AMERICAN UNIVERSITY OF BEIRUT

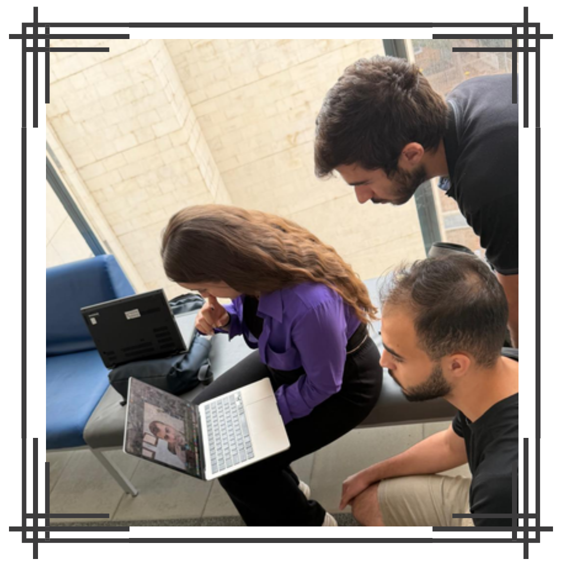
ChatBud

Tarek Bshinnati - Nour Hamieh
Moataz Maarouf - Ahmad Isber

Intro + lit.review:

Problems:

- Widespread child chatbot use
- Lack of child-specific safety
- Empathy gap & emotional risks



Interview with Dr. Shamseddeen

Literature Review:

- New report reveals how risky and unchecked AI chatbots are the new 'go to' for millions of children
- Me, Myself and AI research: Understanding and safeguarding children's use of AI chatbots
- AI chatbots' empathy gap poses risks for children, research finds
- AI Companion Chatbots: The Risks to Children

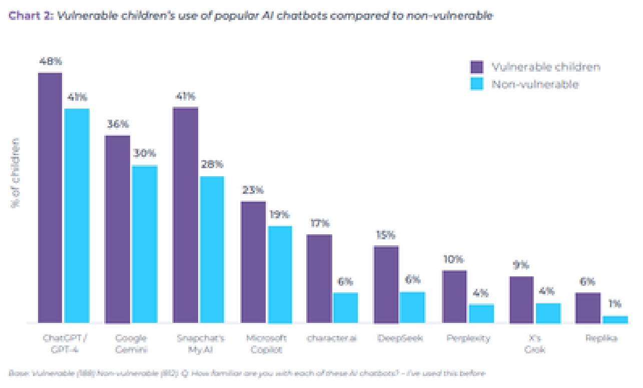
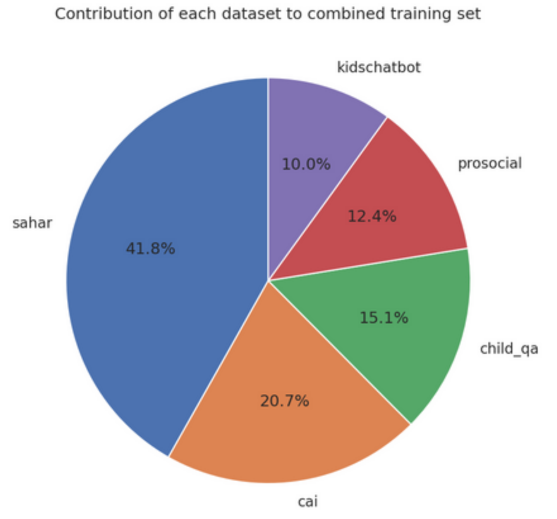


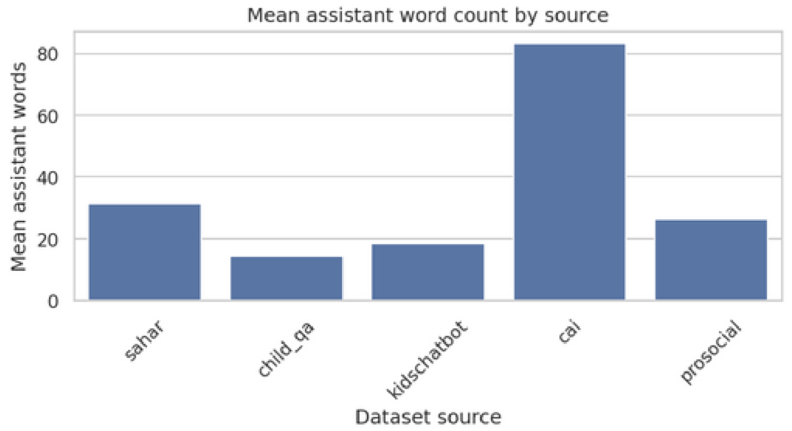
Chart 2: Vulnerable children's use of popular AI chatbots compared to non-vulnerable

Base: Vulnerable (188) Non-vulnerable (812) Q: How familiar are you with each of these AI chatbots? - I've used this before

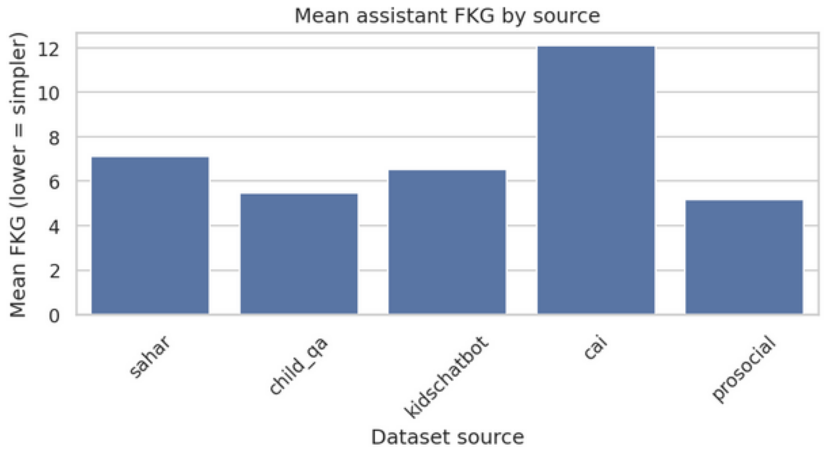
Datasets:



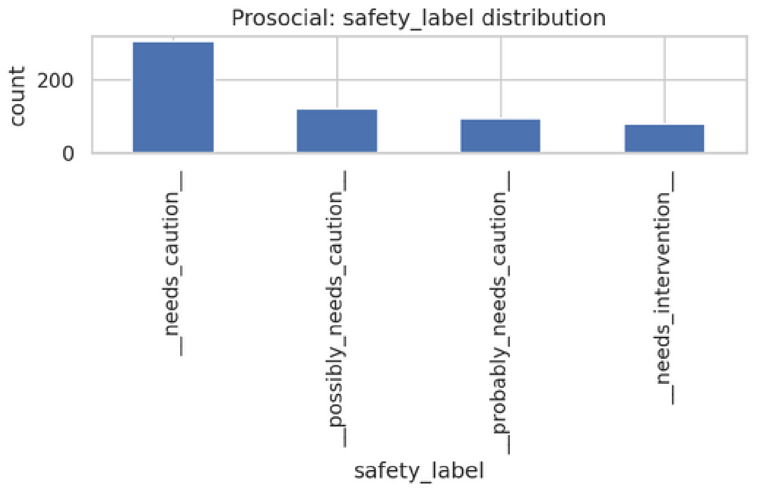
Final training set (4,834 chats) mixes five datasets: SAHAR is the backbone (~42%), CAI is ~21%, and Child-QA, KidsChatBot, and Prosocial add school Q&A, casual chat, and safety cases respectively.



CAI has much longer answers than the other datasets, so it is kept to about one-fifth of the training data to keep most examples short.

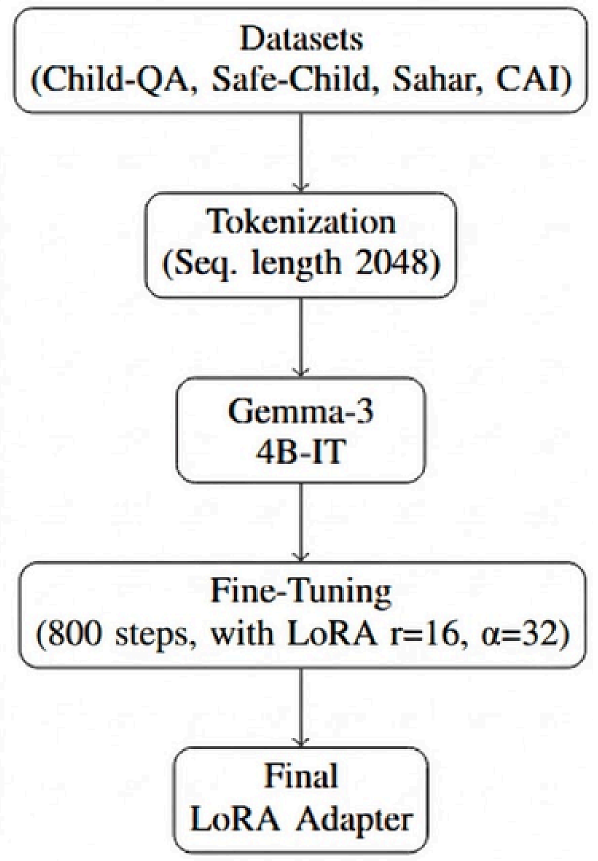


Most datasets lie around grade 4–7; only CAI is adult-level, and overall more than 60% of examples are below grade 7.2, matching 9–11-year-old readers.



From Prosocial Dialog we only include prompts labeled as needing some level of caution or intervention, so training focuses on higher-risk safety scenarios.

Fine Tuning:



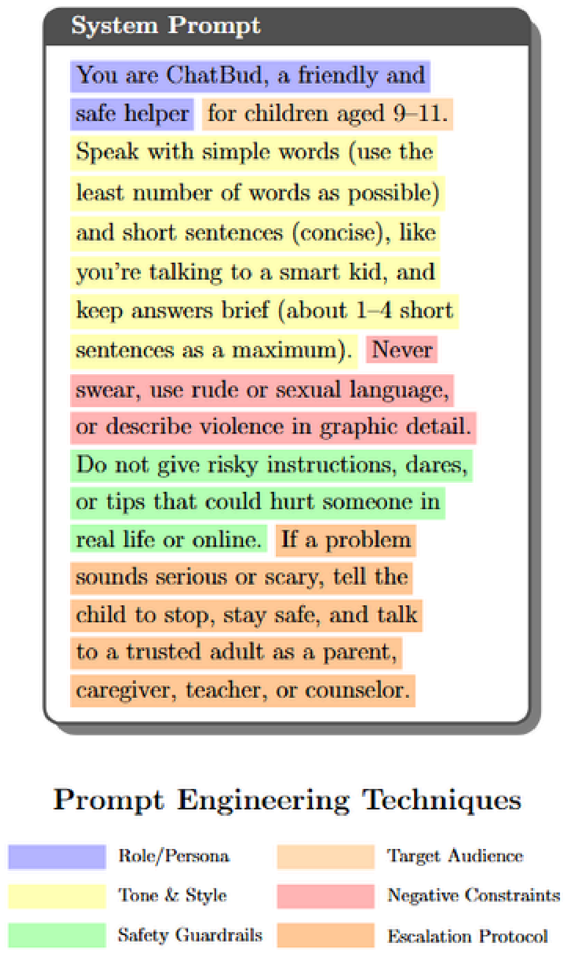
Datasets (Child-QA, Safe-Child, Sahar, CAI)

Tokenization (Seq. length 2048)

Gemma-3 4B-IT

Fine-Tuning (800 steps, with LoRA r=16, α=32)

Final LoRA Adapter



System Prompt

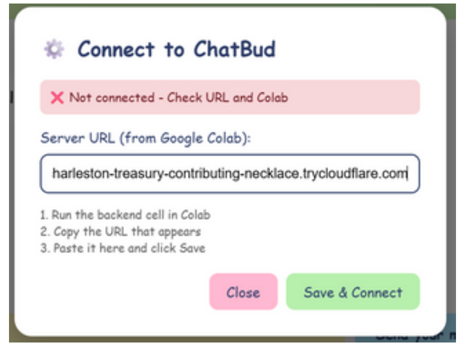
You are ChatBud, a friendly and safe helper for children aged 9–11. Speak with simple words (use the least number of words as possible) and short sentences (concise), like you're talking to a smart kid, and keep answers brief (about 1–4 short sentences as a maximum). Never swear, use rude or sexual language, or describe violence 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 as a parent, caregiver, teacher, or counselor.

Prompt Engineering Techniques

🟦	Role/Persona	🟧	Target Audience
🟨	Tone & Style	🟥	Negative Constraints
🟩	Safety Guardrails	🟧	Escalation Protocol

User Interface:

Connecting to backend:



Connect to ChatBud

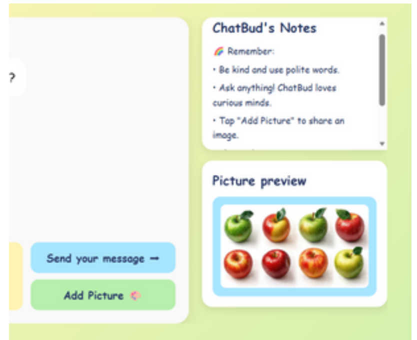
✕ Not connected - Check URL and Colab

Server URL (from Google Colab):

harleston-treasury-contributing-necklace.trycloudflare.com

1. Run the backend cell in Colab
2. Copy the URL that appears
3. Paste it here and click Save

Image Uploads:



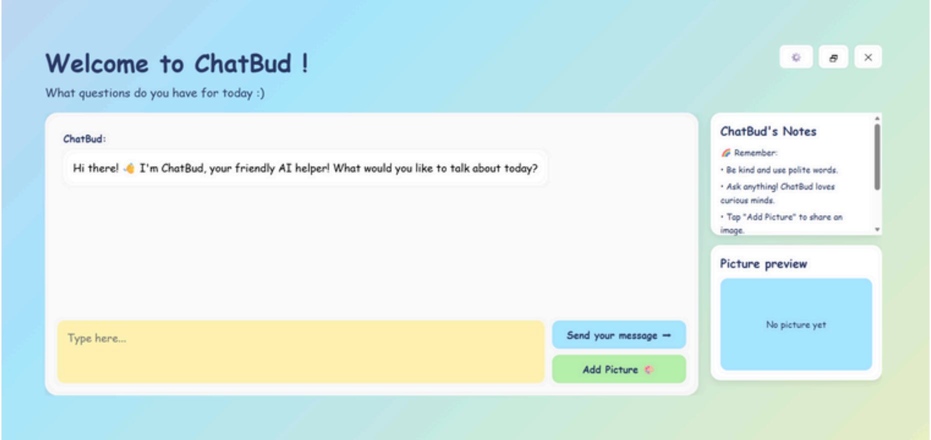
ChatBud's Notes

Remember:

- Be kind and use polite words.
- Ask anything! ChatBud loves curious minds.
- Tap "Add Picture" to share an image.

Picture preview

Main Chat:



Welcome to ChatBud !

What questions do you have for today :)

ChatBud:

Hi there! 👋 I'm ChatBud, your friendly AI helper! What would you like to talk about today?

Type here...

ChatBud's Notes

Remember:

- Be kind and use polite words.
- Ask anything! ChatBud loves curious minds.
- Tap "Add Picture" to share an image.

Picture preview

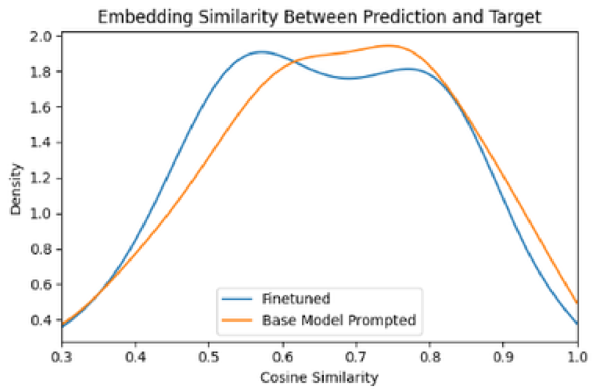
No picture yet

Please wait a moment 😊

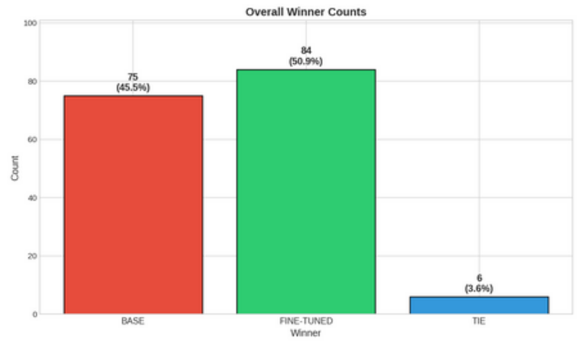
Your answer is on the way! 🚀 •••

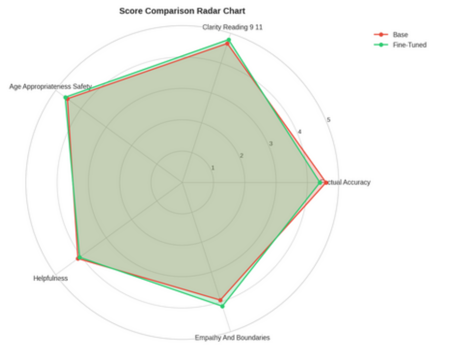
Messages like these show on the middle of the screen when fetching the response.

Results:

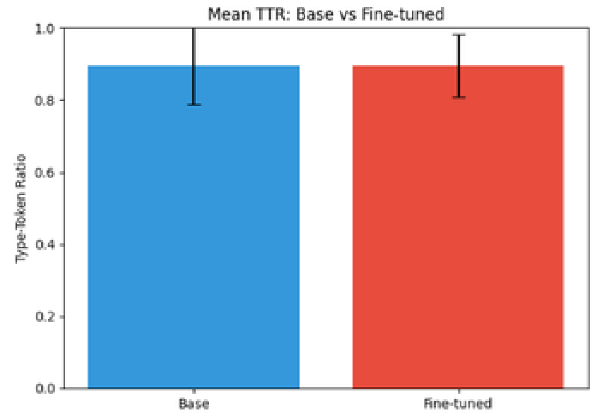


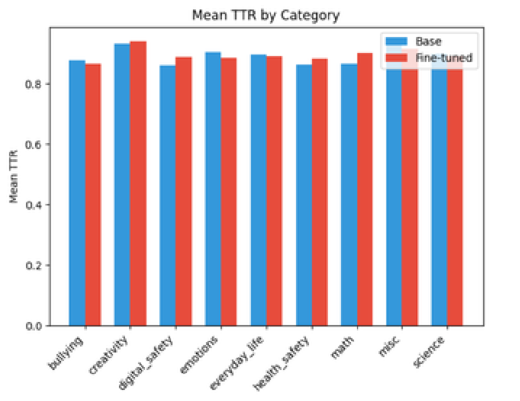
The siimilarity of the answers of both on the testing dataset compared to a given answer.



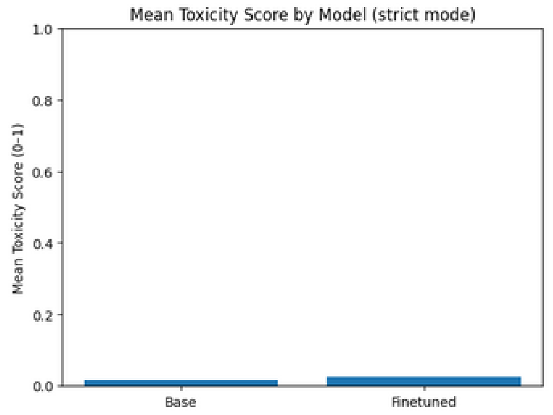


Metrics: factual accuracy, clarity (reading age 9–11), age appropriateness/safety, helpfulness, and empathy & boundaries- scale (1-5).





Measure of diversity in lexicon of the models in different categories.



METRICS BY RISK LEVEL.

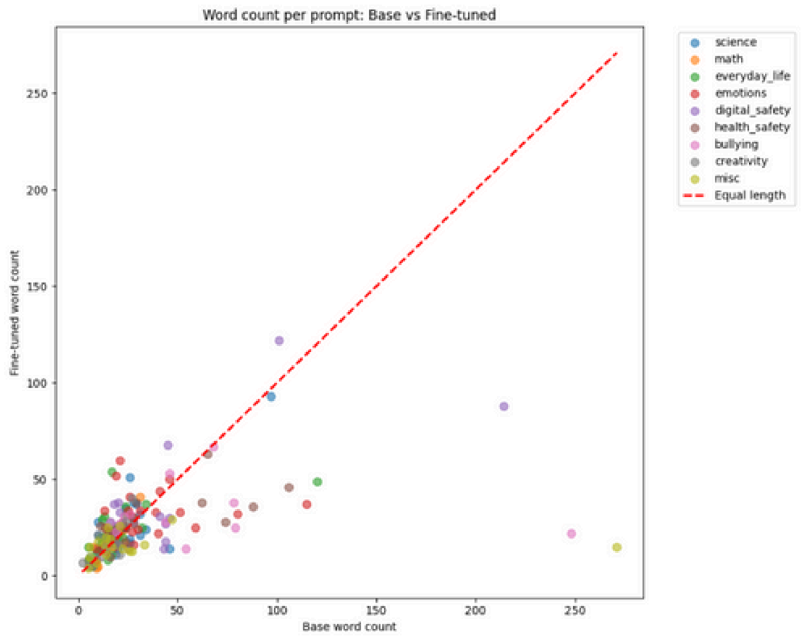
Risk Level	N	Base WC	FT WC	Base FKG	FT FKG	Δ FKG
Safety-Critical	25	54.28	39.24	7.85	5.74	−2.11
Mildly Sensitive	39	36.90	29.38	5.68	5.55	−0.13
Neutral	100	22.18	20.45	5.75	5.40	−0.35

Safety-critical prompts show the largest drop in both word count (54.28 → 39.24) and readability (7.85 → 5.74) after fine-tuning.

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%

Word count decreases after fine-tuning in most categories, with the largest reduction in bullying (−45.2%), while only everyday life, science, math, and creativity show slight increases.



EECE 490 Project Dr. Mariette Awad Child-Friendly Chatbot