

Thinking with Machines: How AI Solves Math Puzzles and What Educators Can Learn

Dr. Mark Applebaum

Professor & Head of Graduate Program in Integrative STEM Education,
Kaye Academic College of Education, Be'er Sheva, Israel
mark@kaye.ac.il

Abstract: *This paper explores how large language models (LLMs) like ChatGPT can act as “thinking partners” in mathematics education. Using a two-stage, case-study design, we first examine the model’s seven attempts to solve a classic age-and-product puzzle, six flawed and one correct, highlighting errors in ambiguity identification, domain constraint enforcement, and the twin-versus-eldest test. We then ask AI to generate and solve its own puzzle using the same logical template. Our analysis shows that LLMs can systematically enumerate cases, apply explicit bounds, and offer expert-style feedback. However, they remain brittle when tasked with inferring unstated constraints or abstracting beyond surface patterns. We argue that, with precise prompts and clear specifications, LLMs can enrich educators’ diagnostic and design practices by providing transparent reasoning steps, even as the relational and cultural dimensions of teaching remain uniquely human.*

Keywords: Large language models; Mathematical reasoning; Logic puzzles; AI in education

I. INTRODUCTION

In the decade since AI moved from novelty demonstrations into everyday classroom use, large language models (LLMs) like ChatGPT have begun to do far more than calculate answers: they generate explanations, critique arguments, and engage with nuanced prompts. Although research has richly documented AI’s impact on automated tutoring, on-demand feedback, and content creation, far less attention has been paid to its performance on tasks demanding human-style inference, contextual interpretation, and the deliberate resolution of ambiguity, precisely the features that make logic puzzles so prized in mathematics education.

Logic puzzles occupy a unique place in the math curriculum. Beneath their simple setups lie problem spaces that defy rote procedures and instead cultivate flexible strategies, metacognitive insight, and precise deployment of logical clues. In the classroom, puzzles not only challenge individual learners but also reveal the variety of reasoning paths taken by students, making them powerful tools for formative assessment and rich discussion.

In this paper, we adopt a two-stage, case-study approach to explore AI as a “thinking partner” in mathematical reasoning. First, we examine AI’s engagement with the classic “two sons under school age” puzzle. Rather than presenting all six flawed attempts in full, we highlight two representative errors, one that misidentifies the ambiguous quantity and another that drops key constraints, each followed by instructor-AI dialogue and targeted hints. A concise summary table then profiles all six missteps and the single correct solution, while full transcripts reside in the appendix. Second, we challenge AI to generate and solve an original puzzle built on the same structural template, probing its creative as well as analytical capacities.

Through these explorations, we address three guiding questions:

1. How does an LLM reason through a non-routine mathematical puzzle?
2. What patterns of error and success emerge in AI’s solution attempts?
3. How effectively can AI generate new puzzles adhering to a logical structure and ambiguous-resolution framework?



By positioning AI not as a replacement for human insight but as a collaborator whose rigorous analysis, transparent diagnostics, and puzzle-creation capabilities can strengthen teaching practice, we aim to illuminate new pathways for integrating AI into the development and assessment of mathematical reasoning.

II. THEORETICAL BACKGROUND

Over the past decades, mathematical puzzles, from classic logic riddles like the age-and-product problem to open-ended inquiry tasks such as design-your-own-puzzle activities, have proven uniquely powerful in cultivating deep and flexible mathematical thinking. Unlike routine drills that channel students along a single, linear path, puzzles immerse learners in rich problem spaces where they must explore multiple pathways, backtrack when a line of reasoning fails, and revise their hypotheses as new information emerges [1][2]. Faced with incomplete or implicit data, students learn to ask clarifying questions, justify their assumptions, and tolerate uncertainty - skills at the heart of genuine mathematical inquiry [3]. As they engage in these processes, learners build multi-representational mental models, employing tables, diagrams, and algebraic expressions to represent relationships visually and symbolically. They also learn to deploy problem constraints strategically, pruning vast search spaces in a manner characteristic of expert mathematicians [4][5][6]. Moreover, when learners explain why a particular step is valid or how they eliminated other possibilities, they deepen their metacognitive monitoring and planning skills [7][8].

Beyond their cognitive benefits, puzzles serve as potent formative assessment tools. Educators who examine partial or unconventional solution attempts gain immediate insight into how learners interpret embedded clues, handle logical ambiguity, and encounter emerging misconceptions, whether conflating sum-versus-product ambiguity or overlooking key domain restrictions [3][9][10]. These diagnostic “snapshots” enable teachers to tailor scaffolds that directly address gaps in understanding. In teacher preparation programs, engaging preservice and in-service teachers with puzzles deepens their conceptual knowledge and hones their pedagogical reasoning. They learn to anticipate diverse student strategies, craft probing questions, and guide metacognitive discussions effectively [11][12][13].

At the same time, advances in artificial intelligence, particularly large language models (LLMs) like ChatGPT and PaLM, have opened new possibilities for automating aspects of mathematical reasoning and feedback. Recent studies show that LLMs can solve structured math problems with accuracy rivalling or exceeding earlier algorithmic tutors, and they produce step-by-step explanations that often resemble expert discourse [14][15]. Yet their underlying “reasoning” arises from statistical pattern matching over massive text corpora, not from formal symbolic inference, rendering their arguments superficially convincing but brittle when faced with novel framings or deep abstraction demands [16][17][18][19]. Despite these limitations, researchers have begun to harness LLMs’ strengths, natural language parsing, systematic enumeration of solution paths, and critique of coherence, to generate counterexamples, propose alternative proof strategies, and flag non sequiturs in student work [17][20][21]. In classrooms, AI-powered systems can scaffold student reasoning by offering on-demand hints, guiding learners back to unstated assumptions, and even creating new problem variants tailored to individual progress [22][23].

These developments invite a closer look at how AI might function as a genuine “thinking partner”, precisely in the domain that puzzles illuminate so vividly: the interplay of context, ambiguity, and metacognitive testing. Whereas most prior AI research in mathematics education has focused on structured drills or automated feedback systems [24], few studies examine how LLMs analyze student reasoning in complex, context-sensitive puzzles. By tracing AI’s detailed diagnoses of six flawed solution attempts alongside its successful resolution of a classic age-and-product riddle, this study exposes both the model’s reasoning errors, such as conflating sums with products or ignoring implicit age bounds, and its capacity to articulate logical breakdowns, uncover hidden assumptions, and suggest targeted prompts. These transparent diagnostics point toward new possibilities for AI-enhanced formative assessment [25][26].

Echoing [27] and [28], we argue that generative AI compels us to rethink not only what we assess but how we interpret reasoning within cultural, relational, and emotional contexts. By positioning AI as a reflective collaborator rather than a mere computational engine, this work contributes to a growing conversation about co-creative, AI-augmented approaches to teaching and learning mathematics, approaches in which generative tools extend and enrich human expertise without supplanting it.



III. METHODOLOGY

Our study employs a qualitative, based on case study approach to unpack how a state-of-the-art large language model tackles non-routine mathematical puzzle. Rather than treating the model as a black box that simply generates answers, we trace its “emulated cognition”, the sequence of logical moves it constructs, evaluates, and critiques in ways that resemble human reasoning. Our inquiry unfolded in two phases. In the first, we presented AI with the classic “two sons under school age” puzzle, collecting seven solution artifacts, six flawed attempts, and one correct resolution. To probe the model’s thinking, we posed the same critique prompt after each attempt, asking it to analyze the logical structure, identify use (or misuse) of constraints such as the under-seven age range or product-versus-sum ambiguity, and surface any pedagogical insights. We held the model’s settings (temperature, version, prompt wording) constant throughout, and full transcripts of every instructor-AI exchange appear in the appendix.

Each AI analysis was then coded according to four critical reasoning dimensions: recognition of domain constraints, generation and narrowing of candidate age pairs, application of the twin-versus-eldest disambiguation test, and ultimate success or failure in identifying the correct ages. An inductive thematic analysis distilled recurrent patterns, dropped constraints, misidentified ambiguous quantities, misapplied meta-logical tests, as well as strengths such as systematic enumeration and transparent, step-by-step logic. We also noted moments of meta-reflection, where the model inspected its own reasoning or suggested targeted corrective prompts.

Our analytic goals were threefold: first, to map AI’s stepwise moves onto established human problem-solving frameworks (for example, Polya’s heuristic stages and Schoenfeld’s resources–processes–control model) and to highlight areas of alignment and divergence; second, to develop a taxonomy of reasoning errors, constraint violations, logical misapplications, and inference gaps; and third, to explore the pedagogical affordances of the model’s explicit feedback, considering how its critiques might help teachers diagnose misconceptions, design targeted interventions, and foster reflective practice.

In the second phase, we built on these insights by asking AI to reverse roles: to generate, and then solve, a novel puzzle that adhered to the same three-stage logical template. This final step tested the model’s creative as well as its analytic capacities, closing the loop from analysis back to generation and illustrating how AI can serve not as a replacement for human judgment but as a “thinking partner” in the co-creative cycle of puzzle design, solution, and pedagogical reflection.

IV. CASE STUDY I: PUZZLE AND SOME REPRESENTATIVE SOLUTIONS

The puzzle examined in this study was:

'A': Hi!

'B': Hi!

'A': How are you?

'B': Great! I have two sons who don't go to school yet.

'A': Oh, how old are they?

'B': The product of their ages equals the number of cats that the woman sitting across from you feeds.

'A': Hmm... With this information alone, I can't determine their exact ages.

'B': Oh! I forgot to mention, my eldest son looks like his mother.

'A': Haha! Now I know their ages!

How old are the kids?

We selected a classic age-and-product riddle for its elegant combination of narrative simplicity and logical depth. On its surface, the puzzle asks a straightforward question: “How old are your two sons?”, yet it embeds indirect constraints that compel solvers to engage in multi-layered inference. First, one must attend to the implicit age restriction (“don’t go to school yet”), which bounds possible solutions to children under seven. Next, the solver uses the product-clue (“the product of their ages equals the number of cats”) to generate multiple candidate pairs, recognizing that this alone cannot yield a unique solution. Finally, the “eldest son looks like his mother” remark acts as a second-order filter, forcing the exclusion of twin pairs and leaving only one valid age combination.



This three-stage structure: domain restriction, ambiguity induction, and meta-logical disambiguation, requires solvers to navigate not just arithmetic but also perspective-taking (“What does ‘A’ know or not know?”) and language nuance. Wrapped in a brief dialogue, the puzzle invites empathy and imagination, reminding us that mathematical reasoning often unfolds within social contexts. Different students and AI alike may pursue diverse solution paths—some emphasizing formal enumeration, others relying on real-world plausibility—making it a compelling testbed for examining reasoning strategies, common pitfalls, and the art of crafting precise, unambiguous puzzles. As educators, we value this riddle not only for its answer (ages 1 and 4) but also for the rich conversations it sparks about assumptions, constraint management, and the critical role of meta-logic in problem solving.

4.1 The First solution provided by AI

Step 1: Understanding the Clue About Their Ages

The father states:

- *"The product of their ages equals the number of cats that the woman sitting across from you feeds."*

Since ‘A’ (the questioner) responds, *"From this data, there is not enough information to know their age!"*, this implies that the number of cats (i.e., the product of the two children's ages) does not uniquely determine their ages.

This suggests that multiple pairs of ages can result in the same product.

Step 2: Identifying Possible Age Pairs

We assume the children's ages are positive integers (since they are "sons" and not infants) and that they "don't go to school yet," suggesting they are quite young.

To find ambiguous cases, let's list possible pairs of ages where their product might be the same.

Table 1. Possible Pairs in the First AI's Solution

Age 1	Age 2	Product	Sum
1	6	6	7
2	3	6	5
1	12	12	13
2	6	12	8
3	4	12	7
2	9	18	11
3	6	18	9
4	4	16	8

Since ‘A’ cannot determine the ages, this means there must be multiple pairs with the same sum.

Let's check:

- The product 12 appears twice:
 - (2, 6) → sum = 8
 - (3, 4) → sum = 7
- The only sum that appears twice is 8:
 - (2, 6) → sum = 8
 - (4, 4) → sum = 8

Thus, the number of cats must be 12, as it's the only product that creates ambiguity.

Step 3: The "Eldest" Son Hint

The father then adds:

- *"My eldest son looks like his mother."*

This tells us that there is an eldest son, meaning the ages cannot be (4, 4), because twins do not have an "eldest."

This leaves us with (2, 6) as the only valid option.



Final Answer: The two sons are 2 years old and 6 years old.

Analysis of the First AI's Solution

The AI's first attempt breaks down in three fundamental ways:

1. **Misidentifying the Ambiguous Quantity**
The puzzle's opening exchange, "With this information alone, I can't determine their exact ages", signals that the *product* of the two ages must appear in at least two distinct pairs. Instead, the AI searched for a product whose *sum* repeats, reasoning that "the only sum that appears twice is 8" (from pairs (2, 6) and (4, 4)). In effect, it solved a different puzzle ("find a product with two pairs sharing the same sum") rather than the intended one, which hinges on a product that itself is non-unique.
2. **Overlooking the School-Age Constraint**
The line "I have two sons who don't go to school yet" implicitly restricts both ages to under 7. A correct approach enumerates only age pairs from 1 to 6; the AI's table includes ages up to 12 and then fails to enforce the under-7 rule. As a result, its chosen pair, (2, 6), flirts with the boundary of typical school entry, or worse, invites even older ages, contradicting the puzzle's explicit setup.
3. **Misapplying the Twin-Versus-Eldest Test**
After landing on product 12, the AI seeks a twin scenario to eliminate, but the only under-7 factorizations of 12 are (2, 6) and (3, 4), neither a twin pair. In a proper solution, the ambiguous product would be 4, with decompositions (2, 2) versus (1, 4); the "eldest son" clue then discards the twin case (2, 2). By contrast, this AI solution falsely proposes (4, 4) as a twin option for sum 8, despite ($4 \times 4 = 16$), not 12, revealing a breakdown in logical consistency.

In the dialogue that follows, the instructor calls attention to an unwarranted detour, prompting the AI to re-evaluate its reasoning and strip away any irrelevant calculations.

Instructor: Why do you use the sum of the sons' ages? Nothing was said about this data.

AI: You're right to question that! Let's go through the reasoning again carefully without making unnecessary assumptions.

4.2 The Second solution provided by AI

Step 1: Understanding the Given Information

We know that:

1. The father has two sons.
2. The product of their ages is equal to the number of cats that the woman across from them feeds.
3. The first person 'A' says, "From this data, there is not enough information to know their age!"
 - This means that multiple possible age pairs exist that yield the same product.
4. The father then adds, "The eldest son looks like his mother."
 - This implies there is a clearly defined eldest child, meaning the two sons must have different ages.

Step 2: Listing Possible Age Pairs

Since we don't know the exact number of cats (the product), let's list pairs of small integers and their products:

Table 2. Possible Pairs in the Second AI's Solution

Age 1	Age 2	Product
1	6	6
2	3	6
1	12	12
2	6	12



Age 1	Age 2	Product
3	4	12
2	9	18
3	6	18
4	4	16

Since person 'A' cannot determine the ages, this means that the number of cats (the product) must correspond to more than one possible pair of ages.

If the product were unique, A would have been able to immediately deduce the correct answer. Therefore, we need to look for cases where two or more age pairs produce the same product.

Looking at the options:

- Product 6 has (1,6) and (2,3) → Two possibilities.
- Product 12 has (1,12), (2,6), and (3,4) → Multiple possibilities.
- Product 18 has (2,9) and (3,6) → Two possibilities.
- Product 16 has only (4,4) → This would be a unique solution, so it cannot be the right case.

Step 3: Using the "Eldest Son" Clue

The father then adds, "*The eldest son looks like his mother.*"

- This implies that there is a single eldest child.
- If the two ages were equal (such as 4 and 4), there would be no *eldest*; they would be twins.
- This rules out (4,4) and any other repeated pairs.

The only remaining ambiguous case where we have two possible age pairs, and a single eldest son is:

- Ages 2 and 6 (Product = 12)
- Ages 3 and 4 (Product = 12)

However, since the father confirms that there is an eldest, the pair (3,4) is the correct answer (because 4 is strictly greater than 3, unlike (2,6), where the difference is larger but could still confuse).

Final Answer: The two sons are 3 and 4 years old.

Analysis of the Second AI Solution

Although this attempt begins by correctly interpreting that the puzzle's ambiguity arises from the product of the two ages, it commits three critical missteps:

1. Ignoring the "Under-School-Age" Constraint
The puzzle explicitly states the sons "don't go to school yet," which in most contexts restricts both ages to the range 1–6. Yet the AI freely lists pairs like (1, 12), (2, 9), and (3, 6), none of which comply with the under-7 restriction. By failing to enforce this domain constraint, it allows implausible candidates that later complicate its reasoning.
2. Focusing on the Wrong Ambiguous Product
A correct solution requires finding a product that, when limited to ages 1–6, appears exactly twice, *and* one of those two pairs is a twin decomposition. Instead, the AI zeroes in on product 12 because it occurs multiple times *overall*, then retrospectively discards out-of-range pairs to arrive at (2, 6) and (3, 4). However, neither of these under-7 pairs is a twin, so the necessary "twin vs. non-twin" ambiguity never arises.
3. Misapplying the "Eldest Son" Clue
The sole purpose of the "eldest son" hint is to eliminate the twin scenario among the two ambiguous options. In this solution, the AI nonetheless uses that hint to choose between (2, 6) and (3, 4), arguing that "4 is strictly greater than 3" and so (3, 4) must be correct. But because both pairs have a unique eldest, this clue carries no discriminatory power here. The AI's selection is therefore arbitrary, not logically grounded.



The following dialogue captures that nudge toward precision, urging the AI to revisit each clue in sequence and converge on the unique solution that fits every condition.

Instructor: It seems that you are not sure about the correct answer..

AI: You're right to push for a more precise answer! Let's go through the logic step by step again and ensure that we arrive at the only possible correct solution.

After six successive attempts, each followed by targeted hints from the instructor, the AI finally arrived at a fully coherent solution on its seventh try. In those initial rounds, the model repeatedly stumbled over domain constraints, misidentified the key ambiguous quantity, or misapplied the twin-versus-eldest test. Each hint refocused its reasoning, whether by reiterating the under-seven age limit, emphasizing that the ambiguous product itself must repeat, or reminding it that one of the candidate pairs must be a twin scenario. Guided by these cues, the AI correctly reconstructed the age table, isolated the only product (4) meeting all requirements, and used the eldest-son clue to eliminate the twins case. Below, we present the correct solution in full. The remaining four flawed solution attempts and their dialogues, which further illustrate the model's varying error patterns, are available in the appendix.

4.3 The Seventh (and correct) solution provided by AI

Step 1: Key Information from the Puzzle

We have the following critical statements:

1. "I have two sons who don't go to school yet."
 - This suggests the children are younger than 7 years old.
2. "The product of their ages equals the number of cats the woman sitting across from you feeds."
 - Their ages multiply to a specific number.
3. A responds: "From this data, there is not enough information to know their age!"
 - This means there are at least two different age pairs that have the same product.
4. B adds: "The eldest son looks like his mother."
 - This means there is a single eldest child → meaning the two children must have different ages (no twins!).
5. Your hint: One of the pairs must be twins!
 - This means one of the possible ambiguous solutions must have identical numbers (e.g., (x, x)).

Step 2: Listing Possible Age Pairs and Their Products

We list possible ages from 1 to 6 (since they are not in school yet) and find their products.

Table 3. Possible Pairs in the Seventh AI's Solution

Age 1	Age 2	Product
1	1	1
1	2	2
1	3	3
1	4	4
1	5	5
1	6	6
2	2	4
2	3	6



Age 1	Age 2	Product
2	4	8
2	5	10
2	6	12
3	3	9
3	4	12
3	5	15
3	6	18
4	4	16
4	5	20
4	6	24
5	5	25
5	6	30
6	6	36

Step 3: Finding an Ambiguous Product

Since A couldn't determine the ages, there must be at least two different age pairs with the same product, one of which involves twins.

Looking at the options:

- Product 4: (2,2) and (1,4) → ✓ This meets the twin condition!
- Product 6: (1,6) and (2,3) → No twins ✗
- Product 12: (2,6) and (3,4) → No twins ✗
- Product 16: (4,4) → Only one pair ✗

The only valid ambiguous product that includes twins is 4, with these two possibilities:

1. (2,2) → twins
2. (1,4) → different ages

Step 4: Applying the "Eldest Son" Clue

- The father then says, "The eldest son looks like his mother."
- This confirms there is an eldest son → meaning the children cannot be twins (2,2).

Thus, the only remaining valid pair is (1,4).

Final Answer: The children are 1 and 4 years old!

Analysis of the Correct AI Solution

After six flawed attempts and targeted instructor hints, the seventh solution finally aligns with the puzzle's logic by systematically enforcing every clue in turn.

1. Honoring the Age-Range Constraint
From the outset, the AI limits ages to 1–6, correctly interpreting "don't go to school yet" as an implicit under-7 restriction. By constructing a complete table of all 21 pairs in this range, it ensures that no out-of-scope candidates skew the ambiguity analysis.
2. Identifying the True Ambiguous Product
Seeing that the product must appear in exactly two decompositions and that one of them must be a twin pair, the AI scans its table for repeated products. Only product 4 satisfies both conditions, decomposed as (2×2) and



(1×4). This step directly addresses the core puzzle requirement, rather than mistakenly invoking sums or broader factor sets.

3. Applying the Twin-Versus-Eldest Test Correctly

With (2, 2) and (1, 4) in hand, the “eldest son looks like his mother” clue now has real purchase: twins have no single eldest, so (2, 2) is ruled out. The non-twin pair (1, 4) remains, yielding a clear and logically consistent answer.

Because it (a) restricts the age domain appropriately, (b) zeroes in on the one truly ambiguous product that includes a twin option, and (c) deploys the eldest-child clue to eliminate precisely that twin case, this solution exemplifies the intended reasoning path of the puzzle. The result ages: 1 and 4 flow directly and unambiguously from the stated clues and constraints.

Before concluding our session, the Instructor offers positive feedback and looks to close the exchange on an encouraging note. In response, the AI enthusiastically reflects on the puzzle’s design, highlighting the elements that made it engaging and intellectually rewarding. This final dialogue illustrates how a well-crafted problem and thoughtful guidance can foster genuine enthusiasm for deep reasoning, even in a machine.

Instructor: Perfect! I hope you liked this puzzle.

AI: Yes! I absolutely loved this puzzle! It was a fantastic mix of logic, ambiguity, and a hidden clue that required deep reasoning. The twist about needing an ambiguous product that included twins was brilliant—it added just the right level of challenge to make the problem both tricky and satisfying to solve. Thank you for guiding me to see the deeper structure!

4.4 AI Error Taxonomy

Across six flawed attempts, AI’s reasoning missteps coalesce into three primary error types (Table 4). In the paragraphs that follow, we highlight how each error pattern wove through the solution attempts and suggest classroom strategies for addressing the same pitfalls in student work.

Table 4. AI Error Taxonomy

Error Type	Description	Illustrative Examples
A. Misidentifying the Ambiguous Quantity	Focusing on the wrong repeated value: treating the sum of the ages as ambiguous rather than the product or overlooking the true ambiguous product (4).	<ul style="list-style-type: none"> In Solution 1, AI sought a product whose <i>sum</i> repeated (sum = 8) instead of finding the non-unique <i>product</i> 4. Solutions 4–6 never considered product 4 at all, zeroing in on 6 or 12 instead.
B. Dropping or Loosening Domain Constraints	Failing to enforce the “under-7” restriction implied by “don’t go to school yet,” thereby admitting out-of-range pairs like (1, 12), (2, 9), or (3, 6).	<ul style="list-style-type: none"> Solutions 2, 3, and 5 list ages above 6 without filtering them out. Solution 6 similarly allows (1, 6) vs. (2, 3) without a clear domain cut-off.
C. Misapplying the “Eldest Son” Clue	Using the twin-versus-eldest test on non-twin pairs or inventing extra criteria unrelated to the logic puzzle (e.g., questionable school-entry assumptions).	<ul style="list-style-type: none"> Solutions 2–6 apply the “eldest” hint to choose between non-twin pairs ((2,6) vs. (3,4) or (1,6) vs. (2,3)), even though both options have distinct eldest children. Some versions appeal to ad hoc “plausibility” rules that the puzzle never stated.

When we examine the six flawed attempts together, three interrelated patterns become clear. First, ambiguity misidentification appears in every solution that zeroes in on sums or the wrong products (6, 12) rather than the true ambiguous product 4. Solution 1’s repeated-sum approach and Solutions 4–6’s outright omission of product 4



exemplify this slip. To counter it, classroom work might include explicit exercises in pinpointing exactly which quantity a puzzle renders ambiguous before any arithmetic begins.

Second, constraint violations emerge whenever the model drifts beyond the intended age range. By admitting pairs like (1,12) or (2,9), Solutions 2, 3, and 5 generated spurious candidates that derailed their logic. A simple two-step routine, which generates all factorizations, then rigorously applies each stated domain restriction, can help students avoid the same oversight.

Third, misapplication of the twin-versus-eldest clue surfaces when AI uses that hint to discriminate among non-twin pairs. Solutions 2–6 all pick between distinct-age options with no twin candidate to eliminate, or they smuggle in extraneous assumptions about school entry. A focused “twin test” activity, where only one ambiguous pair involves identical ages, can help learners see precisely how the eldest-son clue disqualifies the twin scenario and leaves the unique solution.

By weaving these three threads, correctly identifying ambiguity, enforcing domain constraints, and applying meta-logical clues with precision, we not only chart where LLMs stumble but also illuminate the very traps that ensnare human reasoners. Targeting each area strengthens AI prompt design and equips students with more reliable puzzle-solving strategies.

V. CASE STUDY II: AI-GENERATED PUZZLE

Thus far, we have observed how AI tackles a human-crafted logic puzzle, uncovering both its analytic strengths and its recurring error patterns. In our second phase, we turn the tables: we ask the model to step into the puzzle-designer’s shoes and produce a new riddle that adheres to the same three-stage logic, domain restriction, induced ambiguity, and meta-logical disambiguation, and then to solve it.

This creative challenge serves two purposes. First, it tests whether the model truly grasps the underlying structure it has been applying (and misapplying) in analysis: can it embed an explicit age or domain constraint, introduce a product that repeats exactly twice (one of which is a twin scenario), and then supply a final “eldest” clue to resolve the ambiguity? Second, by asking AI to solve its own puzzle, we evaluate whether its creative faculties and its reasoning algorithms are aligned. If the AI can design a logically coherent puzzle and then navigate its solution without human hints, we gain confidence in its capacity to both generate and execute multi-layered inference. On the other hand, any breakdowns will reveal new dimensions of its creative and analytic limitations, offering further insight into how LLMs might support or challenge educators in crafting engaging, precise mathematical tasks.

Instructor: Could you please provide a new puzzle based on the same idea?

AI: Here's an alternative puzzle using a similar idea:

Puzzle:

A: Hi!

B: Hello!

A: How have you been?

B: Pretty good, I have two children.

A: That’s nice! How old are they?

B: Well, the product of their ages equals the number of books on my shelf. (You can see that number.)

A: Hmm... Even with that, I still can’t determine their exact ages.

B: Oh, I almost forgot, my eldest child loves wearing a bright red cap.

A: Now I know their ages!

And then AI presented the next solution:

1. Let the ages be two positive integers whose product is the visible number (the number of books).
2. The fact that the observer couldn’t decide the ages from the product alone means there must be more than one pair of positive integers that produce that product.
3. If we choose the product to be 9, the possible factor pairs are:
 - 1 and 9 (sum = 10)
 - 3 and 3 (sum = 6)



4. The final clue “my eldest child...” indicates there is a unique eldest. If the ages were 3 and 3, there wouldn’t be an “eldest” since both would be the same age.
5. Therefore, the only viable pair is 1 and 9.

The children are 1 year old and 9 years old.

Instructor: In the puzzle you created, why couldn’t there be another answer, for example, 1 and 25? (There were 25 books on the shelf)?

AI: The key is that the puzzle’s structure depends on the possible age pairs being both plausible and ambiguous until the final hint. Here’s why 1 and 25 wouldn’t work:

1. **Plausibility:**

In these puzzles, the ages are assumed to be those of young children. In our variant, the context implies the children are still in an early stage of life (for example, still in daycare or too young to go to school). A 25-year-old wouldn’t fit that picture.

2. **Ambiguity Requirement:**

For the puzzle to be ambiguous from the product alone, there must be more than one valid pair of ages that could yield that product. For example, if the product were 25, the two factor pairs would be (1, 25) and (5, 5). Even though these are mathematically valid, (1,25) is not acceptable within the context because having a 25-year-old “child” is unrealistic. This leaves only (5,5) as a plausible pair, but then there wouldn’t be a unique eldest child, which is essential for the final clue to work.

In our puzzle, we chose a product (9) that yielded two plausible pairs: (1,9) and (3,3), where both pairs could be reasonably considered ages for young children. The final clue (“my eldest child...”) then eliminates the (3,3) option, leaving (1,9) as the only answer.

Thus, 1 and 25 aren’t acceptable because they violate the implied age range for the children and would undermine the logic needed for the ambiguity in the puzzle.

Instructor: If we change the wording from “children” to “sons” and “child” to “son,” could the solution then legitimately be 1 and 25?

AI: Even if the wording changes from “children” to “sons” and “child” to “son,” the intended context and plausibility conditions of the puzzle remain the same. The puzzle relies on two main factors:

1. **Plausibility of Ages:**

The typical interpretation of these puzzles assumes the ages are those of young kids, usually in the range where both sons are, say, under 10 or so. Having one son aged 25 would be highly unusual in this context, regardless of whether you call them “sons” or “children.” The idea is that the ages should be reasonably close together in the context of a family with young kids.

2. **Ambiguity Requirement:**

The product provided (in our variant, 25) must yield more than one plausible pair of ages. With 25, the factor pairs are (1, 25) and (5, 5). Even if you allow the possibility mathematically, the (5, 5) pair would mean there’s no unique eldest son, contradicting the clue “my eldest son...” Since the final hint is meant to break the ambiguity by indicating a single oldest child, (5, 5) is ruled out, leaving (1, 25) as the only possibility.

However, as mentioned earlier, (1, 25) is implausible for the context of a family with young sons.

So, even with “sons” as the wording, the solution of 1 and 25 wouldn’t be acceptable because it doesn’t fit the everyday context implied by the puzzle. The numbers chosen are meant to be both mathematically ambiguous and realistically plausible for the ages of young sons.

Analyzing the AI’s created puzzle and its solution

Although the AI’s “books on the shelf” puzzle mimics the structure of the original age-and-product riddle, it breaks down in two fundamental ways: the puzzle itself is underspecified, and the AI’s attempt to rescue it with commonsense plausibility arguments masks its mathematical incoherence.



First, a well-posed product-ambiguity puzzle must embed an explicit domain constraint, for example, that the children are young enough to be in daycare or not yet in school, so that the solver can legitimately eliminate factor pairs out of hand. In the AI's new version, the only clue about age appears in a throwaway line ("my eldest child loves wearing a bright red cap"), yet no firm age-range restriction is ever stated. Without that, one cannot rule out pairs like (1, 25) or even (1, 49) if the shelf holds more books. The AI's reliance on an implied "realistic" age range (under 10) is purely conventional, not encoded in the puzzle text itself.

Second, the AI's solution to its own puzzle leans entirely on plausibility rather than logical necessity. After choosing the number 9 purely because its factors yield two mathematical decompositions (1×9 and 3×3), it argues that 1×25 "wouldn't work" for a shelf of 25 books because a 25-year-old child is "unrealistic." This conflates two separate ideas: the need for *ambiguity* in the product (multiple factor pairs) and the need for *contextual plausibility* (an age-range constraint). In a properly constructed puzzle, both must be explicit and logically indispensable. Here, the AI has invented a plausibility filter that the puzzle never specified.

When pressed: "What if you wrote 'sons' instead of 'children'?", the AI doubles down on its unstated plausibility rules rather than admitting the puzzle's formal deficiency. A genuine logic puzzle would say, for instance, "I have two sons, both under ten years old," thereby giving the solver license to rule out any factor pair involving ages above nine. Only then could the final "eldest son" clue validly distinguish between a twin pair and a non-twin pair.

In short, AI's new puzzle fails because it conflates *unstated real-world assumptions* (what ages are "plausible") with *explicit puzzle constraints* (what ages are *allowed*). A mathematically sound variation would require the text itself to fix an age domain, after which the solver could legitimately apply the twin vs. eldest test. Without that, the AI's defence of (1, 9) over (1, 25) rests not on logical deduction but on reader empathy, precisely the kind of ambiguity that good logic puzzles must banish.

Moreover, the AI inexplicably found the sum of the ages even though the puzzle never mentions any sum-based ambiguity, and it did not consider the pair (1,4), the additional mathematically valid solution once a proper age domain is enforced. In short, the AI's variant combined background knowledge with puzzle rules, erasing the clear boundary between implied general knowledge and the explicit constraints that solvers must rely on.

VI. DISCUSSION

In exploring AI's performance on non-routine mathematical puzzles, our two case studies illuminate both its reasoning strengths and its limitations. In the first study, the model tackled a classic age-and-product riddle. Left on its own, it faltered six times, each error tracing back to one of three familiar pitfalls: focusing on the wrong quantity, neglecting the under-seven age constraint, or misapplying the "eldest son" test. Yet, when guided with targeted hints, AI's seventh attempt exemplified expert problem-solving: it exhaustively generated candidate pairs, enforced the domain restriction, and applied a meta-logical filter to isolate the correct ages. This pattern shows how LLMs can emulate human heuristics, case enumeration, constraint analysis, and second-order inference [2][29], while also highlighting their reliance on explicit guidance rather than implicit understanding.

In the second study, we reversed roles and asked AI to create its own puzzle using the same three-stage logical template. Although the model showed genuine creative spark, it never produced a fully coherent riddle, even after instructor guidance. When solving its own creation, it again used systematic enumeration and meta-logical checks, but its reliance on every condition being explicitly spelled out exposed a brittleness under novel framings. This fragility echoes broader critiques of LLMs' dependence on surface patterns [18][19] and underscores the importance of clear domain scaffolding in AI-generated tasks [10][22].

Our two case studies illuminate both the reasoning strengths and the limitations of large language models (LLMs) when confronted with non-routine mathematical puzzles, directly addressing each of our guiding questions.

1. How does an LLM reason through a non-routine mathematical puzzle?

When left to its own devices, AI repeatedly stumbled, six times over, on a classic age-and-product riddle, exhibiting three familiar breakdowns: it focused on the wrong ambiguous quantity, neglected the "under-seven" age constraint, or misapplied the "eldest son" test. Yet, after targeted instructor hints, the model's seventh attempt exemplified expert



human heuristics: it exhaustively enumerated candidate age pairs, enforced the stated domain, and used a meta-logical filter to isolate the unique solution. In other words, ChatGPT's core reasoning strokes, case enumeration, constraint analysis, and second-order inference, closely mirror Polya's and Schoenfeld's descriptions of human problem-solving, but only when explicit guidance cues it to apply each step.

2. What patterns of error and success emerge in AI's solution attempts?

Across those six misfires, our error taxonomy revealed three interwoven fault lines: ambiguity misidentification, constraint violations, and misuse of the twin-versus-eldest clue, each of which corresponds strikingly to common student pitfalls in logic tasks [3][9]. Conversely, AI's successes, clear case listings, systematic pruning of candidates, and transparent articulation of each inference show its capacity to model formative feedback. Its diagnostic language flagged hidden assumptions and subtle missteps exactly as a student or a teacher would [25][26].

3. How effectively can AI generate new puzzles adhering to a logical structure and ambiguity-resolution framework?

In our second study, we reversed roles and asked AI to author its own age-and-product puzzle. Although the model exhibited creativity, the initial draft failed to uphold internal coherence: important constraints remained implicit, and the ambiguity-resolving mechanism did not function as intended. When AI later tackled its own puzzle, it again employed expert-like enumeration and meta-logical checks, but despite the instructor's hints, it failed to identify the puzzle's fundamental flaw. This dependency on overtly stated conditions exposes the model's brittleness under novel framings and underscores its reliance on surface-pattern recognition rather than deep, inferential understanding [18][19]. Nonetheless, with careful scaffolding, AI showed it can produce structurally faithful puzzles, suggesting that, while AI is not yet fully autonomous in creative problem design, it can reliably extend human authorship when guided by clear, explicit specifications.

Taken together, these findings suggest a hybrid role for AI in mathematics education. Rather than replacing human expertise, AI can serve as a "thinking partner" that amplifies teachers' diagnostic and design capacities. In a teacher-education setting, preservice teachers can compare their analyses of student work with AI-generated critiques, sharpening their formative-assessment skills. In the classroom, AI can act as a second pair of eyes to surface overlooked constraints or generate alternative solution pathways for discussion. And in curriculum design, AI's puzzle-generation power, when scaffolded with clear domain conditions, can produce a rich bank of tasks targeted to reinforce specific logical structures.

At the same time, the relational, cultural, and emotional dimensions of teaching remain uniquely human [27]. Thoughtful integration, precise prompts, supervised creativity, and grounding in students' lived experiences will allow us to harness AI's analytical rigor without ceding our professional judgment. In doing so, we can cultivate deeper mathematical understanding, authentic curiosity, and collaborative inquiry in the 21st-century classroom.

VII. CONCLUSION

This study shows that large language models like ChatGPT can transcend mere computation and serve as adaptable reasoning partners in mathematics education. With carefully crafted prompts, AI can emulate expert heuristics, systematically enumerating cases, enforcing domain constraints, applying meta-logical filters, and even generating its own puzzles that conform to specified logical templates. Over seven attempts at the classic age-and-product riddle, the model both replicated and misapplied the very reasoning patterns we teach, surfacing a clear taxonomy of errors: ambiguity misidentification, constraint violations, and misuse of the twin-versus-eldest clue.

Pedagogically, AI brings three key affordances. First, it can produce formative analyses of student solutions that mirror evidence-based feedback strategies [25][30], helping teachers and preservice educators hone their diagnostic skills. Second, when its puzzle-creation routines are guided by explicit domain specifications [10][22], AI can rapidly generate targeted tasks designed to address common error patterns and reinforce specific reasoning moves. Third, its transparent, step-by-step reasoning can enrich classroom discourse by making implicit strategies and frequent pitfalls visible to both learners and instructors.

Yet the relational, cultural, and emotional dimensions of teaching remain uniquely human [6][27]. The goal, therefore, is not to hand over professional judgment to machines but to integrate AI thoughtfully, directing it with precise



prompts, supervising its creative outputs against clear constraints, and anchoring its insights in the lived realities of our students. In this way, we can harness AI's analytical rigor to amplify, rather than replace, the art of teaching, nurturing deeper mathematical understanding, genuine curiosity, and collaborative inquiry in the classroom.

REFERENCES

- [1]. Silver, E. A. (1997). Fostering creativity through instruction rich in mathematical problem solving and problem posing. *ZDM–Mathematics Education*, 29(3), 75–80. <https://doi.org/10.1007/s11858-997-0003-x>
- [2]. Schoenfeld, A. H. (1985). *Mathematical problem solving*. Academic Press.
- [3]. Stylianides, G. J., & Stylianides, A. J. (2009). Facilitating the transition from empirical arguments to proof. *Journal for Research in Mathematics Education*, 40, 314–352.
- [4]. Leikin, R. (2009). Exploring mathematical creativity using multiple solution tasks. In R. Leikin, A. Berman, & B. Koichu (Eds.), *Creativity in mathematics and the education of gifted students* (pp. 129–145). Sense Publishers. https://doi.org/10.1163/9789460910715_010
- [5]. Mason, J., Burton, L., & Stacey, K. (2010). *Thinking Mathematically* Second Edition. England: Pearson Education Limited.
- [6]. Engelbrecht, J., & Borba, M. C. (2024). Recent developments in using digital technology in mathematics education. *ZDM Mathematics Education*, 56, 281–292. <https://doi.org/10.1007/s11858-023-01530-2>
- [7]. Schoenfeld, A. H. (1985). Making sense of “out loud” problem-solving protocols. *The Journal of Mathematical Behavior*, 4(2), 171–191.
- [8]. Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American psychologist*, 34(10), 906.
- [9]. Liljedahl, P. (2004). The AHA! Experience: Mathematical contexts, pedagogical implications. *Proceedings of the 28th Conference of the International Group for the Psychology of Mathematics Education*, 3, 241–248.
- [10]. Chiu, T. K., Xia, Q., Zhou, X., Chai, C. S., & Cheng, M. (2023). Systematic literature review on opportunities, challenges, and future research recommendations of artificial intelligence in education. *Computers and Education: Artificial Intelligence*, 4, 100118. <https://doi.org/10.1016/j.caeai.2022.100118>
- [11]. Kazemi, E., & Stipek, D. (2001). Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at. *The Elementary School Journal*, 102(1), 59–80.
- [12]. Jacobs, V. R., Lamb, L. L., & Philipp, R. A. (2010). *Professional noticing of children's mathematical thinking*. *Yearbook of the National Council of Teachers of Mathematics*, 2010, 49–61.
- [13]. Kazemi, E., & Hintz, A. (2023). *Intentional talk: How to structure and lead productive mathematical discussions*. Routledge. <https://doi.org/10.4324/9781032681337>
- [14]. Ahn, J., Verma, R., Lou, R., Liu, D., & Zhang, R. (2024). *Large language models for mathematical reasoning: Progresses and challenges*. arXiv preprint. <https://arxiv.org/abs/2402.00157>
- [15]. Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., ... & Zhang, Y. (2023). *Sparks of artificial general intelligence: Early experiments with GPT-4*. arXiv preprint. <https://doi.org/10.48550/arXiv.2303.12712>
- [16]. Anand, A., Gupta, M., Prasad, K., & Singla, N. (2024). *Mathify: Evaluating large language models on mathematical problem solving tasks*. arXiv preprint. <https://arxiv.org/abs/2404.13099>
- [17]. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35, 24824–24837.
- [18]. Frieder, S., Pinckaers, M., Baracaldo, N., Rao, A., Charette, R., Mirhoseini, A., ... & Ghosh, S. (2023). *Mathematical capabilities of ChatGPT*. arXiv preprint. <https://arxiv.org/abs/2301.13867>
- [19]. Webson, A., Loo, A. M., Yu, Q., & Pavlick, E. (2023). Are Language Models Worse than Humans at Following Prompts? It's Complicated. *arXiv preprint arXiv:2301.07085*.
- [20]. Chen, L., Zaharia, M., & Zou, J. (2024). How is ChatGPT's behavior changing over time? *Harvard Data Science Review*, 6(2). <https://doi.org/10.1162/99608f92.5317da47>



- [21]. Slany, E., Scheele, S., & Schmid, U. (2024, August). Explanatory Interactive Machine Learning with Counterexamples from Constrained Large Language Models. In *German Conference on Artificial Intelligence (Künstliche Intelligenz)* (pp. 324-331). Cham: Springer Nature Switzerland. <https://doi.org/10.1145/3706468.3706488>
- [22]. Awang, L. A., Yusop, F. D., & Danaee, M. (2025). Current practices and future direction of artificial intelligence in mathematics education: A systematic review. *International Electronic Journal of Mathematics Education*, 20(2), em0823. <https://doi.org/10.29333/iejme/16006>
- [23]. Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., ... & Gui, T. (2025). The rise and potential of large language model based agents: A survey. *Science China Information Sciences*, 68(2), 121101. <https://doi.org/10.1007/s11432-024-4222-0>
- [24]. Garrido, A. (2012). AI and Mathematical Education. Retrieved from MDPI.
- [25]. Shin, I., Hwang, S. B., Yoo, Y. J., Bae, S., & Kim, R. Y. (2025). Comparing student preferences for AI-generated and peer-generated feedback in AI-driven formative peer assessment. *Proceedings of the ACM*. <https://doi.org/10.1145/3706468.3706488>
- [26]. Zhai, X., & Nehm, R. H. (2023). AI and formative assessment: The train has left the station. *Journal of Research in Science Teaching*. <https://doi.org/10.1002/tea.21885>
- [27]. Hopfenbeck, T. N., Zhang, Z., Sun, S. Z., & Robertson, P. (2023). Challenges and opportunities for classroom-based formative assessment and AI: A perspective article. *Frontiers in Education*, 8, 1270700. <https://www.frontiersin.org/articles/10.3389/feduc.2023.1270700/full>
- [28]. Wachira, P. W. (2025). Mathematics teaching and assessment in the age of generative AI. In *Mathematics Education Perspectives*. <https://doi.org/10.4018/979-8-3693-6351-5.ch011>
- [29]. Polya, G. (1945). *How to solve it: A new aspect of mathematical method*. Princeton University Press.
- [30]. Daher, W., & Gierdien, F. (2024). Use of Language By generative AI Tools in Mathematical Problem Solving: The Case of ChatGPT. *African Journal of Research in Mathematics, Science and Technology Education*, 28(2), 222-235. <https://doi.org/10.1080/18117295.2024.2384676>



APPENDIX

The Third solution provided by AI

Step 1: Understanding the Clues

We have the following dialogue clues:

1. The product of the children's ages equals the number of cats the woman across from them feeds.
 - This means that the product of their ages is a known number, but we don't explicitly know what it is.
2. A says: "From this data, there is not enough information to know their ages!"
 - This is a crucial clue! It tells us that multiple possible age pairs exist that yield the same product.
 - If there were only one possible pair of ages, A would immediately know the answer.
3. B then adds: "The eldest son looks like his mother."
 - This confirms that there is an eldest son, so the ages cannot be the same (no twins).
 - This single extra piece of information allows A to now determine the ages.

Step 2: Finding Ambiguous Products

To satisfy clue #2, we need to find a product that appears more than once when listing reasonable child ages (keeping in mind they are "not in school yet," so likely between 1-9 years old).

Let's list all possible integer age pairs and their products:

Table 5. Possible Pairs in the Third AI's Solution

Age 1	Age 2	Product
1	6	6
2	3	6
1	12	12
2	6	12
3	4	12
2	9	18
3	6	18
4	4	16

The key observation here is that:

- Product 6 has two solutions: (1,6) and (2,3)
- Product 12 has three solutions: (1,12), (2,6), and (3,4)
- Product 18 has two solutions: (2,9) and (3,6)

Since A originally couldn't determine the ages, the product must be one that has multiple possibilities. This means the product must be either 6, 12, or 18.

Step 3: Applying the "Eldest Son" Clue

Now, let's check which of these ambiguous cases get resolved by the father's additional hint.

- If the children were (2,6), there is a clear eldest son (6 years old).
- If the children were (3,4), there is also a clear eldest son (4 years old).
- If the children were (2,9), there is a clear eldest son (9 years old).

However, the only case where two possible pairs exist AND the eldest son clue resolves the ambiguity is the product 12.

Since the father confirms there is an "eldest son," this eliminates (2,6) (because in (2,6) or (1,12), the statement "eldest son" may still be ambiguous).

Thus, the only valid answer is:

Final Answer: The children are 3 and 4 years old.



Instructor: It seems that you need to use the following data: "I have two sons who don't go to school yet".

AI: You're absolutely right! The statement "*I have two sons who don't go to school yet*" is an important clue that sets a realistic age limit for the children. Let's go through the solution carefully again, incorporating this constraint properly.

The Fourth solution provided by AI

Step 1: Understanding the Clues

1. "I have two sons who don't go to school yet."
 - This suggests that both children are likely younger than 7 years old (as most children start school around age 6 or 7).
2. "The product of their ages equals the number of cats the woman across from you feeds."
 - This means the two ages multiply to some unknown number.
3. "From this data, there is not enough information to know their ages!"
 - This tells us that multiple pairs of ages have the same product.
4. "The eldest son looks like his mother."
 - This means there is a clearly defined eldest son (i.e., no twins with the same age).

Step 2: Listing Possible Age Pairs

Since both children are younger than 7, we list all pairs of numbers under 7 and their products:

Table 6. Possible Pairs in the Fourth AI's Solution

Age 1	Age 2	Product
1	1	1
1	2	2
1	3	3
1	4	4
1	5	5
1	6	6
2	2	4
2	3	6
2	4	8
2	5	10
2	6	12
3	3	9
3	4	12
3	5	15
3	6	18
4	4	16
4	5	20
4	6	24
5	5	25
5	6	30
6	6	36



Step 3: Identifying the Ambiguous Case

- The key clue is that A couldn't determine the ages initially.
- This means the product must appear at least twice in the list (otherwise, A would have known the answer immediately).

Looking at the table:

- Product 6 appears for (1,6) and (2,3).
- Product 12 appears for (2,6) and (3,4).

These are the only products that appear more than once, making them ambiguous.

Step 4: Using the "Eldest Son" Clue

- The father then says, "The eldest son looks like his mother."
- This means there is a clearly defined eldest child (so we must eliminate any pair where the two numbers are the same, like (3,3) or (4,4)).
- Between (2,6) and (3,4), the pair (2,6) has a clear eldest child (6-year-old).
- The pair (3,4) also has a clear eldest child (4-year-old).

However, only (2,6) fits within the "don't go to school yet" constraint, since 6-year-olds often have not started school yet, whereas a 7-year-old likely would have.

Final Answer: The children are 2 and 6 years old.

Instructor: The answer is still wrong. Do you want a hint?

AI: Yes, I'd love a tip! I want to make sure we reach the correct solution. What hint do you have in mind?

Instructor: You wrote the next argument: "no twins with the same age", but you did not use it.

AI: You're absolutely right! I mentioned that the eldest son's clue means there can't be twins (equal ages), but I didn't fully use it in my reasoning. Let's go through the solution one last time correctly and make sure we get the right answer!

The Fifth solution provided by AI

Step 1: Key Information from the Puzzle

1. "I have two sons who don't go to school yet."
 - This likely means both children are younger than 7 years old (since school typically starts at age 6 or 7).
2. "The product of their ages equals the number of cats that the woman across from you feeds."
 - This tells us their ages multiply to a specific number.
3. A responds: "From this data, there is not enough information to know their age!"
 - This means that multiple age pairs have the same product.
4. B adds: "The eldest son looks like his mother."
 - This means there is a clearly defined eldest child → so the two children must have different ages (no twins!).

Step 2: Listing Possible Age Pairs (Younger than 7)

Since the children are younger than 7, we consider only numbers 1 to 6 and list the possible age pairs along with their products:

Table 7. Possible Pairs in the Fifth AI's Solution

Age 1	Age 2	Product
1	1	1
1	2	2
1	3	3
1	4	4



Age 1	Age 2	Product
1	5	5
1	6	6
2	2	4
2	3	6
2	4	8
2	5	10
2	6	12
3	3	9
3	4	12
3	5	15
3	6	18
4	4	16
4	5	20
4	6	24
5	5	25
5	6	30
6	6	36

Step 3: Identifying the Ambiguous Product

Since A couldn't determine the ages, the product must appear at least twice in the table. Checking:

- Product 6: (1,6) and (2,3) → Ambiguous
- Product 12: (2,6) and (3,4) → Ambiguous
- Product 16: (4,4) (Appears only once) ✗
- Product 18: (3,6) and (2,9) (But 9 is too old) ✗

Since the number of cats must be a product that appears in multiple age pairs, the correct number is 12, which corresponds to (2,6) and (3,4).

Step 4: Applying the "Eldest Son" Clue

- The father then says, "The eldest son looks like his mother."
- This means there is a single eldest child.
- If the children were (2,6), the eldest would be 6.
- If the children were (3,4), the eldest would be 4.

BUT, (2,6) is incorrect because (6-year-old children sometimes start school).

AND (3,4) is incorrect because there is NO eldest if both were 4!

The wrong option is (4,4) because there is no eldest, so the correct answer must be (3,4)!

Final Answer: The two sons are 3 and 4 years old! ✓

Thank you for the tip! You helped me see the mistake in my previous reasoning. Hope this is now the 100% correct answer!

Instructor: your explanation: "BUT, (2,6) is incorrect because (6-year-old children sometimes start school). AND (3,4) is incorrect because there is NO eldest if both were 4!" is unacceptable. "6-year-old children sometimes start school" could not be a reason. Try to find the correct answer.



AI: You're absolutely right again! I need to strictly follow the logical clues given in the puzzle without bringing in unnecessary assumptions like school-age rules. Let's solve this puzzle rigorously and correctly.

The Sixth solution provided by AI

Step 1: Key Information from the Puzzle

We have the following statements:

1. "I have two sons who don't go to school yet."
 - This tells us the children are likely 6 years old or younger (but not older).
2. "The product of their ages equals the number of cats that the woman sitting across from you feeds."
 - The ages multiply to a specific number (which is not explicitly given).
3. A says: "From this data, there is not enough information to know their age!"
 - This means the product appears in more than one possible pair of ages.
4. B adds: "The eldest son looks like his mother."
 - This confirms that there is a clearly defined eldest child → meaning the two children must have different ages (no twins).

Step 2: Listing All Possible Age Pairs (Ages ≤ 6) and Their Products

Since the children are at most 6 years old, we list all valid pairs:

Table 6. Possible Pairs in the Sixth AI's Solution

Age 1	Age 2	Product
1	1	1
1	2	2
1	3	3
1	4	4
1	5	5
1	6	6
2	2	4
2	3	6
2	4	8
2	5	10
2	6	12
3	3	9
3	4	12
3	5	15
3	6	18
4	4	16
4	5	20
4	6	24
5	5	25
5	6	30
6	6	36



Step 3: Finding the Ambiguous Product

Since A could not determine the answer at first, the product of the ages must appear in more than one pair. Checking for repetition:

- Product 6: (1,6) and (2,3) → Appears twice ✓
- Product 12: (2,6) and (3,4) → Appears twice ✓
- Product 16: (4,4) → Appears once ✗
- Product 18: (3,6) → Appears once ✗

The only possible products that A found ambiguous were 6 and 12.

Step 4: Using the "Eldest Son" Clue

- The father then says, "The eldest son looks like his mother."
- This means there is a single eldest child → the ages must be different (no twins).
- This eliminates (4,4), (3,3), and (2,2), and (6,6) because those do not have an "eldest."

Now we check the two remaining ambiguous cases:

1. Product 6:
 - (1,6) → eldest is 6 ✓
 - (2,3) → eldest is 3 ✓
 - Since both pairs have an eldest, we need to check if the second statement resolves the ambiguity.
2. Product 12:
 - (2,6) → eldest is 6 ✓
 - (3,4) → eldest is 4 ✓

At this point, we need to think:

- If the product were 6, then (1,6) and (2,3) would have left A confused.
- If the product were 12, then (2,6) and (3,4) would have left A confused.

Now, the "eldest son" clue must remove the ambiguity.

- If the answer were (2,6) or (3,4), A would still be confused after hearing that there was an eldest son.
- However, (1,6) is unique because 6 is clearly the eldest!

Final Answer: The children are 1 and 6 years old. ✓

Why This is the Only Correct Answer

1. The product must be ambiguous, meaning it appeared in at least two different pairs → The product must be 6 or 12.
2. The "eldest son" clue removes the ambiguity.
 - (1,6) is the only valid pair where the eldest son (6) is immediately obvious.
 - (2,3) would still leave ambiguity between the pairs (1,6) and (2,3).
 - (2,6) vs. (3,4) would still leave ambiguity after the eldest son hint.
3. Thus, the correct answer must be (1,6).

Final Answer: The two sons are 1 and 6 years old.

Instructor: The answer is still incorrect. You need to find 2 pairs with the same product, whereas one of the pairs is twins.

AI: Thank you for the tip! Now, let's go through the solution correctly using your guidance.

