

"AI just keeps guessing": Using ARC Puzzles to Help Children Identify Reasoning Errors in Generative AI

AAYUSHI DANGOL, University of Washington, USA

RUNHUA ZHAO*, University of Washington, USA

ROBERT WOLFE[†], University of Washington, USA

TRUSHAA RAMANAN, University of Washington, USA

JULIE A. KIENTZ, University of Washington, USA

JASON YIP, University of Washington, USA

The rapid integration of generative Artificial Intelligence (genAI) into everyday life raises important and pressing questions about the competencies that are required to critically engage with these increasingly powerful technologies. Unlike visual errors that appear in genAI-generated images, textual mistakes are often considerably harder to detect and typically require specific domain knowledge to identify. Furthermore, AI's authoritative tone and highly structured responses can create a powerful illusion of correctness, leading to overtrust, especially among children who may lack the experience to question seemingly authoritative sources. To address this critical challenge, we developed AI Puzzlers, an interactive system based on the Abstraction and Reasoning Corpus (ARC), specifically designed to help children identify and systematically analyze errors in genAI outputs. Drawing on Mayer and Moreno's Cognitive Theory of Multimedia Learning [8], AI Puzzlers uses both visual and verbal elements to reduce cognitive overload and support error detection through dual-channel processing. Based on two participatory design sessions conducted with 21 children (ages 6-11), our findings provide both valuable design insights and an empirical understanding of how children identify errors in genAI reasoning, develop effective strategies for navigating these errors, and critically evaluate AI outputs.

CCS Concepts: • Human-centered computing → Empirical studies in HCI; Interactive systems and tools.

Additional Key Words and Phrases: AI Literacy, Participatory design, Generative AI

ACM Reference Format:

Aayushi Dangol, Runhua Zhao, Robert Wolfe, Trushaa Ramanan, Julie A. Kientz, and Jason Yip. 2025. "AI just keeps guessing": Using ARC Puzzles to Help Children Identify Reasoning Errors in Generative AI. In *Proceedings of Interaction Design and Children (IDC '25)*. ACM, New York, NY, USA, 12 pages. <https://doi.org/10.1145/3713043.3728836>

^{*}denotes equal contribution

[†]denotes equal contribution

Authors' Contact Information: Aayushi Dangol, University of Washington, Seattle, WA, USA, adango@uw.edu; Runhua Zhao, University of Washington, Seattle, WA, USA, runhz@uw.edu; Robert Wolfe, University of Washington, Seattle, WA, USA, rwolfe3@uw.edu; Trushaa Ramanan, University of Washington, Seattle, WA, USA, trushaar@uw.edu; Julie A. Kientz, University of Washington, Seattle, WA, USA, jkientz@uw.edu; Jason Yip, University of Washington, Seattle, WA, USA, jcyp@uw.edu.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.

Manuscript submitted to ACM

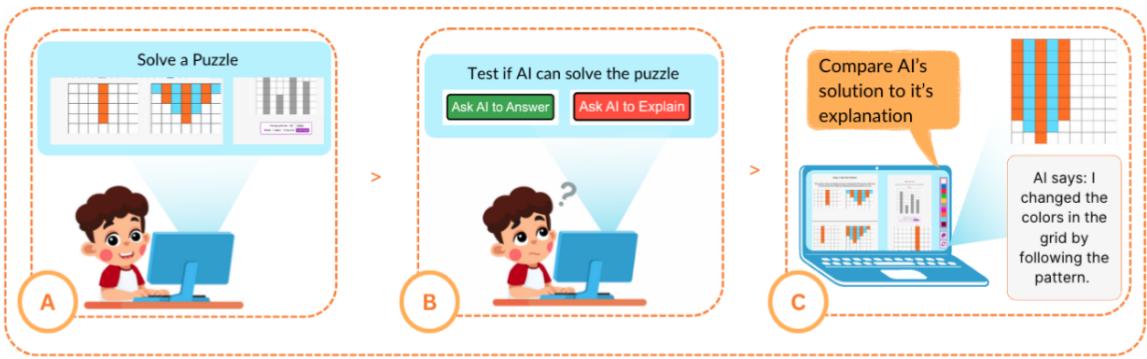
Manuscript submitted to ACM

53 1 Introduction

54 The rapid integration of generative artificial intelligence (genAI) into educational environments presents both opportunities
 55 and challenges for teaching and learning. Recent studies show that a quarter of U.S. teens now use ChatGPT for
 56 schoolwork, with adoption rates doubling in just one year [3]. A pressing concern is students' tendency to uncritically
 57 accept "AI hallucinations"—plausible but factually incorrect information—as fact. Recent reports highlight instances
 58 where students blindly trusted AI-generated claims, revealing gaps in their ability to critically evaluate AI content [5].
 59

60 This underscores the need for AI literacy education, specifically the competency to critically assess AI outputs and
 61 understand when genAI excels (pattern matching, text generation) versus where it falters (multi-step reasoning, novel
 62 problem-solving). However, detecting errors in text-based outputs is inherently difficult. While visual glitches are
 63 immediately perceptible, textual errors often require domain knowledge to identify. This "illusion of correctness" is
 64 exacerbated by AI's authoritative tone, which can mislead even adults [10].
 65

66 To address this need, we introduce *AI Puzzlers*, a game-based system utilizing the Abstraction and Reasoning Corpus
 67 (ARC) [2]. Leveraging Mayer and Moreno's Cognitive Theory of Multimedia Learning (CTML) [8], the system presents
 68 information through visual and verbal channels to reduce cognitive load and support understanding.
 69



85 Fig. 1. Overview of AI Puzzlers: (A) children solve independently, (B) test genAI, and (C) evaluate AI reasoning.

87 Our study systematically explores how children recognize genAI limitations when the errors are made visually
 88 apparent through the puzzle interface. We found that the inherently visual nature of the puzzles allowed even young
 89 children to quickly spot inconsistencies, sparking meaningful discussions about "how AI thinks" and revealing that AI
 90 often appears to "guess" rather than reason through problems systematically.
 91



101 Fig. 2. Comparison of the correct vs AI-generated solutions. The visual nature of AI Puzzlers makes AI errors easy to spot.
 102

105 **2 Related Work**

106 **2.1 Children's Interactions with Generative AI**

108 While children increasingly adopt genAI technologies, their ability to critically evaluate AI-generated content remains
109 inconsistent. Research shows that children can spot errors in familiar domains but tend to overtrust AI outputs in
110 unfamiliar subjects [10]. The polished presentation of AI-generated text creates "aesthetic legitimacy" that can fool
111 users of all ages [1]. Children's limited working memory resources mean they struggle to verify information without
112 structured guidance, particularly when AI presents information coherently but misleadingly [8]. However, recent
113 research highlights that children possess unique cognitive capabilities—such as causal reasoning, innovation, and
114 learning from minimal examples—that current language models lack [14].
115

118 **2.2 Multimedia Learning and AI Literacy**

120 Mayer and Moreno's Cognitive Theory of Multimedia Learning (CTML) [8] posits that humans process information
121 via visual and verbal channels. Distributing information across these channels reduces cognitive load and improves
122 retention and comprehension. This is particularly relevant to AI literacy education, as combining visual outputs with
123 textual explanations helps children process complex AI behaviors without cognitive overload. Previous initiatives using
124 interactive platforms have shown that scaffolded environments improve understanding of AI systems [7].
125

128 **2.3 Learning through Games**

129 Educational games offer low-pressure environments for trial-and-error learning, where mistakes become learning
130 opportunities. Effective games use scaffolding techniques to guide learner attention and structure problem-solving [4].
131 Recent research demonstrates that game-based approaches can effectively foster students' AI literacy development
132 through enhanced knowledge acquisition and engagement [9]. Given that ARC puzzles are visually intuitive for humans
133 yet difficult for AI models [2], they provide an ideal context for children to "outsmart" the system, fostering critical
134 evaluation skills and building confidence.
135

138 **3 AI Puzzlers: System Design**

140 AI Puzzlers is a web-based platform designed to help children (ages 6+) critique AI reasoning using the ARC dataset [2].
141 The system focuses on three design principles: visual comparison to make errors apparent, reducing cognitive load via
142 dual-channel processing [8], and scaffolding through three interaction modes.
143

145 **3.1 System Modes**

147 **Manual Mode** allows children to solve puzzles independently using tools including flood fill, color selection, and grid
148 resizing. This establishes a baseline of human competence before introducing the AI component.
149

150 **AI Mode** integrates GPT-4o capabilities. Children can ask AI to solve puzzles (system converts grids to text, processes,
151 and converts back to visual grids) or ask for explanations providing step-by-step reasoning traces.
152

153 **Assist Mode** empowers children to guide the AI by providing text hints, effectively "debugging" the AI's reasoning.
154 They can adjust parameters like the number of example puzzles or model version, enabling experimentation with
155 factors influencing AI performance.
156

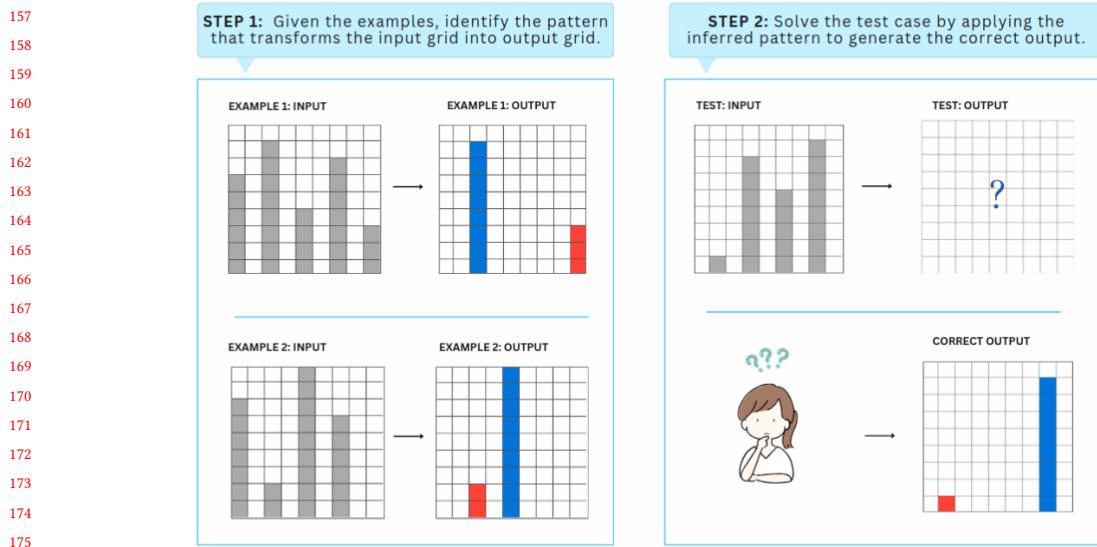


Fig. 3. An example of an ARC puzzle with instructions for solving it.

4 Methods

We utilized Cooperative Inquiry [4], a participatory design method where children and adults work as equitable partners, fostering a dialogic environment suitable for examining how children conceptualize emerging technologies.

4.1 Participants

The study involved 21 children (ages 6-11) recruited from an intergenerational co-design group known as KidsTeam UW. Participants represented diverse ethnic backgrounds and demonstrated varying levels of prior AI experience, ranging from daily users of voice assistants to children with no previous AI exposure.

4.2 Co-Design Sessions

We conducted two 1.5-hour sessions with KidsTeam UW as part of a summer camp. Each session began with a 15-minute informal discussion before hands-on engagement with AI Puzzlers. Children worked in groups of four to five with two adult facilitators, balancing peer-driven exploration with adult guidance.

4.2.1 Session 1. Session 1 began with a warm-up question as an icebreaker. Children were introduced to Manual and AI Modes, divided into five groups, and solved puzzles in Manual Mode to familiarize themselves with the format. Before introducing AI Mode, facilitators asked: (1) "Do you think genAI can solve these puzzles quickly or slowly? Why?" and (2) "Do you think genAI can solve these puzzles without help?" Children then interacted with AI Mode, requesting AI assistance and explanations. The session concluded with a group discussion.

4.2.2 Session 2. Session 2 began with a warm-up question about helping others, connecting to human guidance for genAI. Unlike Session 1, children actively assisted genAI by providing hints in Assist Mode. Facilitators guided exploration with reflection questions. After 50 minutes, groups presented their experiences, reflecting on strategies, challenges, and genAI's limitations.



Fig. 4. Annotated screenshot of Manual Mode in AI Puzzlers.

4.3 Data Collection and Analysis

We collected 927 minutes of video data via Zoom, supplemented with field notes and photographs of physical artifacts. The first, second, and fourth authors created analytical memos using a dual-review process: primary reviewers created narrative summaries at five-minute intervals, and secondary reviewers verified observations and added insights.

We then conducted inductive coding. The first two authors independently coded memos, met to reconcile codes, and developed a codebook with three categories: (1) Perception of AI, (2) Evaluation of AI Performance, and (3) Interaction with the System. After applying codes to the full dataset with a second-pass verification, we organized codes into themes through two refinement rounds.

5 Findings

We present findings from children's interactions with AI Puzzlers, focusing on how they engaged with the system and developed understanding of AI's capabilities and limitations. For consistency with children's language, we use "AI" throughout this section.

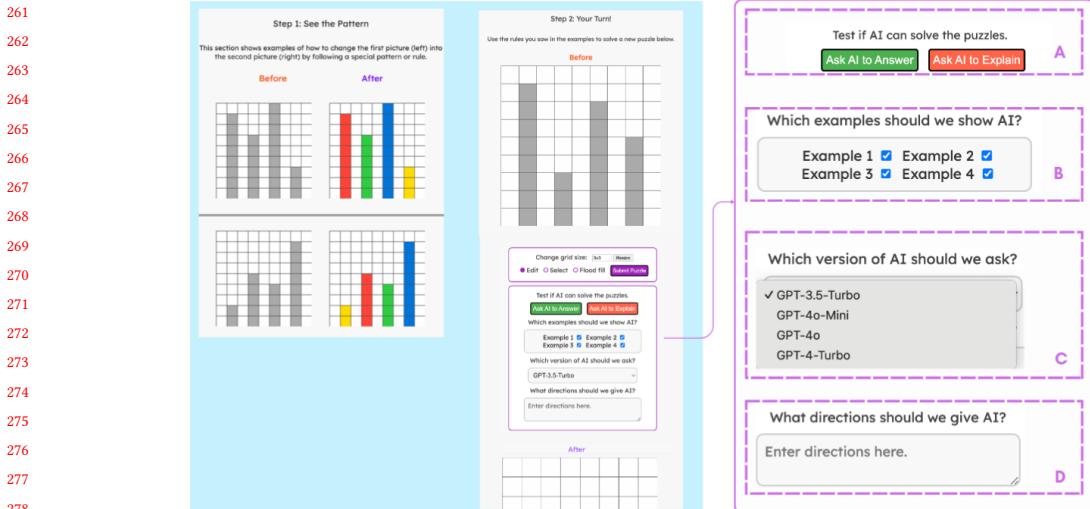


Fig. 5. Screenshot of Assist Mode in AI Puzzlers, highlighting features for testing and guiding the AI.

Table 1. Reported Child Participant Details

Name	Gender	Ethnicity	Age	AI Type	Usage Freq
Kai	Male	Asian/White	8	Voice Assistant	Daily
Lani	Female	Asian/Black	9	None	Never
Juno	Male	Asian	7	Video Game AIs	Daily
Elias	Male	Asian/Black	9	Video Game AIs	Daily
Noa	Female	Asian/White	11	Video Game AIs	Multiple/week
Ren	Male	Hispanic	10	Chatbot	Multiple/week
Matt	Male	Asian/White	9	N/A	N/A
Ivy	Female	White	9	Video Game AIs	Few times/week
Zayn	Male	Asian/Black	9	None	Rarely
Finn	Male	White	10	N/A	N/A
Leila	Female	Asian	8	Voice Assistant	Daily
Mara	Female	Asian/Black	6	Video Game AIs	Few times/week
Emi	Female	Asian/White	8	None	Rarely
Hana	Female	Asian	8	None	Multiple/week
Theo	Male	Asian/White	7	Video Game AIs	Multiple/week
Lucia	Female	Hispanic	6	Video Game AI	Weekly
Rina	Female	Asian	7	Video Game AIs	Monthly
Owen	Male	White	8	Video Game AI	Daily
Nico	Male	Asian/White	6	None	Daily
Selah	Female	Asian/Black	6	None	Never
Elise	Female	Asian/Black	9	None	Never



Fig. 6. Children engaging with AI Puzzlers alongside adult facilitators.

5.1 Children's Interest and Exploration

5.1.1 Surprise, Excitement, and AI's Unexpected Errors. Children initially expected AI to solve the puzzles easily, given their own success and trust in AI's capabilities. However, they quickly noticed AI struggled, reacting with surprise. When one group solved an "alternating between red and grey" pattern and asked AI to solve it, the AI returned an incorrect solution. The group burst into laughter at the failure, with one child commenting, "That is very very wrong." The visual nature allowed them to quickly recognize mistakes.

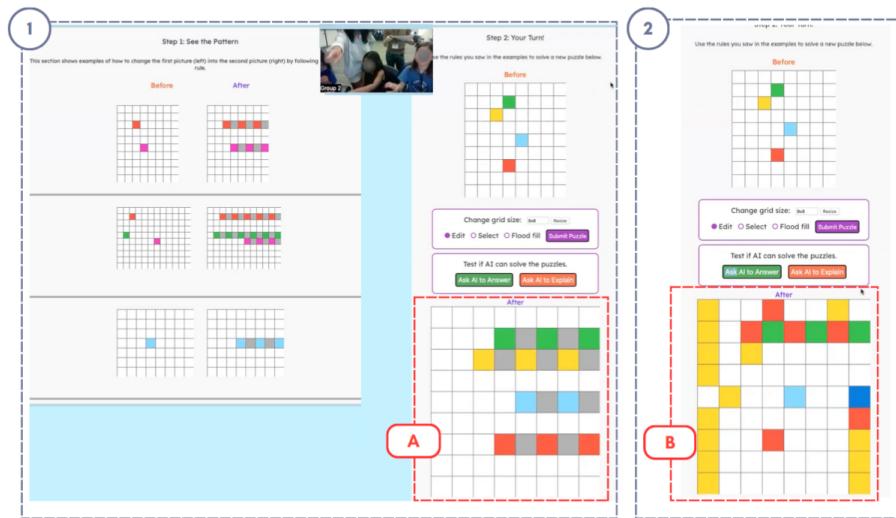
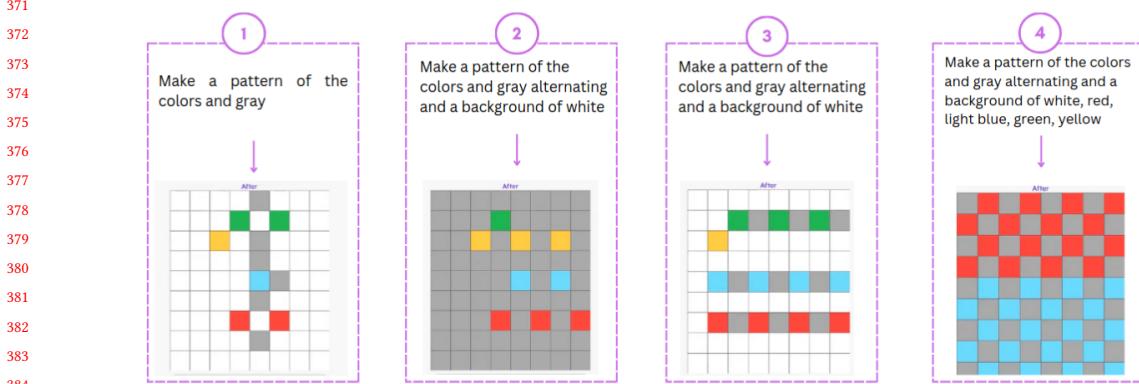


Fig. 7. Children collaboratively solving a puzzle (left) vs. the AI's incorrect attempt (right).

The contrast between AI's solutions and children's correct answers continued to evoke surprise. After seeing AI fail nine times, children still reacted strongly to failures. Moreover, engagement wasn't solely tied to AI's mistakes—children viewed puzzles as problem-solving opportunities. Even as puzzles became difficult, they maintained interest, often describing complex puzzles as "fun." This encouraged them to "outsmart" the AI, turning the activity into an engaging competition. One child remarked with pride, "We can go farther than the AI."

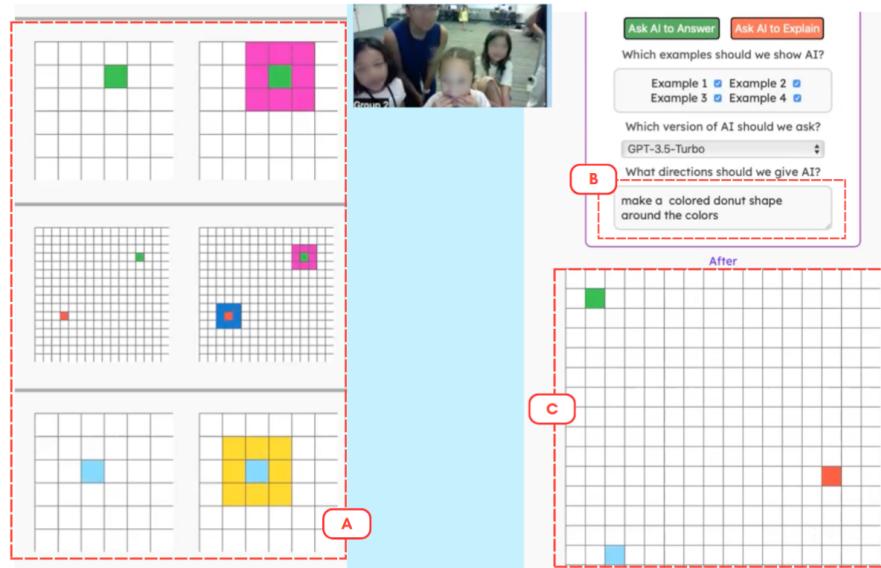
365 5.2 Iterative Debugging

366 In Session 2, children engaged in iterative debugging, systematically refining prompts. One group evolved from "Make
 367 a pattern of gray..." to "Make a pattern of the colors and gray alternating and a background of white, red, light blue,
 368 green, yellow," demonstrating growing understanding of AI communication.
 369



370 Fig. 8. Children iteratively refined their instructions to guide the AI, showing increasing specificity.

371 Children attempted culturally relevant metaphors like "donut shape" to explain visual concepts, but AI often failed to
 372 interpret these correctly, revealing its limitations in understanding human-centered descriptions.



373 Fig. 9. Children using the "donut shape" metaphor in Assist Mode to guide the AI.

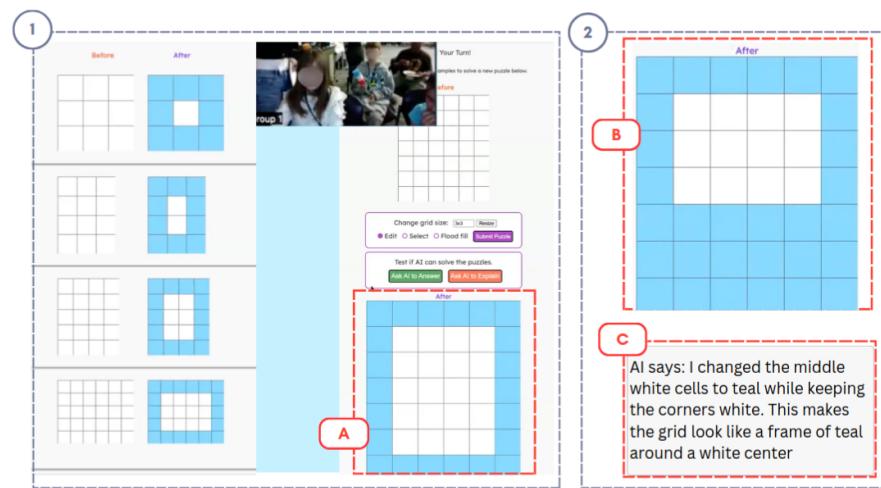
417 5.3 Understanding AI's Limitations Through Observing Inconsistencies

418 5.3.1 *Children Identify Inconsistencies in AI's Reasoning.* Children cross-examined AI's visual solutions with its expla-
 419 nitions, identifying discrepancies between reasoning and outcomes. The visual nature enabled critical evaluation, as
 420 children noted that AI's explanations often sounded "scientific" but didn't match visual results.
 421



441 Fig. 10. Comparison showing the AI's explanation (C) does not match its visual output (B).

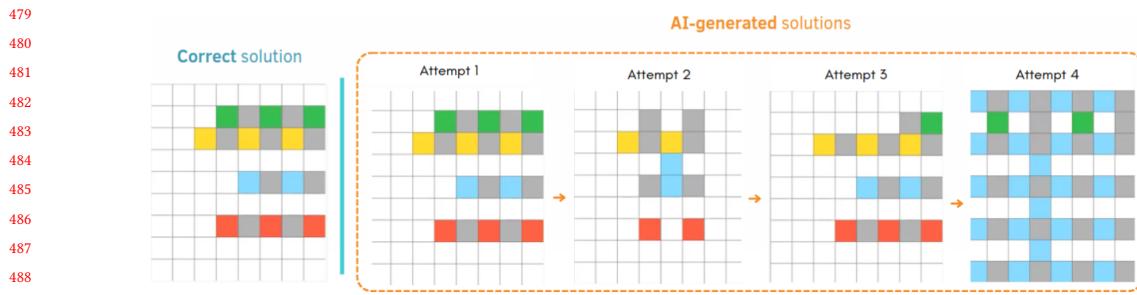
442 One group noted AI claimed to "keep corners white" but the visual grid showed otherwise. A participant remarked,
 443 "It's like someone who is not listening," drawing a parallel to inattentive human behavior.



466 Fig. 11. Another example of a discrepancy between the AI's text explanation (C) and the visual grid (B).

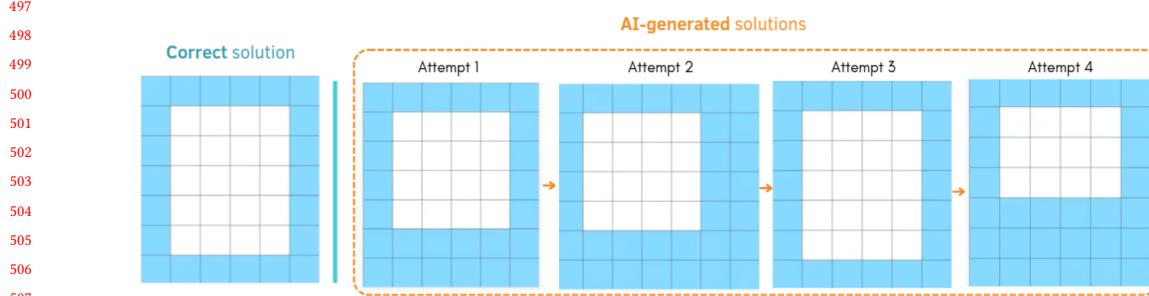
469 Children also critiqued vagueness in AI explanations. One child noted that while explanations sounded "scientific,"
 470 they didn't explain patterns meaningfully, revealing AI was mimicking explanation form without substantive content.
 471 Overall, children actively scrutinized AI's reasoning rather than passively accepting responses.
 472

473 *5.3.2 AI's "Scientific Brain" vs. Human Problem Solving.* Children recognized AI approached problem-solving differently
 474 from humans. While puzzles were "easy" for children, they were "super hard" for AI. Children distinguished between
 475 human cognition (using creativity, experiences, intuition) and AI's reliance on given data. One child described AI as
 476 having only the "internet's mind," limited to provided information without broader experiential knowledge.
 477



478 Fig. 12. Multiple AI attempts showing a lack of learning or improvement over time.
 479
 480

481 The randomness of AI's repeated incorrect answers led children to conclude AI lacked true reasoning capabilities.
 482 Observing AI changed answers randomly with each attempt, one child concluded, "AI just keeps guessing," capturing
 483 AI's trial-and-error approach without systematic learning.
 484



485 Fig. 13. The AI's random guessing behavior across four attempts.
 486
 487

488 Children's reflections demonstrated growing awareness of fundamental differences between human and AI problem-
 489 solving, recognizing AI's rigid parameters and reliance on trial-and-error versus human reasoning and creativity.
 490

511 6 Discussion

512 6.1 Positioning Children as Active Inquirers

513 AI Puzzlers positioned children as active inquirers rather than passive consumers of AI content. The puzzle format
 514 motivated children to systematically analyze AI failures, aligning with established AI literacy frameworks that emphasize
 515 Manuscript submitted to ACM

critical evaluation of AI capabilities and limitations [11]. Three mechanisms fostered this: First, the visual nature made AI errors immediately apparent, sparking critical evaluation. Second, Assist Mode helped children develop schemas of AI capabilities as they refined hints from vague to precise, viewing AI as a system requiring guidance. Third, the game-like nature encouraged competitive problem-solving, reinforcing children's own strengths while revealing AI's reasoning limitations.

6.2 Implications for Design

Future systems should design for interpretability without cognitive overload. This includes visual reasoning traces (flowcharts, decision trees), side-by-side comparisons, and "validity markers" (confidence levels, uncertainty indicators) to guide attention to problematic outputs. While AI can justify outputs, text-heavy explanations may overwhelm young users. Visual traces and opportunities for experimentation—tinkering with parameters and observing outcomes—help children engage in deep reflection on AI systems.

7 Limitations and Future Work

Our co-design sessions engaged 21 children from a single region with prior participatory design experience, which facilitated rich discussions but limits statistical generalization [13]. Future work should examine how children in diverse settings (schools, libraries) and cultural contexts engage with AI Puzzlers to illuminate how different dynamics shape AI literacy development.

While children detected genAI errors within AI Puzzlers' structured environment, we did not assess whether this learning transfers to open-ended genAI interactions. Follow-up studies will investigate transferability to real-world contexts and expand the system to include voice-based interactions representing the range of AI systems children encounter daily.

We selected ARC puzzles for their engaging nature and use as AI reasoning benchmarks [2], but acknowledge their color-based differentiation may pose accessibility challenges. Future work could explore alternative puzzle formats to broaden accessibility. Additionally, as AI capabilities evolve—with newer models like OpenAI o3 showing improved ARC performance [6]—AI Puzzlers could incorporate multiple model versions to help children recognize and interpret genAI's changing capabilities over time.

8 Conclusion

This study presented AI Puzzlers, an educational tool using visual puzzles to help children develop critical evaluation skills for generative AI. Through participatory co-design sessions, we found that visualizing AI errors helps children move from blind trust to informed evaluation. Combining visual and verbal modalities creates an effective environment for recognizing AI's limitations, developing debugging strategies, and understanding distinctions between human and AI reasoning. While generative AI presents significant promises for human learning, it also raises challenges around critical thinking and evaluation [12]. Future work should explore skill transfer to other domains and investigate long-term impacts on children's attitudes toward AI.

References

- [1] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM, New York, NY, USA, 610–623. doi:10.1145/3628516.3655763
- [2] François Chollet. 2019. On the measure of intelligence. *arXiv preprint arXiv:1911.01547* (2019). doi:10.48550/arXiv.1911.01547

- 573 [3] Common Sense Media. 2025. New Report Shows Students Are Embracing Artificial Intelligence Despite Lack of Parent Awareness. <https://www.commonsensemedia.org/press-releases/new-report-shows-students-are-embracing-artificial-intelligence-despite-lack-of-parent-awareness-and>.
- 574 [4] Allison Druin. 1999. Cooperative inquiry: developing new technologies for children with children. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 592–599. doi:10.1145/302979.303166
- 575 [5] Jessica Grose. 2024. What Teachers Told Me About A.I. in School. *The New York Times* (14 August 2024). Opinion.
- 576 [6] Nicola Jones. 2025. How should we test AI for human-level intelligence? OpenAI's o3 electrifies quest. *Nature* 637, 8047 (2025), 774–775. doi:10.1038/d41586-025-00124-w
- 577 [7] Duri Long and Brian Magerko. 2020. What is AI Literacy? Competencies and Design Considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–16. doi:10.1145/3313831.3376727
- 578 [8] Richard E. Mayer and Roxana Moreno. 2003. Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist* 38, 1 (2003), 43–52. doi:10.1207/S15326985EP3801_6
- 579 [9] Davy Tsz Kit Ng, Chen Xinyu, Jac Ka Lok Leung, and Samuel Kai Wah Chu. 2024. Fostering students' AI literacy development through educational games: AI knowledge, affective and cognitive engagement. *Journal of Computer Assisted Learning* 40, 5 (2024), 2049–2064. doi:10.1111/jcal.12984
- 580 [10] Jaemarie Solyst, Amy Ogan, and Jessica Hammer. 2023. Intergenerational Games to Learn About AI and Ethics. In *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 2*. ACM, New York, NY, USA, 1273. doi:10.1145/3545947.3573256
- 581 [11] David Touretzky, Christina Gardner-McCune, Fred Martin, and Deborah Seehorn. 2019. Envisioning AI for K-12: What Should Every Child Know about AI? *Proceedings of the AAAI Conference on Artificial Intelligence* 33, 01 (2019), 9795–9799. doi:10.1609/aaai.v33i01.33019795
- 582 [12] Lixiang Yan, Samuel Greiff, Ziwen Teuber, and Dragan Gašević. 2024. Promises and challenges of generative artificial intelligence for human learning. *Nature Human Behaviour* 8, 10 (2024), 1839–1850. doi:10.1038/s41562-024-01967-8
- 583 [13] Robert K. Yin. 2013. Validity and generalization in future case study evaluations. *Evaluation* 19, 3 (2013), 321–332. doi:10.1177/1356389013497081
- 584 [14] Eunice Yiu, Eliza Kosoy, and Alison Gopnik. 2024. Transmission Versus Truth, Imitation Versus Innovation: What Children Can Do That Large Language and Language-and-Vision Models Cannot (Yet). *Perspectives on Psychological Science* 19, 5 (2024), 874–883. doi:10.1177/17456916231201401
- 585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624