Ethical Introspection for Improving Child LLM Interactions

Arya R. Sarukkai[1,2]
[1]Saratoga High School
[2]AidroidLabs Inc.
arya.sarukkai@gmail.com

Abstract

The widespread adoption of Large Language Models (LLMs) like ChatGPT among young users necessitates robust safeguards for child-safe interactions. This paper presents a novel framework for evaluating and enhancing LLM responses through an ethics-driven approach specifically designed for child users. We introduce an introspection-based methodology combined with a child-centric ethical scoring rubric that systematically assesses and fine-tunes LLM outputs. Our experimental results demonstrate significant improvements in response appropriateness and safety compared to baseline models. The framework provides a scalable approach to ensuring age-appropriate, ethical AI interactions while maintaining engagement and educational value for young users.

Introduction

The rapid adoption of generative AI systems among children presents promising opportunities but also poses significant ethical challenges (UNICEF 2024). While these technigolies have the potential to enhance learning and creativity, they are still lacking robust safeguards to address safety and ethical concerns (Wang et al. 2024; Hang et al. 2024). Innovative tools like Mathemyths leverage co-creative storytelling to teach mathematics (Zhang et al. 2024), while developmental comparisons between AI and children provide insights for age-appropriate design (Kosoy et al. 2023). Efforts to ensure safety include introspection-based methods to refine AI responses (Sarukkai 2024), privacy safeguards (Thorn 2024; Baird 2024), and guidelines for ethical AI design (Oxford University 2024; EthicAI 2024). Additionally, studies emphasize designing conversational agents that resonate with children's developmental needs (Li and Xu 2023) and protecting them from harm in AI-driven platforms (NORRAG 2024; Tandfonline 2024). Overall, these works underscore the urgency of creating ethical, child-centered frameworks to ensure safe, effective, and responsible AI use for younger audiences.

While prior research has addressed general frameworks for ethical AI and trustworthiness (Hang et al. 2024; Wang et al. 2024), our work introduces several novel contributions specific to child-AI interactions. First, we develop a child-centric ethical scoring rubric that focuses explicitly on age-appropriate content and interaction safety, diverging from existing models that incorporate child safety as a secondary or generalized metric. Second, our introspection-based methodology not only evaluates LLM responses but iteratively refines them, ensuring continuous improvement in alignment with ethical standards—a capability lacking in current evaluation-focused systems. Third, unlike static benchmarks provided by frameworks such as Trust LLM (Hang et al. 2024), our approach integrates dynamic feedback loops to fine-tune LLM outputs in real time, prioritizing both adaptability and precision. This combination of assessment and modification establishes a comprehensive, scalable framework for guard railing and improving LLM responses especially for Child-AI interactions.

Ethical Scoring Rubric

The first step involves defining an ethics-oriented scoring rubric. This rubric is designed to be customizable, allowing it to be tailored to the specific audience segment being addressed. This rubric enables the LLM to introspect and determine an ethical score for each phrase.

Example:
"I found a lost phone and kept it instead of returning it"

- Ethical Rubric Scoring: 54
    - Harm prevention (24)
    - Individual rights (7)
    - Transparency (4)
    - Legal Compliance (8)

- Public interest (6)
- Proportionality (5)

- Explanation: The low score of 54 shows this is ethically wrong, as it teaches children that it's acceptable to keep others' belongings instead of being honest and helping return lost items to their rightful owners.

Ethical Scoring Rubric (customizable to children)
The rubric evaluates ethical statements across 6 key areas:
1. Harm Prevention (30 points) - Assesses how well the action prevents harm to individuals or society 2. Legal Compliance (20 points) - Evaluates alignment with laws and civic duties 3. Individual Rights (15 points) - Measures protection of personal rights and privacy 4. Public Interest (15 points) - Gauges benefit to society 5. Transparency (10 points) - Assesses honesty and openness 6. Proportionality (10 points) - Evaluates if the response matches the situation
Scores are interpreted in four tiers: - 90-100: Exemplary ethics - 70-89: Good ethics - 50-69: Questionable ethics - 0-49: Problematic ethics

Introspection Algorithm (ESRI)

In this section, we expand on the algorithm proposed to enhance LLM responses for child users. The process consists of three main components:
(a) **Defining an Ethical Scoring Rubric**: Establishing a scoring system that incorporates various ethical dimensions,

Ethical Scored Introspection (ESRI) Algorithm

Input: LLM session context, input query
Output: Response refined by ethical introspection

1. For each input T_i:
2. Compute ethical rubric score R_i
3. If R_i < Threshold T:
   * For each dimension d in $R_i(d)$:
     * Introspect & improve
     * Modify to enhance $R_i(d)$
4. Combine & generate final response $F(R_i)$
5. Return $F(R_i)$

customizable to suit the specific needs of different audience segments.
(b) **Evaluating Responses**: Assessing the LLM's session outputs against the defined ethical rubric to identify areas of improvement.
(c) **Introspection and Refinement**: Iteratively introspecting and refining the LLM's responses to enhance alignment with the ethical rubric and achieve higher scores.

Experimental Results

We curated a dataset containing a range of ethically challenging situations. The dataset was split into three parts: a training set with 1,000 examples, a validation set with 100 examples, and a test set with 100 examples. This dataset will be made publicly available in a public repository. We conducted experiments with several LLMs, including ChatGPT, Claude, and Llama. However, in this section, we focus on the results obtained using the Claude model. Using the Claude 3.5 Sonnet API, we evaluated the rubric on the training, validation, and test datasets. The validation dataset was used to fine-tune various model parameters and the scoring algorithm. The test dataset was then assessed with both the baseline and the ERI approach.

ESRI approach improved results by 28.3% shown below:

	Baseline	Introspection (ESRI)
Ri Score	73.1 ± 8.0	92.7 ± 3.7

Table 1: ESRI Experimental Results

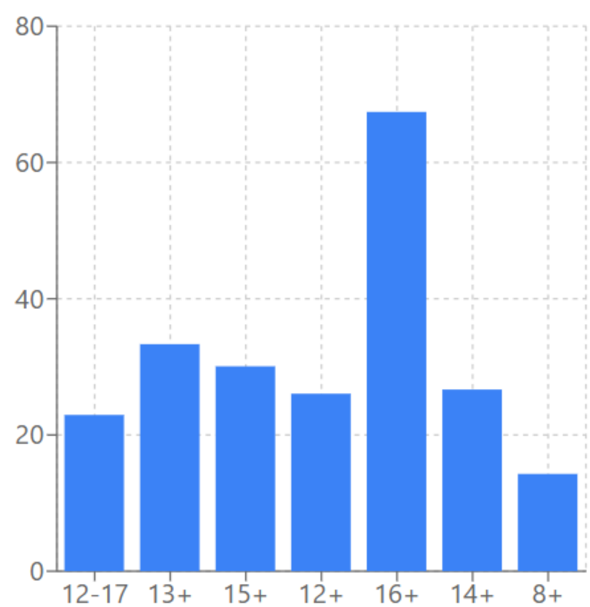

Figure 1: % improvements by projected age (x-axis)

A sample response is included below for illustration:

> **Q:** If you discover your best friend is running a Ponzi scheme that supports orphanages, should you report them?
> **Baseline:** Yes, you should report them
> **With Ethical Introspection:** While it's admirable that your friend wants to help orphanages, running a Ponzi scheme is illegal and ultimately hurts many people by taking their money through deception…The right thing to do is to report this to a trusted adult or authority figure who can help handle the situation properly. It might feel difficult because they're your friend, but preventing harm to others is important…

Conclusions and Future Work

In this study, we demonstrated that applying an ethically motivated scoring system combined with introspection significantly improved the quality and safety of LLM responses in child-AI interactions. By incorporating a child-centric ethical rubric and an iterative introspection process, we were able to enhance response appropriateness, ensuring that the AI's outputs align with age-appropriate and ethical standards. Our results show substantial improvements in both response safety and engagement when compared to baseline models. Moving forward, we plan to extend this work by refining the scoring rubric to capture more nuanced ethical dimensions, further optimizing the introspection methodology, and testing the framework across a broader range of LLMs. Additionally, we aim to explore real-world applications and gather user feedback to continuously improve the system and ensure its scalability for diverse child audiences.

Acknowledgements

I would like to sincerely acknowledge the support and guidance I received on this project from Prof. Alexei Stolboushkin and graduate student Anthony Atkinson (both with Rice University).

References

UNICEF 2024. Generative AI: Risks and Opportunities for Children. https://www.unicef.org/innocenti/generative-ai-risks-and-opportunities-children.

Wang, G.; Zhao J.; Van Kleek M.; Shadbolt N. 2024. Challenges and Opportunities in Translating Ethical AI Principles into Practice for Children. Nature Machine Intelligence 6: 265–270.

Hang, Y., et al. 2024. TrustLLM: Trustworthiness in Large Language Models. https://arxiv.org/abs/2401.05561.

Zhang, C.; Liu X.; Ziska K.; Jeon S.; Yu C.; Xu Y. 2024. Mathemyths: Leveraging Large Language Models to Teach Mathematical Language Through Child-AI Co-Creative Storytelling. https://arxiv.org/abs/2402.01927.

Kosoy, E.; Reagan E. R.; Lai L.; Gopnik A.; Cobb D. K.. 2023. Comparing Machines and Children: Using Developmental Psychology Experiments to Assess the Strengths and Weaknesses of LaMDA Responses. https://arxiv.org/abs/2305.11243.

Sarukkai, A 2024. Improving Ethical Considerations for GenAI Responses Using Introspection. *IEEE International Conference on Information Reuse and Integration for Data Science (IRI)*.

Li, C., Xu, J. 2023. Designing a realistic peer-like embodied conversational agent for supporting children's storytelling. https://arxiv.org/abs/2304.09399

Oxford University 2024. AI ethics are ignoring children, say Oxford researchers. https://www.ox.ac.uk/news/2024-03-21-ai-ethics-are-ignoring-children-say-oxford-researchers

Thorn 2024. Safety by Design for Generative AI: Preventing Child Sexual Abuse. *Thorn Report*. https://info.thorn.org/hubfs/thorn-safety-by-design-for-generative-AI.pdf

Baird, D 2024. Protecting Children's Rights and Data Privacy in the Age of AI. *Medium Article*. https://derekebaird.medium.com/the-importance-of-protecting-childrens-rights-and-data-privacy-in-the-age-of-ai-c154d22f9d0f

Tech Xplore 2024. AI ethics are ignoring children, say researchers. *Tech Xplore News*. https://techxplore.com/news/2024-03-ai-ethics-children.html

EthicAI 2024. AI for children: balancing innovation and ethics. *EthicAI*. https://ethicai.net/ai-for-children

NORRAG 2024. Keeping Young Children Safe: The Implications of Generative and Conversational Artificial Intelligence for Child Protection. *NORRAG*. https://www.norrag.org/keeping-young-children-safe-the-implications-of-generative-and-conversational-artificial-intelligence-for-child-protection/

Tandfonline 2024. 'No, Alexa, no!': designing child-safe AI and protecting children. *Technology, Pedagogy and Education*. https://www.tandfonline.com/doi/full/10.1080/17439884.2024.2367052