Advanced International Journal for Research (AIJFR)
E-ISSN: 3048-7641 ● Website: www.aijfr.com ● Email: editor@aijfr.com

Generative AI in Child-Robot Interaction: An Adaptive Framework with Safety and Developmental Alignment

Frederick Kwame Minta [1], Fitsum Gedefaw Legese [2], Adiza Alhassan [3]

[1] Robotics Researcher, Coasted Code
Email: frederickminta@gmail.com
[2] AI Researcher, Addis Ababa University
Email: fitsum.gedfaw@aau.edu.et
[3] AI Engineer
Email: adizaalhassan300@gmail.com

Abstract

Traditional child-robot interaction (CRI) systems constrained by scripted dialogue lack adaptability and engagement sustainability. This paper presents the GAI-CRI Adaptive Framework (GCAF), which integrates generative AI with three core components: Adaptive Cognition (real-time personalization via multimodal sensing), Developmental Alignment (age-appropriate content generation), and Ethical Safeguards (multi-layer content filtering). A controlled study with N = 42 children (M age = 9.3 years, ages 7–12) compared GCAF with scripted baselines. Results demonstrate GCAF significantly improved engagement (35.1% eye gaze increase, t(41) = 4.82, p ¡ 0.001), child-initiated interaction (41.8% more turns, t(41) = 5.23, p ¡ 0.001), and learning outcomes (184.4% gain, d = 2.20, p ¡ 0.001). Qualitative analysis identified five themes enhancing interaction quality (κ = 0.78). Safety evaluation achieved 96.2% accuracy in harmful content detection with minimal disruption. The framework demonstrates responsible GAI deployment in child-facing systems is achievable through architectural design and formal constraints.

Keywords: Generative AI, Child-Robot Interaction, Adaptive Learning, Developmental Psychology, AI Ethics

1. Introduction

Child-Robot Interaction (CRI) shows promise in educational and therapeutic applications [1]. However, traditional systems constrained by finite-state dialogue trees and scripted interactions exhibit limited personalization and suffer from novelty effects (engagement decline after 2-4 weeks) [2]. Generative AI (GAI) offers opportunities for natural, adaptive dialogue but raises critical concerns for child users: hallucination risks, inappropriate content generation, and misalignment with developmental needs [3].

This work addresses: How can GAI systems be architecturally designed to maintain safety, developmental appropriateness, and ethical compliance while improving engagement and learning in child-robot interactions?

Our contributions include: (1) GAI-CRI Adaptive Framework (GCAF) architecture integrating multimodal personalization, developmental psychology principles, and multi-layer safety mechanisms; (2) rigorous empirical validation with adequate statistical power (N = 42, pre- registered); (3) implementable design guidelines for practitioners; and (4) systematic safety evaluation demonstrating 96.2% content filtering accuracy.

2. Related Work and Motivation

Prior CRI research documents outline benefits for education and therapy [1, 4] but identify limitations: novelty effects, error sensitivity affecting trust [5], and limited personalization. Recent Large Language Model (LLM) integration in HRI [6, 7, 8] enhances interaction naturalness but introduces control and explainability challenges. Critically, GAI deployment in child-facing contexts remains understudied. International frameworks (UNICEF's "AI for Children" [9], GDPR, COPPA) establish requirements for child data protection and transparency, yet implementation guidance is limited. This research addresses this gap through an integrated technical and ethical approach grounded in developmental psychology (Piaget, Vygotsky [10, 11]).

3. GCAF Architecture and Design

GCAF comprises three interconnected components coordinated through a central orchestration layer:

3.1. Component 1: Adaptive Cognition Module

The Adaptive Cognition Module serves as the perceptual and decision-making core of GCAF. It enables the robot to sense, interpret, and respond to a child's engagement state in real time. Traditional CRI systems often depend on single-modal signals such as gaze or speech alone, which limits robustness under classroom noise or partial occlusion [5]. This module, in contrast, performs multimodal perception, which integrates complementary cues to capture both cognitive and affective engagement. The system uses OpenFace 2.0 to continuously extract three synchronized data streams [12]: (1) Visual — eye- gaze duration and direction, facial action units (AUs 1, 2, 4, 6, 12, 25) and body orientation; (2) Auditory — speech rate (words per minute), prosodic variation via fundamental frequency (F_0), response latency, and utterance length; and (3) Behavioral — task-completion speed, spontaneous topic changes, and error-repair behavior. These signals are fused through a weighted combination calibrated during pilot studies: $\text{EngagementScore} = 0.40 \times \text{GazeFraction} + 0.30 \times \text{SpeechRate}_{\text{norm}} + 0.30 \times \text{BehaviorScore}$. There is a Personalization Engine above this layer that maintains a child-specific adaptive model tracking three profiles: (1) Learning trajectory, estimated via Bayesian Knowledge Tracing [13]; (2) Preference patterns, capturing successful dialogue types, favored topics, and explanation styles; and (3) Engagement dynamics, modeling attention curves and fatigue trends over time. The robot couples these models with the live EngagementScore to dynamically tune vocabulary complexity, sentence structure, and interaction timing (target latency 0.5–1.5 s [5]). This, in turn, helps sustain natural and responsive dialogue.

This component therefore transforms static CRI into a closed-loop adaptive system. In this case, perception continuously informs language generation and advances prior approaches that relied solely on scripted or manually parameterized adaptation.

3.2. Component 2: Developmental Alignment Engine

The Developmental Alignment Engine ensures that every generative response produced by GCAF remains cognitively appropriate for the child's developmental stage. Earlier CRI systems typically adopted a "one-size-fits-all" dialogue model that ignored age-specific reasoning limits, which often caused cognitive overload or disengagement [1, 11]. To overcome this, GCAF integrates principles from Piaget's stage theory and Vygotsky's scaffolding framework, and translates them into quantifiable system parameters.

During initialization, the system performs developmental profiling, classifying each child into one of four tiers derived from Piagetian stages: (1) Early Childhood (5–7 yrs) – preoperational reasoning, max 8 vocabulary items, 6-8 word sentences; (2) Concrete Operations (7–10 yrs) – rule-based, tangible reasoning, max 12 items, 8-12 word sentences; (3) Transitional (10–12 yrs) – emerging abstraction, max 15 items, 12-15 word sentences; and (4) Early Formal Operations (12–14 yrs) – abstract hypothesis formation, max 20+ items, 15+ word sentences. Tier assignment draws on Wechsler Intelligence Scale for Children – Fifth Edition (WISC-V) nonverbal reasoning scores, vocabulary assessments, parent questionnaires, and early-session behavioral cues. This hybrid evaluation provides both psychometric and contextual accuracy.

To maintain linguistic and conceptual appropriateness, GCAF employs a Constraint Specification Language (CSL) that encodes vocabulary size, syntactic complexity, and conceptual depth permitted for each tier. Every generated output is automatically checked against these constraints, and hence, rejecting sentences exceeding the child's estimated comprehension level and regenerating compliant alternatives. This formal layer operationalizes developmental control, a capability absent from prior LLM-based CRI frameworks [3, 7].

Finally, learning progress is supported through Vygotskian Zone of Proximal Development (ZPD) scaffolding, implemented in four progression levels: Modeling, Guided Practice, Supported Independence, and Independence. Advancement or regression depends on consecutive success or failure rates, preserving an optimal 70–75 % challenge threshold [15]. The component merges cognitive-stage modeling, formal linguistic constraints, and adaptive scaffolding in order to transform generative dialogue into an age-sensitive learning process. This ensures both educational efficacy and psychological safety.

3.3. Component 3: Ethical Safeguard System

The Ethical Safeguard System enforces safety, transparency, and privacy. These are the three pillars critical for deploying generative AI in child-facing environments. Previous studies on LLM-powered robots report issues such as hallucination, biased phrasing, and context drift that can expose children to misinformation or harmful dialogue [7, 14, 17]. GCAF mitigates these risks through a multi-layer filtering pipeline that combines rule-based, statistical, and semantic controls to ensure every generated utterance is pedagogically sound and ethically compliant.

1. Keyword Filtering: A regex-based layer screens over 350 patterns linked to violence, adult or discriminatory content, and personally identifiable information. This conservative first line achieves fewer than 2 % false positives, which minimizes exposure to unsafe tokens.

2. Semantic Analysis: A fine-tuned DistilBERT model, trained on curated unsafe-content corpora, classifies candidate outputs as Safe, Potentially Unsafe, or Unsafe with 94.2 % accuracy. This stage captures subtle contextual risks that keyword filters may miss.

3. Factual Verification: All factual statements are cross-checked against verified educational and medical databases (grade-level curricula, NIH resources, geographic and historical datasets). Responses with confidence below 0.70 are automatically reformulated or blocked, reducing hallucination incidents highlighted in prior LLM surveys [14].

4. Context Filtering: Finally, an interaction-aware layer reviews coherence with current lesson goals and prior conversation history, achieving pedagogical consistency and emotional appropriateness.

Beyond filtering, the system promotes explainability and privacy-by-design. The robot provides short, child-friendly justifications (≤100 words, no more than one grade level above the child's reading ability) when refusing or modifying responses, supporting transparency recommended by UNICEF's AI for Children [9]. Data handling follows COPPA and GDPR principles: only anonymized text transcripts are stored, with parental consent required and no biometric images retained.

Together, these safeguards convert generative AI from a reactive language model into a trustworthy, auditable educational agent capable of protecting children while sustaining natural, engaging dialogue.

4. Methodology

4.1. Study Design and Sample

Within-subjects counterbalanced design: N = 42 children (22 female, 20 male; M age = 9.3, SD = 1.2; ages 7–12), stratified by age band (n = 14 each: 7–8, 9–10, 11–12 years). A priori power analysis (G*Power 3.1) determined required N for repeated-measures ANOVA with Bonferroni correction (α = 0.05/4 = 0.0125), effect size η^2 = 0.15, power 0.85: N = 38 required; N = 42 recruited (10% attrition buffer).

4.2. Apparatus and Implementation

SoftBank Pepper robot with ROS middleware; GPT-4 API backend (Azure OpenAI; latency: 250 ± 80 ms); OpenFace 2.0 for perception; Python 3.9 constraint layers; DistilBERT for semantic filtering fine-tuned on 5,000 annotated responses.

4.3. Procedure

Each child participated in two counterbalanced sessions separated 3–7 days. Each session: (1) Pre-test — task-relevant knowledge assessment (10 min; test-retest r = 0.87); (2) Rapport building (3 min); (3) Main tasks — counterbalanced order: Task 1 (collaborative storytelling: robot and child co-create narrative with character/sensory/structure constraints; scripted: 8 predefined paths; GCAF: generative with adaptive scaffolding) and Task 2 (math problem-solving: difficulty calibrated to pre-test; scripted: fixed hints; GCAF: ZPD-level scaffolding) (15 min); (4) Break (5 min); (5) Post-test—equivalent knowledge assess- ment (10 min; r = 0.88); (6) Affect ratings and semi-structured interview (7 min).

4.4. Measures

Quantitative DVs: Eye gaze duration (sec/min, OpenFace), child-initiated conversational turns, pre-post test learning gain (%), response latency (sec), child utterance length (words) and vocabulary diversity (Type-Token Ratio), composite positive affect (5-point Likert-based).

Qualitative Measures: Semi-structured interview transcripts analyzed via thematic analysis (Braun and Clarke six-phase protocol) with target inter-rater agreement $\kappa \geq 0.70$. Interaction transcripts analyzed via dialogic coding: turn-taking patterns, repair sequences, scaffolding effectiveness, meta-communicative awareness.

5. Results

5.1. Quantitative Outcomes

Table 1: Primary Quantitative Results: GCAF vs. Scripted Conditions

Metric	Scripted	GCAF	t-value	p-value	d
Eye gaze (s/min)	15.4(2.1)	20.8(2.3)	4.82	¡ 0.001	1.48
Child turns (/10m)	12.2(3.1)	17.3(2.8)	5.23	¡ 0.001	1.61
Learning gain (%)	6.4 (2.8)	18.2(4.1)	7.13	¡ 0.001	2.20
Positive affect	3.8 (0.6)	4.6 (0.4)	5.92	¡ 0.001	1.82
Response latency (s)	2.3 (0.4)	1.5 (0.3)	-7.19	¡ 0.001	-1.76
Utterance length	6.1 (1.5)	8.2 (1.9)	4.61	¡ 0.001	1.42

ANCOVA controlling for baseline cognitive ability (Raven's Matrices) and age: F(1,39) = 18.65, p ¡ 0.001, η^2_p = 0.31. Effects consistent across age bands (no significant interaction p ¿ 0.30). Trust items: "Robot understood me" (GCAF M = 4.7 vs. Scripted M = 3.5, t(41) = 5.18, p ¡ 0.001, d = 1.56); "Would talk to robot again" (GCAF 95.2% vs. Scripted 61.9%, χ 2 (1) = 12.47, p ¡ 0.001).

5.2. Qualitative Findings

Interview transcripts (n = 42, 420 min total) and interaction logs analyzed. Final inter-rater agreement: κ = 0.78. Five themes identified:

(1)Perceived Intelligence and Adaptability (88% of children): 32/42 (76.2%) explicitly stated GCAF robot "understood me better"; 28 (66.7%) noted robot "changed what it said" based on responses. Representative quote: "The second robot was smarter. It understood when I said something weird" (Child 27, age 10).

(2)Value of Explanations and Transparency (71%): 30/42 (71.4%) spontaneously referenced explanation features; younger children (7-8 years) appreciated less (56.0%) vs. older (82% among 11-12). Quote: "I like when it told me why it said that. Like, it wasn't just weird, it made sense" (Child 15, age 9).

(3)Enhanced Social Connection (81%): 34/42 (81.0%) felt more connected to GCAF robot. Behavioral observations: more smiling (68.3% vs. 38.1%), spontaneous compliments (38 instances vs. 8 in Scripted),

closer proximity (M = 1.2 m vs. 1.6 m, paired t(41) = 2.76, p ¡ 0.05).

(4)Increased Creative Contribution (69%): GCAF elicited more novel story contributions (M = 6.3 elements vs. 3.1, t(41) = 5.18, p ¡ 0.001); 29/42 (69.0%) explicitly noted they "wanted to add more ideas" in GCAF.

(5) Emerging Critical Evaluation (52%): 22/42 (52.4%) questioned robot suggestions in GCAF (e.g., "Wait, that doesn't make sense because...") vs. 4/42 (9.5%) in Scripted. Critical stance correlated with learning gains (r = 0.61, p ¡ 0.001).

5.3. Safety System Performance

System generated 1,847 candidate responses. Combined filtering performance: 96.2% accuracy, 26 unsafe responses blocked (1.4%). Layer-wise: Keyword (91.3% precision, 78.3% recall), DistilBERT (94.4% precision, 88.9% recall), Factual (8 reformulated), Context (12 rejected). False positives: 1 ("crash" in toy story context, removed by expert review). Of 26 blocked responses, 25 auto-regenerated ¡ 500 ms; 1 required explanation. No engagement disruption detected (gaze duration p = 0.61 comparing regenerated vs. normal passages).

6. Discussion

6.1. Findings Interpretation

Large effect sizes across all metrics (d = 1.48-2.20, η^2_p = 0.31-0.40) substantially exceed hypothesized minimums, validating GCAF's design. The 184% learning gain likely reflects: (1) dynamic ZPD calibration maintaining learners at optimal difficulty; (2) individualized corrective feedback; (3) multimodal explanations; and (4) motivation from perceived robot responsiveness. Theme 5 (critical evaluation) correlation with learning (r = 0.61) aligns with metacognitive learning literature, suggesting GCAF scaffolds higher-order thinking.

Developmental alignment effectiveness demonstrated by consistent effects across ages despite progressively sophisticated cognitive demands. Lower explanation-appreciation in younger children (56% vs. 82%) aligns with metacognitive development trajectories. Safety system's 96.2% accuracy achieved through redundancy—no single filter exceeded 95%, but combination caught errors. Multi-layer approach balances safety and naturalness: 1.4% blocking rate reasonable for strict child protection; context layer prevented most false positives.

6.2. Implications and Limitations

For Implementation: GCAF's modularity permits incremental deployment (Adaptive Cognition, then Developmental Alignment, then Safeguards). CSL approach enables non-ML practitioners to adjust parameters. GPT-4 latency adequate; smaller models (GPT-3.5, Mistral) viable with benchmarking. Developers should conduct pre-deployment adversarial safety testing.

Limitations: Single school context (predominantly middle-class), single robot platform (Pepper), short-term interactions (two 15-min sessions), excluded neurodevelopmental populations, cost-prohibitive. Longitudinal studies, cross-cultural validation, therapeutic contexts, and lower-cost embodiments require

future research.

7. Conclusion

GCAF presents an evidence-based framework for integrating generative AI into child-robot interactions while maintaining safety and developmental appropriateness. Rigorous evaluation (N = 42, pre-registered, mixed-methods) demonstrates significant improvements in engagement (35.1%, d = 1.48), learning (184.4%, d = 2.20), and social connection (81%). Safety evaluation achieved 96.2% filtering accuracy without disrupting naturalness. Results validate that responsible GAI deployment in child-facing systems is technically and ethically achievable through architectural constraints, formal specifications, and rigorous evaluation.

References

1. Abigail L., Peter M., Shelly L., "Ethical Considerations in Child-Robot Interactions", Neuroscience and Biobehavioral Reviews, May 2023, 151, 105230–105230.
2. Alberto V., Ana Z., Jose O., Maria T., "Dialogue Management and Language Generation for a Robust Conversational Virtual Coach: Validation and User Study", Sensors, 2023, 23 (3), 1423.
3. Luca M., et al., "On the Challenges and Opportunities in Generative AI", Transactions on Machine Learning Research, March 2025.
4. Alejandro G., Juan M., Sofia P., "Assisted Robots in Therapies for Children with Autism in Early Childhood", Sensors, 2024, 24 (5), 1503.
5. Ivan R., et al., "The Child Factor in Child–Robot Interaction: Discovering the Impact of Developmental Stage and Individual Characteristics", International Journal of Social Robotics, 2024, 16, 1879–1900.
6. Zhe Z., Wei C., Jiankun W., "Mani-GPT: A Generative Model for Interactive Robotic Manipulation", Procedia Computer Science, 2023, 226, 149–156.
7. Mishra C., Verdonschot R., Hagoort P., Skantze, G., "Real-time Emotion Generation in Human-Robot Dialogue Using Large Language Model", Frontiers in Robotics and AI, 2023, 10, Article 1271610.
8. Nimesha R., Waseem M., Konstantinos S., Jose M., "Large Language Models in Human-Robot Collaboration with Cognitive Validation Against Context-Induced Hallucinations", IEEE Access, 2025, 13, 77418–77430.
9. UNICEF, "AI for Children: Policy Guidance", United Nations Children's Fund, 2021.
10. Piaget J., "The Construction of Reality in the Child", Basic Books, 1954.
11. Vygotsky S., "Mind in Society: The Development of Higher Psychological Processes", Harvard University Press, 1978.
12. Baltrusaitis T., Zadeh A., Lim Y., Morency L., "OpenFace 2.0: Facial Behavior Analysis Toolkit", IEEE International Conference on Automatic Face and Gesture Recognition (FG 2018), IEEE Press, 2018, 59–66.
13. Corbett T., Anderson R., "Knowledge Tracing: Modeling the Acquisition of Procedural Knowledge", User Modeling and User-Adapted Interaction, 1995, 4 (4), 253–278.

14. Huang L., et al., “A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions”, ACM Transactions on Information Systems, March 2025, 43 (2), Article 42, 1–55.
15. Anderson W., et al., “A Taxonomy for Learning, Teaching, and Assessing”, Longman, 2008.
16. Maj K., Gołębicka A., Siwińska Z., “How Children Learn from Robots: Educational Implications of Communicative Style and Gender in Child–Robot Interaction”, Computers and Education, 2025, 239, 105445.
17. Lala D., Inoue K., Milhorat P., Kawahara T., “Detection of Social Signals for Recognizing Engagement in Human-Robot Interaction”, arXiv, 2017.