TOWARDS AI-DRIVEN SOCIALLY ASSISTIVE VIRTUAL ROBOTS FOR PERSONALIZED EARLY CHILDHOOD EDUCATION: AN INVESTIGATION OF FEASIBILITY, USABILITY, AND PARENT-SUPERVISED CONFIGURATION

By

IBRAHIM ABBAS

A dissertation submitted in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY IN ELECTRICAL AND COMPUTER ENGINEERING

2025

Oakland University
Rochester, Michigan

Doctoral Advisory Committee:
Subramaniam Ganesan, Ph.D., Chair
Osamah Rawashdeh, Ph.D.
Debatosh Debnath, Ph.D.
Harvey Qu, Ph.D.

AKNOWLEDGMENTS

To Syria, France, and the United States of America, each country shaped me, taught me, and carried me forward on this journey.

For the Syrian people, may peace replace suffering and hope replace despair.

To my wife and children, with deep love and gratitude to my wife for giving me the time and conditions to focus on my studies.

To my parents, who always believed in me and gave me constant motivation.

To my sisters, especially my oldest sister, and my brother-in-law, who helped me greatly at the beginning and guided me to this opportunity, making this dream possible.

And finally, to Oakland University, the Department of Electrical and Computer Engineering, my supervisor Dr. Subra, Dr. Louie and Dr. Osamah for their invaluable guidance, support, and encouragement throughout this journey. A special thank you also to Dr. Debnath, Dr. Zou, and Dr. Qu for their support and advice.

ABSTRACT

TOWARDS AI-DRIVEN SOCIALLY ASSISTIVE VIRTUAL ROBOTS FOR PERSONALIZED EARLY CHILDHOOD EDUCATION: AN INVESTIGATION OF FEASIBILITY, USABILITY, AND PARENT-SUPERVISED CONFIGURATION

By Ibrahim Abbas

Defense Committee Chair: Subramaniam Ganesan

This paper presents an extensive investigation into the adoption of AI-driven Socially Assistive Virtual Robots (SAVRs) in simulation-based environments for early childhood education. By integrating child-focused feasibility data (engagement, speech recognition, success rates) with parent-focused usability data (task configuration challenges, speech misinterpretations, AI clarity), we analyze how eight children (ages 3–4) and ten parents used a ChatGPT-powered simulation framework to practice counting and letter recognition at home.

Child results show that 4-year-olds achieved near-perfect task completion with minimal frustration, while 3-year-olds faced more difficulties due to incomplete articulation, shorter attention spans, and repeated speech recognition errors. Parent findings reveal moderate setup complexity (40% described it as “tedious”), frequent manual overrides for speech errors (70% cited speech recognition as a major frustration), and a strong preference (80%) for cost-effective simulation over expensive physical SARs, provided usability improvements are made.

We explore literature on Socially Assistive Robots (SARs), child-specific speech recognition challenges, AI-based adaptive learning, and prior human-robot interaction (HRI) studies, situating our work in the context of both physically embodied SARs and simulation-based agents that increasingly fulfill similar pedagogical roles [1], [2], [3], [4]. The methodology details the CoppeliaSim environment, ChatGPT-based real-time adaptation, multi-threaded speech orchestration, and the parent configuration interface. The results section includes tables (I–III) summarizing child performance, supplemented with in-depth analysis. We then present a discussion linking child engagement patterns to parent usability concerns, culminating in recommendations for child-specific ASR, multimodal input, short session durations, wizard-style interfaces, and context-aware AI prompts.

Future directions explore hybrid physical-virtual usage, specialized acoustic modeling for 3-year-olds, and scaling to larger, more diverse samples. Overall, the study addresses child feasibility and parent usability in a unified manner, underscoring how speech recognition and user-friendly design can support wide-scale implementation of AI-driven SAVRs at home.

Table of Contents

LIST OF ABBREVIATIONS

AI – Artificial Intelligence

API – Application Programming Interface

ASR – Automatic Speech Recognition

CASA – Computers Are Social Actors

EdTech – Educational Technology

GUI – Graphical User Interface

HRI – Human-Robot Interaction

LED – Light Emitting Diode

LLM – Large Language Model

NAO – NAO Robot

PC – Personal Computer

SAR – Socially Assistive Robot

SAVR – Socially Assistive Virtual Robot

STT – Speech-to-Text

TAM – Technology Acceptance Model

TTS – Text-to-Speech

UI – User Interface

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1 Background and Context

Early childhood, typically ranging from 3 to 5 years old, remains one of the most pivotal and influential phases in human development. During these formative years, children not only acquire fundamental skills in speech, thinking, and social interaction, but they also develop the cognitive, emotional, and social foundations that profoundly shape their long-term academic and life trajectories [5], [6], [7]. Research consistently demonstrates that high-quality, play-based, and adaptive educational interventions during early childhood can lead to significant improvements in numeracy, emergent literacy, self-regulation, problem-solving, and cooperative behaviors [8], [9]. These early competencies are critical, as they often serve as the foundation for later academic success, reducing the risk of future learning difficulties and enabling smoother transitions into formal schooling.

Moreover, early childhood is characterized by rapid social and emotional development. Young children are particularly sensitive to interactive, engaging experiences that provide immediate, positive reinforcement. Studies show that when children receive prompt feedback and dynamic support, they are more likely to remain focused, motivated, and eager to learn [6], [10], [11], [12]. This immediacy not only reinforces learning but also bolsters a child’s self-esteem and willingness to engage in further challenges. On the other hand, delays or inconsistencies in feedback can quickly lead to disengagement, thereby curtailing the learning process and undermining developmental progress.

In this context, Socially Assistive Robots (SARs) have emerged as an innovative tool to support early education. Over the past decade, physical SAR platforms, such as NAO, Pepper, and other humanoid or semi-humanoid devices, have been introduced into classrooms and therapeutic settings. These robots utilize dynamic gestures, expressive verbal communication, and customized feedback to create an interactive learning environment that captures children's attention and stimulates active participation [13], [14]. Previous research indicates that children often perceive these robots as friendly, peer-like companions, which encourages them to engage in educational activities such as counting, letter identification, storytelling, and even social-emotional learning exercises [3], [7]. The embodied nature of physical robots with features like arm waving, head nodding, and expressive LED eyes have been shown to enhance children's curiosity and sustain their engagement over time.

Despite these promising outcomes, the widespread deployment of physical SARs is limited by several significant factors. The high cost of advanced robotics hardware, often amounting to thousands or tens of thousands of dollars, poses a significant barrier to large-scale adoption. Additionally, these physical systems require ongoing maintenance, including battery management, sensor calibration, and periodic software updates, as well as the need for specialized technical staff to manage and troubleshoot the devices [3], [7]. Such requirements render physical SARs impractical for many typical families and under-resourced educational institutions, restricting their use largely to research labs or short-term pilot programs.

In response to these challenges, there has been a growing shift toward simulation-based SARs. These systems are entirely software-driven, replicating the core “social” components of physical robots, such as speech, gestures, emotional expressions, and interactive adaptivity, on standard computing devices like PCs or tablets [12], [15]. By eliminating the financial and logistical burdens associated with physical hardware, simulation-based SARs offer a scalable, cost-effective alternative that can bring advanced tutoring experiences into a much wider range of home and classroom settings. Families can deploy a virtual “humanoid” avatar that engages children in brief, interactive, voice-based tasks, thus democratizing access to cutting-edge educational technology.

The integration of advanced artificial intelligence further enhances the potential of simulation-based SARs. In particular, large language models such as ChatGPT can be harnessed to interpret children’s spoken responses in real time, generate context-aware prompts, and dynamically adjust the difficulty of tasks to suit each child’s performance level [8], [16], [17]. This AI-driven adaptivity allows a “virtual NAO” to operate as a digital tutor, a responsive, peer-like companion that can offer clarifications, provide tailored feedback, and even expand the content of the lesson as needed. The integration between simulation technology and AI not only addresses issues of cost and scalability but also opens new avenues for personalized learning experiences that are finely tuned to individual developmental needs.

However, two major obstacles must be overcome to ensure the success of these systems in home settings:

1. **Child Engagement:** A central concern is whether a purely on-screen robot can engage young children as effectively as a physical robot. The lack of tangible presence might affect how "real" the robot feels, potentially diminishing a child's willingness to interact. Moreover, conventional speech recognition systems, often optimized for adult speech, may struggle with the unique articulation patterns and "kid talk" of 3–5-year-olds (e.g., mispronunciations such as "fwee" or "fo-wuh"). Repeated misrecognitions can lead to negative feedback, causing children to lose confidence and disengage from the learning activity [6], [18], [19], [20].

2. **Parent-Supervised Usability:** In a home context, parents play a critical role as facilitators. They are responsible for installing and configuring the system, selecting appropriate tasks (e.g., counting or letter recognition), calibrating speech thresholds, and intervening when the system fails to understand the child's speech. If the interface is too complex or if the AI prompts are ambiguous, parents may become frustrated and ultimately abandon the technology. On the other hand, a user-friendly, wizard-based or template-driven interface can empower parents to tailor the system to their child's needs, effectively bridging gaps in articulation and cognitive readiness [21], [22], [23]. This ease of use is essential for ensuring that technology can be incorporated into everyday routines without imposing excessive burdens on busy families.

By integrating insights from both child engagement metrics and parent usability feedback, this paper aims to reveal how the success of an AI-driven, simulation-based SAR system is deeply intertwined with both its technological capabilities and the active role of the parent as a facilitator. Our analysis highlights that improved speech recognition and intuitive, user-friendly design are indispensable for scaling these systems for home-based use, ensuring they are both cost-effective and accessible for long-term educational support.

1.2 Main Objectives

Previous research on SARs has often addressed either the child feasibility dimension, evaluating how effectively children interact with and benefit from SAR-based interventions, or the parent usability dimension exploring the challenges and limitations parents face when setting up and maintaining such systems. In this study, we integrate these two perspectives to provide a comprehensive analysis of how a ChatGPT-powered virtual NAO environment can function in home settings with effective parental supervision.

Our primary objectives are to:

- **Demonstrate High Performance in Older Preschoolers:** Show that children around 4 years old can achieve near-seamless performance with minimal speech-based issues, exhibiting high engagement and success rates in short, interactive sessions. These children typically interact with the system independently, requiring little intervention and thereby validating the feasibility of the simulation-based approach for this age group.

- **Highlight Challenges for Younger Preschoolers:** Illustrate that younger children, approximately 3 years old, face considerable obstacles such as incomplete articulation, shorter attention spans, and a higher risk of frustration or meltdown when the system does not quickly adapt to their speech patterns. These challenges highlight the need for additional support measures, such as more robust speech recognition tailored to young children or alternative input methods to ensure continued engagement.

- **Examine Parent Usability Issues:** Document the practical difficulties that parents encounter, including moderate setup complexity (with an average rating of around 3.0 on a 1–5 scale), the need for frequent manual overrides due to speech misinterpretations, and the challenge of interpreting ambiguous AI feedback like generic prompts that lack actionable guidance. These factors are especially critical for families with limited time or multiple children, where ease of use is a determining factor in technology adoption.

- **Propose Targeted Design Strategies:** Offer a comprehensive suite of design improvements to bridge these gaps. Recommended strategies include the development of specialized child speech models to accurately interpret common mispronunciations (e.g., "fwee" as "three"), the use of domain-limited vocabularies for tasks like counting, the implementation of short sessions enriched with celebratory animations to maintain engagement, and the design of wizard-based or template-driven interfaces that simplify setup and configuration. Furthermore, integrating actionable, clickable AI prompts that allow for real-time adjustments without ambiguous instructions is essential for seamless operation.

By merging child-level engagement data with parent-level usability feedback, our study clarifies the interdependencies between technological performance and the human factors that determine system success. The insights gained will guide the development of more robust, scalable, and user-friendly simulation-based SARs that can deliver significant educational benefits while reducing operational complexity for everyday families.

1.3 Structure of the Paper

This paper is organized into several comprehensive sections that collectively provide a holistic understanding of the study:

- **Section 2: Literature Review** – Offers an in-depth exploration of the evolution of SARs in educational settings, the transition from physical robots to simulation-based systems, the essential role of parents in home-based EdTech, the challenges of child-specific speech recognition, and the transformative potential of AI-based adaptive learning. This section lays the theoretical groundwork for the study.
- **Section 3: Research Questions** – Outlines the core research questions that guide our investigation. These questions link quantitative and qualitative metrics of child performance (accuracy, engagement, error rates) with parent usability measures (setup complexity, speech recognition issues, clarity of AI feedback).
- **Section 4: Methodology** – Details the research design, including participant demographics (eight children and ten parents), the configuration of the CoppeliaSim + ChatGPT simulation environment, the design of counting and letter recognition tasks, the parent configuration interface, and the procedures used for data collection and analysis. This section ensures the replicability and reliability of our study.

- **Section 5: Results** – Presents the study's findings, incorporating quantitative data summarized in Tables I–III, qualitative observations from video recordings and system logs, and survey responses from parents. This section clarifies how child engagement and speech recognition performance correlate with parent usability.
- **Section 6: Discussion** – Interprets the results in depth, comparing the experiences of older and younger preschoolers, examining the impact of speech recognition challenges, and discussing the broader implications for educational technology. It addresses the critical need for user-friendly interfaces and adaptive AI that can bridge the gap between technological capability and real-world usage.
- **Section 7: Recommendations** – Offers practical design improvements aimed at overcoming the identified challenges. These include enhancements in child-specific speech recognition, the development of wizard-based setup processes, multimodal input options, and integrated, actionable AI prompts.
- **Section 8: Conclusion** – Summarizes the key findings, discusses the limitations of the study (such as sample size and demographic homogeneity), and outlines potential directions for future research, including hybrid physical-virtual approaches and advanced acoustic modeling for preschool speech.

By weaving together detailed analyses of both child engagement and parent usability, this paper provides a comprehensive assessment of how AI-driven, simulation-based SARs can be effectively implemented in home settings. Addressing both technological and human factors is crucial for developing scalable, cost-effective, and interactive tutoring tools that have the potential to transform early childhood education.

2. LITERATURE REVIEW

This section offers a significantly more detailed exploration of the research themes that support the development and implementation of simulation-based, AI-driven Socially Assistive Virtual Robots (SAVRs) in early childhood education. Although the child engagement perspective and the parent usability perspective form our central points of focus, the relevant literature spans multiple domains: the historical evolution and inherent obstacles of physical SARs, the rising influence of simulation as a tool for cost-effective robotic education, the pivotal role of parents in home-based EdTech adoption, the distinctive issues posed by child-specific speech recognition, and the promising potential of AI, especially large language models (e.g., ChatGPT), to deliver adaptively responsive tutoring. By examining these five interlocking areas, we clarify how advanced technology, and user-friendly design must converge to make AI-driven SARs truly effective in typical family environments.

2.1 Socially Assistive Robots (SARs) for Early Childhood Education

2.1.1 Background and Evolution of Physical SARs in Education

Socially Assistive Robots (SARs) are robotic platforms intended to help or assist human users primarily through social engagement, spoken dialogue, gestures, and supportive feedback, rather than by performing mechanical tasks such as manipulation or heavy lifting [12], [6]. This distinction differentiates SARs from industrial or service robots, reframing them as interactive agents that can stimulate or mentor users, especially in educational or therapeutic contexts. Within early childhood, physical SARs like NAO or Pepper have been studied in an array of teaching scenarios: basic numeracy activities like counting objects, recognizing shapes, building foundational vocabulary in a first or second language, and even fostering social-emotional learning (turn-taking, empathy, self-awareness) for preschoolers [11], [19], [22].

A repeatedly emphasized advantage of physical SARs are their embodied nature and tactile presence. By physically existing in the same space as a child, they can deploy gestures arm waving, head nodding, expressive LED eyes, that prompt curiosity and more sustained involvement. Researchers have found that many children spontaneously humanize these robots, perceiving them as friendly peers or pet-like companions [9], [20]. This sense of social rapport has been shown to promote deeper attentiveness, repeated attempts at challenging tasks, and natural verbal interplay. Moreover, certain physical SARs have proven beneficial for children with developmental or learning differences, such as autism spectrum disorder, by providing consistent, predictable social cues that can support and make otherwise difficult social interactions easier [24].

However, the successful use of physical SARs presents a fundamental challenge: scalability. The high cost of purchasing humanoid robots, often ranging in the thousands of dollars per unit, along with the need for specialized staff or robotics-savvy researchers to maintain and program them, creates significant practical barriers [6], [13]. Battery management, sensor calibration, software updates, and mechanical wear demand ongoing attention, which few ordinary preschools or homes can provide. Consequently, even though physical SARs display notable potential in short-term research or pilot contexts, they typically fail to transition into daily use for average families, leaving many promising educational interventions either limited to specialized labs or overshadowed by simpler, cheaper technologies that may not be as socially engaging. This feasibility gap has encouraged researchers to investigate alternative approaches capable of emulating core SAR features, speech-based tutoring, interactive feedback, without the excessive hardware overhead [7], [17].

2.1.2 Simulation-Based SAVRs: A Cost-Effective Alternative

Simulation-based SAVRs attempt to reconstruct the speech outputs, gestures, and adaptive behaviors characteristic of physical robots but solely in a software environment [15], [12], [16]. By loading a 3D scene, often in engines like Unity, Unreal Engine or CoppeliaSim users can see a robot avatar performing animated motions, speaking via text-to-speech, or changing expressions in response to child interactions. This approach requires only a standard PC or tablet, entirely sidestepping the high acquisition and maintenance costs of real humanoid units. Moreover, removing the mechanical dimension allows for seamless integration of AI-based decision-making, since the simulation can be updated or patched as new modules become available [17], [21].

Early pilot studies highlight that well-animated and short on-screen sessions can often sustain preschoolers' interest, particularly for cognitively oriented tasks like counting or letter recognition, which depend on spoken dialogues and visual feedback [3], [14]. Critics, however, question whether a purely on-screen avatar can replicate the social novelty and tangibility of a physical robot. Children might lose interest if they expect to physically touch or move around with a robot. Researchers also note that for certain interventions (e.g., sensorimotor-based therapies), a virtual presence may be insufficient [25], [20]. However, for straightforward counting or reading tasks reliant on speech-based or screen-based cues, many of the “social presence” attributes that define SAR-based pedagogy, like real-time feedback, personalizing difficulty, praising correct answers, can be fully retained in a simulation.

Hence, for everyday home usage, simulation-based SAVRs hold significant advantages: they drastically lower cost, avoid mechanical failures, and can be downloaded anywhere, giving more children access to advanced tutoring experiences [10], [17]. Simulation-based learning environments are increasingly recognized for their scalability and capacity to replicate real-world scenarios for early childhood education. They enable controlled, repeatable experiences that support both cognitive and social development. However, these platforms present notable challenges, including limitations in simulation fidelity, maintaining learner engagement, and ensuring usability for both children and non-expert adult supervisors. High-fidelity simulations have been shown to increase realism and engagement but often lack essential sensory experiences such as haptic feedback, which is critical for immersive learning. Furthermore, significant discrepancies in interaction timing between young children and simulated Socially Assistive Virtual Robots (SAVRs) have been observed, emphasizing the need for adaptive dialogue systems and enhanced natural language understanding, especially tailored for preschool-aged children [26].

Users do not require specialized robotics knowledge to set up the system, expanding scalability to typical families, daycares, and under-resourced educational settings. However, achieving consistent child engagement over repeated sessions demands thoughtful design, ensuring short, visually stimulating tasks and robust speech interfaces. If the system can handle child articulation and keep sessions engaging, a simulation-based SAR can rival many benefits once exclusive to expensive physical machines [16], [21].

2.1.3 Conceptual Unification of Virtual and Physical SARs

The core definition of Socially Assistive Robotics emphasizes how assistance is delivered, through social interaction, rather than what material form the agent takes [1]. In traditional terms, a “robot” implies a physical embodiment, but conceptually an interactive virtual agent can meet the same criteria if it engages users via speech, gestures, and responsive feedback. Researchers have noted that modern virtual agents can replicate many of the tutoring functions of physical robots without the logistical overhead of hardware [2]. Accordingly, several recent projects have treated on-screen robot avatars as bona fide SAR platforms, reconstructing humanoid robot’s behaviors entirely in software. A simulated SAVR, for example, a 3D animated character in Unity or CoppeliaSim, can speak, display facial expressions, and adapt its prompts just as a physical robot would, thereby fulfilling the social and assistive roles central to the SAR paradigm. In this unified framework, the distinction between a “virtual” and “physical” robot is one of embodiment but not of fundamental purpose: both aim to assist learners via social engagement and personalized interaction.

Adopting this inclusive view does not ignore the known effects of physical embodiment on human-robot interaction. Studies consistently find that physically present robots often evoke stronger social responses and learning gains than their on-screen counterparts. Young children tend to find a tangible robot more novel and attention-grabbing, leading to increased eye contact, compliance with the robot’s requests, and sustained engagement

[2], [3], [4]. These advantages are largely attributed to the robot's co-location and corporeal presence, being a visible, touchable agent in the child's real space [3]. Indeed, some critiques maintain that a purely screen-based character cannot fully replicate the "social spark" of a physical machine: children might lose interest if they expect to hug or physically play with a robot, and certain intervention types (e.g. motor skills training) inherently benefit from a tangible partner.

However, initial findings from our pilot study suggest that, when a virtual agent is designed with rich social cues and interactivity, children may perceive it as a social partner and respond with surprising naturalism [3]. The distinction between virtual and physical agents in this context appears to be primarily quantitative, though further research is needed to confirm this pattern. For many educational tasks that revolve around dialogue and on-screen feedback (counting games, storytelling, letter recognition), researchers have shown that the critical "social presence" features of SAR, real-time turn-taking, personalized encouragement, mirroring of the child's actions, can be preserved in a virtual robot. In other words, a well-designed SAVR can leverage the same principles of social reinforcement and adaptive tutoring as a physical unit, even if the degree of engagement might differ. This perspective is supported by the CASA paradigm (Computers Are Social Actors) in human-computer interaction, which finds that people naturally treat communicative media agents as social beings. The implication is that virtual robotic tutors and physical robots lie on a continuum: they share common goals and interaction mechanics, with embodiment as a variable influencing effectiveness rather than a binary divider of "robot or not."

The relevance of this unified framework is directly reflected in the approach and findings of the present study. By treating the simulated virtual robot such as the NAO robot, in our experiments as a true SAVR platform, we acknowledge that it operates under the same pedagogical and social principles as a hardware robot. This thesis demonstrates that an AI-driven virtual robot can replicate many benefits of physical SARs in practice: for instance, our results showed that 4-year-old children achieved nearly 100% task success with the on-screen robot, a level of performance comparable to outcomes reported with physical robot tutors [4]. Such evidence reinforces the idea that physical embodiment is not a strict requirement for achieving the educational and engagement objectives of SAR.

At the same time, our work also highlights what is missing or different when the robot is virtual, such as the greater challenge in holding the attention of some 3-year-olds or the absence of tactile feedback. Rather than undermining the SAR paradigm, these differences inform it: they pinpoint where additional design measures (more expressive animations, enriched audio-visual cues, or even occasional physical-robot interactions) are needed to fully realize the potential of simulation-based agents.

In summary, viewing virtual and physical SARs within one conceptual framework allows us to leverage the strengths of both. SAVRs offer unrivaled accessibility and scalability, while physical robots offer palpable presence, together, they define a spectrum of socially assistive technologies. By unifying these in theory, this research advocates for a broad, inclusive notion of what a “socially assistive robot” can be, ultimately arguing that the defining feature of SAR is the social assistance itself, not the robot’s material instantiation. This theoretical integration is pivotal for scaling up SAR solutions: it means that insights from purely software-based systems (like those in this study) are applicable to the wider SAR field, and that future early education interventions can strategically choose virtual, physical, or hybrid implementations without leaving the core scientific framework of Socially Assistive Robotics.

2.1.4 Effect of Avatar Design and Visual Embodiment in SAVRs

The visual embodiment of a Socially Assistive Virtual Robot (SAVR), that is, how the agent appears on screen, plays a pivotal role in shaping children's engagement, attention, and perception of agency. Although the underlying artificial intelligence, speech recognition, and adaptive logic may remain identical, research shows that modifying only the avatar's appearance can significantly alter the child's social and emotional response during learning tasks [27], [28], [29].

Humanoid or robot-like avatars (e.g., virtual NAO or Pepper) are often perceived by children as "social partners" or "tutors" rather than as entertainment figures [30], [31]. This perception aligns with the Computers Are Social Actors (CASA) paradigm, which suggests that human users, especially young children, naturally attribute social intention and credibility to agents that display anthropomorphic cues such as gaze, posture, and gesture [32]. Such cues enhance social presence, the sense that the virtual agent is an intentional being capable of understanding and responding meaningfully. This heightened presence contributes to longer attention spans, improved task persistence, and stronger affective bonds during repeated sessions [12], [33].

By contrast, purely cartoon-based or abstract avatars, such as animated animals or shapes, may attract initial attention through bright visuals but often elicit weaker perceptions of intelligence or intentionality [34], [35]. As a result, children may interpret them as part of a game or television show rather than as an interactive peer or tutor. Research in educational technology indicates that animated pedagogical agents can enhance learning outcomes and cognitive engagement when they are designed to support, rather than distract from, instructional goals. Human-like pedagogical agents have been shown to improve comprehension and increase brain activation during learning tasks [35]. However, such avatars may still offer value for reducing anxiety among very young learners or for tasks emphasizing creativity and imagination [12].

In simulation-based SAVRs, the choice of avatar thus becomes a pedagogical variable, not merely a design preference. A humanoid robotic avatar strengthens alignment with the theoretical foundations of Socially Assistive Robotics by sustaining a perception of agency and interactivity similar to physical robots [30], [31]. Conversely, non-robotic avatars may simplify visual design but risk diluting the social-assistive framing, thereby shifting user perception from learning companion to animated software. This distinction underscores that visual embodiment functions as a bridge between traditional educational media and the social dynamics characteristic of SAR-based learning. Future studies should systematically compare avatar morphologies to determine optimal forms for balancing engagement, trust, and cognitive outcomes in early childhood contexts [12].

2.2 Parent Involvement in Home-Based Educational Technology (EdTech) and SAR Solutions

2.2.1 Parents as Essential Facilitators in Early Childhood

In the home setting, the success or failure of any educational technology for preschoolers often depends on parent attitudes and actions [23], [36]. Unlike older children who might independently explore apps or manage device setups, 3–4-year-olds rely heavily on adults to find suitable tools, install software, define tasks, and intervene when confusion arises. This dependency intensifies for simulation-based SAVRs or advanced AI solutions, as parents must also handle microphone calibration, potential speech recognition errors, and child meltdown if repeated "incorrect" feedback emerges [22], [37].

Studies on home-based robotics confirm that parental acceptance is crucial from the outset: if parents anticipate frequent speech failures or a steep learning curve, they may simply not attempt repeated usage [19], [38]. Conversely, if parents perceive immediate positives, like fun, accurate child recognition, or minimal setup time, they are more likely to incorporate daily or weekly short sessions, thereby reinforcing a child's incremental learning. This "parent-as-gatekeeper" dynamic is particularly pronounced for emergent literacy or numeracy tasks, where consistent practice can make a sizeable difference in early skill mastery [21], [11]. In short, even the best AI-driven SAR technology means little if parents find it too demanding, confusing, or unreliable to manage in their daily routines.

2.2.2 Key Usability Frictions for Parents

Experience with robotics and AI-based educational technologies reveals recurring usability bottlenecks that deter non-expert parents [6], [7], [13]:

1. **Speech Recognition Calibration**: Adjusting "confidence thresholds" or "noise gating" for child speech can be mysterious, leading to minimal improvements or unintentionally increasing error rates [18]. Parents lacking technical knowledge typically leave defaults alone, occasionally resulting in repeated misinterpretations of preschool voices.

2. **Task Setup Guesswork**: Without pre-built templates (e.g., "Beginner Counting 1–5"), parents must guess numeric or letter ranges, risking tasks that are too easy or too advanced [17], [20]. Some might choose 1–10, discovering mid-session that the child only truly handles 1–3, prompting meltdown.

3. **Ambiguous AI Prompts**: If ChatGPT or a related system says, "Reduce difficulty," parents may be uncertain which specific control to adjust or by how much. Coupled with time pressure from a restless child, this confusion stymies timely adaptation.

4. **Time Constraints**: Busy schedules leave little room for repeated "tweaking" or reboots. If each session demands 10+ minutes of overrides or advanced tinkering, parents may abandon the system, overshadowing the potential AI benefits [21], [37].

Hence, user-centric design must simplify parent workflows with wizard-based interfaces, integrated "click to apply" AI suggestions, and robust defaults that minimize speech-related challenges. Absent these elements, families might only run the system once or twice, concluding it's too cumbersome to maintain. Another key challenge lies in the complexity of configuring these environments. Many parents and educators face technical barriers due to the specialized knowledge required to set up and operate SAVR-based simulations. Studies have found that up to 40% of parents reported difficulties in system setup, particularly when initial configurations were complex or when interpreting AI-generated feedback. Speech recognition reliability further compounds usability issues, with 70% of parents citing difficulties stemming from errors in automatic speech recognition (ASR), especially for younger children whose articulation differs significantly from adult patterns [39].

2.2.3 Parental Acceptance and Adaptation to Educational Technology

Parents play a pivotal role as gatekeepers in the adoption of educational technologies for young children. In early childhood home-learning, technology is typically parent-mediated; parents must perceive value and safety before integrating tools like educational apps or socially assistive robots into family routines [3], [36]. Existing literature shows that emerging technologies such as socially assistive robots (SAVRs) can strongly capture young children's attention and interest, offering personalized, engaging learning experiences [3]. For example, SARs have been used to support early literacy and numeracy through interactive play, achieving notable gains in children's engagement and learning outcomes [3], [21]. However, traditional physical SARs face practical adoption barriers, such as high cost, maintenance needs, and limited scalability, which constrain their widespread use [3]. This has led to the exploration of simulation-based SAVRs, or virtual robot agents that deliver similar interaction benefits without costly hardware, thereby making home implementation more feasible [3]. Ultimately, parents' acceptance of any educational technology depends on their confidence that the tool is beneficial, user-friendly, and aligns with their child's needs and family values [36], [40]. Theories of technology acceptance, such as the Technology Acceptance Model (TAM), emphasize that perceived usefulness and ease of use drive adoption, along with social and contextual factors [40], [41]. In the context of home learning, this means parents must trust that a given technology will effectively support their child's learning and that they, as non-expert end-users, can manage and supervise its use without undue difficulty [40]. Recent global survey research found that parents engaged more with educational technology when well-structured tools were recommended by schools or peers, which boosts credibility, and disengaged when tools were overly complex or beyond their technical skills [40].

Parents are generally motivated to adopt home-learning technologies when they perceive clear educational benefits for their child. A consistent driver is the expectation of improved learning outcomes or future advantages for the child [42]. For instance, in language learning contexts, parents readily invest in AI-driven tutoring robots if they believe it will give their child a competitive academic edge [42]. Existing studies have documented that social robots and adaptive learning software can indeed enhance children's performance, increase their enthusiasm for learning, and sustain long-term engagement [3], [21]. Such evidence of effectiveness strengthens parents' utilitarian motivation, that is, the conviction that technology is useful for skill development [42]. At the same time, hedonic motivation, or the idea that the learning tool makes education fun and enjoyable for the child, is another key factor positively influencing parents' intention to adopt [42]. Parents often observe that interactive educational games or robots captivate their children more than traditional study methods; for example, a storytelling robot was seen by parents as more engaging than a tablet screen for story time, keeping the child's attention better [43]. Another practical drive is technology's potential to assist or augment the parent's own role in teaching. Busy or resource-constrained parents may welcome a tool that can tutor or entertain the child when the parent is unavailable or needs a break [38], [44]. Recent studies show that parents appreciate using robots as collaborators rather than replacements, allowing the robot to take over repetitive instructional tasks or answer questions in domains where the parent feels less confident, while parents remain in a supervisory role [38]. This collaborative use relieves some pressure from parents and increases their willingness to integrate the robot into home learning [38], [44]. In essence, parents are motivated by technologies that complement their parenting, including tools that extend their capabilities, reinforce their efforts, and validate their teaching rather than compete with or replace them [38]. Social influence also plays a role, as other parents, teachers, or pediatricians who endorse a particular educational app or robot tutor provide external validation that can reassure caregivers of its value [40].

Despite the potential benefits, parents often have significant concerns that can hinder the adoption of educational technologies for young children. One major category of concerns revolves around the child's health, safety, and development. Parents worry about excessive screen time or device use harming their child's physical health, such as eyesight and sleep patterns, and socio-emotional development [45]. In interviews, some parents express fear that if a robot or app occupies too much of the child's time, the child might "lose interpersonal skills and stop knowing how to interact with humans" [45]. This underscores a prevalent concern that technology could displace valuable human interaction or reduce the child's engagement in imaginative play and real-world socialization. Many parents are therefore cautious about over-reliance on digital or robotic learning tools and insist on moderation and balance with non-digital activities [38]. Ensuring that the technology does not become a surrogate caregiver is crucial for trust, as families need to see it as a supportive tool rather than an "electronic babysitter" that might erode parent–child bonds or the child's social growth [46].

Another common barrier is concern over content appropriateness and values alignment. Parents must trust that educational technology will expose their young child only to safe, high-quality content. Unsupervised internet connectivity or AI-generated content can raise alarms about inappropriate material or misinformation [47]. Relatedly, parental control features are often demanded. Parents feel more at ease if they can monitor or restrict what the child is doing, such as disabling certain features or requiring approval for new content [38]. This extends to time control as well, since many parents prefer technologies that allow session time limits or automatic "screen locks" after a set duration [47]. Privacy and safety risks constitute another barrier, especially for socially assistive robots or smart toys with cameras and microphones. Parents may fear data collection or hacking, especially in the intimate setting of the home [38]. Research on technology adoption indicates that risk perception can significantly dampen users' intention to adopt; for example, perceived physical risks or the possibility of technical malfunctions tend to lower acceptance levels for novel gadgets [42].

Design strategies that encourage parental uptake include promoting parental involvement and oversight, intuitive and simple interfaces, customizable content and settings, and transparency about AI actions and privacy [3], [41], [44]. Technologies that are adaptable and provide personalized feedback, but also give parents clear control and understanding, are most likely to gain parental trust [3], [44]. Designers should communicate the role of technology as a supportive, not replacement, tool for parents, and ensure strong safety, privacy, and age-appropriate content features. When educational technologies are developed with these considerations, parents are more likely to adopt them and integrate them consistently into the home learning environment [44], [46].

2.3 Child-Specific Speech Recognition

2.3.1 Mismatch of Preschool Speech and Adult-Optimized Speech-to-Text (STT)

A longstanding issue in child–robot and child–AI interactions is the voice interface mismatch [18], [19], [48]. STT is a technology that converts spoken language into written text, commonly used in voice interfaces to transcribe user input. While modern STT tools excel for adult voices, they rarely handle the irregular phonetics, higher pitch, or incomplete articulation common among preschoolers (3–5). Children might say "tee" for "three," "fo-wuh" for "four," or produce partial syllables or unclear endings. These divergences cause general-purpose STT to misinterpret or skip words, leading to incorrect feedback from the system. For older kids (4–5) with clearer enunciation, mild re-speaking can salvage some queries, but 3-year-olds often lack that capacity or patience [17], [21].

The upshot is that "state-of-the-art" STT can register a 20–40% error rate for younger preschoolers, meaning repeated "That's not correct!" flags even when a child's response was essentially right [7], [22]. Pilot data revealed that 70% of participating parents reported significant difficulties with the system's speech recognition when working with 3–4-year-olds, often due to pauses, unclear articulation, and higher pitch. This resulted in frequent misinterpretations of correct answers and increased frustration, sometimes requiring repeated manual intervention. These findings underline the urgent need for child-adapted ASR and robust multimodal feedback in early childhood SAVR environments [39]. This not only disrupts the child's sense of success but also leads to meltdown or refusal to continue. Parents may step in with typed overrides, but doing so repeatedly breaks session flow and raises adult frustration levels, undermining the notion of an autonomous or semi-autonomous robot tutor [16], [37].

2.3.2 Proposed Approaches: Domain Limits, Child Speech Corpora, or Alternative Input

Multiple solutions exist to ameliorate these child-speech recognition hurdles:

1. **Domain-Limiting**: Restrict recognized vocabulary to a small set, e.g., "one, two, three, four, five." This approach drastically reduces confusion for counting tasks, though it must be reconfigured for letters or other domains [23], [20].

2. **Phonetic Dictionary**: Incorporate typical child mispronunciations or partial words into the recognition dictionary [18], [48]. This requires some knowledge of common "kid talk" patterns but can significantly lower error rates.

3. **Child Speech Corpora**: Corpora (plural of corpus) are large, structured collections of spoken language data. Child speech corporation specifically refers to datasets compiled from recordings of children's speech, which are used to train speech recognition systems to better understand the unique phonetic patterns and articulation of young children [14], [16]. Retraining or fine-tuning STT on these audio samples helps to account for pitch range and articulation quirks, improving recognition accuracy for younger speakers. Although more robust, this can be resource-intensive and not readily available to all developers.

4. **Alternate Input**: Provide a fallback if the system misinterprets a child repeatedly, such as on-screen numeric or letter buttons the child or parent can tap [37], [49]. By bridging repeated speech errors, this method reduces meltdown risk.

Implementing these features can create a child-friendly environment where the AI logic consistently interprets a preschooler's speech, thus enabling stable, adaptive feedback. Without them, younger children remain vulnerable to repeated misinterpretations and sudden session termination [6], [16].

2.4 AI-Based Adaptation: ChatGPT for Real-Time Tutoring

2.4.1 Child-Facing Versus Parent-Facing AI

Large language models (LLMs), including ChatGPT, have revolutionized contextual language processing, enabling systems to parse recognized text, evaluate correctness, and generate relevant, natural-sounding responses [8], [17]. In early childhood tasks, this adaptivity means a child can receive immediate praise ("Great job!") for correct answers, partial hints or step-by-step counting for near misses, and full demonstrations for repeated errors. This dynamic closely mimics how a patient human tutor might respond, offering direct scaffolding that can keep children engaged [20], [21].

Simultaneously, parent-facing suggestions can help non-expert adults adjust tasks mid-session. If a child repeatedly fails counting above 5, ChatGPT might prompt: "Your child is having difficulty with items 6–10. Would you like to reduce the item count?" This approach spares parents from having to guess how to correct or adapt. However, studies indicate that if these AI prompts remain purely textual, many parents won't fully utilize them [19], [37]. Clear, clickable instructions, like a button to "reduce to 1–5", are often needed to seamlessly integrate AI insights into real-time session management.

2.4.2 Opportunities and challenges in AI-Driven Adaptation

In principle, a well-tuned AI can auto-level tasks, ensuring the child operates in an "optimal challenge zone." If correct answers come easily, the system can escalate to slightly harder prompts; if multiple consecutive errors surface, it might revert to simpler tasks [16], [50]. This real-time adaptivity is essential for sustaining interest without overloading or boring the child [9]. Yet, the AI logic depends critically on the accuracy of recognized speech: if the system misinterprets "three" as "tree," the child might be unfairly flagged as incorrect, leading to undeserved difficulty downgrades or negative feedback. Moreover, unbounded large language models can occasionally produce tangential or inappropriate content, necessitating domain-limited or child-safe prompt engineering [51], [52].

Despite these hazards, integrating ChatGPT or comparable LLMs with a simulation-based SAVR remains highly appealing. The integration of dynamic, context-aware dialogue, immediate feedback, and short, visually engaging tasks can replicate or surpass the educational benefits historically associated with physical SARs, minus their cost or logistical demands [17], [21]. Achieving that integration, however, requires robust speech solutions, user-friendly UI designs for parents, and domain constraints that keep the AI's responses consistently helpful for preschool tasks.

2.5 Risks and Limitations of SARs, SAVRs, and Other Digital Tools in Early Childhood Education

Early childhood educators and researchers have raised significant concerns about the downsides of Socially Assistive Robots (SARs), including their virtual counterparts (SAVRs), as well as learning apps, animated educational videos, and other computer-based learning tools. While these technologies offer novel educational opportunities, a growing body of literature highlights psychological, behavioral, privacy, and educational risks associated with their use in the preschool and early elementary years. For example, scholars have identified issues ranging from children forming emotional attachments or dependencies on artificial agents, uncritically obeying robot instructions, and reduced human-to-human interaction, to ambiguity in accountability when something goes wrong [53]. There are also broad concerns that heavy use of digital media in early childhood may limit physical activity and real social engagement and could even violate young children's privacy and data security (e.g. when classroom robots record personal data) [53], [54].

2.5.1 Screen Time and Behavioral Effects

Excessive screen time from digital learning tools can displace essential activities such as free play and face-to-face interaction [55]. Developmental experts warn of possible sleep disturbances and childhood obesity linked to high screen exposure [56]. One longitudinal study found that toddlers exposed to multiple hours of screen time daily exhibited more behavior problems and poorer vocabulary growth than peers with limited screen use [57]. Pediatric guidelines emphasize limiting media use in children under 2 and practicing moderation for preschoolers [55].

Young children may develop dependencies on apps and videos, leading to inattention and difficulty with self-regulation [13], [57]. Flashy, fast-paced visuals in apps can condition children to expect entertainment during learning. One study found that 58% of early childhood educational apps were low in quality by developmental standards [58]. While SARs might reduce screen exposure, they may still encourage sedentary behavior, and virtual robots (SAVRs) continue to involve screens, maintaining the same behavioral concerns.

2.5.2 Reduced Human Interaction and Social Development

Children learn socially, and time with screens or robots can reduce meaningful caregiver interaction. Background media like television has been shown to reduce both the complexity and amount of caregiver-child dialogue [55], [59]. Similarly, children may interact less with peers and more with robots, risking social skill development. Prolonged interaction with SARs may lead to impaired emotion recognition and reduced empathy, as found in experimental studies [53], [57].

Concerns have been raised about emotional dependencies on robot tutors, with attachment issues arising due to the robot's limited expressive capabilities and emotional reciprocation [60] . Children have been found to trust robots over their own judgment, a phenomenon known as automation bias [13].

2.5.3 Privacy and Ethical Concerns

Many learning apps and robotic tools collect sensitive data. Without proper safeguards, children's names, recordings, and usage habits can be exposed [54]. Some robots have transmitted recordings to cloud servers without encryption. Ethical concerns also emerge when children do not understand the robot's nature or the monitoring it performs [54], [60]. It has been noted that robotic companions recording a child's actions raise significant psychological and ethical implications [60].

2.5.4 Educational and Developmental Limitations

Digital tools often lack the adaptability and nuanced responsiveness of human educators. Apps may allow guessing or rote drills without deeper comprehension, while videos lack interaction entirely, contributing to the "video deficit" effect [59]. SARs and AI tutors operate within fixed parameters and may miss contextual cues a human would catch. Most educational apps fail to incorporate the four pillars of learning, active, engaged, meaningful, and socially interactive experiences [58].

Additionally, group and imaginative play may suffer when individual tech use dominates the classroom. Despite enthusiasm around robot-led activities, current evidence does not support sustained, long-term learning improvements [54].

2.5.5 Comparative Risk Profiles of Different Technologies

- Screen Exposure: Apps, SAVRs, and videos present higher risks than SARs, though all can encourage sedentary behavior [13].
- Social Development: Non-interactive videos isolate children, while SARs may cause misplaced social bonding. Apps fall in between [13], [53], [57].
- Privacy: SARs and connected apps pose significant data and security risks, while offline videos present minimal privacy concerns [54].
- Educational Efficacy: SARs and advanced apps may outperform videos but still lack the human teacher's flexibility and depth [54], [58], [59].

In conclusion, these tools offer unique advantages but come with distinct risks that must be carefully managed. Experts advocate for adult supervision and complementary, not replacement, use of technology in early childhood education.

2.6 Gaps Addressed by This Integrated Paper

While numerous pilot projects on physical SARs have demonstrated short-term child engagement, their real-world extension is hampered by hardware expenses and specialized needs [6], [14]. Some initial simulation-based efforts show promise but raise questions about whether children truly remain captivated by a virtual on-screen agent and whether parents can easily manage it daily [16], [19]. Additionally, mainstream STT's mismatch with younger children's speech hampers seamless, autonomous interactions. Meanwhile, parents might find advanced AI suggestions too vague or complex, especially if short on time [7], [37].

Hence, the major constraints revolve around:

- **Cost/Scalability** for physical SAR systems, preventing broad adoption [6], [17].
- **Speech Recognition** shortfalls, especially for 3-year-olds, undermining child success and fueling meltdown [16], [18].
- **Parent Usability** complexities, including interface confusion, ambiguous AI advice, and repeated manual overrides, deterring sustained usage [7], [21].

Our paper **directly** tackles these issues by:

1. Presenting integrated child data from a ChatGPT-enabled simulation environment (eight total children, split between older and younger preschoolers).
2. Analyzing parent feedback from ten parents, focusing on usability obstacles, speech recognition frustrations, and suggestions for wizard-based solutions or "click to apply" AI prompts.
3. Drawing on prior scholarships to propose child-specific ASR methods, domain-limited recognition for counting tasks, short session structures with celebratory animations, and direct AI guidance that parents can instantly implement without guesswork.

In merging these multiple domains, child-level speech success, parent-driven setup, AI-based adaptation logic, we highlight how simulation-based SAVRs, if thoughtfully engineered for younger speech patterns and straightforward parent workflows, can indeed produce widely scalable, cost-effective, and engaging early-education platforms. Only by addressing the recognized challenges in speech recognition and user design can we realize the full potential of AI-driven robotic tutoring in ordinary family contexts.

3. RESEARCH QUESTIONS

Based on the prior review, this research proposal tackles four core questions:

1. **Child Engagement & Performance**
   - **Research Question 1**: Do children aged 3–4 remain engaged with short counting/letter tasks in a ChatGPT-driven, simulation-based SAVR setting, and how does age affect success rates and frustration?
   - **Research Question 2**: Can iterative feedback (e.g., re-demonstrations, lowered difficulty) help 3-year-olds partially overcome articulation challenges?
2. **Speech Recognition**
   - **Research Question 3**: How often does mainstream STT misinterpret preschool speech, particularly for 3-year-olds, and does this lead to session interruptions?
   - **Research Question 4**: What strategies, manual overrides, repeated attempts, domain-limited recognition, do parents use, and with what effect on the child's performance?
3. **Parent Usability**
   - **Research Question 5**: How do parents rate setup complexity, speech recognition reliability, and clarity of AI guidance, and do they favor a wizard-based or template-based approach?

- **Research Question 6**: Does AI such as ChatGPT effectively prompt parents to adapt tasks or is it too vague, resulting in confusion?

4. **System Refinements**

- **Research Question 7**: Based on integrated evidence, how can simulation-based SAVR platforms incorporate child-specific ASR, multimodal inputs, short tasks, wizard-based interfaces, and direct “clickable” AI suggestions to optimize child success and parent acceptance?

4. METHODOLOGY

In this section, we outline how eight children (ages 3–4) and ten parents participated in short, home-based sessions with a simulation-based Socially Assistive Virtual Robot (SAVR) environment powered by ChatGPT. This expanded methodology describes the participant selection, the simulation platform (CoppeliaSim + ChatGPT integration), the parent configuration interface, and the multi-step data collection process, concluding with details on how child and parent data were analyzed. The main aim is to ensure both child engagement metrics and parent usability metrics are captured in a systematic, replicable manner.

4.1 Participants

4.1.1 Child Participants

A total of eight children were recruited, divided into two age-based subsets:

1. **4-year-olds (N=5):** Children A, B, C, D, E. According to parental reports, these children already possessed moderate familiarity with basic counting (1–5 or 1–10) and some letter recognition skills (A–D or A–F). Parents described them as reasonably confident in speaking, though not all had prior experience with speech-based technology.

2. **3-year-olds (N=3):** Children F, G, H. These younger preschoolers varied more in language development and articulation clarity. Parents indicated that these children recognized fewer letters or could count only up to 3–5 at best. Notably, the younger group was less accustomed to extended screen-based interactions and typically needed more adult guidance.

No strict exclusion criteria (e.g., socioeconomic factors, second-language backgrounds) were applied. Instead, families self-selected based on willingness to experiment with a home-based "robot tutoring" system. The smaller sample size reflects the pilot nature of the research, aiming to capture both older and younger preschool experiences. These insights are grounded in pilot studies involving 10 parents and 8 children (aged 3–4 years) who participated in configuring and using SARs within a simulated environment. All parent participants had university degrees and demonstrated high digital literacy, which allowed the study to focus specifically on usability and engagement challenges in SAR configuration, rather than on basic technology adoption barriers [39].

4.1.2 Parent Participants

Ten parents oversaw the sessions, each responsible for installing the simulation software, and configuring tasks. All parents reported at least an undergraduate-level education, with moderate computer literacy, though they had varying degrees of comfort with advanced AI or robotics interfaces. This demographic profile means that while the parents were not experts in speech recognition or HRI, they were accustomed to basic digital tools (e.g., typical smartphone apps, computer usage). Each parent consented to:

- **Video Recording** of the child's sessions (for post-hoc engagement coding).
- **System Logging** of recognized speech, speech thresholds, AI suggestions, and typed overrides.
- **Post-Session Surveys** on usability, speech recognition, and AI clarity, plus optional open-ended interview questions regarding overall impressions.

All study procedures received approval from the Oakland University Institutional Review Board (IRB), as documented in the IRB approval letter (Figure 4 IRP Approval Letter). This oversight ensured compliance with established ethical standards for research involving human participants. Written informed consent was obtained from all parents prior to participation, with the right to withdraw at any time. Children were free to discontinue their participation at any point during the sessions.

4.2 Simulation Environment

4.2.1 Platform and Architecture

The simulation environment is built on the CoppeliaSim platform, which serves as the 3D engine hosting a virtual NAO robot within a carefully designed, stylized “classroom” scene. This digital setting is crafted to mimic a real-world learning space, complete with visual elements that capture the essence of an educational environment. The virtual classroom features a projected screen that serves as the primary display for tasks such as counting objects and recognizing letters, enabling the system to deliver dynamic visual feedback that evolves based on the child’s interaction.

This simulation model is designed to support a range of interactive behaviors, including:

- **Animated Gestures:** The NAO avatar is programmed with a series of naturalistic movements, such as waving its arms, head movements, and performing celebratory gestures. These animations are triggered when the child responds correctly, thereby reinforcing positive behavior through immediate visual praise.
- **Dynamic Scene Changes:** The virtual classroom includes a flexible “projection screen” that dynamically displays different sets of items, such as colorful balls for

counting tasks or letters for recognition tasks. These items can change in real time according to the underlying AI logic, ensuring that the presented content is continuously updated based on the child's performance and the task difficulty settings.

A critical component of the system's architecture is a multi-threaded Python application that seamlessly integrates multiple software components and connect to various APIs. This application acts as the central mediator between CoppeliaSim, Google Speech-to-Text (STT), and ChatGPT. The architectural design is carefully engineered to replicate many of the real-time features of a physical SAR, including immediate speech-based prompting, dynamic visual feedback, and adaptive responses tailored to the child's input. By synchronizing these components, the system is able to deliver near-real-time interactivity that is essential for maintaining the illusion of a conversational, socially interactive tutor.

4.2.1.1 Speech Recognition (Google STT)

In our simulation, children's verbal responses are captured via the computer's built-in microphone and sent directly to Google's Speech-to-Text (STT) service for transcription. The STT service converts spoken language into text, capturing common responses such as "three," "four," or even individual letters like "C." Recognized text is then passed on to ChatGPT for further interpretation and decision-making. Given that young children often have variable articulation and may speak in a manner that deviates from standard adult speech patterns, the system includes an option for parents to adjust the STT "confidence thresholds" and enable noise reduction features through the graphical user interface (GUI). Although most parents tend to leave these settings at their default values (typically between 0.7 and 0.75), the option to tweak these parameters is crucial when repeated misinterpretations occur, thereby providing a means to fine-tune the system's sensitivity to child speech.

4.2.1.2 ChatGPT Integration

The integration of the ChatGPT API is central to providing a dynamic, context-aware tutoring experience. Once the STT module converts the child's spoken input into text, this text is transmitted to ChatGPT, which then processes the input to determine whether the response is correct. The AI model applies natural language processing techniques to evaluate the response and, if it matches the expected answer, triggers the robot to deliver immediate praise or celebratory gestures. In cases where the response is ambiguous or incorrect, ChatGPT is programmed to re-ask the question or offer a step-by-step demonstration to guide the child towards the correct answer.

Integration with the simulation environment and the virtual robot is managed through a state machine, which the AI uses to control the sequence of actions and transitions, ensuring the robot responds appropriately to the child's input throughout the session.

Moreover, ChatGPT generates parent-facing prompts in instances where repeated errors are detected. For example, if the child consistently misidentifies numbers above five, the system might display a prompt such as "Your child might be missing items above 5, would you like to reduce the item count?" Such feedback is designed to be both informative and actionable, enabling parents to make real-time adjustments to the task difficulty. This adaptive feedback mechanism is essential for creating a "tutor-like" dynamic, where the system not only evaluates the child's performance but also suggests adjustments that can be immediately implemented to improve learning outcomes.

4.2.1.3 Multi-Threading and Response Timing

To achieve the near-real-time interactivity necessary for effective tutoring, the Python application employs a multi-threaded architecture. Separate threads are allocated for critical functions:

- **Audio Capture/Recognition:** This thread is responsible for continuously capturing the child's speech and transmitting the audio data to Google STT & TTS services without interruption.
- **ChatGPT Inference:** A dedicated thread processes the recognized text, determines its correctness, and generates an appropriate response based on pre-defined educational logic and dynamic adaptation rules.
- **Coppelia Sim Control:** Another thread manages the animation of the NAO robot, ensuring that the robot's gestures, head movements, and other visual cues are synchronized with the feedback generated by ChatGPT.

This parallel processing design ensures that the entire recognition and response cycle occurs within approximately 1 to 2 seconds, preserving the immediacy of the interaction. Maintaining low-latency feedback is critical for mimicking the natural flow of a live tutoring session, where conversational responsiveness is key to keeping the child engaged and the learning experience fluid. The efficiency of this architecture plays a central role in maintaining the illusion of a conversational, socially interactive system, similar to an in-person tutor who responds in real time to a child's cues.

4.2.2 Tasks

The simulation environment supports two primary tasks that are designed to match the cognitive and developmental levels of preschoolers:

1. **Counting Task:**

   In the counting task, the system displays a set of objects, ranging anywhere from 1 to 10 items, often colorful balls, on the virtual projection screen. The child is prompted with a question such as "How many balls do you see?" Once the child responds verbally, the speech is captured by the system, transcribed via STT, and then analyzed by ChatGPT for correctness. If the response is correct, the robot executes a celebratory animation (such as clapping or waving), reinforcing the child's effort. If the response is incorrect or ambiguous, the system offers a step-by-step demonstration or re-displays the objects to provide a clearer reference, thereby giving the child another opportunity to learn and respond accurately.

2. **Letter Recognition Task:**

   For the letter recognition task, the system presents individual letters (from A to Z) or short words on the projection screen. The child is asked to identify the letter or spell out the word, with prompts such as "Which letter is this?" or "Can you tell me the letters of this short word?" The procedure mirrors that of the counting task: correct answers yield immediate praise and animated feedback, while incorrect or incomplete responses result in clarifications and repeated prompts.

Parents have the flexibility to define which letters or ranges of letters are presented, tailoring the difficulty of the task to the child's current skill level.

Both tasks are carefully designed to be brief, typically lasting around 2 to 3 minutes to align with the short attention spans of preschool-aged children. The brevity of the sessions is intended to maximize engagement while minimizing the potential for fatigue or frustration. Additionally, parents are given the authority to adjust the maximum range of items or letters based on their child's individual capabilities. If repeated failures occur during a session, ChatGPT automatically prompts the parent to lower the difficulty, ensuring that the learning experience remains positive and appropriately challenging. Parents were asked to configure SAR tasks such as object counting and letter recognition, tailoring the difficulty and activity type to their child's developmental level. Structured questionnaires and thematic analysis of qualitative feedback were used to assess usability, engagement, and AI feedback clarity, alongside quantitative measures such as task completion rates and frequency of speech recognition errors. These findings highlighted the need for features like setup wizards, structured onboarding, and context-aware AI prompts to improve system usability [39].

4.3 Software Architecture of the Experimental System

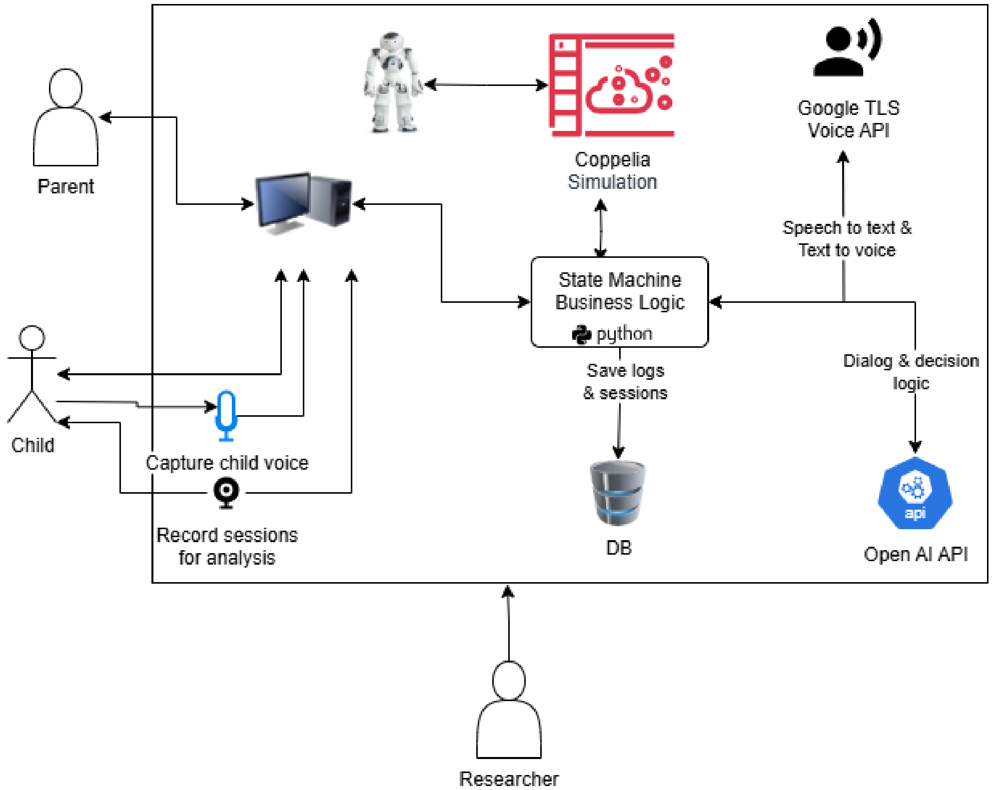

Figure 1 System Architecture

Figure-1 represents Overall system architecture of the AI-driven simulation-based SAR for early childhood education, illustrating key components and data flow.

The system architecture integrates a virtual robotic platform with a robust multi-threaded software stack and leverage the AI decision-making core, creating an interactive virtual learning environment that mirrors a physical SAR. The design centers on a virtual-simulated NAO humanoid robot within a 3D environment, controlled by a Python-based application that orchestrates speech recognition, dialogue generation, gestural animation, and real-time pedagogical logic. This section details the major components of the architecture, including the ChatGPT-based decision module, speech interfaces (ASR and

TTS/STT), the concurrent software framework, the Coppelia Sim simulation for the NAO robot, and the parent configuration interface, and explains how they interconnect to deliver an adaptive educational experience.

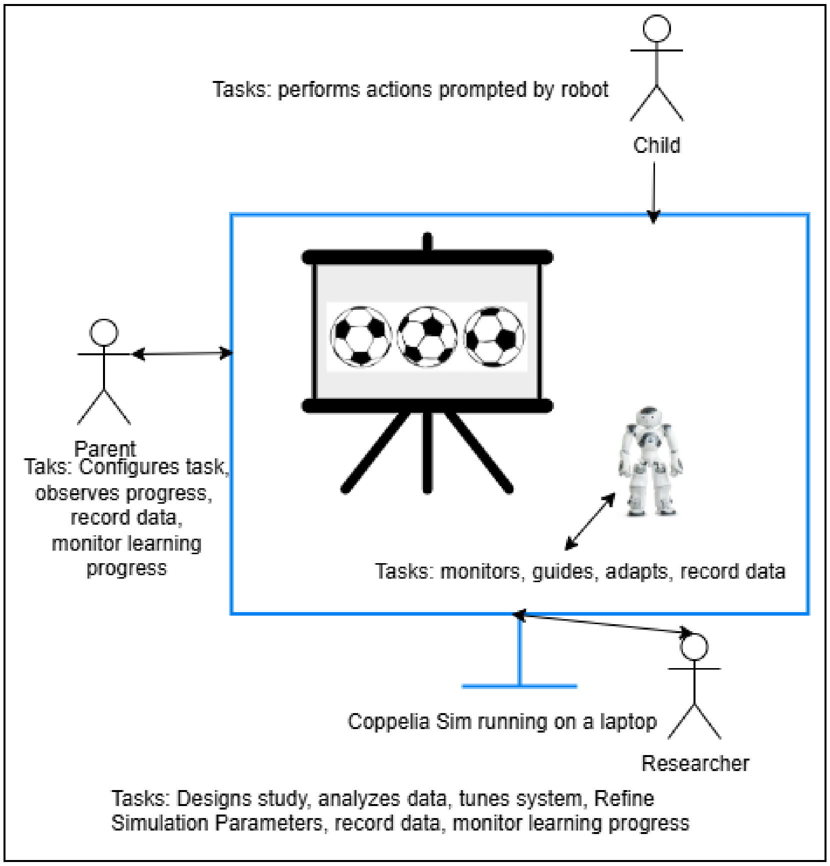

Figure 2 System Use Case

4.3.1 Simulation Platform and Robotic Avatar (Coppelia Sim and NAO Robot)

At the heart of the physical architecture is the NAO robot avatar simulated in Coppelia Sim. NAO is a child-sized humanoid robot model (58 cm tall) with multiple degrees of freedom (e.g. 2-DOF neck, 5-DOF arms) that enable realistic head, arm, and leg movements. The NAO model is placed in a virtual classroom scene rendered by Coppelia Sim, which provides a physics-enabled 3D environment for robot animation and sensor emulation. The virtual classroom includes a large projected display (screen) used for instructional visuals, for example, showing items to count or letters to identify, thereby emulating a real-world learning setting. The NAO avatar is programmed with a library of animated gestures (waving, clapping, head nodding, etc.), including celebratory motions triggered as positive reinforcement when the child answers correctly. A combination of pre-scripted movement sequences and dynamic motion commands is used to ensure the robot's responses appear smooth and natural.

Coppelia Sim is configured to accept external control commands from the main Python application via its remote API, allowing real-time manipulation of the robot's joints and the scene content. For instance, when a child gives a correct answer, the AI feedback pilot and control software immediately to send a command to NAO's simulation to perform a "celebration" gesture (like a hand wave or arm raise) accompanied by a happy animation on the screen. Likewise, if the answer is incorrect, the AI feedback through the state machine of the software, can direct the NAO to perform a gentle corrective gesture (e.g. a guiding hand motion) and update the on-screen content (such as highlighting the correct count) to assist the child. The virtual projector screen in the scene is dynamically updated by the system's logic, it can display different numbers of objects or letters in real time based on the task progression, ensuring that visual stimuli align with the learning context and the child's performance. This seamless integration of the virtual-simulated robot and environment with the control software provides an interactive experience analogous to a physical SAR, without the constraints of hardware (e.g. wear-and-tear or sensor noise).

4.3.2 Multi-Threaded / Layered Control Software and AI Decision Engine

All system functionality is coordinated by a custom multi-threaded / layered Python application, which serves as the central controller connecting the speech interfaces, the AI engine, the simulation environment, and the software logic. The software is architected into concurrent modules, each handling a distinct aspect of the interaction pipeline. Key modules include:

- **Automatic Speech Recognition (ASR)**: Listens to the child's speech via a microphone and transcribes it to text in real time. This is implemented using Google's Speech-to-Text API, chosen for its robust accuracy and ability to handle continuous speech input. The ASR module runs in its own thread, continuously monitoring the audio stream so that the child's responses are captured promptly and sent for processing.

- **AI Decision Module (ChatGPT Integration)**: This module uses OpenAI's ChatGPT API to provide real-time decision making. It receives the transcribed child words (along with context from the current state of the game/task) and generates an appropriate response or next action decision. The ChatGPT integration provides the system with advanced natural language understanding and generation capabilities. Crucially, ChatGPT analyzes the child's answer in context, for example, determining if the answer was correct, partially correct, or incorrect, and decides how the robot should respond (praise, encouragement, correction, hint, etc.).

- **System Logic and Session Manager**: This component maintains the overall state of the educational game and enforces the pedagogical rules defined for the task. It determines what question or prompt to give next, checks the child's answer against the expected result, and interfaces with ChatGPT to adjust difficulty or provide feedback. The system logic thread triggers ChatGPT queries with relevant context (e.g., the correct answer, the child's previous performance, any adaptive hints to include) and receives from ChatGPT a decision on how to respond. For example, if the child is learning to count, the system logic knows the target number of objects in the current round and can tell ChatGPT this information to help it decide if the child's spoken number is correct. The system logic also updates the virtual scene via the simulation API (to show the next set of objects or letters) and keeps track of engagement metrics like the number of attempts or time taken. It effectively ties together the recognition, the AI reasoning, and the robot's actions into a coherent instructional sequence, all done through a state machine design pattern.

- **Text-to-Speech (TTS) and Audio Output**: To vocalize the robot's dialog, the system uses a text-to-speech engine (for example, a standard TTS library or NAO's built-in voice) to convert ChatGPT's textual response into spoken words with an encouraging tone. This voice playback module runs in its own thread, enabling the system to prepare the next speech output even as the AI decision is being finalized. The use of multithreading here ensures there is a minimal lag between deciding on a response and the robot actually speaking it. The TTS module outputs audio through speakers, giving the illusion that the virtual NAO is talking to the child in real time.

- **Gesture and Animation Controller**: This module is responsible for synchronizing the NAO robot's movements and the on-screen visuals with the dialog. Upon receiving the AI decision from ChatGPT (e.g., "child was correct, give praise" or "child is incorrect, provide help"), the controller triggers the appropriate pre-defined animation in Coppelia Sim via API calls. For instance, if the decision is to congratulate, it might command NAO to clap its hands and display celebratory graphics on the virtual screen. If the decision is to correct the child, it could initiate a head-shake animation or point to objects while counting aloud. Because this runs on a separate thread, the system can begin a gesture immediately when the child finishes speaking, even while the speech recognition or AI processing might still be wrapping up. This parallelism helps the interaction feel fluid and lifelike, as the virtual robot can nod or gesture during the conversation without noticeable delay.

The multi-threaded architecture is critical for achieving real-time interactivity. It allows voice recognition, language understanding, physical animation, and speech synthesis to occur concurrently. For example, as soon as the child starts answering, the ASR thread captures and transcribes it; the moment the text is ready, the ChatGPT thread begins formulating a response; in parallel, the gesture controller readies a corresponding animation; and once the AI response text is returned, the TTS thread immediately voices it while the NAO animation plays. This design minimizes idle time between a child's action and the system's reaction. During testing, the end-to-end latency from a child's spoken answer to the robot's feedback was typically on the order of 1 to 1.5 seconds, which closely mimics natural conversation timing. Such low latency – facilitated by fast API calls and parallel processing, ensures the child remains engaged and perceives the interaction as immediate, preventing frustration that could arise from long pauses. (One bottleneck noted was the speech transcription speed; while ChatGPT's processing was near-instantaneous, Google STT sometimes introduced slight delays and other accuracy errors, so future optimizations may include streaming or prefetching techniques.)

4.3.3 Parent Configuration Interface and Task Customization

In addition to the child-facing components, the system includes a parent-side interface for configuring and personalizing the learning tasks. This interface, essentially a GUI front-end to the system logic, allows non-expert users (parents or educators) to set up the session parameters before the child interacts with the robot. Through the configuration panel, a parent can set the type of activity (e.g. counting objects or letter recognition), adjust the difficulty level (such as the range of numbers or the number of items to show), and input any custom content or rules (for instance, the child's name for personalized greetings, or specific learning objectives to emphasize).

To assist parents during setup, the system leverages the same ChatGPT AI engine as a configuration assistant.

Once the parent confirms the setup, these configurations (task type, content, rules for progression, etc.) are passed to the system logic module in the core application. The session manager then uses these parameters to drive the interaction. For example, the rules defined might include how many consecutive correct answers trigger a level-up in difficulty, or whether the robot should switch to a different game after a certain time. Parents' ability to tailor these rules is important for personalization; however, it was noted that some parents struggled with defining AI rules precisely, occasionally leading to unintended robot behaviors. This insight is being used to further refine the interface (such as incorporating drag-and-drop rule building and clearer explanations) so that the AI-driven adaptability remains under easy-to-understand human control.

4.3.4 System Workflow and Data Flow Sequence

With the above components in place, the end-to-end operation of the SAR system can be understood as a loop of interaction cycles between the child and the robot tutor, moderated by the AI logic. The typical flow of a single question-response exchange is as follows:

1. **Robot Prompt & Session Introduction:** The session begins with the virtual NAO robot greeting the child by name, creating a friendly and welcoming atmosphere. Next, the robot introduces the learning activity, such as counting objects or naming colors, using both speech and visual display. For example, the robot might say, "How many balls do you see?" while the screen shows a set of colored balls. This prompt is generated by the system logic (based on the chosen task and difficulty level) and delivered through the text-to-speech (TTS) module for voice, as well as Coppelia Sim for any accompanying gesture or pointing motion.

2. **Child's Response Capture:** The child responds by speaking (e.g., "three balls"). The system's microphone captures the spoken answer, and the ASR module immediately transcribes it to text. Google Speech-to-Text is used for this transcription; it sends the audio stream to the cloud and returns the recognized text. In cases where the child's speech is unclear or contains childish mispronunciations, the transcription result may be imperfect.

3. **AI Evaluation (ChatGPT Decision):** The recognized text (the child's answer) is forwarded to the ChatGPT API, along with context from the system logic (for instance, the correct answer or the next step). ChatGPT processes the input in real time, interpreting the child's intent and accuracy. Based on this evaluation, ChatGPT classifies the response as correct, partially correct, or incorrect, and decides on the appropriate feedback strategy. This decision is essentially a piece of generated text or a structured command that the system logic can interpret (e.g., a message like "Correct answer – praise and increase difficulty" or "Incorrect – provide a hint and repeat the question").

4. **Feedback and Robot Reaction:** Given the AI's decision, the system executes the corresponding feedback actions. If the child was correct, the system logic signals a positive reinforcement sequence: the NAO robot enthusiastically acknowledges the success (for instance, *"Correct, great job!"*) and performs a celebratory animation (clapping, cheering, virtual confetti on screen). This involves sending the ChatGPT-generated praise text to the TTS module for speech output and commanding Coppelia Sim for the gesture and any visual effects. If the child's

answer was wrong or incomplete, the system adopts a step-by-step corrective approach: the NAO might say encouragingly, *"Not quite. Let's count together,"* and then demonstrate the correct answer by counting the objects aloud while pointing, before giving the child another tries. For partially correct answers, the AI can direct the robot to give a subtle hint or ask a follow-up question rather than an outright correction, thereby encouraging the child to self-correct. Throughout this feedback step, the voice, gesture, and visual display are tightly coordinated to maintain the child's attention, for example, the robot's head will turn toward the objects as it counts them, aligning verbal and visual cues.

5. **Adaptive Task Adjustment:** After providing feedback, the system's logic updates the task state. The child's performance on that question (success or error, response time, need for hints) is logged and used to adjust the upcoming content. The architecture supports adaptive difficulty modulation: if the child has answered several questions correctly in a row, the AI may decide to increase the challenge (e.g., showing more items to count, or moving to a higher number or a new letter). Conversely, if the child is struggling (multiple errors or needing a lot of help), the system can simplify the task or give additional practice on easier items. This dynamic adjustment is done in real time by ChatGPT in conjunction with preset pedagogical rules. For instance, ChatGPT might be prompted by the system logic with something like, *"Child has made 3 errors on counting to 5. Suggest next action,"* and it might respond, *"Repeat with 3 balls and guide the counting."* Thus, the AI-driven logic ensures the session remains neither too hard nor too easy, matching the child's learning pace. The visual content on the screen

is then updated accordingly (e.g., new number of objects, next letter, etc.) for the next question.

6. **Engagement Monitoring:** In parallel with academic performance, the system monitors the child's engagement level through indirect signals (since it's a simulation, direct gaze or face recognition is not used, but timing and interaction patterns are). The system tracks if the child hesitates too long or stops responding. If a certain threshold of inactivity or incorrect attempts is reached, the system can prompt an engagement strategy, typically with the help of the researcher. For example, the NAO might say a motivating phrase or display a fun animation to recapture attention, but these interventions are usually triggered manually or semi-automatically by the researcher rather than autonomously by the AI. This approach helps maintain an effective learning interaction, similar to how a human tutor would notice and address a student losing focus. By introducing supportive prompts or playful moments as needed, the system aims to keep the child involved throughout the session.

7. **Loop Continuation or Termination:** The next question or task is then presented, and steps 1–6 repeat. The session continues in this cyclic manner until a stopping condition is reached (for example, a fixed number of tasks completed, a time limit, or a parent/child terminating the session). At the end of the session, the robot provides a brief conclusion or summary (e.g., *"Well done! You counted a lot today. See you next time!"*), and the system may log a performance report. The data logged (responses, accuracy, hints given, etc.) can later be reviewed by

parents or researchers to evaluate the learning outcomes and the system's effectiveness.

Throughout this process, real-time coordination among components is what enables a smooth user experience. The concurrency of threads means the child is rarely kept waiting, while one question is being spoken by the TTS, the next visual is already loading; while the child is viewing an animation, the ASR is ready to capture their next response. This design proved essential for young children's engagement, as any noticeable lag could lead to distraction. Indeed, during evaluations with 3-4-year-olds, the responsiveness of the simulated SAR was highlighted as a key factor in sustaining their attention. By leveraging an AI-driven pipeline for understanding and response generation, the system can handle the unpredictability of child input in a flexible manner, much like a human tutor would, interpreting mistakes generously and keeping the interaction flowing.

In summary, the system architecture of the AI-driven, simulation-based SAR consists of tightly integrated modules for speech, vision, and action that operate under the guidance of a powerful language model. ChatGPT serves as the cognitive core, interpreting the child's responses and deciding the robot's pedagogical moves in real time. Google's ASR provides the hearing ability, and the TTS voice provides the robot's speech, enabling natural spoken dialogue. The Coppelia Sim NAO provides a body for the AI to inhabit, conveying social cues through motion and creating a sense of presence. Finally, the parent system configuration wraps the system in a user-friendly setup layer, ensuring that the activities can be tailored to each child and that even non-technical users can deploy the SAR effectively at home. This architecture demonstrates how modern AI (large language models) can be combined with interactive simulations to create a Socially Assistive Robot tutor that is both adaptable and accessible, potentially delivering personalized early education on a scale.

The overall interaction cycle, from parent configuration through robot–child exchange and adaptive feedback, is summarized in the sequence diagram in Figure-3:

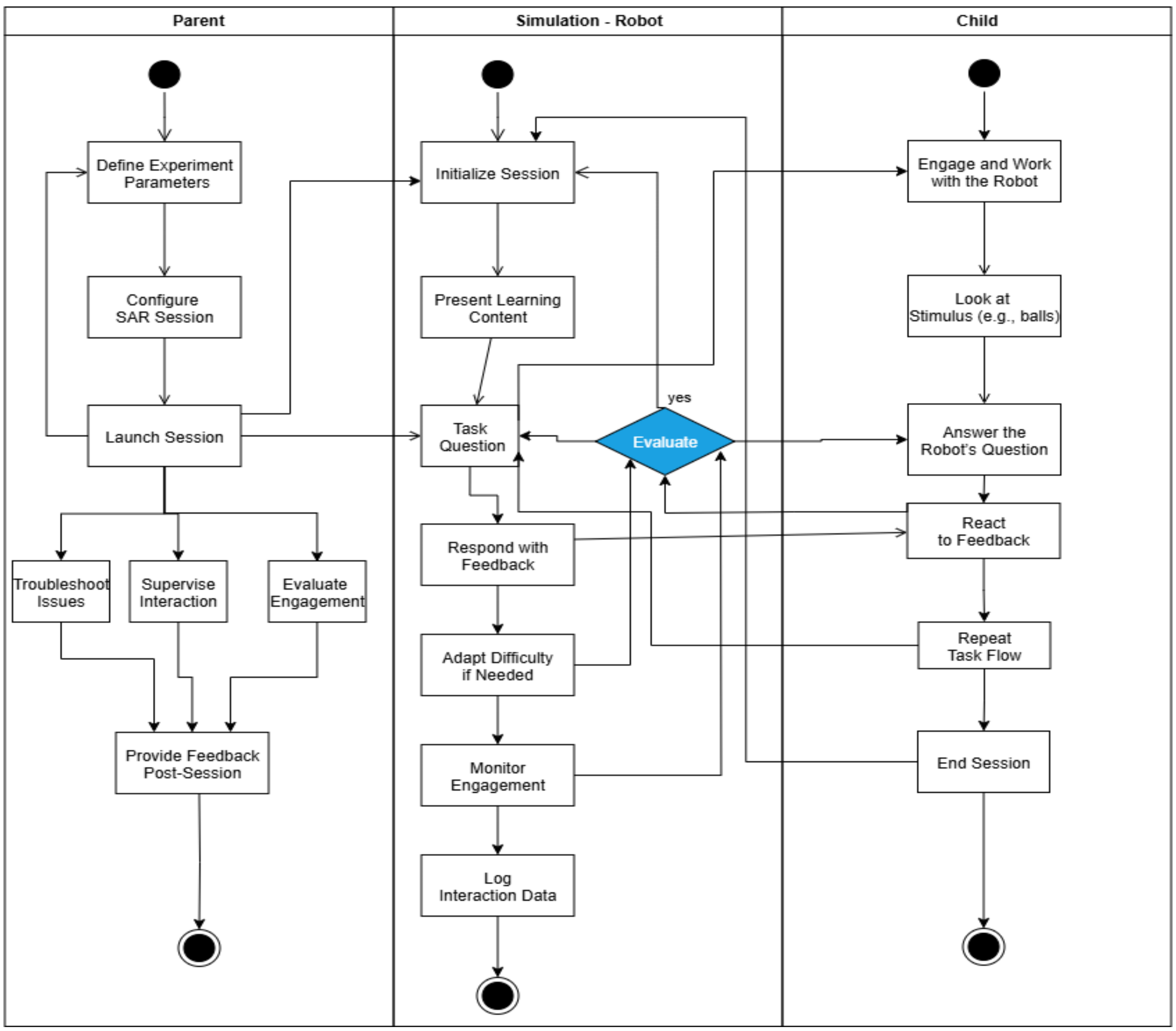

Figure 3 System Sequence Diagram

4.4 Procedure

4.4.1 Onboarding and Tutorial

Each parent received a short, printed tutorial papers detailing:

- **Simulation Setup & starting:** Explanations of how starting the simulation environment.
- **Task Range Configuration**: Suggested defaults (e.g., 1–5 items for a younger child; 1–10 for an older child who can count well).
- **AI Suggestions**: Illustrations of typical ChatGPT prompts, encouraging parents to watch for and interpret them.

Parents were free to do a "test run" alone, verifying that the robot avatar responded correctly to adult speech. Once comfortable, they introduced the child.

4.4.2 Child Session

Each child typically attempted one or two tasks (counting, letter recognition), each lasting ~2–3 minutes if the child remained engaged. Steps in a typical session:

1. **Robot Introduction**: The virtual NAO greeted the child by name, briefly explaining the task.
2. **Prompt**: The system displayed objects (for counting) or letters, then asked a question (e.g., "How many apples do you see?").
3. **Child Response**: The child answered verbally. The recognized text was fed to ChatGPT, which determined correctness or partial correctness.

4. **Feedback**: Correct replies, triggered positive TTS lines ("Fantastic job!") and celebratory animations; incorrect replies led to partial demos, re-asking, or simpler variations if repeated errors were detected.

5. **Session Completion**: If the child completed ~4–6 successful attempts or lost interest, the parent concluded the session. The entire process was video recorded for post-analysis.

For younger children (3-year-olds), some sessions ended after 1–2 prompts if the child disengaged. Parents could attempt re-engagement, but if repeated speech errors persisted, researcher will help to process to the next question / task.

4.4.3 Adaptation and Parent Involvement

When repeated child errors occurred, ChatGPT might order the system using its state machine to produce parent-facing pop-ups: "Would you like to reduce the number of objects to 5?" or "Consider repeating the counting steps together." Parents who enabled these suggestions were free to accept or ignore them. Meanwhile, if the recognized text was clearly incorrect, the researcher manually intervened to ensure the activity reflected the child's actual response.

This dynamic created a **"human-in-the-loop"** approach for the false-positive answers, especially for younger children who frequently needed help if speech recognition flagged correct answers as "wrong." Parents were also free to up-level difficulty if the child answered swiftly and wanted more challenge.

4.4.4 Data Logging

To enable comprehensive analysis, multiple data sources were captured:

- **Video Recordings**: Observed child posture, facial expressions, signs of frustration or disinterest, and approximate session length.

- **System Logs**: Timestamped recognized speech, final correctness decisions, number of AI suggestions displayed, and any manual overrides.

- **Survey + Open-Ended Responses**: Post-session, parents answered Likert-scale items on speech recognition difficulty, setup complexity, AI clarity, and child engagement, plus optional free-text prompts for more detailed feedback.

Parents completed these questionnaires after each child session or at the end of the experiment day, ensuring fresh recollections of the session experience.

4.5 Data Analysis

4.5.1 Child Performance

1. **Accuracy / Success Rate**: We computed each child's correct answers over total attempts in counting, or letter recognition tasks. For older children, we also noted if they completed tasks consistently or expressed disinterest from ease. For younger children, we recorded all attempts, noting partial improvements when repeated demonstrations eventually led to correct answers.

2. **Engagement**: From video, we coded whether the child-maintained eye contact with the screen, responded promptly, or showed restlessness. We also noted if a session ended early due to child frustration or disinterest.

3. **Session Duration**: The approximate time from the robot greeting to session conclusion (usually 2–3 minutes for successful runs, less if a child disengaged).

4.5.2 Parent Usability

1. **Likert Ratings**: On a 1–5 scale, parents rated overall setup difficulty, speech recognition difficulty, and AI clarity. We computed means (M) and standard deviations (SD) across all parents.

2. **Open-Ended Feedback**: We thematically coded remarks about repeated speech errors ("child said 'three,' recognized as 'tree'"), interface confusion, or how ChatGPT suggestions were used. These codes allowed us to identify patterns like "Wanted a wizard" or "Manual override was too frequent."

3. **Correlations**: We examined whether higher levels of speech recognition frustration were associated with early session termination by children.

4.5.3 Integration of Child and Parent Data

Finally, we merged:

- **Child engagement and success logs**: To determine whether children maintained consistent performance, became frustrated, or ended sessions early.

- **Parent interventions**: Whether the parent overcame speech errors with researcher help or adjusting task difficulty when necessary.

- **Session outcomes**: Assessing whether the session was completed successfully or terminated early, and how this related to both the child's skill level and the parent's ability to configure and adapt the system.

This integrated approach allowed us to confirm which design elements (short tasks, domain-limited recognition, parent overrides, AI suggestions) contributed to a smoother experience for older preschoolers and partially successful experiences for younger ones, insights later expanded upon in the Discussion and Recommendations sections.

5. COMPARATIVE ANALYSIS OF SAVRS AND EDUCATIONAL SOFTWARE

5.1 Educational Outcomes and Learning Effectiveness

Socially Assistive Robots (SARs), including both physical robots and Socially Assistive Virtual Robots (SAVRs), have consistently demonstrated effective learning outcomes comparable to traditional tutoring methods, significantly improving cognitive and affective skills [12], [61]. Physical SARs capable of dynamically adapting instructions based on student performance show notable improvements in learning gains, supporting personalized education [12]. Recent research, including this study, confirms that children can also learn effectively from well-designed SAVRs (virtual agents), achieving outcomes comparable to those reported for human teachers or physical robots [3], [16], [62]. Longitudinal studies with physical SARs indicate sustained improvements and retention of concepts due to repeated social interactions [61], [63]. Emerging research further suggests that SAVRs can foster similar benefits, particularly when enhanced by adaptive AI and rich interactive feedback [3], [16], [21]. These agents effectively support deeper understanding through personalized guidance and timely feedback, which encourages active learning and reflective thinking [63]. In particular, SAVRs have shown promise in enhancing language acquisition, mathematics skills, and problem-solving abilities in young children, aligning closely with key educational standards [3], [16], [21]. Math-focused educational applications, such as DragonBox and Mathletics, significantly enhance numeracy skills, mathematical problem-solving, and student engagement through gamification [64] [65]. A randomized controlled study found substantial improvements in algebra proficiency after short interventions using these applications, illustrating their efficacy in targeted mathematical contexts [64]. Research highlights significant increases in analytical thinking and conceptual understanding, promoting deeper learning experiences beyond mere rote memorization [66]. Additionally, math-focused apps effectively combine intrinsic and extrinsic motivational strategies, greatly enhancing student perseverance and engagement with challenging content [66]. These applications have demonstrated considerable effectiveness in promoting self-directed learning and improving students' confidence and attitudes towards mathematics, contributing to overall academic improvement [65] [66].

Cartoon-based learning platforms such as Khan Academy Kids and ABCmouse have demonstrated substantial gains in foundational literacy and numeracy skills through interactive storytelling and engaging animations [67]. These platforms effectively foster school readiness and sustain engagement for young learners, particularly when usage is consistent and well-supported [67]. Studies specifically note improvements in phonemic awareness, basic arithmetic skills, and narrative comprehension [68]. These platforms employ immersive narratives and familiar characters, significantly reinforcing learner motivation, cognitive retention, and facilitating the transfer of learned concepts to new contexts [68]. Additionally, cartoon-based educational platforms actively promote creativity and imagination, further enhancing cognitive development and educational outcomes [68].

5.2 Engagement and Sustained Attention

SAVRs enhance student engagement through social interactions and adaptive behaviors, significantly reducing distractions and maintaining attention through personalized interactions [12] [69]. Studies emphasize that children interacting with SARs display fewer off-task behaviors, higher persistence rates, and increased willingness to engage in challenging tasks [69] [70]. These robots also effectively utilize social presence and interactive feedback to cultivate a highly engaging learning environment, significantly enhancing sustained attention spans in young learners [70]. The interactive and personalized nature of SARs makes learning experiences more compelling, directly influencing students' intrinsic motivation and sustained interest in educational activities [70].

Math-focused educational applications successfully utilize gamification to maintain attention and motivation, particularly through competitive and reward-driven mechanisms. However, without intrinsic social interaction, they may not sustain long-term attention and engagement equally effectively [64] [65]. Research underscores the critical role of game-based elements such as points, leaderboards, and badges in motivating sustained student participation in educational apps [64] [71]. Such gamified interactions effectively maintain learner interest by continuously challenging users with escalating tasks and rewarding incremental achievements, thus significantly boosting learner motivation and persistence [71]. Furthermore, math-focused applications that integrate adaptive difficulty levels effectively sustain student interest by balancing challenge and skill, promoting optimal learning engagement [71].

Cartoon-based platforms utilize engaging narratives and appealing visual animations to sustain attention effectively in short bursts, ideal for younger children. However, engagement may wane over time without substantial variability or external motivational factors [67]. Studies suggest that the presence of familiar animated characters significantly increases emotional connection and sustained interaction compared to traditional app interfaces [67] [72]. This engagement is often maintained through repeated interactions with beloved characters and storylines that provide a comforting yet stimulating learning context [72]. Additionally, these platforms often use humor and playful scenarios to retain learner interest, enhancing emotional engagement and sustaining attention over extended periods [72].

5.3 Adaptivity and Responsiveness

SAVRs provide advanced adaptive learning experiences by personalizing instructional content and offering immediate social and emotional feedback, enhancing engagement and learning efficacy [12] [61]. High responsiveness and ability to adapt based on emotional and cognitive states significantly surpass typical educational applications. Cutting-edge systems incorporate deep learning techniques to refine interaction strategies dynamically, enhancing responsiveness [73]. Advanced adaptivity fosters real-time personalized learning, significantly enhancing instructional content relevance and promoting deeper cognitive engagement [73]. SARs equipped with emotion detection capabilities adapt interactions, providing tailored emotional support and feedback aligned with learners' immediate emotional states, thus significantly improving educational experiences [73]. Furthermore, continuous real-time data analysis and predictive modeling allow SARs to anticipate learning difficulties, proactively providing support before frustration or disengagement occurs [73].

Math-focused applications employ mastery-based progression adaptivity, adjusting problem difficulty based on student performance but lack nuanced responsiveness to emotional states or complex interactive behaviors [64] [65]. Recent research highlights potential improvements in emotional adaptability, significantly reducing learner frustration and improving retention rates [74]. Enhanced adaptivity strategies improve student engagement through individualized feedback mechanisms, supporting emotional and academic needs simultaneously [74]. Innovations such as adaptive hint systems and scaffolded learning support reduce anxiety and boost learner confidence, enhancing educational outcomes [74]. The integration of predictive analytics into math-focused applications could further enhance their adaptivity, allowing preemptive identification of struggling students and provision of tailored instructional support [74].

Cartoon-based platforms adapt primarily through curriculum adjustments based on student progression but provide limited real-time responsiveness or nuanced adaptivity based on immediate learner states [67]. Periodic assessment checkpoints effectively adapt future content delivery, yet real-time adaptive feedback during interactions remains limited [67] [75]. Incorporating immediate adaptivity measures significantly enhances real-time interaction quality and educational efficacy [75]. Adaptive storytelling and branching narratives could substantially enrich the educational impact of cartoon-based learning platforms, creating immersive and personalized learning experiences [75]. Moreover, leveraging real-time interaction data to dynamically adjust narrative paths and difficulty levels would enhance the personalization and educational value of these platforms, addressing diverse learner needs and preferences [75].

5.4 Simulation-Based SARs vs. Traditional Early Childhood Educational Apps

5.4.1 Interaction Modalities in Early Childhood Learning

Based on our findings, simulation-based learning with socially assistive virtual robots (SAVRs) enables speech-based, real-time interactions, allowing children to engage in natural dialogue and social behaviors that mimic human tutoring. In contrast, traditional computer-aided educational apps, such as ABCmouse or Khan Academy Kids, rely primarily on visual and touch input modalities [21], [76]. Unlike most tablet apps, which rely primarily on touch inputs and pre-recorded prompts, SAVRs create opportunities for peer-like turn-taking and social engagement [21]. For instance, studies have observed that preschool children interacting with a robot tutor tend to respond verbally, follow the agent's gestures, and exhibit spontaneous social behaviors rarely seen with standard educational software [21], [74]. The presence of real-time, adaptive dialogue in SAR systems enhances social presence, prompting children to treat the agent as an interactive peer rather than a static program [19], [21]. These socially contingent interactions are crucial for capturing and sustaining young learners' attention, especially in early childhood [74].

In our SAVR system, multimodal feedback, including speech, visual cues, and simulated gestures, was used to enhance interactivity and engagement. For example, a robot tutor may use cheerful vocal encouragement, expressive facial animations, and digital stickers to reinforce correct answers and motivate learners [74]. Recent IEEE research on inclusive early education has shown that such multimodal, interactive features significantly improve engagement and learning outcomes, especially for children with learning challenges [74]. In contrast, traditional educational apps, despite their use of bright visuals and sound effects, lack genuine interactivity and tend to offer static, non-contingent feedback [19]. This difference in feedback depth and adaptivity explains why SAR-based platforms tend to hold a child's attention longer and support diverse learning preferences [21], [74].

5.4.2 Real-Time Adaptation and Responsiveness

A key strength of SAR-based systems is their ability to deliver real-time adaptation and immediate feedback. In our SAVRs system, we leverage artificial intelligence to monitor user input, including speech recognition timing and interaction flow, and adapt responses accordingly. While prior studies have explored adaptive strategies in educational apps [74], [76], our approach uniquely integrates these mechanisms into a virtual robot-based simulation. For example, in our SAVRs system, if a child hesitates or answers incorrectly, the robot tutor can quickly offer additional support, a tailored hint, or encouragement within the same interaction cycle. While prior work has demonstrated that adaptive apps can boost motivation and learning persistence through personalized feedback [74], [76], such strategies are rarely integrated with embodied or speech-driven interactions. By contrast, most traditional apps rely on predetermined responses or fixed content sequences, adjusting little in real time to the child's needs or engagement levels [19].

Notably, some SARs, especially those enhanced with affective computing, have shown potential to detect signs of distraction or emotional disengagement and trigger re-engagement strategies, such as using the child's name, changing tone, or offering playful prompts. This type of moment-to-moment adaptivity, largely absent in typical early childhood software, enables a more personalized and supportive learning environment, especially valuable for preschoolers with limited attention spans [25].

5.4.3 Comparative Learning Outcomes and Child Responses

While prior research on educational apps has demonstrated positive learning outcomes and engagement [21], [74], [76], this study adds evidence that simulation-based Socially Assistive Virtual Robots (SAVRs) can achieve comparable or enhanced effects in early childhood learning environments. In our simulation-based pilot study, children interacting with the SAVR system, demonstrated learning gains and were able to answer correctly following repeated, adaptive support, which is comparable to outcomes reported for high-quality educational apps [74], [76].

However, the process of learning and child response patterns appeared qualitatively distinct. In our pilot study, children using the SAVR simulation often showed visible enthusiasm, required minimal adult prompting, and remained focused across task sequences. While we did not conduct a direct comparison with app-based systems, these observations are consistent with patterns reported in SAR literature. Additionally, children occasionally reacted emotionally to the robot's behavior or engaged with it socially, such as smiling, repeating its words, or responding with expressive gestures. These kinds of social engagement behaviors are characteristic of interactions with socially contingent agents and, according to prior research, are less commonly observed in touchscreen-based educational apps [19].

The study findings further suggest that the SAVR approach can promote higher behavioral and emotional engagement, as reflected in more frequent child-initiated questions, longer sustained attention, and a greater willingness to persist through challenges. While some adaptive apps have shown positive motivational effects [21], [74], the results of this study indicate that simulated robot interaction, especially when paired with tailored feedback and gestures, provides a more immersive and socially responsive experience.

Finally, we noted that several children requested repeated sessions with the virtual robot and showed signs of intrinsic motivation without parental prompting. Although traditional apps have demonstrated short-term engagement [58], they often struggle to maintain attention once the novelty fades [19]. In contrast, the SAVR simulation-based environment, maintained children's interest across multiple tasks, reinforcing its potential as a viable and engaging alternative to static educational tools.

5.4.4 Summary and Implications

While both simulation-based SAVR systems and traditional educational software can effectively support foundational skills, our pilot study and the broader literature indicate qualitative advantages for SARs in terms of social presence, adaptivity, multimodal feedback, and child engagement [19], [21], [25], [74], [76]. Although most prior research has focused on either physical SARs or adaptive apps, these features are especially important in early childhood education, as they capitalize on children’s social instincts, increase attention and motivation, and foster positive learning behaviors.

As SAVR technology matures and becomes more accessible, its potential to democratize high-quality early learning at scale is significant. Ongoing research continues to explore these systems' effects on socio-emotional development and their integration with hybrid physical-virtual learning environments [25]. The findings summarized here reinforce the rationale for further research and development of simulation-based SAVRs as a promising tool in early childhood educational innovation.

6. RESULTS

In this section, we present and analyze the data from both child performance (accuracy, engagement, session completion) and parent usability (speech recognition challenges, perceived interface difficulty, AI feedback clarity). By integrating these two perspectives, child outcomes and parent experiences, our aim is to offer a comprehensive understanding of how ChatGPT-powered, simulation-based SAVRs perform in real home contexts.

6.1 Child Performance and Engagement

We begin by examining how the children performed in terms of counting, focusing on overall success rates, session durations, and behavioral observations. Notably, the eight participating children are split by age group: five 4-year-olds (A, B, C, D, E) and three 3-year-olds (F, G, H). This division illuminates how speech recognition mismatches and task complexity vary with developmental readiness.

6.1.1 Overview of Task Success (Tables I–III)

Table I summarizes initial vs. post-assistance success rates for each child, highlighting whether repeated demos, task simplification, or parent interventions improved performance. Meanwhile, Tables II and III detail engagement metrics, such as average time to completion, attention levels, and motivational cues, derived from video recordings and system logs.

Table I Task Performance Summary

Participant	Initial Success Rate	Post-Assistance Success	Improvement	Key Observations
Child A (4y)	100%	N/A	Consistent	Zero guidance needed; responded rapidly; stable performance
Child B (4y)	100%	N/A	Consistent	Confident, few misrecognitions; tasks felt "easy"
Child C (4y)	100%	N/A	Consistent	Engaged, sometimes preempted robot's question; liked visuals
Child D (4y)	100%	N/A	Consistent	Enthusiastic, repeated tasks "for fun," minimal frustrations
Child E (4y)	100%	N/A	Consistent	Completed quickly, rarely misinterprets; no sign of disinterest
Child F (3y)	20%	40%	+20%	Improved with repeated demos & lowered difficulty; overcame some mispronunciations
Child G (3y)	0%	N/A	No engagement	Shy, refused to speak, parents ended session early
Child H (3y)	40%	Partial	N/A	Answered 2 prompts, then lost focus; left mid-session

Interpretation

- **4-year-olds** (A–E): Each reached 100% success from the initial session onward, reflecting minimal speech misinterpretation and strong motivation. They never needed repeated demos or parental overrides.

- **3-year-olds** (F, G, H): Showed widely varied engagement and success. Child F improved from 20% to 40% across repeated attempts, while Child G declined to participate verbally. Child H began well (40% success) but quit after a few prompts due to speech errors or general disinterest.

Quantitative analysis revealed that all participating 4-year-olds completed 100% of their assigned tasks with high engagement and minimal corrective feedback, indicating strong cognitive readiness. In contrast, 3-year-olds demonstrated more variable outcomes: some required additional multimodal cues and simplified instructions to maintain focus and achieve partial task completion. These findings underscore the need for age-appropriate adaptation and robust AI support in early childhood simulations [26].

Engagement Metrics: Tables II and III

TABLE II (4-year-olds) and TABLE III (3-year-olds) expand on average times, response patterns, observed motivation, and any distinctive behaviors. These details derive from video coding and system logs, providing a qualitative dimension to the quantitative success rates above.

Table II Engagement Metrics and Observations– 4-Year-Old Group

Metric	**Child A**	**Child B**	**Child C**	**Child D**	**Child E**
Avg. time to complete	~2 min 40 sec	~2 min 45 sec	~2 min 55 sec	~2 min 48 sec	~2 min 50 sec
Response pattern	Often preempted prompts	Confident, rarely hesitated	Engaged but selectively attentive to visuals	Mostly correct, occasionally asked for re-	Prompt, high accuracy

				prompts if bored	
Attention level	Very high, consistent	Generally focused, minimal drifting	High for colorful tasks, responded quickly	Engaged, mild wandering but returned promptly	Fully attentive, rarely looked away
Motivation	Very excited, found tasks "fun"	Enthusiastic about correct answers, minimal confusion	Enjoyed variety, disliked repeated identical tasks	Mixed interest but never refused to continue	Strong motivation, no sign of frustration
Key behavior	No manual overrides needed; speech recognized well	Rare overrides, overcame small misfires by re-speaking	Quickly overcame minor speech issues, responded before the robot finished sometimes	Found tasks "easy," repeated them for fun gestures	Breezed through tasks, occasionally asked for "harder" questions

Table III Engagement Metrics and Observations – 3-Year-Old Group

Metric	**Child F**	**Child G**	**Child H**
Avg. time to complete	~3 min 15 sec	N/A (session ended early)	N/A (exited after partial)
Response pattern	Initially hesitant, overcame some articulation with repeated demos	Refused to respond verbally, ended session	Answered 2 prompts, lost focus, left session
Attention level	Partial at first, improved after the first correct attempt	Zero engagement; turned away from monitor, showed no interest	Began curious, quickly lost interest following a misinterpretation

Motivation	Gained confidence after hearing “Great job!” for partial success	Displayed shyness or discomfort, minimal willingness to speak	Early excitement but meltdown or disinterest after minimal errors
Key behavior	Overcame “fwee” → “three” mismatch with parent overrides; improved to 40% success	Did not talk beyond initial greeting; parent ended the session	Possibly overshadowed by environment stimuli; child walked away mid-activity

Interpretation

- **4-year-olds** consistently completed tasks in ~2–3 minutes, rarely encountering speech-based disruptions. They showed curiosity or asked to do more, indicating the simulation environment was sufficiently “fun” and immediate.
- **3-year-olds** each displayed different patterns: F overcame some mispronunciations with repeated, slower demos from the robot; G never engaged, presumably too shy or discouraged by early speech issues; H began well but left after a short time, possibly due to multiple speech misfires or short attention.

6.1.2 Qualitative Observations on Child Engagement

From the video analysis and post-session parent interviews, we learned that older preschoolers viewed the virtual NAO robot as a “friendly, talking character,” responding with excitement at each correct praise. They seemed unbothered by the lack of physical embodiment, finding the animations convincing enough for short tasks (2–3 minutes). In contrast, younger children required more adult mediation, particularly Child F, whose initial frustrations improved only after the parent repeatedly typed manual overrides.

Child G refused speech; this observation suggests that, for some shy children, a misinterpreted first or second speech attempt may discourage further attempts at re-speaking.

Some parents also noted that 3-year-olds sometimes needed more “tactile” or “physical” stimuli to remain engaged, referencing the child’s usual preference for tangibility. These remarks highlight the possibility that an exclusively on-screen environment may not suffice for certain younger children unless supplemented by highly engaging animations, short session times, and minimal frustration from speech errors.

6.2 Parent Usability Findings

We now turn to the parent experiences configuring the environment, handling speech recognition difficulties, and interpreting ChatGPT prompts. This perspective is critical because, in home contexts, parents act as both system operators and child facilitators.

6.2.1 Overall Setup Difficulty

On a 1–5 scale (1 = Very Easy, 5 = Very Hard), parents averaged **3.0** (SD = ~0.82), indicating a “moderate” level of complexity. Specific feedback patterns include:

- **40%** found it “somewhat challenging,” largely due to uncertain numeric range choices or confusion about speech threshold adjustments.
- **60%** considered it “manageable,” stating they followed the PDF tutorial or used trial-and-error to locate feasible tasks.

One parent wrote:

“I wanted a single button that said, ‘Count 1–5.’ Instead, I typed in numeric fields and guessed. My child struggled at first, so I had to switch to fewer items after multiple errors”.

Such remarks underscore the desire for wizard-based or template approaches, consistent with prior educational technology, research highlighting that non-expert users prefer simple, guided setups rather than open-ended configuration.

6.2.2 Speech Recognition as Major Frustration

70% rated speech recognition difficulty as 4 or 5 on the 1–5 scale for younger children, with 3-year-old speech especially problematic. For 4-year-olds, parents encountered far fewer issues, typically rating it 2–3. Overall, the mean hovered around 3.4–3.5, reflecting moderate to high annoyance. Common misinterpretations included:

- “Three” → “Tree,” “Free,” or “We”.
- “Four” → “For,” “Fore,” or occasionally random words.
- Child partial speech recognized as silence or incomplete text.

Parents often typed manual overrides to correct recognized text. During post-session interviews, one parent indicated that repeated override cycles disrupted the flow and forced them to be constantly on alert. These speech-based frictions especially undermined younger children’s readiness to continue, consistent with the meltdown observed in Child G and Child H. Meanwhile, older kids overcame minor errors in recognition by re-speaking slowly.

6.2.3 AI Feedback Clarity

When asked about ChatGPT's parent-facing prompts, parents averaged 2.8 on clarity (SD = 0.9), indicating mild dissatisfaction or confusion. While they appreciated the concept of "the AI noticing my child is stuck," they found suggestions like "Try simpler tasks" too vague. Without direct, clickable UI links, parents had to interpret the instructions and manually access the settings file to reduce item count sets. Some parents effectively used the suggestions, while others ignored them, complaining that the instructions did not reference specific controls:

"The AI said, 'Lower difficulty.' I had no idea if that meant 1–5 or 1–3. A direct link would have helped."

Real-time, context-aware prompts delivered by the AI were particularly valuable for supporting parents as they modified tasks or troubleshot system errors. Parents who engaged with these prompts reported increased confidence in adapting to their children's needs, ultimately leading to higher engagement and improved learning outcomes [39].

This mismatch between advanced AI logic and user-friendly interface design remains a recurring theme in human-centered AI research.

6.2.4 Time Constraints and Manual Overrides

Many parents, especially those with multiple children or busy schedules, commented on how repeated speech errors or guesswork in configuration consumed more time than anticipated. One parent described typing an override "almost every 30 seconds" for a 3-year-old who consistently said "fwee" for "four," frustrating both child and adult. Another parent found that once they discovered an optimal difficulty (e.g., 1–5 items), tasks flowed smoothly, but it took “several sessions of trial and error.” This dynamic implies that if parents do not find workable settings relatively quickly, they might abandon usage altogether.

6.3 Speech Recognition Analysis

In addition to parent feedback, the system-logs provided insight into how often recognized text matched child speech. For older children, the correct recognition rate approximated 85–90%. Minor misfires typically occurred with short/unclear speech, but children corrected themselves. For 3-year-olds, recognized text was correct on the first try only ~60–70% of the time. In tasks with 4–6 prompts, that level of error can lead to repeated negative feedback. The manual override logs show that in sessions for Child F, the parent manually corrected speech around 4–6 times in a ~3-minute window. For Child G, presumably, no data exists beyond the initial attempts, as the child refused to proceed. Meanwhile, Child H’s logs indicate 2–3 overrides in a short 2-prompt session, after which the child left.

These results echo parent narratives, confirming that mainstream STT remains inadequate for many 3-year-olds unless domain-limited or phonetically expanded. In counting tasks that used a narrower set of numeric words (1–5 or 1–10), misrecognition rates dropped slightly, but children who spoke softly or in “baby talk” still faced repeated rejections.

6.4 Integrating Child and Parent Perspectives

6.4.1 Younger Children + High Speech Recognition Failures = Early sessions termination

When misinterpretations arose repeatedly, younger children quickly lost motivation. Parents had to intervene or typed overrides. If the parent was uncertain how to reduce difficulty or disliked the overhead of constant corrections, sessions ended prematurely. This interaction explains the “0% engagement” for Child G, who never truly participated after initial speech attempts failed, and partial success (Child H) who quit after minimal progress.

6.4.2 Older Children + Minimal Overrides = High Completion sessions

On the other hand, older children overcame small STT errors with self-correction, rarely needing parental assistance. This independence not only improved success rates but also minimized adult overhead. Parents frequently reported these sessions as “easy,” “fun,” or “short,” praising the cost-effectiveness of a purely software-based robot (compared to expensive hardware).

6.4.3 Parent-Enabled Adaptations

Many parents overcame initial struggles if they promptly changed item ranges (e.g., from 1–10 to 1–5), while researcher provided typed overrides for known mispronunciations. Child F’s improvement from 20% to 40% exemplifies how repeated demos and direct parent involvement can salvage a younger child’s session. Still, these interventions are time- and effort-intensive. Some parents embraced the AI suggestions, but the lack of direct user-interface links prevented immediate adoption.

6.5 Summary of Key Findings

1. **4-year-olds** consistently achieved 100% success, demonstrating that older preschoolers handle short simulation-based tasks well, with minimal speech confusion or parental input.

2. **3-year-olds** encountered frequent recognition issues, short attention spans, and higher meltdown risk unless parents vigorously adjusted tasks or overrode misinterpretations. One child refused to make a speech altogether, ending the session.

3. **Parents** rated setup complexity around 3.0, speech recognition difficulty ~3.4–3.5, and AI clarity ~2.8, suggesting moderate complexity and repeated frustrations. They welcomed the low hardware cost but called for child-specific speech solutions, wizard-based interfaces, and "click to apply" AI prompts.

4. **System Logs** confirm that younger children's recognized text accuracy, was only ~60–70% initially, leading to repeated negative feedback loops or forced manual corrections.

5. **Adaptation** worked best when parents quickly responded to ChatGPT's suggestions or recognized mispronunciations. Child F's partial success improvement underscores that repeated demos and simpler tasks can somewhat mitigate initial struggles.

Overall, these pilot studies demonstrate that simulation based SAVRs can offer personalized and engaging learning experiences comparable to those with physical robots. However, persistent challenges, including speech recognition errors, configuration complexity, and the need for more nuanced, age-specific adaptation, highlight critical areas for ongoing improvement in AI-driven early childhood education platforms [26], [39].

7. DISCUSSION

This section synthesizes core findings from both the child engagement and parent usability perspectives, focusing on how simulation-based Socially Assistive Virtual Robots (SAVRs), powered by ChatGPT, operate in home contexts with preschool-aged children. We address the interplay of child speech recognition issues, developmental readiness, parent facilitation, and potential improvements for achieving accessible, AI-driven early childhood tutoring. By integrating these aspects, we highlight key insights for applying advanced robotics research in everyday family settings.

7.1 Revisiting the Feasibility of Simulation-Based SAVRs for Preschoolers

7.1.1 Confirming Strengths for Older Preschoolers

One of the clearest outcomes is that 4-year-olds generally excel with short counting or letter recognition tasks delivered through an AI-powered (ChatGPT-driven) simulation environment. Each older participant (Children A–E) achieved near-100% success, rarely encountering speech-driven frustrations or requiring parental overrides. Their sessions typically lasted between 2–3 minutes, aligning with previous evidence that older preschoolers (around 4–5 years old) can handle quick “micro” tasks with minimal adult input [6], [21]. In interviews, parents reported that children quickly understood the system’s prompts, showed enthusiasm to respond, and asked for additional rounds of interaction, indicative of a self-driven curiosity also noted in prior work on child-robot “peer” relationships [7], [12]. Some parents specifically described these sessions as “fun,” with children spontaneously wanting to do more or “try again,” suggesting that a well-animated virtual robot can replicate much of the social presence and novelty that physical robots have traditionally offered [11]. These finding echoes broader research in child-robot interaction: older preschoolers exhibit greater autonomy, better speech clarity, and a willingness to persist, especially when the agent provides lively, animated feedback [13]. Moreover, the minimal number of misinterpretations meant parents had fewer interruptions for speech overrides, freeing them to observe rather than constantly step in.

It is valuable to compare this success with prior research on physical SARs in classroom contexts, where 4–5-year-olds often adapt quickly to short interactive lessons [11]. In those cases, the limiting factors included hardware cost, staff training, or scheduling constraints. With simulation-based approaches, these obstacles largely disappear, leaving only the question of whether children remain sufficiently motivated in the absence of a physically embodied robot. The results here confirm that, for older preschoolers, the "lack of tangibility" does not significantly degrade motivation, at least over short, 2–3-minute interactions. Some children even found the tasks "too easy," implying that difficulty progression or more advanced modules (for instance, early spelling or basic arithmetic) could further sustain interest [17].

In our small-sample pilot, older preschoolers completed sessions with high rates of success and minimal need for adult intervention. Their performance appeared to benefit from clear initial articulation and greater task familiarity, reducing the need for constant parent overrides. These preliminary results suggest that, as an independent or semi-autonomous learning tool, simulation-based SAVRs may be viable for 4-year-olds as suggested by the present pilot results.

7.1.2 The Additional challenges for Younger Preschoolers

For the 3-year-old group (F, G, H), however, repeated speech confusion and short attention spans significantly undercut the experience. While Child F demonstrated partial improvement (20% to 40%), Children G and H either refused speech entirely or left the session early [21], [22]. Parents often highlighted how these children became quickly discouraged when correct answers were flagged as wrong or were not recognized at all. This resonates with broader child development literature showing that at age 3, articulation can still be incomplete, while frustration tolerance is lower, particularly when faced with technology that "does not understand" them [9], [20].

Additional feedback from parents reveals that some younger children struggled to remain engaged beyond the first or second misinterpretation. In real-world contexts, repeated speech errors demand continuous adult facilitation, parents must override speech outputs or simplify tasks at the slightest sign of meltdown [19], [48]. Our findings illustrate that if parents do not swiftly correct or adapt tasks (for instance, shifting from 1–10 objects to 1–5), younger children stop responding altogether. The friction is especially high for shy participants (like Child G), who will not attempt to re-speaking if the first or second attempt is misunderstood. Thus, while older preschoolers can succeed with minimal friction, younger ones face a far less stable scenario, typically needing more advanced modifications (e.g., domain-limited recognition, extremely short sessions, or multimodal/touch-based input) to mitigate meltdown risk [23], [49].

Some parents also noted that 3-year-olds exhibited more reliance on tangible elements, such as physical toys or pictures, to sustain interest. This insight suggests that a purely screen-based SAR might not suffice unless supplemented by physical objects (countable toys, letter blocks) or integrated into the environment in ways that allow tactile reinforcement. Researchers such as in [25], have similarly proposed hybrid strategies that combine real materials with digital or robotic guidance, thereby catering to children who still expect physical interaction as part of their learning routine.

7.2 Persistent Speech Recognition Barriers

7.2.1 Consequences of Mismatch for 3-Year-Olds

A major cross-cutting obstacle is the mismatch between mainstream STT (trained largely on adult corpora) and preschool articulation patterns. Our results confirm that 3-year-olds typically see recognized text accuracy well below that of older peers, forcing parents to intervene with multiple manual overrides. Many parents described this as “frustrating,” a finding reinforced by system logs showing 4–6 overrides in a single 2–3-minute session for a child with incomplete speech. This mismatch leads directly to negative child feedback (“That’s not correct!”) even when the child is correct, undermining trust and prompting repeated meltdown scenarios if the parent is not on hand to correct it [21], [52].

Moreover, from a developmental psychology standpoint, repeated negative feedback can create a cycle of discouragement, where the child internalizes the notion that they are “not good at counting” or “not speaking correctly.” This emotional impact is as significant as any technical shortcoming: a 3-year-old may not realize that the system is the one misinterpreting them rather than vice versa [77]. As a result, some children refuse to speak or leave the session after just a few failed attempts, diminishing the potential benefits of the SAR interaction.

7.2.2 Potential for Domain-Limited or Phonetic Approaches

One recognized solution is to heavily constrain recognized vocabulary to match the child's task domain [23], [20]. For instance, if counting only from 1 to 5, the STT can be biased heavily toward those five words, reducing confusion between "four" and "for." Similarly, a phonetic dictionary mapping known child mispronunciation ("fwee" → "three") can drastically cut error rates [18], [48]. Both methods maintain better session flow, especially for 3-year-olds, who cannot easily re-pronounce words or wait out repeated system errors.

Despite these strategies, domain-limiting can require significant programming overhead: the system must be aware of exactly which words to expect. Phonetic expansions also depend on carefully assembled child-speech corpora, which might be expensive or difficult to assemble [16]. Still, the potential gains are considerable: if misrecognition rates approach 80–90% accuracy for younger preschoolers, the environment becomes much less frustrating, requiring fewer manual overrides. As the data show, older preschoolers already hover near that threshold, which reduces the need for adult intervention. By narrowing or customizing the speech domain for 3-year-olds, developers can reduce frustration and expand the range of families willing to adopt simulation-based SAVRs.

7.3 Parent Usability, Task Configuration, and AI Prompt Clarity

7.3.1 Setup Difficulty and the Desire for Wizard-Based Interfaces

Many parents rated the system "3.0" for difficulty, describing it as "not overly complicated but far from intuitive", typical of research prototypes that lack polished UI design [3], [17]. In open-ended surveys, they mentioned the absence of a guided wizard or template-based approach as a major pain point. Instead, they had to type numeric, or letter ranges manually, discovering through trial-and-error that 1–10 was too advanced for a child who can only count to 3. Such guesswork consumes valuable time, especially if parents must simultaneously monitor a frustrated child.

A "one-click" interface, like "Choose skill: beginner counting (1–5), advanced counting (1–10)", could drastically reduce confusion [37], [50]. Moreover, having the system suggests an optimal difficulty based on observed performance would preempt meltdown by automatically lowering or raising item counts. This feature is seen in some advanced EdTech platforms, but it remains less common in child-robot interactions, where standardization across multiple tasks is still evolving.

Furthermore, if AI/ChatGPT suggestions appear only as text, for example, "Try simpler tasks", parents may still not know how or where to implement that advice [36]. The system's potential for rapid adaptation is limited by this user experience gap, highlighting that advanced AI logic must be paired with direct, integrated UI actions. Summarily, parents need simpler interfaces that reduce guesswork and provide immediate, context-aware instructions.

7.3.2 ChatGPT's Parent-Facing Suggestions in Practice

Parents recognized the value of AI prompts, particularly those highlighting repeated errors or encouraging a difficulty downgrade. Nevertheless, they criticized these text-based suggestions as too vague or "buried" in on-screen logs. Some parents effectively used them, improving outcomes for children like Child F (20% → 40%), but others ignored or overlooked them, uncertain how to proceed. This scenario highlights a broader issue in human-centered AI: generative models can produce helpful text, but unless it is cleanly integrated into the interface (e.g., clickable "apply" buttons), many users will not adopt the recommended changes [16], [20].

For instance, if the environment displayed a straightforward pop-up: "Your child has failed 3 times above 5. Click here to reduce objects to 1–3," parents could seamlessly apply that adaptation. Several participants specifically mentioned that such direct prompts would reduce confusion and avoid retyping or scanning multiple screens. This design principle aligns with prior EdTech frameworks emphasizing user-friendly "teacher or parent dashboards," where the system's AI suggestions can be instantly implemented [21], [23].

7.3.3 Overriding Speech Misfires and Time Pressures

Repeated manual overrides, especially for younger children, often become logistically burdensome [18], [7]. In some sessions, parents needed to type frequent corrections to address speech misrecognitions, which could halt the session flow. This not only frustrates the parent but also disrupts child engagement, as each override introduces a break in the conversational rhythm. Busy parents or those with multiple children are likely to abandon the system if it demands ongoing textual edits [6], [37].

Minimizing override needs requires more accurate child speech recognition or a quick, one-click correction method (e.g., a pop-up showing recognized text that the parent can correct with a single tap). This approach preserves the child's sense of autonomy and maintains the momentum of each lesson. However, when parents must repeatedly interpret suggestions or manually intervene, their engagement can decrease, as the system demands greater focus and effort. Ultimately, unless overrides are easy and rare, the advanced features of an AI-driven SAR may go unused because families find it too high-maintenance in real-world contexts.

7.4 Child-Parent Interaction in Task Outcomes

7.4.1 The Interdependence of Child Performance on Parental Action

Observation logs and video recordings confirm that younger children's success or meltdown hinges closely on how actively the parent intervenes. If a parent promptly adjusts item counts (1–10 → 1–5) or overrides "fwee" → "for", "thee" → "three" the child may remain engaged and improve. When suggestions are ignored, or parents feel uncertain, meltdowns frequently follow (Child G, Child H), ending the session prematurely [16], [21], [22].

This underscores that child performance in emergent tasks is co-determined by parent facilitation in simulation-based settings. The parent essentially acts as a "live troubleshooter," bridging the gap between system constraints (speech mismatch, advanced tasks) and the child's developmental level [10], [17]). Without this interaction, younger children often give up, highlighting the significance of creating user-friendly parent controls that minimize guesswork.

7.4.2 Reducing Frustration for Younger Learners

Our analysis reveals that younger children benefit most from tasks that are brief and straightforward, with designs that can handle incomplete speech. If the environment insists on near-perfect adult-like speech, meltdown risk escalates. One way to alleviate frustration is to incorporate child-friendly re-prompts or confirmations, for example, “Did you say three or tree?”, giving children or parents a chance to correct the system before it logs a failure [19], [48]. Another is to design tasks in very small increments (counting only 1–3) with frequent celebratory animations or “positive loops,” so the child experiences success more often, building confidence.

Additionally, some parents suggested multimodal or touch-based backups. If speech is misinterpreted multiple times, the system could display clickable numbers or letters, letting the child confirm the correct answer [49], [37]. This approach also prevents meltdown by providing an alternate route for success when articulation fails.

7.5 Wider Implications for Cost-Effective Home Adoption

7.5.1 Replacing Physical Robots

The cost advantage of a purely software-based SAVR is substantial. Parents in the study consistently indicated that if speech recognition issues are addressed, they would strongly prefer an affordable or free digital platform over a high-priced physical robot [3], [14]. This perspective aligns with a trend in EdTech toward cloud-based or app-based solutions, which are simpler to update and deploy [17], [23]. Real robots often remain out of reach for typical households, confining robot-assisted learning to specialized labs or short-term pilots.

7.5.2 Potential for Continuous Upgrades and AI Enhancements

Because the approach is purely software-based, developers can patch new functionalities, such as additional languages, advanced tasks (like spelling or early arithmetic), or improved child speech recognition [48], [16]. As child-specific corpora grow, the system can be retrained to better handle 3-year-olds, bridging the performance gap between older and younger preschoolers. This iterative development model extends the novelty factor, allowing families to keep using the system as it evolves [50], [37].

7.5.3 Ethical and Privacy Dimensions

As with any AI-driven, voice-based system, privacy is paramount. Some parents voiced concerns about storing audio data or exposing child speech to cloud services [51], [52]. Ensuring robust encryption, local data processing, or transparent usage policies can mitigate these worries. Another ethical aspect is content filtering: large language models can inadvertently generate unexpected statements unless tightly constrained [20]. Considering the age group (3–4 years old), developers need to thoroughly test the system's responses to ensure they are suitable for young children. They should also use prompts that are child-friendly and select datasets that specifically reflect the language and speech patterns typical of preschoolers.

Summative Reflections on the Discussion

In summary, this expanded Discussion spotlights the disparities in speech readiness and attention span between older and younger preschoolers, revealing how a ChatGPT-powered simulation can work well for 4-year-olds but create repeated friction for many 3-year-olds. Parents see cost-savings and flexibility as major strengths but stress the need for improved child speech recognition, wizard-style configuration, and integrated "click to apply" AI suggestions. If these design elements are refined, a purely software-driven SAVR model could reach a wider audience, delivering short, interactive micro-lessons in counting, letters recognition, or more advanced topics, an opportunity previously limited to expensive physical robots. Ultimately, bridging robust child speech solutions, well-structured parent workflows, and advanced AI adaptivity can shift simulation-based SAVRs from a promising prototype into a commonplace early learning tool in ordinary households.

8. RECOMMENDATIONS

Building on these integrated findings, we propose the following enhancements, derived from child data, parent data, and prior robotics/EdTech research:

8.1 Child-Specific Automatic Speech Recognition (ASR) and Domain Constraints

1. **Phonetic Dictionaries**: Incorporate known child pronunciations ("fwee," "fo-wuh," etc.) so the engine interprets them as "three," "four," etc. [18], [19], [48].

2. **Restricted Vocabulary**: If tasks revolve around 1–5, the STT should heavily bias these words, ignoring adult synonyms like "for," "fore," "core," etc. This domain-limited approach significantly cuts error rates [9], [17], [23].

3. **Adaptive Confirmation**: If uncertain, ask "Did you say three or tree?" letting the parent/child quickly confirm. Minimizing mistaken "incorrect" responses can preserve motivation [6], [9], [78].

8.2 Multimodal or Touch-Based Backup

- **On-Screen Selections**: If repeated misinterpretations occur, show numeric or letter buttons to let the child confirm. This ensures progress even for younger or shy children [16], [49].

- **Speech Bubble**: Display recognized text, let parent/child do a quick "Yes, that's correct" or "No, correct it." This shortens typed overrides significantly [6], [22].

8.3 Wizard-Based Setup and Pre-Built Templates

1. **Beginner / Intermediate / Advanced**: Provide easy picks for “Beginner Counting (1–5),” “Intermediate (1–10),” “Advanced (1–15),” or letter sets (A–C, A–F, etc.) [7], [17].

2. **Linear Wizard**: Assessing parents through child’s approximate skill, recommended tasks, and relevant speech threshold. This approach eliminates guesswork.

3. **Interactive Tutorials**: Pop-ups explaining how to override speech, how to interpret AI/ChatGPT’s suggestions, or how to quickly drop from 1–10 to 1–5 if repeated errors arise [21], [23].

8.4 Context-Aware AI Prompts

- **Clickable** Suggestions: Rather than “Try fewer items,” say, “Child has made errors 3 times above 5. Click here to reduce items to 1–4.” This direct link fosters immediate adaptation [16], [77].

- **Configurable Frequency**: Let parents set how often they want AI suggestions or at which threshold (e.g., 2 repeated errors). This tailors the system to novices vs. advanced users.

8.5 Short Sessions and Age-Appropriate Rewards

- **1–2 Minute** Caps for Younger Preschoolers: Provide a cheerful “You did it!” screen after a few correct answers. This prevents meltdown or disinterest [9], [13].

- **Immediate** Animated Celebrations: Confetti, star icons, or cartoon applause each time the child answers correctly, especially for 3-year-olds. Quick, repeated feedback can sustain attention [6], [16].

8.6 Longer-Term Tracking and Hybrid Approaches

- **Multi-Session** Gains: If a child consistently performs well at 1–5, automatically expand to 1–7 next session. Parents see progress charts, reinforcing continued usage [7], [17], [22].

- **Hybrid** Physical-Virtual: Families might occasionally interact with a real NAO at a library or community event, harnessing the novelty, but rely on the simulation for daily cost-free practice [6], [12], [15]. This approach merges best-of-both-worlds without incurring hardware expenses for each household.

Based on these findings, several enhancements are proposed for SAVR simulation platforms. User-friendly configuration interfaces, including guided tutorials, intuitive templates, and drag-and-drop elements, can help non-expert users overcome setup challenges. Implementing AI-driven adaptive feedback as real-time, context-aware prompts can support parents and enhance learning outcomes. Upgrading speech recognition systems to include child-trained models and manual override options will reduce frustration and errors. Finally, integrating cloud and edge computing can help minimize latency and improve system responsiveness, especially for young children with limited attention spans [39] [26].

9. CONCLUSION

9.1 Overall Findings

This integrated paper merges data from eight children (ages 3–4) and ten parents to illustrate how AI-driven simulation-based SAVRs, specifically a ChatGPT-powered virtual SAVR environment, function in real home contexts. The child outcomes confirm that:

- **4-year-olds** adapt well, hitting near 100% success with minimal parental or system overrides, suggesting the approach effectively replicates many benefits of physical SARs for older preschoolers.
- **3-year-olds** face speech articulation challenges, leading to partial improvements (Child F from 20% to 40%) or outright disengagement (Child G at 0% success, refusing to speak) if misinterpretations aren't quickly addressed.

From the parent viewpoint, the concept garners strong support (80% prefer software-based solutions over costly hardware), but repeated references to speech misfires (70% citing moderate/high frustration) and moderate complexity (~3.0 rating) highlight the need for child-specific ASR and wizard-based or template-based configuration. AI suggestions, though valuable in concept, must be more direct and clickable to be broadly useful.

Collectively, the findings of this thesis support a unified framework for Socially Assistive Virtual Robots that encompasses both virtual and physical embodiments. In this pilot study, the simulation-based SAVR successfully engaged preschoolers, with older children achieving educational outcomes similar to those reported in some studies with physical robots. These preliminary results suggest that social interactivity and adaptivity may play a more significant role in assistive effectiveness than tangible form alone [1], [3]. However, larger comparative studies are needed to confirm this pattern. By showing that an AI-powered virtual robot can deliver personalized tutoring and real-time encouragement at home, this work broadens the practical definition of SAR and confirms that "robotic" social assistance need not be confined to expensive hardware. At the same time, the nuances observed (such as the extra support needed to maintain a 3-year-old's attention in the absence of physical presence) shed light on how embodiment influences user experience, reinforcing that our unified approach must account for these differences [2], [4]. In summary, the thesis validates that simulation-based SAVR and physically embodied SARs share a common goal and can be treated as complementary approaches on a continuum of socially assistive technology. This integrated perspective not only underpins the feasibility of cost-effective virtual robot tutors but also lays a conceptual foundation for future designs that seamlessly blend virtual and physical elements, with the aim of maximizing accessibility, scalability, and social engagement in early childhood learning [4], [79].

9.2 Limitations

1. **Sample Size**: Eight children and ten parents are modest, providing indicative patterns rather than broad generalizable results.

2. **Homogeneous** Demographics: All parents were relatively tech-savvy, so families with lower digital literacy might encounter greater difficulties.

3. **Short** Sessions: Typically lasting 2–3 minutes for older kids, 3–4 for younger, capturing immediate engagement rather than extended usage or skill retention.
4. **English-Only** Setting: Speech recognition and tasks were in English; bilingual or non-English contexts might differ significantly [18], [19], [20].

9.3 Future Directions

Potential expansions include:

1. **More Diverse** Participants: Recruiting families from varied socioeconomic, linguistic, and cultural backgrounds to assess universal usability.
2. **Longitudinal** Studies: Monitoring repeated usage over weeks or months to observe whether children significantly improve in counting or literacy, and whether the novelty effect wanes.
3. **Advanced** Child-Specific Acoustic Modeling: Partnering with speech recognition experts to create or adapt specialized corpora for 3–4-year-old voices, drastically reducing misinterpretations [18], [19], [48].
4. **Physical–Virtual** Hybrid: Investigating how periodic encounters with a real robot in community centers can supplement daily simulation usage, maximize novelty while minimizing cost.
5. **Vision-Based** HRI Extensions: Potentially adding gaze/posture detection so the system can sense distraction or disinterest, prompting more vivid re-engagement strategies [22], [51], [78].

In conclusion, integrating child speech challenges, parent usability preferences, and AI adaptivity are essential for successfully scaling simulation-based SAVRs for early childhood. By refining speech solutions, short tasks, wizard-based interfaces, and context-aware AI prompts, developers and researchers can unlock the full potential of software-driven 'robot tutors' at home, providing advanced adaptive learning to families worldwide without the cost burden of physical robotics.

10. APPENDICES

- Appendix A: IRB Approval Letter

Date: 8-1-2025

IRB #: IRB-FY2025-300
Title: Exploring the Feasibility of AI-Driven Educational Robots in Simulation Environments for Early Childhood Learning: A Pilot Study
Creation Date: 2-16-2025
End Date:
Status: Approved
Principal Investigator: Subramaniam Ganesan
Review Board: OU IRB
Sponsor:

Study History

Submission Type Initial	**Review Type** Expedited	**Decision** Approved

Key Study Contacts

Member Subramaniam Ganesan	**Role** Co-Principal Investigator	**Contact** ganesan@oakland.edu
Member Subramaniam Ganesan	**Role** Principal Investigator	**Contact** ganesan@oakland.edu
Member Ibrahim Abbas	**Role** Primary Contact	**Contact** ibrahimabbas@oakland.edu

Figure 4 IRP Approval Letter

- Appendix B: Related Published Work Supporting This Thesis

Table IV Associated Publications Based on Thesis Contributions

Title	Authors	Venue	Year
Exploring the Feasibility of AI-Driven Educational Robots in Simulation Environments for Early Childhood Learning: A Pilot Study	Ibrahim Abbas, Subra Ganesan	IEEE EIT Conference	2025
Challenges in Parent-Supervised Setup of Simulated Socially Assistive Robots for Personalized Early Childhood Learning	Ibrahim Abbas, Subra Ganesan	IEEE EIT Conference	2025
Exploring Challenges in Simulation Environments with Socially Assistive Robots for Early Learners	Ibrahim Abbas, Subra Ganesan	IEEE EIT Conference	2025
Challenges and Opportunities in AI-Driven Simulation-Based Socially Assistive Virtual Robots for Early Childhood Education	Ibrahim Abbas, Subra Ganesan	Submitted to IEEE Transactions on Learning Technologies	2025

References

[1] D. Feil-Seifer and M. J. Matarić, "Defining Socially Assistive Robotics," in *Proc. 9th Int. Conf. on Rehabilitation Robotics (ICORR)*, 2005.

[2] A. S. a. J. Z. S. Thellman, " Physical vs. Virtual Agent Embodiment and Effects on Social Interaction," in *In Intelligent Virtual Agents: 16th International Conference, IVA*, Los Angeles, CA, USA, 2016.

[3] I. Abbas and S. Ganesan, "Exploring the Feasibility of AI-Driven Educational Robots in Simulation Environments for Early Childhood Learning: A Pilot Study," in *IEEE International Conference on Electro/Information Technology (EIT)*, Valparaiso, Indiana, 2025.

[4] Neef, Caterina, and Anja Richert, "Likable or Intelligent? Comparing Social Robots and Virtual Agents for Long-Term Health Monitoring," in *In 2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*.

[5] F. Tanaka, A. Cicourel, and J. R. Movell, "Socialization between toddlers and robots at an early childhood education center," *Proc. Natl. Acad. Sci. U.S.A.,* vol. 104, no. 46, p. 17954–17958, 2007.

[6] O. Mubin, C. J. Stevens, S. Shahid, A. Al Mahmud, and J. J. Dong, "A review of the applicability of robots in education," *Technol. Educ. Learn,* vol. 1, no. 1, pp. 1-7, 2013.

[7] G. Gordon, C. Breazeal, and S. Engel, "Can children catch curiosity from a robot?," *Proc. ACM/IEEE Conf. Human-Robot Interaction,* p. 91–98, 2015.

[8] M. Murtaza, Y. Ahmed, J. A. Shamsi, F. Sherwani and M. Usman, "AI-Based Personalized E-Learning Systems: Issues, Challenges, and Solutions," *IEEE Access,* vol. 10, pp. 81323-81342, 2022.

[9] J. Kory-Westlund and C. Breazeal, "Assessing children's perceptions of social robots," *Int. J. Soc. Robot,* vol. 11, no. 4, p. 555–580, 2019.

[10] Y. Honghu, L. Ting, and G. Lan, "The key artificial intelligence technologies in early childhood education: A review," *Springer AI Systems,* pp. 1-15, 2023.

[11] Chen, Huili, Hae Won Park, and Cynthia Breazeal, "Teaching and learning with children: Impact of reciprocal peer learning with a social robot on children's learning and emotive engagement," *Computers & Education,* vol. 1, no. 150, p. 103836, 2020.

[12] T. Belpaeme, J. Kennedy, A. Ramachandran, B. Scassellati, and F. Tanaka, "Social robots for education: A review," *Science Robotics,* vol. 3, no. 21, 2018.

[13] A. L. Vollmer, R. Read, D. Trippas, and T. Belpaeme, "Children conform, adults resist: A robot group induced peer pressure on normative social conformity," *Science Robotics,* vol. 3, no. 21, p. 1–6, 2018.

[14] K. R. Pradeep, A. S. Manish, A. S. Adithiyaa, N. Sahana and S. T. Abhishek, "Personalized adaptive learning platform empowered by artificial intelligence," in *2024 International Conference on Knowledge Engineering and Communication Systems (ICKECS)*, Chikkaballapur, India, 2024.

[15] Sarkar, Abhinav, Avijit Pandey, Amit Kumar, Andrea Furtado, Kalind Karia, and Kavi Arya, "Teaching complex skills through an online robotics competition during Covid-19 pandemic," *In 2022 International Conference on Advanced Learning Technologies (ICALT),* pp. 199-200, 2022.

[16] Sulaymanova, Sevara, Zulfiya Boboyeva, and Ozodbek Soliyev, "The integration of artificial intelligence towards interactive based higher education system design," in *2024 4th International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, 2022.

[17] Essa, Saadia Gutta, Turgay Celik, and Nadia Emelia Human-Hendricks, "Personalized adaptive learning technologies based on machine learning techniques to identify learning styles: A systematic literature review," *IEEE Access,* vol. 15, no. 11, pp. 48392-48409, 2023.

[18] Kennedy, J., Lemaignan, S., Montassier, C., Lavalade, P., Irfan, B., Papadopoulos, F., Senft, E., & Belpaeme, T, "Child Speech Recognition in Human-Robot Interaction: Evaluations and Recommendations," in *ACM/IEEE International Conference on Human-Robot Interaction*, 2017.

[19] Tolksdorf, Nils F., and Katharina J. Rohlfing, "Parents' Views on Using Social Robots for Language Learning," in *29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, Naples, Italy, 2020.

[20] B. Irfan, A. Ramachandran, S. Spaulding, G. I. Parisi and H. Gunes, "Lifelong Learning and Personalization in Long-Term Human-Robot Interaction," in *17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2022.

[21] M. M. Neumann, L.-C. Koch, J. Zagami, D. Reilly, and D. L. Neumann, "Preschool children’s engagement with a social robot compared to a human instructor," *Early Childhood Research Quarterly,* vol. 64, pp. 182-192, 2023.

[22] S. Vogt, M. de Haas, C. de Jong, P. Baxter, and E. Krahmer, "Child-robot interactions for second language tutoring to preschool children," *Frontiers in Human Neuroscience,* vol. 13, p. 254, 2019.

[23] Lee, Changjae, and Jinjun Xiong., "From Keyboard to Chatbot: An AI-powered Integration Platform with Large-Language Models for Teaching Computational Thinking for Young Children," *arXiv preprint arXiv:2405.00750 (2024)..*

[24] F. Amato, M. Di Gregorio, C. Monaco, M. Sebillo, G. Tortora, and G. Vitiello, "Socially assistive robotics combined with artificial intelligence for ADHD interventions," *IEEE Consumer Communications & Networking Conference (CCNC),* p. 1–6, 2021.

[25] Crompton, Helen, Kristen Gregory, and Diane Burke, "Humanoid Robots Supporting Children's Learning in an Early Childhood Setting," *British Journal of Educational Technology,* vol. 49, no. 5, pp. 911-927, 2018.

[26] I. Abbas and S. Ganesan, "Exploring Challenges in Simulation Environments with Socially Assistive Robots for Early Learners," in *IEEE International Conference on Electro/Information Technology (EIT)*, Valparaiso, Indiana, 2025.

[27] Kim, Yanghee, and Michael Tscholl, "Young children's embodied interactions with a social robot," *Educational Technology Research and Development,* pp. 2059-2081, 2021.

[28] C. Breazeal, Designing sociable robots, MIT press, 2004.

[29] Tazhigaliyeva, Nazgul, Yerassyl Diyas, Dmitriy Brakk, Yernar Aimambetov, and Anara Sandygulova, "Learning with or from the robot: exploring robot roles in educational context with children," *In International conference on social robotics, Cham: Springer International Publishing,* pp. 650-659, 2016.

[30] Kennedy, James, Paul Baxter, and Tony Belpaeme, "The robot who tried too hard: Social behavior of a robot tutor can negatively affect child learning," in *In Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction*, 2015.

[31] Lampropoulos, Georgios, "Social robots in education: Current trends and future perspectives," *Information,* vol. 16, no. 1, p. 29, 2025.

[32] Nass, Clifford, and Youngme Moon, "Machines and mindlessness: Social responses to computers," vol. 56, no. 1, pp. 81-103, 2000.

[33] A. E. P. J. P. C. C. a. M. A. Vrins, "Are you paying attention? the effect of embodied interaction with an adaptive robot tutor on user engagement and learning performance," *In International conference on social robotics, Cham: Springer Nature Switzerland,* pp. 135-145, 2022.

[34] E. V. G. A. M. S. S. a. N. K. Rubegni, "Child–display interaction: Lessons learned on touchless avatar-based large display interfaces," *Personal and Ubiquitous Computing,* vol. 26, no. 3, pp. 667-680, 2022.

[35] W. F. W. R. E. M. a. T. L. Li, "Animated pedagogical agents enhance learning outcomes and brain activity during learning," *Journal of Computer Assisted Learning,* vol. 38, no. 3, pp. 621-637, 2022.

[36] Osorio-Saez, Eliana Maria, Nurullah Eryilmaz, and Andres Sandoval-Hernandez, "Parents' acceptance of educational technology: Lessons from around the world," *Frontiers in Psychology,* vol. 12, p. 719430, 2021.

[37] Okur, E., Sahay, S., Alba, R. F., & Nachman, L, "End-to-End Evaluation of a Spoken Dialogue System for Learning Basic Mathematics," *arXiv preprint arXiv:2211.03455*, 2022.

[38] Cagiltay, B., & Mutlu, B. (2024), "Toward Family-Robot Interactions: A Family-Centered Framework in HRI," *IEEE Transactions on Human-Machine Systems.*

[39] I. Abbas and S. Ganesan, "Challenges in Parent-Supervised Setup of Simulated Socially Assistive Robots for Personalized Early Childhood Learning," in *IEEE International Conference on Electro/Information Technology (EIT)*, Valparaiso, Indiana, 2025.

[40] Lin, Guan-Yu, Ciao-Chun Jhang, and Yi-Shun Wang, "Factors affecting parental intention to use AI-based social robots for children's ESL learning," *Education and Information Technologies,* vol. 29, no. 5, pp. 6059-6086, 2024.

[41] Thunberg, Sofia, and Tom Ziemke, "Pandemic effects on social companion robot use in care homes," in *IEEE international conference on robot & human interactive communication (RO-MAN)*, 2021.

[42] Đan, Nguyễn Ngọc, Nguyễn Thị Minh Thư, Hoàng Phương Vy, Lê Tuấn Phát, and Nguyễn Hoàng Bảo Châu, "Exploring Parental Acceptance of Digital Math Games on Mobile Devices for Primary School Children," *Tạp chí Khoa học Trường Đại học Sư phạm TP Hồ Chí Minh,* vol. 21, no. 6, p. 971, 2024.

[43] Lin, Chaolan, Selma Šabanović, Lynn Dombrowski, Andrew D. Miller, Erin Brady, and Karl F. MacDorman, "Parental Acceptance of Children's Storytelling Robots:

A Projection of the Uncanny Valley of AI," *Frontiers in Robotics and AI,* vol. 8, p. 579993, 2021.

[44] Ho, Hui-Ru, Edward M. Hubbard, and Bilge Mutlu, "It's Not a Replacement: Enabling Parent-Robot Collaboration to Support In-Home Learning Experiences of Young Children," in *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 2024.

[45] Perella-Holfeld, Francisco, Samar Sallam, Julia Petrie, Randy Gomez, Pourang Irani, and Yumiko Sakamoto, "Parent and educator concerns on the pedagogical use of ai-equipped social robots," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies,* vol. 8, no. 3, pp. 1-34, 2024.

[46] Chen, Huili, Anastasia K. Ostrowski, Soo Jung Jang, Cynthia Breazeal, and Hae Won Park, "Designing long-term parent-child-robot triadic interaction at home through lived technology experiences and interviews," in *IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 2022.

[47] Vaiopoulou, Julie, Stamatios Papadakis, Eirini Sifaki, Dimitrios Stamovlasis, and Michail Kalogiannakis, "Parents' perceptions of educational apps use for kindergarten children: Development and validation of a new instrument (PEAU-p) and exploration of parents' profiles," *Behavioral Sciences,* vol. 11, no. 6, p. 82, 2021.

[48] Beneteau, Erin, Olivia K. Richards, Mingrui Zhang, Julie A. Kientz, Jason Yip, and Alexis Hiniker, "Communication breakdowns between families and Alexa," in *In Proceedings of the 2019 CHI conference on human factors in computing systems*, 2019.

[49] Sumak, Bostjan, L. Diego, D. Olga, D. C. Sergio, M. Luisa, L. Secundino, S. Irfan, C. Tuncer, I. Darja, and P. Maja, "AI-Based Education Tools for Enabling Inclusive Education: Challenges and Benefits," in *47th MIPRO ICT and Electronics Convention (MIPRO)*, 2024.

[50] Coninx, Alexandre, Paul Baxter, Elettra Oleari, Sara Bellini, Bert Bierman, O. Henkemans, Lola Cañamero et al, "Towards long-term social child-robot interaction: using multi-activity switching to engage young users," *Journal of Human-Robot Interaction,* vol. 5, no. 1, pp. 32-67, 2016.

[51] Obrenovic, B., Gu, X., Wang, G. et al., "Generative AI and human–robot interaction: implications and future agenda for business, society and ethics," *AI & Soc,* 2024.

[52] A. V. Boteanu and R. T. Vaughn, "Exploring ChatGPT Capabilities and Limitations: A Survey," *2022 IEEE International Conference on AI Applications,* pp. 45-56, 2022.

[53] Steinmetz, Victoria, Andres Rosero, and Elizabeth Phillips., "Socially assistive robots: Considerations for negative effects on empathy and emotion recognition in human interaction." Proceedings of the Human Factors and Ergonomics Society Annual Meeting," *Sage CA: Los Angeles, CA: SAGE Publications,* vol. 67, no. 1, 2023.

[54] Tolksdorf, Nils F., Scarlet Siebert, Isabel Zorn, Ilona Horwath, and Katharina J. Rohlfing, "Ethical considerations of applying robots in kindergarten settings: Towards an approach from a macroperspective," *International Journal of Social Robotics,* vol. 13.2, pp. 129-140, 2021.

[55] Lee, Boh Young, "Technology/Media Use in Early Childhood Education: Publication Trends in the US from 2013 to 2022," *Early Childhood Education Journal ,* pp. 1-43, 2024.

[56] Schwarzer, Clarissa, Nico Grafe, Andreas Hiemisch, Wieland Kiess, and Tanja Poulain, "Associations of media use and early childhood development: cross-sectional findings from the LIFE Child study," *Pediatric research,* vol. 91, no. 1, pp. 247-253, 2022.

[57] Ray, Munni, and Kana Ram Jat, "Effect of electronic media on children," *Indian pediatrics,* vol. 47, no. 7, pp. 561-568, 2010.

[58] Meyer, Marisa, Jennifer M. Zosh, Caroline McLaren, Michael Robb, Harlan McCaffery, Roberta Michnick Golinkoff, Kathy Hirsh-Pasek, and Jenny Radesky, "How educational are "educational" apps for young children? App store content analysis using the Four Pillars of Learning framework," *Journal of children and media,* vol. 15, no. 4, pp. 526-548, 2021.

[59] M. L. Courage, "Screen media and the youngest viewers: Implications for attention and learning," *In Cognitive development in digital contexts,* vol. Academic Press, pp. 3-28, 2017.

[60] Sharkey, Noel, and Amanda Sharkey, "The crying shame of robot nannies: an ethical appraisal," *Machine ethics and robot ethics. Routledge,* pp. 155-184, 2020.

[61] Gordon G, Spaulding S, Westlund JK, Lee JJ, Plummer L, Martinez M, Das M, Breazeal C, "Affective personalization of a social robot tutor for children's second

language skills.," *InProceedings of the AAAI conference on artificial intelligence,* vol. 30, no. 1, 2016.

[62] Westlund JK, Dickens L, Jeong S, Harris P, DeSteno D, Breazeal C, "A comparison of children learning new words from robots, tablets, & people," in *InProceedings of the 1st international conference on social robots in therapy and education*, 2015.

[63] S. Serholt, "Serholt S. Interactions with an empathic robot tutor in education: students' perceptions three years later," *In Artificial Intelligence and Inclusive Education: Speculative Futures and Emerging Practices,* pp. 77-99, 2019.

[64] Decker-Woodrow LE, Mason CA, Lee JE, Chan JY, Sales A, Liu A, Tu S, "The impacts of three educational technologies on algebraic understanding in the context of COVID-19," in *AERA open*, 2023.

[65] Chan, Jenny Yun-Chen, et al, "How and for Whom Does DragonBox12+ Support Math Learning? The Relation Between In-Game Progress and Math Performance.," in *Proceedings of the 17th International Conference of the Learning Sciences-ICLS*, 2023.

[66] Sailer, Michael, and Lisa Homner, "The gamification of learning: A meta-analysis," *Educational psychology review,* vol. 32, no. 1, pp. 77-112, 2020.

[67] Sanders, Jackie, Robyn Munford, and Joseph Boden, "Improving educational outcomes for at-risk students," *British Educational Research Journal,* vol. 44, no. 5, pp. 763-80, 2018.

[68] Yu, Zhonggen, Mingle Gao, and Lifei Wang, "The effect of educational games on learning outcomes, student motivation, engagement and satisfaction," *Journal of Educational Computing Research,* vol. 59, no. 3, pp. 522-546, 2021.

[69] Vagnetti, Roberto, Alessandro Di Nuovo, Monica Mazza, and Marco Valenti, "Social robots: a promising tool to support people with autism. a systematic review of recent research and critical analysis from the clinical perspective.," *Review Journal of Autism and Developmental Disorders,* pp. 1-25, 2024.

[70] Chung, Eva Yin-han, Kenneth Kuen-fung Sin, and Daniel Hung-kay Chow, "Effectiveness of robotic intervention on improving social development and participation of children with autism spectrum disorder–a randomised controlled trial," *Journal of Autism and Developmental Disorders,* vol. 55, no. 2, pp. 449-456, 2025.

[71] Hamari, Juho & Shernoff, David & Coller, Brianno & Asbell-Clarke, Jodi & Edwards, Teon., "Challenging games help students learn: An empirical study on

engagement, flow and immersion in game-based learning," *Computers in Human Behavior,* vol. 54, pp. 170-179, 2016.

[72] Richard M. Ryan, Edward L. Deci, "Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions," *Contemporary Educational Psychology,* vol. 25, no. 1, pp. 54-67, 2000.

[73] A. I. Károly, P. Galambos, J. Kuti and I. J. Rudas, "Deep Learning in Robotics: Survey on Model Structures and Training Strategies," *IEEE Transactions on Systems, Man, and Cybernetics: Systems,* vol. 51, no. 1, pp. 266-279, 2021.

[74] M. H. Al Omoush, S. E. Salih, S. Kishore and T. Mehigan, "Interactive Multimodal Learning: Towards Using Pedagogical Agents for Inclusive Education," in *IEEE International Humanitarian Technology Conference (IHTC)*, 2023.

[75] Mejeh, M., Sarbach, L. & Hascher, T, " Effects of adaptive feedback through a digital tool – a mixed-methods study on the course of self-regulated learning," *Educ Inf Technol,* vol. 29, pp. 1-43, 2024.

[76] Kim, Yanghee, Jaejin Hwang, Seongmi Lim, Moon-Heum Cho, and Sungchul Lee, "Child–robot interaction: designing robot mediation to facilitate friendship behaviors," *Interactive Learning Environments,* vol. 32, no. 8, pp. 4169-4182, 2024.

[77] J. M. Kory-Westlund and C. Breazeal, "A long-term study of young children's rapport, social emulation, and language learning with a peer-like robot playmate," *Frontiers in Robotics and AI,* vol. 6, p. 81, 2019.

[78] A. D. Rossi, A. Zamfirescu-Pereira, and J. Jones, "Human-centered AI Technologies in Human-Robot Interaction for Education," *ACM/IEEE International Conference on Human-Robot Interaction (HRI),* 2023.

[79] Reeves, Byron, and Clifford Nass, "The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places," *Cambridge, UK 10,* vol. 10, pp. 19-36, 1996.