STEAM education teacher, she is very nice, funny and she connects emotionally with her students. The student is curious and young, he will ask a lot of follow-up questions. SAHAR should thoroughly answer the student's questions with short, simple and STEAM inspired answers while trying to spur his/her curiosity. Make the conversation as creative as possible, and remember, short answers with a lot of mind-stimulating follow up questions. SAHAR should say words of encouragement from time to time. SAHAR and the student know each other, no need for introductions. SAHAR doesn't have access to the internet and she cannot send the student any links or documents. The student hears about a STEAM topic and s/he's curious to know more about it. The conversation should include a scenario on how the student knew about this specific topic. The topic of the conversation is: $<< Topic >>$.

Figure 3: The prompt for generating STEAM content in the Sahar dataset.

are not limited to fossils, space, AI, biomimicry, origami, mosaic art, and climate change. The second step was to generate multi-turn dialogues that simulate the child-chatbot interaction regarding those STEAM topics. The chatbot had to be kind, emotionally connect with the child, encourage his/her curiosity, and answer any follow up questions the child may have. Also, since the child is in elementary school, the questions and answers are expected to be short. To keep the child safe, the chatbot should not provide the student with any links to the internet. Figure 3 presents the prompt used to generate the STEAM dialogues between a child and the chatbot. The $<< TOPIC >>$ token is replaced each time with one of the 210 curated topics and the gpt-3.5-turbo API is called to generate a new dialogue.

4.3 The Empathetic Content

A similar approach to that presented in the previous section was used to curate the empathetic content. ChatGPT was asked to compile a list of situations where children would want to express themselves, discuss their achievements or failures, or seek advice for a problem they are facing. A list of 71 situations was compiled. The list includes situations such as being subjected to bullying, failing a test, being asked to read or perform in front of class, making a new friend, getting praised for good behavior, or having their academic achievements recognized. Figure 4 presents the prompt used to generate the empathetic content of the Sahar dataset. The $<< SITUATION >>$ token is replaced by one of the 71 situations curated list of empathetic topics.

4.4 Formatting the dialogues

Prompt

Generate a conversation between SAHAR and an elementary school student. SAHAR is a compassionate chatbot, that connects deeply with the students, gives them the floor to express their feelings on a specific situation, and will sometimes help them find a solution to a specific problem. SAHAR will, if needed offer emotional support to the student. The student will talk to SAHAR about the following: $<< SITUATION >>$. SAHAR and the student know each other, no need for introductions. Limit the conversation to 1000 tokens.

Figure 4: The prompt for generating the empathetic content of the Sahar dataset

Given the context from previous questions and their corresponding answers, the Chatbot needs to be trained to generate a response aligned with the child's last question . However, feeding the entire dialogue to the model is not optimal. As such, we divided each multi-turn dialogue into sub-dialogues. The input and the labels for the sub-dialogues we used are as follows: 1. Input: the input sample is comprised of a student's question, preceded by the relevant context from previous questions and answers in the same dialogue. The context and questions are enclosed with beginning and end of question tokens ($< BOQ >$ and $< EOQ >$). 2. Label: the target output is the response to the student's question given the previous questions and answers as context. The target output will be enclosed with beginning and end of answer tokens ($< BOA >$ and $< EOA >$).

The $< BOQ >$, $< EOQ >$, $< BOA >$, and $< EOA >$ tokens are used for the following reasons. First, the $< BOQ >$ and $< EOQ >$ tokens bound the question and context and help the model understand where the question finishes so that it can start generating the response. Without the $< EOQ >$ token, the model may decide to generate a response that adds to the student's question before answering it. Second, the $< EOA >$ token helps the model understand where their response needs to be terminated so that the model does not generate text beyond the needed response. Third, even if the model generated text after the $< EOA >$ token, only text before the $< EOA >$ can be presented to the user in post-processing. Applying this methodology allowed for generating several data points from a single dialogue. Therefore, the final dataset is comprised of about 2000 samples.

5 Dataset Evaluation

To validate the Sahar dataset's suitability for elementary school children, we obtained IRB (AUB 2025) approval for human evaluation. Fourteen reviewers with diverse engineering backgrounds (Computer, Electrical, and AI both Bachelor, Masters) and varying levels of English proficiency participated in the study. All reviewers were verbally informed about the study, approved the online consent forms, and agreed to have their data aggregated in the final analysis.

We asked the reviewers questions that are common to the STEAM and empathetic parts of the datasets and ad-

Partition	Question	Average Score
STEAM	The Content is Factual for an Elementary School Level (%)	90.19
	The Conversation is Informative (%)	99.38
	The Conversation is Redundant (%)	3.75
	The Conversation is on Topic (%)	99.75
	What minimum school grade level can find reading this conversation easy?	5
Empathy	The Conversation is Informative (%)	94.05
	The Conversation is Redundant (%)	5.12
	The Conversation is on Topic (%)	99.29
	What minimum school grade level can find reading this conversation easy?	5
	Sahar suggests viable solutions to the child's problem. (% Agree)	92.90
	Sahar validates the child's feelings in the situation. (% Agree)	100.00
	Sahar's suggestions put the child in danger. (% Agree)	0.95

Table 5: Reviewers' evaluation of the language used in the Sahar dataset

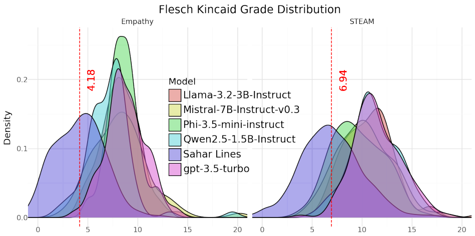


Figure 5: Density plot of the Sahar dataset is more child-centric than mainstream LLMs.

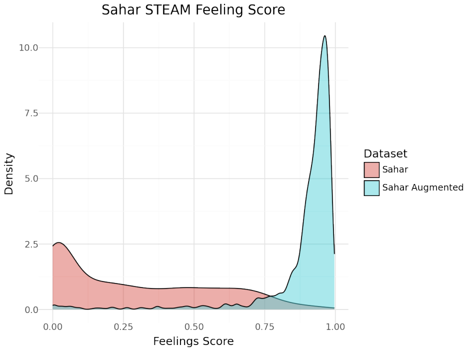


Figure 6: Feeling score distribution of the STEAM portion of the Sahar dataset.

ditional questions that are specific to each part. The common questions aimed to assess if the dialogues remained on topic, had any redundancy, and were simple to read. For the STEAM content, we were also interested in knowing if the answers provided by the chatbot were factual for an elementary school level. Regarding empathetic content, our questions targeted the chatbot's ability to validate the child's feelings, and suggest feasible and safe solutions. We also asked the reviewers to give suggestions for chats that they rate poorly. These suggestions were then used to amend the poor chats to ensure the quality of the dataset. Table 5 lists the questions that were asked to the reviewers about the dataset.

Table 5 also summarizes the average scores per question. The scores are overall positive in favor of the Sahar dataset. We notice high ratings for the informativeness and coherence and an average redundancy less than 5% for the STEAM and empathetic content. Moreover, the human evaluations estimate that on average a 5th grade student should be able to understand with ease the dialogues for both the STEAM and empathetic parts of the dataset. Furthermore, more than 90% of the STEAM content was considered factual by the evaluators, and the questions specific for the empathetic content show that the dialogues provide viable and safe solutions, while validating the child's feelings. To confirm the raters' alignments, we calculated the Fleiss' Kappa (Fleiss 1971) inter-raters' agreement measure, and as expected the result was near zero, and this is due to the low standard deviation between the raters' responses.

We also compared the dataset to the vanilla responses of popular LLMs. For each topic in the Sahar dataset, we asked five different LLMs to generate a brief paragraph. The models used were Llama 3.2 3B Instruct, Mistral 7B Instruct, Phi 3.5 mini Instruct, Qwen 2.5 1.5B Instruct, and GPT 3.5 turbo. We then calculated the Flesch-Kincaid Grade readability score for the outputs for all models and Sahar's responses STEAM and empathetic from the Sahar dataset. The results in Figure 5 show that Sahar's language, for both STEAM and Empathetic content, is significantly easier for a child to read compared to the other models.

Additionally, we evaluated how frequently Sahar mentions or refers to feelings in its dialogues within the STEAM content. In other words how much it ”nudges” the child to switch from factual to emotional dialogues. To do this, we first augmented all Sahar lines in the STEAM content by adding a statement related to feelings. This was done by prompting GPT-4 for each individual line. Next, we calculated a score for how much each line referenced feelings, both for the original and augmented versions. No nudging-specific models exist, so, to calculate the score, we used Hugging Face's zero-shot classification pipeline to perform binary classification on whether the line asks about feelings or not. We utilized the ”facebook/bart-large-mnli” model, and the model's confidence in detecting feelings was used as the score. The results, shown in Figure 6, indicate that it is unlikely for Sahar to bring up feelings in factual dialogues since the model showed confidence levels below 50% for most of the data. This is contrary to the model's high confidence levels for the data that was augmented with emotional statements. This shows that the chosen model is capable of detecting emotional statements, and it was unable to find emotional statements for most of Sahar's factual dialogues.

6 Finetuning Llama 3.2 1b

To demonstrate the usefulness of the generated dataset, we finetuned the small Llama 3.2 1b model on a 80%-20% train-test split of the data. The results in figure 7 show that the responses of the finetuned model are more similar to the targets than those of the prompted base model.

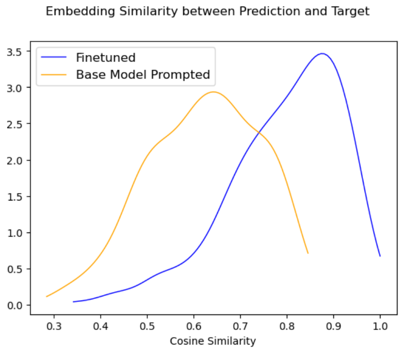


Figure 7: Embedding similarity results

7 Limitations

While this study provides a human-evaluated dataset that can help train a factually accurate and empathetic chatbot, it has several limitations. First, the dataset is not exhaustive to every STEAM and empathetic topic a child may want to discuss. Therefore, a chatbot trained on this dataset is not guaranteed to be hallucination-free when asked about topics it was not exposed to. Second, the generation and evaluation process are not scalable since it requires human evaluators to proofread and amend the ChatGPT generated text to guarantee factuality and safety. This limited the size of our dataset. Third, the dataset is small and our evaluators are all AUB students limiting the study of cultural and linguistic generalization. However, it should be noted that the raters come from different cultural backgrounds, which provides some cultural and linguistic generalization. Fourth, while the human evaluation results are promising, synthetic data overlooks nuances in real interactions, and may contain biases that we did not investigate. Finally, none of our evaluators came from literary or childcare backgrounds. Perhaps having reviewers from such backgrounds may shed light on further improvements that can be made to the dataset.

8 Conclusion

The future of AI in education and emotional intelligence hinges on relevant dataset curation. Ensuring that LLMs are trained on ethically sourced and pedagogically rich data is not just a technical challenge but a societal responsibility—one that will define how future generations interact with AI, not only as learners but as emotionally intelligent individuals in an increasingly AI-integrated world. Driven to make STEAM education more accessible to elementary school students, we developed the Sahar dataset. This dataset is designed to finetune an LLM-based chatbot that offers both STEAM educational and emotional support. The chatbot could collaborate with children in a compassionate and informative tone while ensuring its responses are accurate. To that end, we used ChatGPT to curate the Sahar dataset. Comparing our dataset to an available instruction dataset, our dataset is more readable for children, and has more conversational and empathy components. This makes it more suitable for finetuning an LLM to become a STEAM and empathetic based companion that is child centric. Future works will focus on discovering more scalable ways of constructing such datasets to guarantee a wider topic coverage.

A Classification

We encoded text embeddings for the Sahar dataset using the "all-MiniLM-L6-v2" sentence transformer model. The embeddings are then projected into a 2D space using UMAP. The results are shown in Figure 8 and 9 for both the STEAM and empathetic content respectively.

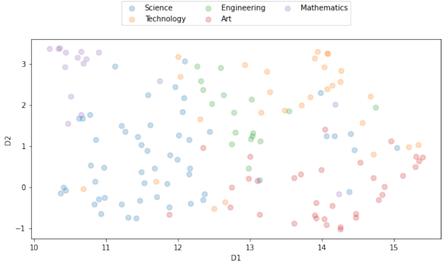


Figure 8: The classification UMAP of the embeddings of the STEAM dialogues of the Sahar dataset.

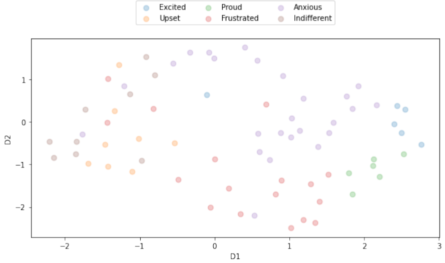


Figure 9: The classification UMAP of the embeddings of the empathetic dialogues of the Sahar dataset.

Acknowledgments

The authors would like to acknowledge the support of the University Research Board, the AI, Data Science and Computing Hub, the VIP and the Maroun Semaan Faculty of

Engineering and Architecture at the American University of Beirut for funding this research and providing the essential infrastructure needed to conduct this study.

References

2016. Hugging Face – The AI community building the future. https://huggingface.co/datasets?other=instruction-finetuning&sort=downloads. [Online; accessed 2023-12-28].

Aarts, T.; Markopoulos, P.; Giling, L.; Vacaretu, T.; and Pillen, S. 2022. Snoozy: A Chatbot-Based Sleep Diary for Children Aged Eight to Twelve. In *Interaction Design and Children*, IDC '22, 297–307. New York, NY, USA: Association for Computing Machinery. ISBN 9781450391979.

AUB. 2025. Human Research Protection Program (HRPP) — aub.edu.lb. https://www.aub.edu.lb/irb/Pages/default.aspx. [Accessed 17-03-2025].

Axelbrooke, S. 2017. Customer Support on Twitter.

Cooper, A.; and Ireland, D. 2018. Designing a Chat-Bot for Non-Verbal Children on the Autism Spectrum. *Studies in health technology and informatics*, 252.

Ding, N.; Chen, Y.; Xu, B.; Qin, Y.; Zheng, Z.; Hu, S.; Liu, Z.; Sun, M.; and Zhou, B. 2023. Enhancing Chat Language Models by Scaling High-quality Instructional Conversations.

Fleiss, J. L. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5): 378–382.

Kim, H.; Yu, Y.; Jiang, L.; Lu, X.; Khashabi, D.; Kim, G.; Choi, Y.; and Sap, M. 2022. ProsocialDialog: A Prosocial Backbone for Conversational Agents. In *EMNLP*.

Kincaid, J. P.; Fishburne Jr, R. P.; Rogers, R. L.; and Chissom, B. S. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.

Köksal, A.; Schick, T.; Korhonen, A.; and Schütze, H. 2023. LongForm: Optimizing Instruction Tuning for Long Text Generation with Corpus Extraction.

Kurian, N. 2024. 'No, Alexa, no!': designing child-safe AI and protecting children from the risks of the 'empathy gap' in large language models. *Learning, Media and Technology*, 0(0): 1–14.

Lee, A. 2023. What Are Large Language Models Used For and Why Are They Important? | NVIDIA Blog. https://blogs.nvidia.com/blog/2023/01/26/what-are-large-language-models-used-for/. [Online; accessed 2023-02-18].

Li, Y.; Su, H.; Shen, X.; Li, W.; Cao, Z.; and Niu, S. 2017. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In *Proceedings of The 8th International Joint Conference on Natural Language Processing (IJCNLP 2017)*.

Liu, C.-C.; Liao, M.-G.; Chang, C.-H.; and Lin, H.-M. 2022. An analysis of children' interaction with an AI chatbot and its impact on their interest in reading. *Computers & Education*, 189: 104576.

Liu, C.-Y.; and Wu, C.-J. 2022. STEM without art: A ship without a sail. *Thinking Skills and Creativity*, 43: 100977.

Mageira, K.; Pittou, D.; Papasalouros, A.; Kotis, K.; Zangogianni, P.; and Daradoumis, A. 2022. Educational AI Chatbots for Content and Language Integrated Learning. *Applied Sciences*, 12(7).

Rajwal, S. 2023. Design of a Chatbot for Four- to Ten-Year-Old Children Based on Emotional Intelligence. In Gupta, D.; Khanna, A.; Bhattacharyya, S.; Hassanien, A. E.; Anand, S.; and Jaiswal, A., eds., *International Conference on Innovative Computing and Communications*, 675–683. Singapore: Springer Nature Singapore. ISBN 978-981-19-2821-5.

Rohan, T.; Ishaan, G.; Tianyi, Z.; Yann, D.; Xuechen, L.; Carlos, G.; Percy, L.; and B., H. T. 2023. Alpaca: A Strong, Replicable Instruction-Following Model. https://crfm.stanford.edu/2023/03/13/alpaca.html. [Online; accessed 2023-12-09].

Santos, K. A.; Ong, E.; and Resurreccion, R. 2020. Therapist vibe: Children's expressions of their emotions through storytelling with a chatbot. In *Proceedings of the Interaction Design and Children Conference, IDC 2020*.

Seo, K.; Tang, J.; Roll, I.; Fels, S.; and Yoon, D. 2021. The impact of artificial intelligence on learner–instructor interaction in online learning. *International Journal of Educational Technology in Higher Education*, 18(1).

Song, S.; Xu, H.; Ma, J.; Li, S.; Peng, L.; Wan, Q.; Liu, X.; and Yu, J. 2025. How to Alleviate Catastrophic Forgetting in LLMs Finetuning? Hierarchical Layer-Wise and Element-Wise Regularization. arXiv:2501.13669.

Ubani, S.; Polat, S. O.; and Nielsen, R. 2023. ZeroShotDataAug: Generating and Augmenting Training Data with ChatGPT.

Xu, L.; Xie, H.; Qin, S.-Z. J.; Tao, X.; and Wang, F. L. 2023. Parameter-Efficient Fine-Tuning Methods for Pretrained Language Models: A Critical Review and Assessment.

Xu, Y.; Zhu, J.; Wang, M.; Qian, F.; Yang, Y.; and Zhang, J. 2024. The Impact of a Digital Game-Based AI Chatbot on Students' Academic Performance, Higher-Order Thinking, and Behavioral Patterns in an Information Technology Curriculum. *Applied Sciences*, 14(15).

Yakman, G. 2008. STEAM Education: An Overview of Creating a Model of Intergrative Education. In *Pupils Attitudes Towards Technology (PATT-19) conference : Research on Technology, Ionnovation, Design & Engineering Teaching*, volume 53.

Zhang, Z.; Xu, Y.; Wang, Y.; Yao, B.; Ritchie, D.; Wu, T.; Yu, M.; Wang, D.; and Li, T. J.-J. 2022. StoryBuddy: A Human-AI Collaborative Chatbot for Parent-Child Interactive Storytelling with Flexible Parental Involvement. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22. New York, NY, USA: Association for Computing Machinery. ISBN 9781450391573.

Zhou, C.; Liu, P.; Xu, P.; Iyer, S.; Sun, J.; Mao, Y.; Ma, X.; Efrat, A.; Yu, P.; Yu, L.; Zhang, S.; Ghosh, G.; Lewis, M.; Zettlemoyer, L.; and Levy, O. 2023. LIMA: Less Is More for Alignment.