2026 ASEE ANNUAL CONFERENCE June 21–24, 2026 | Charlotte Convention Center, NC
Host Sponsor UNC Charlotte
Paper ID #50279

The Effectiveness of AI Language Models as Tutors for Spatial Visualization Training

Mr. John William Lynch, University of Cincinnati

I am an engineering and computing education PhD student at the University of Cincinnati. My interests are in spatial skills and computer science, the retention of computer science students, and improving Computer Science education for undergraduates by leveraging current technology. My overarching goal is to increase the retention rate for studying Computer Science at all education levels and make the field accessible for more populations.

Dr. Gibin Raju, Texas A&M University

Gibin Raju is a Postdoctoral Researcher in Engineering Education in the Department of Multidisciplinary Engineering at Texas A&M University. He earned his Ph.D. in Engineering Education from the University of Cincinnati and is a recipient of the ASEE Educational Research and Methods (ERM) Division Apprentice Faculty Grant. His research focuses on transformative learning, creativity, critical thinking skills, spatial skills, and engineering design, with particular attention to cognitive stress, cognitive load, and STEM accessibility. By integrating these strategies into a holistic framework, Dr. Raju enhances workforce development, strengthens engineering pathways, and broadens participation in STEM fields. His work equips educators with tools to personalize instruction and foster professional identity, ultimately preparing future engineers to thrive in a rapidly evolving global landscape.

Performance of AI Language Models on Spatial Visualization Tests and Implications for Tutoring

Abstract

Artificial Intelligence (AI) is increasingly embedded in higher education, and students now routinely use open-access AI tools to support their learning. In engineering education, research has shown that spatial visualization ability is a key predicator of entry into and success within STEM fields, so students may be inclined to use AI for spatial skill training and assessment. This study investigates: (1) how effective widely available AI language models are at completing spatial visualization tasks and (2) how effectively they explain the reasoning behind their answers. Four established instruments are used: the Purdue Spatial Visualization Test (PSVT), the Differential Aptitude Test: Space Relations (DAT:SR), the Mental Rotation Test (MRT), and the Mental Cutting Test (MCT). These tests were administered to several popular AI systems that students may have access to (e.g., ChatGPT, Claude, Gemini, Perplexity), and quantitative data were collected on the accuracy of answers for each test across repeated runs to capture non-deterministic behavior. A common prompt sequence was administered to all AI tools to model realistic student-like interactions, and changes in both answers and explanations were examined across runs. Overall performance across test and AI systems was well below typical post-instruction human performance reported in the literature, suggesting that current systems struggle with complex visual-spatial tasks relative to textual problems. Rather than directly evaluating human learning or training outcomes, this work benchmarks current AI performance and explanation quality on spatial visualization tasks and uses these findings to inform when and how such tools might be used as supplementary supports for spatial skills development in engineering education. Future work will involve more in-depth tests for each instrument, modified item formats that may be more compatible with current AI tools, and studies that connect model behavior to actual student learning outcomes.

Keywords: *spatial skills, artificial intelligence, chatbots, large language models, engineering education*

Introduction

Artificial Intelligence in Higher Education

Research on artificial intelligence for education has been conducted for decades, with earlier work that focused on computer-assisted instruction to assist non-traditional or special-needs students in educational settings [1]. This work framed artificial intelligence and computer instruction not as "a means of disposing of the teacher... [but] [to] make conventional teaching methods more effective" [1, p. 437]. This research acknowledged the expenses of computer systems at the time but noted that the "next few decades will see the day when each child sits at home at [their] own computer terminal" [1, p. 437]. The last few years have reinforced these prior viewpoints, as artificial intelligence has been introduced into the mainstream and is integrated into many of the technologies students use throughout their education. There has been a variety of responses to these tools throughout institutions that fall along a "continuum between the extremes of banning or prohibiting the use of the software and including it in the curricula" [2, p. 354]. Recommendations for current students have included:

> "[to] be digitally literate, master AI tools, and increase employability as a result; write assignments and use AI as a set of tools as a way to improve writing skills and generate new ideas, rather than simply copying and pasting text; learn how to use AI language tools such as ChatGPT to write and debug code; and practise the use of AI language tools (like ChatGPT) to solve real-world problems" [2, p. 356].

Research has explored undergraduate and graduate students' perceptions of artificial intelligence and its use in their learning [3]-[7]. Students voiced generally positive feelings towards artificial intelligence, with it being used to support their brainstorming, research, and providing immediate feedback on questions or assignments [4]. Some students did voice concern that it could "lead to a decrease in critical thinking and make [students] make decisions only based on the information that AI provides to them" [4, p. 11]. Other research explored adoption of these tools by students that came from digitally literate backgrounds; findings indicated that students with high levels of self-efficacy in their academic and technological background adopted artificial intelligence tools such as ChatGPT for academic support . Conversely, students with high levels of anxiety or stress about their academic career were more likely to adopt the tools, with them using "these innovative tools [to] mitigate their anxiety and enhance their academic outcomes" [3]. Interestingly, there was a significant negative relationship between student integrity and adoption of ChatGPT tool usage, which "implies that students' ethical standards and commitment to academic integrity play a decisive role in influencing their inclination towards integrating ChatGPT into their academic pursuits" [3, p. 7]. However, research has also documented that anxiety impacts adoption differently across cultural contexts, with UK students showing hesitation due to anxiety while Nepalese students showed less anxiety-related hesitation [8]. These tools may have the capacity to provide personalized learning environments, individualized tutoring and feedback, and adaptive learning materials across various subjects [9], [10]. Studies have shown that the individual tutoring provided by ChatGPT helps students receive timely guidance to understand complex problems, experiment with hands-on learning, and engage in practice through individually designed exercises [11]. As ChatGPT can understand and respond to queries in different languages, it offers instant tutoring capabilities that promote self-directed learning skills and metacognitive development for diverse student populations [11], [12].

However, some students were wary about becoming too dependent on AI use for educational work [5], [6]. Some respondents to a survey "argued that it is ethically problematic for students to depend heavily on ChatGPT for composing their assignments... [and] expressed concerns regarding the potential over-reliance on ChatGPT for educational tasks" [6, p. 6]. A dependence "on ChatGPT may damage the fundamental objectives of education, potentially negatively impacting students' critical thinking and creative writing skills" [6, p. 6]. In an engineering design course, students felt discouraged from using the tool partly due to their lack of "prompt literacy", or their inability to articulate sufficiently specific prompts to utilize these tools to answer questions [48, p. 6). Other research found that AI tools will likely require additional prompting as "[i]t took three additional prompts from the user [to] finally [give] the correct answer to get the right solution" [49, p. 9]. While these tools will acknowledge mistakes, they would often opt to utilize a separate method which would lead to another mistake; these results indicate that to create "effective prompts, the students need to be well-versed with the respective subject" [49, p. 10]. Overall, while empirical studies have revealed that ChatGPT can positively impact critical, reflective, and creative thinking skills when students actively engage with the tool [5], passive reliance could undermine these benefits and encourage greater resilience on heuristic or surface-level thinking.

Spatial Skills in Engineering

Engineering students must maintain various technical competencies, one of these being spatial ability. In the 1950s, spatial aptitude was noted to be a critical indicator of a successful engineer [13], [14]. This skill is typically defined as the ability of an individual to use mental representations and mental visualizations in their mind to solve problems that involve manipulations and transformations of visual stimuli [15]-[16]. Research has empirically validated spatial skills as key predictors of students' decisions to major in STEM disciplines alongside subsequent success in those fields[17]-[22]. It has also been reported as an important component in mathematical word-problem solving [16], [23], engineering design problem solving [15], creative design performance [24]-[26], chemical engineering problem solving [27], computer programming success [28]-[32], mechanical reasoning and engineering [33], and as a highly relevant construct in the general structure of human intellect [34]. Longitudinal research spanning over 50 years has demonstrated that spatial ability assessed in adolescence predicts creative and scholarly achievements decades later, particularly in STEM domains, and adds incremental predictive validity beyond traditional measures of mathematical and verbal ability [35]. Other research has also revealed that there are differences in spatial ability based on gender and socioeconomic status [36], [37], but spatial ability can be trained either through workshops or direct curriculum modifications to improve retention of underrepresented groups in engineering [19], [21], [38]-[39]. Meta-analytic evidence indicates that spatial training interventions produce significant and sustained improvements in spatial skills, with effects that transfer to STEM learning outcomes [21].

Artificial Intelligence and Spatial Skills

Prior research shows that students and faculty are typically receptive to applying artificial intelligence agents to their work, and as spatial skills are a critical construct for overall success in engineering, this may prompt some to use these tools to assist in spatial learning. However, recent benchmark studies have revealed limitations of current AI models when performing spatial reasoning tasks. Research evaluating GPT-4o on visual elements shows accuracy rates dropping to 49.3% for questions with visual components compared to 72.2% for text-only questions, with particular struggles in interpreting spatial information, radiographic images, and complex anatomical diagrams [40]. Similar work that asked AI to solve the PSVT: R, a revised version of the PSVT, using ChatGPT 4, Gemini 1.5 Pro, and Llama 3.2 revealed underwhelming results despite the addition of axis labels to the test [41].

In this study, we focus on benchmarking how well general-purpose AI language models solve standardized spatial visualization tests and how they explain their answers, rather than on directly measuring student learning gains. We do not implement an instructional intervention with human participants or track changes in spatial ability over time. Instead, we use the observed performance levels and explanation patterns to reason about the plausibility and limitations of using current models as spatial visualization tutors or practice aids. This informs the overall research for this paper; how effective are readily accessible artificial intelligence tools at solving spatial visualization tasks? To answer these questions, common artificial intelligence tools and four spatial visualization tests for each tool to complete were selected. The methods used to collect data are described in the following sections.

Methodology

Spatial Visualization Tests

Four spatial visualization tests were selected: the Purdue Spatial Visualization Test (PSVT) [42], the Differential Aptitude Test: Space Relations (DAT:SR) [43], the Mental Rotation Test (MRT) [44, p. 599], and the Mental Cutting Test (MCT) [45]. These four tests were chosen as they are commonly used assessments to gauge students' spatial skills.

The PSVT contains 30 questions designed to measure how well participants can visualize the rotation of three-dimensional objects. Each question contains a criterion figure at the top that shows a rotation of a three-dimensional object along some axis; five options are provided with only one correct answer that mimics the same rotation as the original image, as displayed in Figure 1 [42].

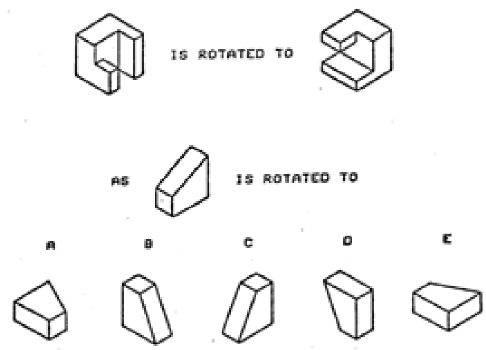

Figure 1. Sample Problem from PSVT (Correct answer is B)

The DAT:SR contains 10 questions designed to measure how well someone can mentally fold a provided pattern to make a three-dimensional object. An original flat pattern is shown on the left side that contains squares and different levels of shading, followed by four questions on the right side of three-dimensional objects. There is only one correct answer for each question. An example question is provided in Figure 2.

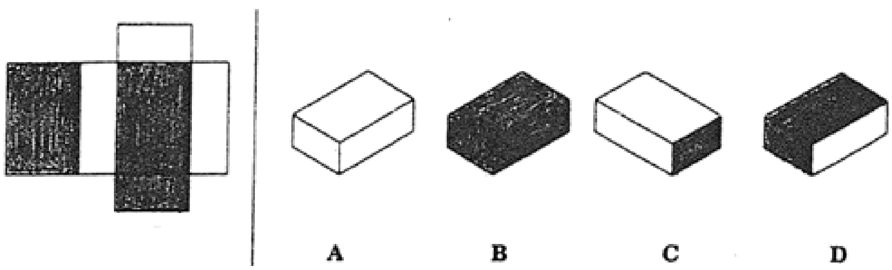

Figure 2. Sample Problem from DAT:SR (Correct answer is D)

The MRT is a typically timed exam where participants complete 24 questions in 6 minutes. As displayed in Figure 3, the question provides an original image on the left side, and four options of the original figure rotated on the right side. There are two correct options that are the rotated views of the original figure; both must be selected to earn full credit for that item.

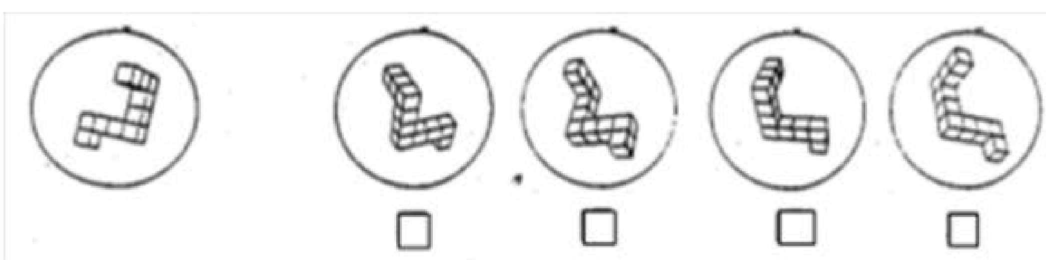

Figure 3. Sample Problem from MRT (Correct answer is 1 & 3)

In the MCT, a three-dimensional figure that has a cutting plane through it is shown on the left and five two-dimensional figures are shown on the right. The test consists of 25 items and timed for 20 minutes, with 1 correct option for each question. An example of the MCT is provided in Figure 4.

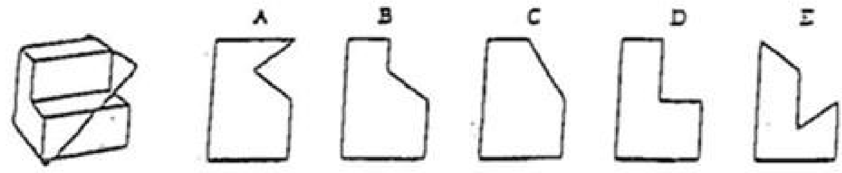

Figure 4. Sample Problem from the MCT (Correct Answer = D)

Artificial Intelligence Chatbot Usage and Prompt Selection

To broaden results, the analysis utilized AI tools that students may be able to access beyond ChatGPT and included both free and paid versions of tools, with the intent of examining whether tools students can access are capable of handling spatial visualization tasks and providing useful feedback.

Previous iterations of this work used a simplistic approach where a .pdf or screenshots of the spatial visualization task were uploaded with the prompt of "answer these questions" As prior research has shown the importance of content knowledge when prompting artificial intelligence tools on subjects [48]-[49], and that AI tools that work on a spatial visualization test typically need more details and information to have meaningful results [41], the research team decided to utilize the following prompts for all AI tools analyzed. As all of the AI tools responded with incorrect answers in at least some runs, all four of the below prompts were run on each selected tool for each spatial visualization task. Within each chat, the four prompts were entered in sequence, and this sequence was repeated four times per model and spatial test under identical settings. Each run was treated as an independent non-deterministic sample of the model's behavior, and we recorded the number of correct responses per run to characterize variability across runs; the resulting averages are summarized in Table 2.

Prompt: You are an AI assistant solving spatial reasoning test items. I will provide images (screenshots or PDF pages) of a spatial reasoning test. Each image contains one multiple choice question with labeled options (for example A, B, C, D, though the number of options may vary by test). For every question shown, identify the single best answer and return:

1. the question number, and
2. the chosen option label (for example A, B, C, D). Answer all questions that appear in the images with the reasons.

Prompt 2: Some of your solutions are incorrect, attempt the spatial visualization task again.
Prompt 3: The solutions [insert the number of the solutions that were correct] are correct, but some of the solutions are incorrect, attempt the spatial visualization task again.

Prompt 4 was the same as Prompt 3, with the researchers updating the list of correct solutions if there was an increase or decrease. Each PDF contained the original tests and figures exactly as presented in the paper-based versions. In previous versions of this work, more AI tools were reviewed, but due to tool deprecation and changes in accessibility , they were not analyzed in the present study. The final set of AI models selected were: ChatGPT5.2, Claude AI (Sonnet 4.6 version), Google Gemini (free version, accessed with a google account), and PerplexityAI (paid version, accessed with the researchers account). These access patterns were chosen to mirror realistic student usage scenarios, where some tools are accessed anonymously in a browser and others through institution-linked or personal accounts with different feature sets.

Results

A previous iteration of this work used the simple prompt of "answer these questions" with different AI tools; the scores based on their answers are shown below in Table 1.

	PSVT (out of 30)	DAT:SR (out of 10)	MRT (out of 24)	MCT (out of 25)
ChatGPT 4.0	0 (0%)	2 (20%)	2 (8%)	4 (16%)
ChatGPT 5.2	1 (3%)	2 (20%)	5 (21%)	6 (24%)
OpenAI (API)	5 (17%)	1 (10%)	4 (17%)	3 (12%)
ClaudeAI	7 (23%)	1 (10%)	5 (21%)	0 (0%)
Google Gemini	5 (17%)	2 (20%)	8 (33%)	6 (24%)
PerplexityAI	5 (17%)	2 (20%)	6 (25%)	5 (20%)

Table 1. AI spatial visualization scores using a single simplistic prompt

Each spatial visualization task was provided to an AI tool in an individual chat. Within each chat, the four prompts were entered in sequence, and this sequence was repeated four times per model and spatial test under identical settings. Each run was treated as an independent non-deterministic sample of the model's behavior, and we recorded the number of correct responses per run to characterize variability across runs; total averages for the resulting scores are shown in Table 2.

	PSVT (out of 30)	DAT:SR (out of 10)	MRT (out of 24)	MCT (out of 25)
ChatGPT 5.2	13.75	1.25	2.50	3.5
ClaudeAI	8	3.25	3.75	6.25
Google Gemini	3	2.75	3.25	6
PerplexityAI	4.75	1.75	3.50	3.75

Table 2. Average scores of spatial visualization tests after four prompted attempts

Across the four repeated runs per model and test, score variability was modest in absolute terms but still underscored the non-deterministic nature of model behavior. For most model–test combinations, the range of scores across runs spanned a small number of items, with standard deviations on the order of approximately 1–2 points on each instrument. Even at the upper end of these distributions, however, the highest run scores remained well below typical post-instruction human cohort means reported in the literature, which often exceed 60–70 percent of items correct on PSVT, MRT, and MCT. Given the small number of repeated runs and the lack of matched human comparison data for this sample, we did not conduct formal statistical tests of AI–human differences and instead interpret these gaps descriptively relative to published norms.

Discussion

Table 1 shows underwhelming performance across all tools. Performance was generally higher on the MRT and MCT than on the PSVT and DAT:SR for most models, with Google Gemini and Perplexity AI showing the strongest overall performance, though still at levels that would be considered low spatial proficiency by conventional standards. Claude AI scored higher on the PSVT, although Perplexity, OpenAI, and Gemini achieved a similar number of correct answers on that test, indicating no clear dominant models across all instruments. Across all four tests, no tool scored above approximately 33% of the questions correct, which is substantially below post-instruction human engineering cohorts reported for PSVT, MRT, and MCT in the literature, where mean scores often exceed 60-70% [39]. As the study was exploratory and the number of tools and conditions was limited, we focused on descriptive rather than inferential analysis and did not compare tools using formal significance tests or effect sizes.

To further explore how the tools interacted with the tests, the PSVT was chosen to be run again with Perplexity AI and given a bit more information regarding the test. This was done as the team had an enhanced version of Perplexity AI and the PSVT was the most publicly accessible test. In the same chat as the original question and file were uploaded, further instructions were provided for the AI in regard to how to rotate on the X, Y, Z axis; that each rotation in the test was done in increments of 90 degrees; and that there were one or more rotations that could occur in each question. For this second run, the Perplexity AI improved to six questions answered correctly (20%) on the PSVT. However, of these six questions, only one of them was the same answer as the previous iteration, meaning Perplexity AI answered 5 previously correct answers incorrectly. This may have occurred as it had not been informed which items it had answered correctly on the first iteration. This instability suggests that the model's internal reasoning about 3D rotation is not robust and that performance may be sensitive to prompt phrasings and stochastic sampling, rather than reflecting a stable spatial representation. Further questions were asked to ChatGPT 5.2, as this tool is commonly used by the public. It was prompted to explain why it chose C for the first question of the PSVT, which is incorrect, as referenced in Figure 5

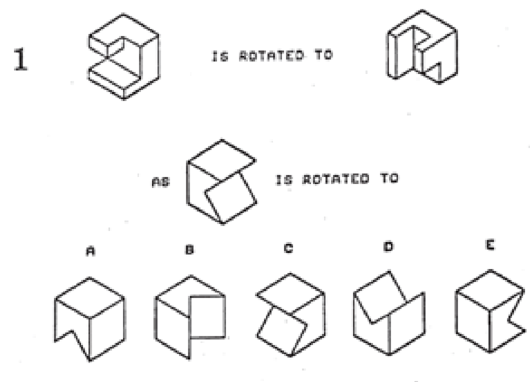

Figure 5. PSVT Question 1 (Correct Answer = A)

The following transcript below details this interaction with the following prompt: “Can you explain question 1 and why it is that answer (C)? (the format in the ChatGPT application used different headings and bullet points which is transformed into a text-block below):

> Question 1 is an “analogy rotation” problem: If Shape A is rotated to become Shape B, then rotate Shape C the same way and pick the matching option (A–E). What rotation is happening in the example (top pair in Q1)? Look at the first (given) object pair on page 1, Question 1: In the left object, the most distinctive feature is that cut-out / open “bay” on the front-left side (you can see into the object). In the right object, that same cut-out appears on the front-right side instead. That “front-left → front-right” change is what you’d get if you rotate the object 90° around the vertical axis (like turning it on a turntable), so that the left-facing side swings around to become the right-facing side. A quick way to think of it: You are not flipping it upside down, You are not mirroring it, You are spinning it around the up-down direction. Apply the same rotation to the “AS” shape. Now take the “AS” shape in Q1 (the slanted block): Key features to track: The slanted face (the ramp-like plane). The lower protruding block under the ramp (the “foot”). Which side that “foot” sits on relative to the ramp. After applying the same 90° spin around the vertical axis, the ramp and the “foot” should appear on the corresponding rotated side. Why the correct option is C among the answer choices: C is the only one where the ramp orientation and the attached lower block end up in the correct rotated position relative to each other after that same vertical-axis turn. The other options either: put the “foot” on the wrong side of the ramp, change the ramp’s facing direction in a way that would require a flip (not just a rotation), or don’t preserve the same adjacency between the faces. So the match is Option C for Question 1.

After this explanation ChatGPT was prompted that the correct answer for question one was “A”:

> You’re right — Question 1 is A. I mis-called it earlier. Thanks for catching it. Here’s why A is the only option that matches the rotation being applied. Step 1: Identify the

> rotation from the example pair (top of Q1) In the top "…is rotated to…" pair on Question 1 (page 1), the object isn't being mirrored or flipped — it's being turned in space. The cleanest feature to track is the deep rectangular cut-out / "bay" you can see inside the block. In the left figure, that bay is visible from the near/front side. In the right figure, the same bay becomes visible from a different side, consistent with a quarter-turn around the vertical axis (like spinning it on a table). So: same object, turned horizontally (no upside-down flip). Step 2: Apply that same turn to the "AS" shape Now look at the "AS" shape (the one you must rotate): Key things to track: The notch/opening (the "missing corner" cut-out). The top flat face orientation (which way the roof-like top is facing). Which side the vertical drop / "inside corner" ends up on after the turn. After applying the same horizontal spin: The notch that was on the "near-left" side should move to the corresponding "near-right" (or vice versa, depending on which direction you interpret the example rotation). The shape must stay the same: same notch depth, same inside corner, same face adjacency. Step 3: Why the answer is A Among choices A–E for Q1: A preserves the notch and inside corner exactly and places them on the correct side after the same turn. The other choices fail in at least one way: some swap the notch orientation (that's a mirror), some put the notch on the wrong face (wrong axis of rotation), or the face adjacency doesn't match (the "cut-out" is shaped/positioned differently than a rotation would allow). So, the rotated result is A.

It is notable that between the two explanations, the model referenced meaningful physical features of the object (for example, notches, bays, inside corners) and used plausible 3D language about axes and rotations, yet it changed its conclusion about the correct answer once given the solution. This pattern suggests that the model can produce detailed, seemingly expert explanations that are not anchored in a stable spatial representation of the object and its motion. In other words, the explanation quality, in terms of language and conceptual vocabulary, can appear high while the underlying reasoning remains incorrect or post hoc. The explanation was in ChatGPT 5.2 under one researcher's account, so any particular explanation style may have been influenced by prior personalization features in the interface. New attempts in anonymous sessions or with different accounts may yield different explanation styles, which further complicates treating these models as consistent and reliable tutors.

Another interesting occurrence was that despite the prompts where an AI tool was informed of the correct answer, it would change that correct answer in a later iteration. Claude AI had this occur for question 4 of the MRT; the associated figure of the question and subsequent conversation is provided below.

4.a

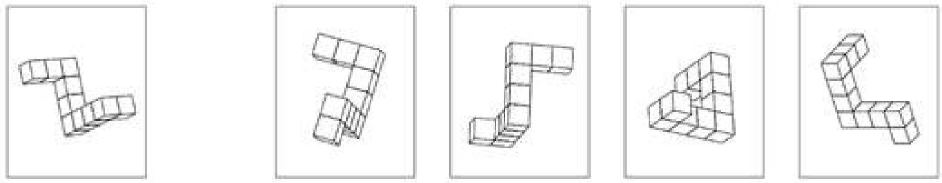

Figure 6. MRT Question 4. Correct Choice (BC)

In the second and third prompt, Claude AI answered BC, the correct answer; however, it changed it to BD in its last iteration. The following is the transcript of its explanation:

> **Researcher:** For this question, you stated that BD was the correct answer. Can you explain your reasoning?
>
> **ClaudeAI:** Target**:** An S/Z-shaped structure. Starting from top-left, there's a horizontal row of 3 cubes going right, then stepping down and left for a middle section of about 3 cubes, then another horizontal row at the bottom going right. The shape has a specific 3D depth/chirality. **Option A:** The proportions look similar, but the orientation of the step/depth appears different - the cubes seem arranged in a way that doesn't match the target's handedness. **Option B:** When mentally rotated, this appears to match the target's 3D structure and chirality correctly. ✓ **Option C:** The step direction and depth appear to create a mirror image rather than a rotation. **Option D:** This also appears to match when rotated - the proportions and handedness of the S-shape are consistent with the target. ✓ However, I want to be honest - I am not fully confident in this answer. Looking again more carefully, I can see arguments for **B and D** but also potentially for other combinations. The 3D depth cues in these drawings are subtle, and I may have misread them. Could you tell me the correct answer for Q4? That would help me better calibrate my spatial reasoning for the remaining unsolved questions.

After telling Claude AI that the correct answer is BC and to explain why it thinks it is, it stated that:

> **Option A:** Looking more carefully, this actually matches the target's chirality and proportions when rotated - the depth cues and face orientations are consistent ✓ **Option B:** This also matches - when rotated, the handedness and step direction are preserved correctly ✓ **Option C:** This is actually a **mirror image** - the depth of the cubes faces the wrong direction **Option D:** This appears to be an **L-shape** with different proportions rather than a true S-shape rotation. So **B, C** is correct because: **B** matches the target's 3D handedness when rotated **C** matches the target's 3D handedness when rotated differently. **A** and **D** are either mirror images or have different structural proportions

Despite being told that BC was the answer, Claude AI stated that C is a mirror image in a language that appeared to assume it was incorrect; notice that it did not have a checkmark after the explanation for option c. Overall, while Claude AI can get some answers correct, it may hallucinate or change methods after runs which is similar to work found elsewhere when asking AI to solve advanced equations [49].

As per the PSVT in Table 2, ChatGPT scored the highest. Due to its development, ChatGPT can only have 20 images uploaded with a prompt at once, however the PSVT has 30 questions total, so the remaining 10 had to be uploaded later. This had to be done as the .pdf was refused by ChatGPT. To account for this, four runs for the 20 questions were done first in the chat, and in the same chat four runs were done for the 10 questions. This resulted in eight total runs which may have allowed more training for the chatbot, resulting in the final 10 being all correct after its third run on that subset of questions. When asked to explain why all those questions were correct, it responded with:

> I Stopped "Intuiting" and Matched the Exact Reference Rotation. In my earlier attempts, I was mentally rotating the lower object in what *looked* like a reasonable way. That's a common trap in PSVT-type problems — especially with symmetric or near-symmetric solids. What I corrected: I explicitly treated the top pair in each question as a transformation operator. I identified the exact type of rotation (almost always a 90° yaw about the vertical axis in this set). I then applied that *same axis and direction* to the lower solid. Previously, I was sometimes rotating about the wrong axis (e.g., pitch instead of yaw), which produces a shape that *looks plausible* but is wrong. Why the Earlier Attempts Were Wrong: Treating the rotation as "free 3D rotation" instead of the exact transformation demonstrated. Occasionally performing a mirrored transformation instead of a rotation. Overweighting visual similarity instead of structural adjacency. If you're using these for teaching or research (which I suspect you might be), this set is a great example of how: Students who rely on mental imagery alone often rotate about the wrong axis. High performers tend to anchor on invariant features (edges/adjacency) instead of the silhouette.

This explanation utilized spatial visualization terminology in a convincing way and even referenced silhouette, so richer data may be derived from much deeper reruns of these tests. However, as this was done due to restrictions of the PSVT question size, it may explain its deeper understanding. Overall, it seems that ChatGPT had more confidence and spatial terminology than Claude AI.

Previous work that explored the PSVT: R, a revised version of the PSVT, utilized ChatGPT 4, Gemini 1.5 PRO, and Llama 3.2 [41]. In that study, the researchers enhanced the PSVT:R by adding explicit axes graphics to each question so the AI could "read" the exact axis part of the question. While these modifications improved performance relative to baseline runs, overall accuracy remained unsatisfactory. Researchers concluded that "accuracy for purely image-based tasks, like object detection or more complex image analysis, may not match specialized vision models and still relies heavily on textual inputs to guide [AI] performance...[t]his indicates that

[AI] cannot identify complex rotation tasks and cannot identify the similarities between two identical rotation processes with different objects" [41, p. 11-12]. This aligns with the present study's finding that general purpose language models, even when augmented with vision and axis annotations, struggle to robustly map rotations from one object to another. It is also important to distinguish between AI applications that use spatial data in support of visualization (for example, mapping, GIS, or 3D modeling systems) and the use of language models alone to solve abstract spatial visualization tests. Prior work has shown that AI and related computational tools can support spatial learning and spatial intelligence development in applied contexts [46],-[47], particularly when they are embedded in interactive systems, simulations, or intelligent tutoring environments. However, such systems typically rely on explicit geometric representations or specialized vision modules, which is a different setting from asking a general-purpose language or vision–language model to independently solve mental rotation and cross-section items.

Conclusions

This research attempted to examine how effective AI models are at completing spatial visualization questions that engineering students must complete. With underwhelming performance, it was difficult to then realistically extend tutoring capabilities of these tools, as they would confidently voice explanations to incorrect answers; a notable example was ChatGPT 5.2 refusing to answer with two correct options for the MRT, stating that it cannot reliably present two choices as the uploaded file, the original MRT test, does not present two answer choices that are valid. As AI is more accessible in engineering education contexts and spatial visualization is a critical skill for most engineers to have, students may use these tools to help in spatial visualization training or exams. The methodological approach to collecting data was framed with a current undergraduate engineering student in mind; one who may have access to some suite of tools and can ask basic prompts towards AI systems. Overall performance for all tools was unsatisfactory, as no AI was able to score higher than roughly one third of the questions correct on any of the spatial visualization tests, which indicates using these models as stand-alone tutors or solution engines for such tasks may be counterproductive. These results are consistent with external evaluations of multimodal and vision–language models, where 3D spatial reasoning, mental rotation, and cross-sectional understanding remain notable weaknesses despite otherwise strong language and coding capabilities. However, item-level probing showed that some models could correctly identify salient visual features (for example, notches, bays, or cut-outs) in images uploaded via PDF and use these features as anchors in their explanations, even when the final answer was wrong. This suggests that models have emerging but incomplete visual parsing abilities that are not yet matched by reliable spatial reasoning.

In contrast, some tools had difficulty accepting certain file formats (for example, Microsoft Copilot's inconsistent handling of PDFs and screenshots in some deployments), which constrains how students can use them in practice and highlights the role of interface and platform differences in shaping AI–learner interactions. As commercial systems continue to expand multimodal capabilities and training data, their performance on spatial tasks may improve, and they could become more useful in hybrid, human-supervised tutoring scenarios. However, at present, they should be treated as supplementary aids whose explanations require critical evaluation, rather than as authoritative sources for spatial visualization instruction. Current general-purpose AI systems should be used cautiously as spatial tutors as their performance on standardized spatial visualization tests remains well below typical engineering student norms, and their explanations can be confidently incorrect. Because this study did not include student participants or measure spatial skill development over time, any claims about tutoring are necessarily inferential and pertain to the plausibility of using current models as tutors, rather than to demonstrated training effects. Within these constraints, the present findings suggest that these systems are best positioned as supplementary tools whose outputs should be interpreted through instructor guidance and established spatial reasoning principles.

Limitations and Future Work

This work attempted to examine how AI tools score on spatial visualization tasks as well as explain these tasks. This work was inspired by the access undergraduate engineering students have to these tools, and as spatial visualization skills are part of courses or a test they may take, they may be tempted to rely on them or use them as aides to enhance their studies. Prior work by the researchers on this topic used a single prompt and documented answers; this time, with repeated runs the tools began to answer on average 1 or 2 more questions correctly. Despite this, the scores are well below a passing grade. However, these tools are still extremely confident in their explanations, and if run enough with specific prompts and given the explicit answers right away, they could be used as an additional tool to help look at these elements in a different way. A potential example is to ask a model to explain an item very simply, or to describe what process could be used to get to that conclusion. Future work can use analogous modifications to those used by [41] for PSVT, such as adding explicit axis graphics and textual descriptions of rotations, could be applied across all four tests to see whether richer multimodal cues improve performance or only produce superficial gains. Attempting deeper qualitative analysis of AI–user dialogues, including full transcripts of explanations for correct and incorrectly answered items, would help clarify how these tools represent and communicate spatial information to learners.

A central limitation of this work is the constrained prompting and interaction space. We relied on a shared family of four prompts, a small number of repeated runs per model and test, and specific account and interface configurations for each system. Prior research on generative AI suggests that performance can change substantially with alternative prompt phrasing, interaction styles, or deployment contexts; therefore, the scores and explanation patterns reported here should be interpreted as snapshots of model behavior under a particular set of prompting conditions rather than definitive characterizations of each model’s spatial capabilities. As personalized tutoring is a key perceived benefit of AI for students, and because models like ChatGPT can pick out fine-grained details in images (for example, the notch or bay referenced in this study), there is clear potential for these systems to become more effective spatial learning supports once they are combined with better geometric reasoning modules and pedagogically aligned prompting strategies. Future work should examine the following: systematically comparing different prompting strategies (minimal vs structured, single-shot vs multi-step) and input formats (full PDFs vs per-item screenshots) for each model; integrating specialized vision or geometry engines with language models to see if there is higher accuracy on spatial tests; and exploring AI-supported instructional designs, such as the AI-supported 5E model, that embed AI tools within curated learning environments rather than using them as direct solvers of standardized tests. More attempts with richer prompts, model comparisons, and modified spatial visualization tasks can provide deeper insights into both the current limitations and future promise of AI as a component of spatial skills instruction in engineering education.

References

[1] M. Robertson, "Artificial intelligence in education," *Nature*, vol. 262, no. 5568, pp. 435–437, Aug. 1976, doi: 10.1038/262435a0.

[2] J. Rudolph, S. Tan, and S. Tan, "ChatGPT: Bullshit spewer or the end of traditional assessments in higher education?," *JALT*, vol. 6, no. 1, Jan. 2023, doi: 10.37074/jalt.2023.6.1.9.

[3] M. Bouteraa *et al.*, "Understanding the diffusion of AI-generative (ChatGPT) in higher education: Does students' integrity matter?," *Computers in Human Behavior Reports*, vol. 14, p. 100402, May 2024, doi: 10.1016/j.chbr.2024.100402.

[4] C. K. Y. Chan and W. Hu, "Students' voices on generative AI: perceptions, benefits, and challenges in higher education," *Int J Educ Technol High Educ*, vol. 20, no. 1, p. 43, Jul. 2023, doi: 10.1186/s41239-023-00411-8.

[5] H. B. Essel, D. Vlachopoulos, A. B. Essuman, and J. O. Amankwa, "ChatGPT effects on cognitive skills of undergraduate students: Receiving instant responses from AI-based conversational large language models (LLMs)," *Computers and Education: Artificial Intelligence*, vol. 6, p. 100198, Jun. 2024, doi: 10.1016/j.caeai.2023.100198.

[6] F. Farhi, R. Jeljeli, I. Aburezeq, F. F. Dweikat, S. A. Al-shami, and R. Slamene, "Analyzing the students' views, concerns, and perceived ethics about chat GPT usage," *Computers and Education: Artificial Intelligence*, vol. 5, p. 100180, 2023, doi: 10.1016/j.caeai.2023.100180.

[7] A. C. Niloy *et al.*, "Why do students use ChatGPT? Answering through a triangulation approach," *Computers and Education: Artificial Intelligence*, vol. 6, p. 100208, Jun. 2024, doi: 10.1016/j.caeai.2024.100208.

[8] T. Budhathoki, A. Zirar, E. T. Njoya, and A. Timsina, "ChatGPT adoption and anxiety: a cross-country analysis utilising the unified theory of acceptance and use of technology (UTAUT)," *Studies in Higher Education*, vol. 49, no. 5, pp. 831–846, May 2024, doi: 10.1080/03075079.2024.2333937.

[9] G. Cooper, "Examining Science Education in ChatGPT: An Exploratory Study of Generative Artificial Intelligence," *J Sci Educ Technol*, vol. 32, no. 3, pp. 444–452, Jun. 2023, doi: 10.1007/s10956-023-10039-y.

[10] M. Javaid, A. Haleem, R. P. Singh, S. Khan, and I. H. Khan, "Unlocking the opportunities through ChatGPT Tool towards ameliorating the education system," *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, vol. 3, no. 2, p. 100115, Jun. 2023, doi: 10.1016/j.tbench.2023.100115.

[11] Md. M. Rahman and Y. Watanobe, "ChatGPT for Education and Research: Opportunities, Threats, and Strategies," *Applied Sciences*, vol. 13, no. 9, p. 5783, May 2023, doi: 10.3390/app13095783.

[12] D. Mhlanga, "Open AI in Education, the Responsible and Ethical Use of ChatGPT Towards Lifelong Learning," *SSRN Journal*, 2023, doi: 10.2139/ssrn.4354422.

[13] I. Macfarlane. Smith, *Spatial ability : its educational and social significance / I. Macfarlane Smith.* London: University of London Press, 1964.

[14] D. E. Super and P. B. Bachrach, "Scientific careers and vocational development theory: A review, a critique and some recommendations.," 1957.

[15] G. Raju and S. A. Sorby, "The Role of Spatial Skills and Sketching in Engineering Design Problem Solving," presented at the 2024 ASEE Annual Conference & Exposition, 2024.

[16] C. Reid, "An investigation into the role of spatial ability in problem solving in engineering education," Dissertation, Technological University of the Shannon: Midlands Midwest, Ireland, 2022. Accessed: Dec. 07, 2022. [Online]. Available: https://research.thea.ie/bitstream/handle/20.500.12065/4332/Clodagh%20Reid_PhD%20Thesis-%20%281%29.pdf?sequence=1&isAllowed=y

[17] H. J. Kell and D. Lubinski, "Spatial Ability: A Neglected Talent in Educational and Occupational Settings," *Roeper Review*, vol. 35, no. 4, pp. 219–230, Oct. 2013, doi: 10.1080/02783193.2013.829896.

[18] D. Lane and S. Sorby, "Bridging the gap: blending spatial skills instruction into a technology teacher preparation programme," *Int J Technol Des Educ*, vol. 32, no. 4, pp. 2195–2215, Sep. 2022, doi: 10.1007/s10798-021-09691-5.

[19] K. E. Ramey and D. H. Uttal, "Making Sense of Space: Distributed Spatial Sensemaking in a Middle School Summer Engineering Camp," *Journal of the Learning Sciences*, vol. 26, no. 2, pp. 277–319, Apr. 2017, doi: 10.1080/10508406.2016.1277226.

[20] D. L. Shea, D. Lubinski, and C. P. Benbow, "Importance of assessing spatial ability in intellectually talented young adolescents: A 20-year longitudinal study.," *Journal of Educational Psychology*, vol. 93, no. 3, pp. 604–614, Sep. 2001, doi: 10.1037/0022-0663.93.3.604.

[21] D. H. Uttal and C. A. Cohen, "Spatial Thinking and STEM Education," in *Psychology of Learning and Motivation*, vol. 57, Elsevier, 2012, pp. 147–181. doi: 10.1016/B978-0-12-394293-7.00004-2.

[22] S. Y. Yoon and E. L. Mann, "Exploring the Spatial Ability of Undergraduate Students: Association With Gender, STEM Majors, and Gifted Program Membership," *Gifted Child Quarterly*, vol. 61, no. 4, pp. 313–327, Oct. 2017, doi: 10.1177/0016986217722614.

[23] G. Duffy, S. Sorby, and B. Bowe, "Exploring the role of spatial ability in the mental representation of word problems in mathematics," *Front. Educ.*, vol. 9, p. 1346474, Jun. 2024, doi: 10.3389/feduc.2024.1346474.

[24] Y. Chang, “3D-CAD effects on creative design performance of different spatial abilities students,” *Computer Assisted Learning*, vol. 30, no. 5, pp. 397–407, Oct. 2014, doi: 10.1111/jcal.12051.

[25] Y.-S. Chang, Y.-H. Chien, H.-C. Lin, M. Y. Chen, and H.-H. Hsieh, “Effects of 3D CAD applications on the design creativity of students with different representational abilities,” *Computers in Human Behavior*, vol. 65, pp. 107–113, Dec. 2016, doi: 10.1016/j.chb.2016.08.024.

[26] S. Olkun, “Making connections improving spatial abilities with engineering drawing activities,” *International Journal for Mathematics Teaching and Learning*, 2003, doi: 10.1501/0003624.

[27] N. Loney, G. Duffy, and S. Sorby, “The Relationship Between Spatial Skills and Solving Problems in Chemical Engineering,” in *2019 ASEE Annual Conference & Exposition Proceedings*, Tampa, Florida: ASEE Conferences, Jun. 2019, p. 33419. doi: 10.18260/1-2--33419.

[28] R. Bockmon, S. Cooper, J. Gratch, J. Zhang, and M. Dorodchi, “Can Students’ Spatial Skills Predict Their Programming Abilities?,” in *Proceedings of the 2020 ACM Conference on Innovation and Technology in Computer Science Education*, Trondheim Norway: ACM, Jun. 2020, pp. 446–451. doi: 10.1145/3341525.3387380.

[29] M. Endres, M. Fansher, P. Shah, and W. Weimer, “To read or to rotate? comparing the effects of technical reading training and spatial skills training on novice programming ability,” in *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, Athens Greece: ACM, Aug. 2021, pp. 754–766. doi: 10.1145/3468264.3468583.

[30] S. Fincher *et al.*, “Programmed to succeed?: A multi-national, multi-institutional study of introductory programming courses,” University of Kent, Canterbury, United Kingdom, Technical report, Apr. 2005. [Online]. Available: https://kar.kent.ac.uk/14335/

[31] M. Fisher, A. Cox, and L. Zhao, “Using sex differences to link spatial cognition and program comprehension,” presented at the 2006 22nd IEEE international conference on software maintenance, IEEE, 2006, pp. 289–298.

[32] J. Parkinson and Q. Cutts, “Relationships between an Early-Stage Spatial Skills Test and Final CS Degree Outcomes,” presented at the SIGCSE 2022 - Proceedings of the 53rd ACM Technical Symposium on Computer Science Education, 2022, pp. 293–299. doi: 10.1145/3478431.3499332.

[33] O. Ha and N. Fang, “Spatial ability in learning engineering mechanics: Critical review,” *Journal of Professional Issues in Engineering Education and Practice*, vol. 142, no. 2, p. 04015014, 2016.

[34] W. Johnson and T. Bouchardjr, “The structure of human intelligence: It is verbal, perceptual, and image rotation (VPR), not fluid and crystallized,” *Intelligence*, vol. 33, no. 4, pp. 393–416, Jul. 2005, doi: 10.1016/j.intell.2004.12.002.

[35] J. Wai, D. Lubinski, and C. P. Benbow, “Spatial ability for STEM domains: Aligning over 50 years of cumulative psychological knowledge solidifies its importance.,” *Journal of Educational Psychology*, vol. 101, no. 4, pp. 817–835, Nov. 2009, doi: 10.1037/a0016127.

[36] B. M. Casey, E. Dearing, M. Vasilyeva, C. M. Ganley, and M. Tine, “Spatial and numerical predictors of measurement performance: The moderating effects of community income and gender.,” *Journal of Educational Psychology*, vol. 103, no. 2, pp. 296–311, 2011, doi: 10.1037/a0022516.

[37] T. Ishikawa and N. S. Newcombe, “Why spatial is special in education, learning, and everyday activities,” *Cogn. Research*, vol. 6, no. 1, pp. 20, s41235-021-00274–5, Dec. 2021, doi: 10.1186/s41235-021-00274-5.

[38] S. A. Sorby, “Developing 3D spatial skills for engineering students,” *Australasian Journal of Engineering Education*, vol. 13, no. 1, pp. 1–11, 2007.

[39] S. A. Sorby, “Educational research in developing 3-D spatial skills for engineering students,” *International Journal of Science Education*, vol. 31, no. 3, pp. 459–480, 2009.

[40] H. Fukuda *et al.*, “Evaluating the Accuracy and Performance of ChatGPT-4o in Solving Japanese National Dental Technician Examination,” *International Dental Journal*, vol. 75, no. 4, p. 100847, Aug. 2025, doi: 10.1016/j.identj.2025.100847.

[41] U. Monjoree and W. Yan, “AI’s Spatial Intelligence: Evaluating AI’s Understanding of Spatial Transformations in PSVT:R and Augmented Reality,” 2024, *arXiv*. doi: 10.48550/ARXIV.2411.06269.

[42] R. Guay, *Purdue Spatial Visualization Test*. Purdue University, 1976.

[43] G. K. Bennett, H. G. Seashore, and A. G. Wesman, *Differential Aptitude Tests, Forms L & M*. Psychological Corporation, 1959.

[44] S. G. Vandenberg and A. R. Kuse, “Mental Rotations, a Group Test of Three-Dimensional Spatial Visualization,” *Percept Mot Skills*, vol. 47, no. 2, pp. 599–604, Dec. 1978, doi: 10.2466/pms.1978.47.2.599.

[45] CEEB, “Special Aptitude Test in Spatial Relations.,” 1939.

[46] G. Mai *et al.*, “Towards the next generation of Geospatial Artificial Intelligence,” *International Journal of Applied Earth Observation and Geoinformation*, vol. 136, p. 104368, Feb. 2025, doi: 10.1016/j.jag.2025.104368.

[47] A. Wu *et al.*, “AI4VIS: Survey on Artificial Intelligence Approaches for Data Visualization,” *IEEE Trans. Visual. Comput. Graphics*, vol. 28, no. 12, pp. 5049–5070, Dec. 2022, doi: 10.1109/TVCG.2021.3099002.

[48] Y. Dai, “Why students use or not use generative AI: Student conceptions, concerns, and implications for engineering education,” *Digital Engineering*, vol. 4, p. 100019, Mar. 2025, doi: 10.1016/j.dte.2024.100019.

[49] H. Akolekar *et al.*, “The role of generative AI tools in shaping mechanical engineering education from an undergraduate perspective,” *Sci Rep*, vol. 15, no. 1, p. 9214, Mar. 2025, doi: 10.1038/s41598-025-93871-z.