From Thinking to Delegating: Cognitive Offloading in LLM-based play

Søren Emil Skaarup (SES), AU689717, 202207285@post.au.dk &

Sophia Inge Soewarta (SIS), & AU690636, 202104119@post.au.dk

Cognition and Communication, Cognitive Science, Aarhus University

Jens Chr. Skous Vej 2, 8000 Aarhus C, Denmark

Supervisor: Marc Malmdorf Andersen

19/05/2025

Keywords: *Constructive Play, Spatial reasoning, Mental Rotation, Large Language Models, Human-Computer Interaction*

Abstract

As artificial intelligence tools, particularly Large Language Models (LLMs), become increasingly integrated into educational and gaming environments, a need for critical assessment of their cognitive impacts on the users is imminent. This exploratory study investigates how LLM-assisted gameplay in Minecraft affects children's spatial reasoning, specifically mental rotation abilities, compared to traditional, non-assisted gameplay. Twenty children participated in a pre-post between subject design, completing a mental rotation task before and after a one-hour gameplay session. One group played standard Minecraft, while the participants of the experimental group had to build by prompting the language model in-game. Results showed that the group playing traditional Minecraft exhibited greater improvements in reaction time on the mental rotation task, whereas accuracy improvements were not significantly different between the groups. These findings suggest that self-directed gameplay may promote deeper cognitive engagement in spatial reasoning tasks, while LLM-assisted play may lead to cognitive offloading.

Table of contents

1.Introduction

In recent years, artificial intelligence (AI) has had an increased presence in educational and recreational settings. One of the leading developments is the rise of Large Language Models (LLMs), such as ChatGPT, which assist with writing, problem-solving, and even gaming. While these systems can provide guidance and efficient solutions, heavy use might promote the potential for cognitive offloading - where users rely on external systems to perform tasks that would otherwise require more mental effort (Gerlich, 2025). This trend is particularly relevant for children, whose cognitive capacities are still in development and may be shaped by their interactions with AI tools.

Spatial reasoning, and specifically mental rotation, is a critical component of cognitive development. It is closely tied to success in STEM-related fields and has been shown to be shapeable through gameplay (Newcombe & Frick, 2010). One domain that has attracted attention for its cognitive benefits is construction-based digital environments like Minecraft. Previous research has shown that playing Minecraft, or other block-building games, can enhance spatial abilities by engaging children in tasks that require mental rotation (Zhang et al., 2020), meaning that these games encourage players to rotate and “assemble” virtual objects.

However, the recent integration of Large Language Models (LLMs) into digital games such as Minecraft introduces a new challenge for cognitive development. When gameplay is augmented with AI-generated solutions, such as real-time construction, the demand for independent problem-solving may be reduced. While these tools can make games more accessible and engaging, especially for novices, they also risk encouraging a shift in how users approach tasks—relying more on the system’s output than on their own cognitive resources. The increasing availability of AI tools can lead to a form of cognitive offloading that undermines the development of critical thinking and problem-solving skills, particularly when individuals defer to AI guidance instead of actively engaging with challenges(Gerlich, 2025). In the context of spatial reasoning, where deep mental processing is essential for skill acquisition, this shift raises an important concern: does AI assistance in games support or hinder the development of cognitive abilities like mental rotation?

While the potential for cognitive offloading has been discussed theoretically and in adult populations (Risko & Gilbert, 2016), there is limited empirical research investigating its effects on children in interactive environments. In particular, there is a lack of studies that directly compare traditional gameplay with LLM-assisted gameplay to examine the potential trade-offs in skill acquisition.

This exploratory study addresses this gap by testing the effects of LLM-assisted versus regular Minecraft gameplay on children's mental rotation abilities. Twenty children participated in a between-subjects design, where one group played standard Minecraft and the other received assistance from an integrated LLM during gameplay. All participants completed a mental rotation task both before and after the intervention. We hypothesized that regular gameplay would promote greater cognitive engagement and result in stronger improvements in mental rotation compared to LLM-assisted play. This implies that we have assumed two null hypotheses; one for reaction times and one for accuracy. Both of them state that there are no significant differences between the groups, and thereby belong to the same population. The interest of our study is to assess whether or not these hypotheses can be rejected or not.

The aim of this study is to investigate whether LLM-assisted gameplay in Minecraft supports or hinders the development of children's mental rotation abilities compared to traditional, non-assisted gameplay. We hypothesize that children who engage in self-directed play without AI guidance will show greater improvements in mental rotation performance, as active cognitive involvement is more beneficial for spatial skill acquisition than externally guided problem-solving.

2. Methods

2.1 Ethics

This experiment was conducted as part of the Applied Cognitive Science course at the School of Communication and Culture, Aarhus University. The participants in this study were boys and girls between nine and thirteen years of age who regularly attend the Musvågevej after-school club. Prior to the commencement of the study, ethical approval was obtained from the pedagogical leader of the institute of child care at Musvågevej. This study was conducted in accordance with the ethical guidelines for research involving children and digital technologies. Our research protocol involved handling all data collection locally and anonymously. The study was designed to ensure a high standard of participant protection and data security. Informed consent was obtained from both parents/guardians and children, with age-appropriate forms for the participants. The consent process included information about the study's purpose and the procedure. Participants were informed of their right to withdraw from the study at any time without penalty, and no personal data beyond the general age range of the children at the after school club was collected. The study maintained transparency throughout the entire process, with no deceptive practices in the research design. A break was scheduled during gameplay sessions, and was taken if needed, to ensure the comfort of all participants.

The experiment was designed to pose no risks beyond those typically encountered in everyday life. The interactions were designed to minimize stress and discomfort for the participants, and debriefing sessions were held after participation in case any questions or concerns arose during the gameplay session. These were held to ensure transparency and understanding. We gathered the children in a comfortable setting, and they were reminded of the goals of the study and clarified the roles they had played. If the participants were interested, we explained the ideas behind the MindCraft project, including how the technology works, what data we had collected, and how their input contributes to the broader research goals in accessible terms. We also encouraged open discussion, giving the children a space to ask questions, express thoughts, and reflect on their experiences, as this also has potential to serve as valuable data. Their curiosity was met with straightforward and respectful answers. We believe this dialogue was crucial in reinforcing a sense of contribution and ownership in the research process. Importantly, we reassured the children that there was no right way to answer the tasks they completed. The session ended with gratitude for their involvement and an invitation to share any remaining concerns with the team of Musvågevej or their guardians.

2.2 Design

This study used a between-subjects pretest-posttest design to investigate the effect of LLM-assisted gameplay on children's spatial reasoning skills. Participants were randomly assigned to either a traditional Minecraft group or an LLM-assisted Minecraft group. Their spatial reasoning, and the change in said skills, was assessed before and after gameplay using a mental rotation task. This design was chosen to minimize potential biases by ensuring that each group experienced only one type of gameplay, while the pretest-posttest structure allowed for assessment of changes in spatial reasoning. The use of random assignment helps control for individual differences, making it possible to draw more robust statistical conclusions about the impact of LLM-assisted gameplay in Minecraft on mental rotation abilities.

Each group engaged in approximately one hour of gameplay. Before and after the intervention, participants completed the same mental rotation task, designed to assess spatial processing through a visual multiple-choice format.

2.2.1 Mindcraft

The primary difference between the two conditions was the use of a custom-built AI assistant in the experimental group. This version of Minecraft was powered by the open-source project “Mindcraft”, developed by Kolby Nottingham and publicly available on GitHub. Mindcraft integrates large language models into Minecraft Java Edition using the Node.js-based Mineflayer library. In our implementation, we used the OpenAI API key and specified the ChatGPT-4o model to power the in-game assistant, “Andy.” This setup enabled natural language interaction through typed chat commands in Minecraft, with the AI executing actions in the Minecraft environment.

The Minecraft world was hosted locally using Java Edition version 1.20.4, and “Andy” was launched using a custom script within the Mindcraft environment. The assistant’s behavior was defined using a profile andy.json file, and API access was managed through a keys.json file. Participants gave commands such as “build a house”, “build a church”, “dig a pool” in the game’s chat-line, which was processed using the ChatGPT-4o model and translated into in-game actions.

All commands, including the subcommands- and goals generated by the LLM itself, were visible in a parallel script in the computer terminal that ran alongside the game. This made the LLM’s decision-making process transparent and traceable, and allowed for the monitoring of how language instructions were interpreted and executed in real time, which also made it easier to identify when a participant’s input needed to be clarified. This allowed us to guide the participants upon errors, and to reset and troubleshoot the code, when they encountered errors.

2.2.2 Mental rotation task

To assess spatial reasoning in our participants, we used a mental rotation task inspired by the classic paradigm developed by Shepard and Metzler (1971). Their work demonstrated that people mentally rotate three-dimensional objects in order to determine if they are congruent, a process still regarded as a central mechanism in spatial cognition. In our adaptation of this task, each trial presented one grey target figure at the top of the screen and two red figures below it. Participants were instructed to decide which of the red shapes was the same as the grey one, merely rotated in space (*figure 1*). Only one of the two red alternatives was a correct match; the other was a distractor that differed subtly in structure. Using a computer mouse, participants selected the red shape they believed matched the target by mentally rotating it to align with the grey stimulus.

The mental rotation task used in this study was custom-built using the PyGame library, a Python package designed for writing interactive applications and games. PyGame allowed us to create an environment where stimuli could be precisely timed and user input could be accurately recorded. Participants responded using keyboard inputs, and both reaction times and accuracy were recorded for each trial and stored in CSV (comma-separated value) files. This ensured consistency across sessions and reduced the need for data cleaning.

The task began with five untimed practice trials to ensure participants understood the format. All instructions were given verbally and supported with clear visual examples. The task required no reading skills, making it suitable for a broad range of ages. Its design emphasized visual clarity and simplicity, with large, color-contrasted figures and intuitive interaction to make the process as engaging and accessible as possible.

For this study, we chose to use specially designed figures from a test developed by Gijsbert Stoet and made available through PsyToolkit, with the intention of the stimuli being more accessible to children than the classic Centicube-style images typically used in traditional spatial reasoning tasks. The smoother contours, clearer layouts, and reduced visual complexity of the Stoet stimuli made them more suited for young participants, especially those who might struggle with more abstract or detailed 3D representations like those used in Centicube-based formats.

Although our version of the task did not include explicit gamification features such as point scoring or levels, the experience was intentionally designed to feel like a small interactive game. Shapes appeared on screen with dynamic transitions, inviting participants to "solve the puzzle" by rotating them mentally and clicking the correct one. This structure, with stimuli “popping up” and requiring an active choice, created a light, game-like rhythm that is especially engaging for children (Habgood & Ainsworth, 2011). Research shows that even low-level gamification, or task designs that mimic aspects of gameplay, can increase intrinsic motivation and task performance (Mekler et al., 2017). By enhancing engagement in this way, we likely encouraged participants to perform at their best, which in turn improved the quality of the data collected.

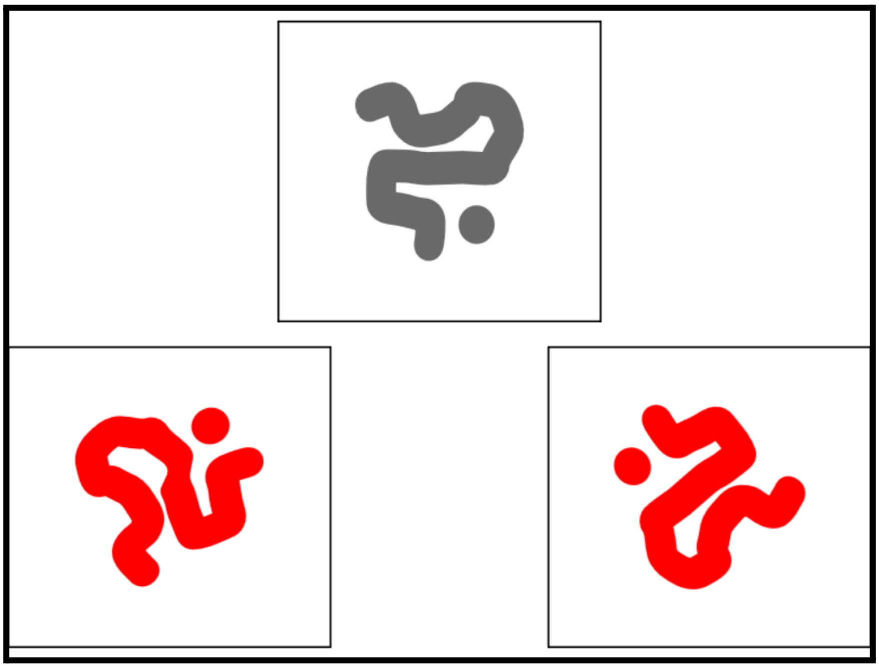

Fig. 1. Mental rotation task

In conclusion, our study employed a randomized between-subjects pretest-posttest design to examine whether our intervention produced a measurable effect on spatial reasoning. Participants in the intervention group played Minecraft with the support of an AI-powered building assistant, while the control group engaged with the standard version of Minecraft in creative mode. To assess potential cognitive changes, we administered a mental rotation task. This design allowed us to isolate the specific impact of offloading the building process by LLM assistance on participants’ performance in mental rotation, thereby probing how active- versus assisted spatial construction influences spatial cognition.

2.3 Procedure

Following the pre-test, participants engaged in approximately one hour of Minecraft gameplay. For both of the groups, we had created a completely flat minecraft world that would not give out any inspiration - a blank canvas for them to paint. The control group played the standard version of Minecraft in creative mode, focusing on free exploration and construction activities. The experimental group also used the standard version of Minecraft in creative mode, but with a twist. We hosted a local LAN server, and connected our LLM to the server via the Mindcraft code. The bot was then able to be communicated with, and respond through making API calls to OpenAI's servers. It replied to the user with human-like language, expressing statements that resemble classical outputs from language models. These replies would often include a consideration of the prompt it had been given, and a phrase that it would be executing it. Following this, it executed these javascript commands, which allowed the agent to move around and change blocks based on this dynamic, building simple structures and altering features. As a result, some participants in this group required additional support to enter commands, particularly when spelling or phrasing became a barrier. Assistance was provided as needed to ensure participants could engage meaningfully with the game, without influencing their decision-making or in-game behavior. After the session, the children completed the same mental rotation task they had taken before the gameplay, allowing for a direct comparison of their spatial reasoning skills before and after the intervention.

The procedure was designed to maintain as much consistency as possible across participants, with all participants engaging in the same gameplay duration and completing the same task at both time points. This structure allowed for an assessment of how the LLM-assisted Minecraft environment influenced spatial reasoning abilities in comparison to traditional gameplay.

2.4 Sampling plan

Participants were recruited from “Klubben Musvågevej” and ranged in age from 9 to 13 years. A total of 20 children took part in the study. Participants were assigned to one of two groups: the traditional Minecraft group or the LLM-assisted Minecraft group, with 10 children in each.

Although this is a relatively novel research topic, and no prior studies have examined the effects of LLM-assisted gameplay on children's mental rotation abilities specifically, related studies exist in similar domains. When conducting the power analysis, we referred to *RUNNING HEAD: LEGO® & MATHS IN CHILDREN* (McDougal et al., 2023), which compared children's spatial and mathematical reasoning after engaging with digital versus physical LEGO-based construction tasks. That study reported a Cohen’s *d* = 0.6. Based on this, required sample size was calculated using the *pwr* package in RStudio (R version 4.4.2), with α = .05 and desired power = .80. The analysis indicated that a minimum of 45 participants per condition would be necessary to detect an effect of comparable magnitude. Due to practical limitations and the exploratory nature of this research, the target sample size could not be achieved. This study should therefore be considered a pilot investigation into the cognitive effects of AI-driven support in educational gaming, and although the data collected are not sufficient to draw a definitive conclusion, the study offers a foundation for future research and draws attention to how different game designs may impact spatial reasoning in children.

2.5 Analysis plan

2.5.1 Data cleaning

Data were collected automatically using a custom Python/PyGame application that presented each participant with a mental-rotation task and recorded response accuracy and reaction time (RT) on ten trials. Each participant completed one pre-test and one post-test, yielding forty CSV files in total. A log recorded the mapping between each participant’s pre- and post-test filenames. A Python script then read each paired CSV, matched trials by index, and computed the signed reaction time change for each trial as RT_pre – RT_post. Positive values indicate faster responses after the intervention (improvement), whereas negative values indicate slower post-test performance. The trial-level differences for the ten control participants were concatenated into a single dataset (100 observations) and likewise for the ten experimental participants. The same pipeline extracted the binary correctness flag from each file, calculated overall accuracy (correct trials / 10) for pre- and post-tests, and computed per-participant accuracy change as accuracy_post – accuracy_pre, resulting in ten accuracy-change scores per group. The cleaned RT-change and accuracy-change datasets were then exported for statistical analysis in RStudio (R version 4.4.2).

2.5.2 Statistical tests

All analyses were conducted in R (R version 4.4.2) with the standard level of significance $\alpha = .05$. We analyzed two types of data: reaction-time differences across trials and overall accuracy differences per participant. For each dataset, we first tested for normality using the Shapiro–Wilk test. If the data passed the normality test, we would proceed to assess the equality of variances between groups using Levene's test. When both assumptions were satisfied, we would apply a two-tailed independent-samples Student's t-test. However, if either the normality or homogeneity assumption were violated, we would use the non-parametric Wilcoxon rank-sum test as an alternative to compare the control and experimental groups.

3. Results

3.1 Results on reaction time

The analysis of RT differences between the control and experimental groups revealed several key findings. Firstly, we worked with the dataset that contained the differences in reaction times. Initial assumption checks indicated that the RT difference data were not normally distributed for either group (Shapiro-Wilk test: $p < 0.001$ for both groups), though the variances were homogeneous (Levene's test: $p = 0.22$). These findings suggested the use of a non-parametric statistical test. The Mann-Whitney U test was used to compare RT differences between the control and experimental groups. The test revealed a significant difference between the groups ($U = 6010.50$, $p = 0.0136$), with the experimental group exhibiting smaller RT differences overall. However, the effect size was small ($r = 0.175$), indicating that while the experimental manipulation had a statistically significant impact, the magnitude of this effect was mild. Visualizations of the data further supported these findings (see figure 2). The experimental group's RT differences were more concentrated around lower values, while the control group showed greater variability and the presence of outliers. These patterns were consistent across different pairs, as illustrated in the additional visualizations (see figure 3). Conclusively, the results indicate that the experimental manipulation reduced RT differences compared to the control condition, though the effect was small.

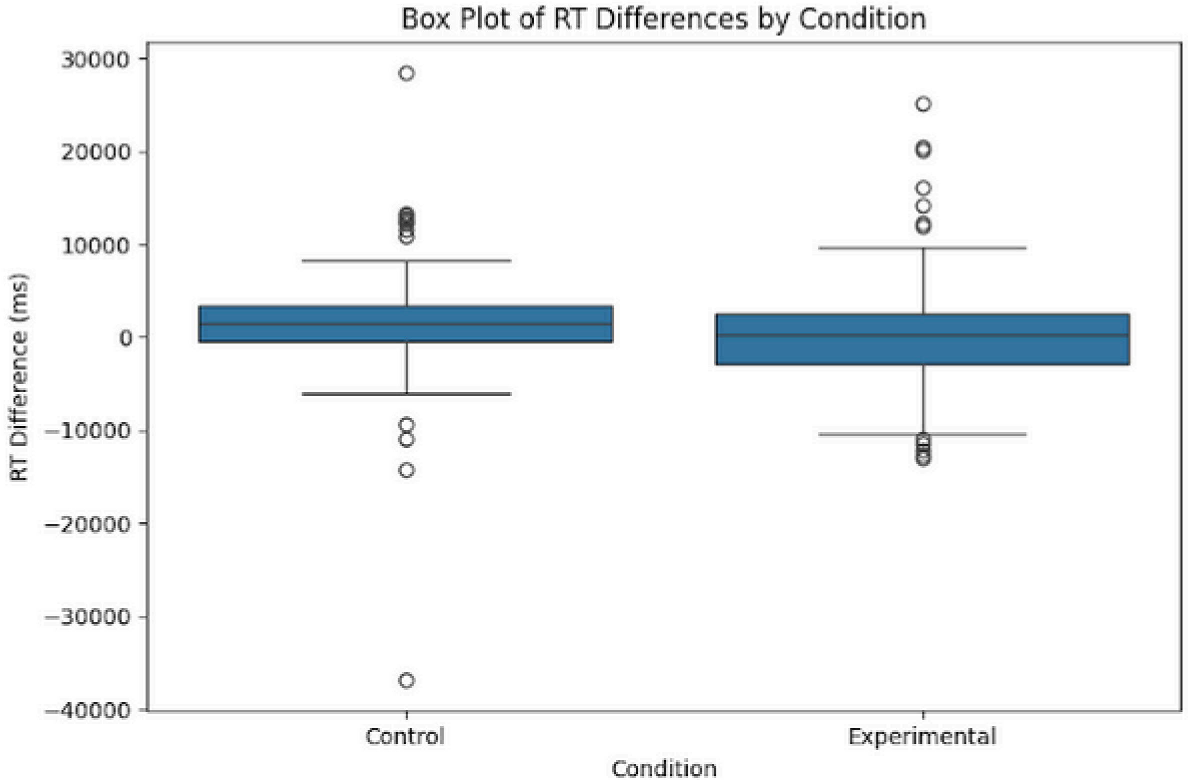

Fig 2. Boxplot of RT differences by condition

Figure 2 plots the pre-minus-post reaction-time (RT) differences for each condition, where positive values indicate faster responses after the intervention. The control group's median lies clearly above zero, while the experimental group's median sits close to (or slightly below) zero, signalling that children who played regular Minecraft tended to improve their speed more than those who played LLM-assisted Minecraft. Although the interquartile ranges overlap, and both conditions display substantial spread and numerous outliers, the overall shift of the control distribution toward positive values is consistent with the statistical result that regular play produced the larger gain in mental-rotation RT.

3.2 Results on accuracy

To assess whether the type of Minecraft children play affected accuracy in the mental rotation task, we compared the change in accuracy scores between the experimental and control groups. Shapiro-Wilk tests indicated that the change scores were normally distributed for both the experimental group (*p* = .648) and the control group (*p* = .521). An F-test showed no significant difference in variances, *F*(9, 9) = 3.45, *p* = .079, justifying the use of a Student's *t*-test. The t-test found no significant difference in mean change in accuracy between the experimental (M = -0.10) and control (M = -0.10) groups, *t*(18) = 0.00, *p* = 1.00, 95% CI [-2.42, 2.42].

4. Discussion

4.1 Evaluation of results

The results of this study provide an insight to the effects of playing LLM-assisted games on children's development of spatial abilities. Reaction time (RT) differences between the experimental (LLM-assisted Minecraft) and control (traditional Minecraft) groups were found to be statistically significant, with the experimental group showing smaller RT differences overall (U = 6010.50, p = 0.0136). This suggests that the LLM-assisted gameplay led to quicker responses, though the small effect size (r = 0.175), indicates that the practical significance of this effect is mild. Figure 3 depicts the distribution of pre-minus-post RT differences for both conditions. The control group (blue) clusters just to the right of zero, with the bulk of observations in the 0–5 000 ms range, indicating that most children were noticeably faster after unassisted play. The experimental group (orange) is centred closer to zero and extends further into negative territory, showing that many LLM-assisted participants either improved only marginally or actually slowed down. Both curves are right-skewed and exhibit long tails, but the overall shift of the control distribution toward positive values reinforces the conclusion that self-directed Minecraft yielded the greater gain in mental-rotation speed. This illustrates that the experimental group had more concentrated RT differences around lower values, while the control group displayed greater variability and outliers, reinforcing the idea of a modest, yet significant effect.

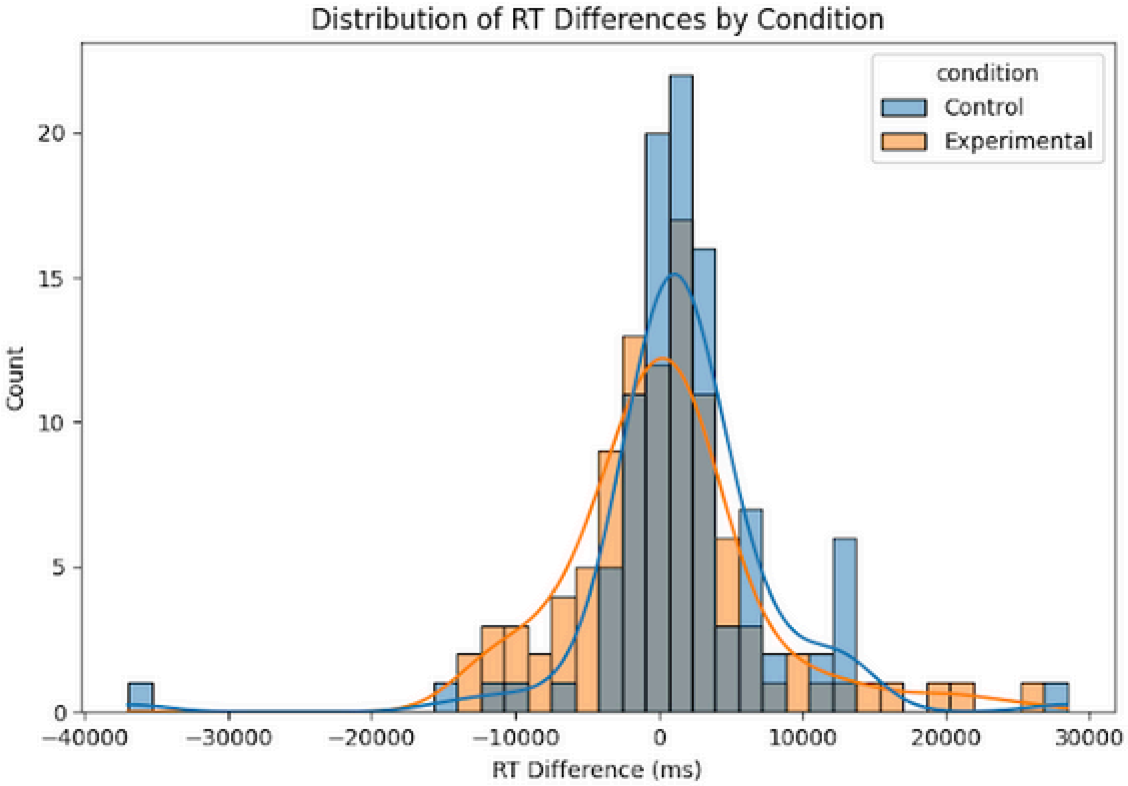

Fig 3. Histogram of RT differences by condition

In contrast, when examining accuracy, no significant difference emerged between the two groups. Both the experimental and control groups exhibited an equal mean change in accuracy (-0.10), and a Student's t-test found no significant difference in the mean change (t(18) = 0.00, p = 1.00). This suggests that despite the differences in reaction times, the type of gameplay did not lead to any measurable change in the accuracy of participants' mental rotation task performance. The results here imply that the intervention, although influencing RT, did not affect how accurately children were able to complete the task.

4.3 Limitations

While this study provides insights into the potential effects of LLM-assisted gameplay on children's spatial reasoning, several key limitations should be considered when interpreting the results. These limitations suggest areas for improvement and future research.

One of the primary limitations of this study is the small sample size, consisting of only twenty participants divided equally between the control and experimental groups. Furthermore, the participants in our study belong to the “WEIRD” demographic: Western, Educated, Industrialized, Rich, and Democratic, which has been shown to be unrepresentative of the global population, both in general and in research. Studies relying heavily on WEIRD samples may yield findings that do not generalize well to children from other cultural, economic, or educational backgrounds. Henrich et al. (2010) argue that WEIRD populations are often outliers on many psychological measures, and thus findings derived from such samples should be interpreted with caution when making broader claims about human cognition and development. This restricted the generalizability of the findings to a broader population of children. Additionally, the study’s small sample size may have resulted in a lack of statistical power, meaning that smaller or more subtle effects might not have been detected. A more diverse sample, drawn from multiple schools or communities would have been ideal. A limited sample size reduces the statistical power of this studies’ calculations, increasing the risk of issues such as Type II errors and potentially leading to overestimation of effect sizes. This concern is well-documented in cognitive science research, where small sample sizes have been associated with reduced reliability and reproducibility of findings (Button et al., 2013).

Another limitation was the one-hour session in which participants played Minecraft. A single, short session may not have provided enough time for the LLM-assisted gameplay to produce lasting or significant effects on the children’s cognitive abilities. For instance, studies involving extended cognitive training sessions over several weeks or months have demonstrated significant enhancements in targeted cognitive functions (Willis et al., 2006). A longer intervention or repeated sessions over multiple days would have yielded stronger or more consistent results as seen in studies of gaming behavior (Newcombe, N. S et al., 2012).

An additional limitation of our study is the potential for practice effects, as the same mental rotation task was administered both before and after the gameplay intervention. Repeated exposure to identical tasks can lead to performance improvements due to familiarity rather than the intervention itself (Meneghetti et al., 2017). This effect may be more pronounced in children who learn quickly, potentially skewing results in favor of those with higher cognitive processing speeds (Hambrick et al., 2014). Prior experience with similar tasks or gaming environments can thereby significantly influence learning outcomes. Children who were already familiar with Minecraft or video games in general may have found the tasks easier to understand and complete. In contrast, novices may have struggled with the game mechanics or interface, leading to disparities in performance improvements (Myhill & Brackley, 2004). However, experienced players are also subject to what is known as the expertise reversal effect: a phenomenon in which methods that are effective for novices become less effective, or even counterproductive, for individuals with more prior knowledge (Kalyuga et al., 2003). This effect is often explained in terms of cognitive schema theory. Novices lack well-developed “schemas” for approaching a task, so explicit instructions helps them build foundational understanding. In contrast, experts already possess task-relevant “schemas”, and additional guidance may interfere with or contradict their existing mental models.

The children who participated in this study had varying levels of familiarity with Minecraft and video games more broadly. While some participants had years of prior experience with Minecraft, others had never played the game or were unfamiliar with basic video game mechanics on a computer. This variation in gaming background presents a significant limitation, as it likely influenced how quickly participants adapted to the gameplay and how effectively they engaged with the task. Prior research has shown that previous gaming experience enhances learning and performance in new game environments, as experienced gamers are better at transferring skills and strategies across contexts (Green & Bavelier, 2012). As a result, children who were already comfortable with Minecraft may have found it easier to navigate and complete the tasks, while less experienced participants may have been disadvantaged by unfamiliarity with the controls, environment, or objectives. These individual differences in baseline proficiency introduce variability that complicates the interpretation of our findings, making it more difficult to isolate the effect of the LLM-assisted intervention.

Furthermore, gender differences in gaming habits also present a potential source of bias. Boys are generally more likely to engage in video gaming than girls, both in terms of frequency and familiarity with game mechanics (Van Rooij et al., 2014). This discrepancy may result in uneven starting points between genders, potentially influencing performance on both the intervention and the mental rotation task.

The study was conducted at an after-school club, which inherently caused a range of distractions. Participants were in a social environment, surrounded by peers who were engaged in various types of activities. As a result, interruptions such as noise or peer interactions, may have affected the children's ability to concentrate fully on the mental rotation task. These tasks require attention and cognitive engagement, and even minor distractions could have affected their performance. Previous research has demonstrated that distractions can impair performance on complex cognitive tasks, particularly those involving problem-solving and spatial reasoning (Sanders & Baron, 1975). Moreover, children are generally more susceptible to environmental distractions than adults, due to developmental differences in attentional control (Hoyer et al., 2021). These environmental factors likely contributed additional noise to our data, potentially masking or distorting the effects of the intervention.

The implementation of the LLM-assisted Minecraft system was constrained by financial limitations, primarily due to the cost structure associated with token-based access to large language models. Each API call incurs a charge based on the number of tokens used, which includes both the prompts sent to the model and the generated responses. In interactive applications like ours, where frequent and responsive dialogue is essential, these costs accumulate rapidly. More critically, more capable LLMs are significantly more expensive to use than smaller models, and their superior performance on standard language benchmarks has been well documented (OpenAI, 2023). Due to budgetary constraints, we were unable to deploy the most advanced available models or provide continuous, rich

AI support throughout the intervention. As a result, the AI guidance offered to participants may have lacked the nuance, relevance, or responsiveness that a more powerful (and costly) model could have delivered. This limitation likely impacted the quality of the LLM-assisted experience and may have diminished its potential cognitive benefit.

A more subtle limitation was the physical behavior of the children during the mental rotation task. Some children were observed to rotate their heads while completing the task, likely in an attempt to align the images with their visual perception. This behavior, though not directly related to the cognitive task itself, may have resulted in faster completion times, thus affecting the reaction time data. Similarly, research found that children performing mental rotation tasks while simultaneously rotating a handle experienced greater difficulty when the handle's rotation direction did not align with the mental rotation required by the task (Jansen, P., & Kellner, J., 2015). This suggests that manual and mental rotations may share common cognitive processes, and when these processes are not aligned, performance is hindered. It highlights the fact that children may engage in external actions (like head movements or manual tasks) to assist in completing the task, potentially clouding the interpretation of the data, especially because the children engaged in this behavior to varying degrees. This behavior underscores the complexity of measuring mental rotation in children, as it blurs the line between internal mental processes and external physical actions.

A final limitation of this study is the absence of a more comprehensive control condition. While we compared traditional Minecraft gameplay to an LLM-assisted version, we did not include a “no-intervention” control group—participants who completed the mental rotation tasks before and after the study without engaging in any gameplay. Such a group would help determine whether observed improvements stemmed from the interventions themselves or from practice effects, expectancy, or increased familiarity with the testing procedure. As Boot et al. (2013) argue, failing to include a passive control group leaves open the possibility that improvements are not due to the intervention, but to placebo-like effects or other uncontrolled factors associated with repeated testing.

4.4 Outlook

This study provides insights into the effect of LLM-assisted gameplay on children's spatial reasoning, though future research could expand on these findings in multiple ways. First and foremost, increasing the sample size would enhance the statistical power and generalizability of the findings. A larger and more diverse participant pool would help to reduce random variability and allow for more robust analyses, including gender-based comparisons. Extending the duration of the intervention is likewise critical. A single one-hour session is likely too short and unable to capture cognitive changes; and future studies should consider designs with repeated gameplay sessions over several weeks or months to better capture potential learning effects and cognitive development. A focus on long-term effects is crucial, as understanding whether the cognitive benefits of LLM-assisted gaming persist over time would provide more clarity on the intervention's lasting impact. Additionally, the inclusion of a "no-intervention" control group would help isolate the effects of gameplay from natural improvements due to repeated testing. Furthermore, the use of identical pre- and post-tests should be reconsidered. Employing parallel versions of the mental rotation task could reduce practice effects and allow clearer attribution of performance changes to the intervention rather than task familiarity. Environmental factors such as distractions in the after-school setting should be better controlled. Testing in a more quiet and structured environment would help reduce noise in the data and result in better measurement precision. Other factors such as differences in gaming experience across participants should be accounted for as ensuring a more uniform level of familiarity with the game would allow a clearer interpretation of the effects of the intervention itself. Exploring how specific game design elements influence cognitive outcomes could help optimize future LLM-assisted gameplay interventions. By addressing these factors, future research can deepen our understanding of how LLM-assisted computer interactions can be used effectively to support children's cognitive development.

4.5 AI is causing cognitive off-loading

Research has demonstrated that engaging in activities requiring spatial reasoning, and becoming good at them, are directly linked to success in STEM fields (Uttal et al., 2013). In this study, the lack of improvement in spatial reasoning in the LLM-assisted group might indicate that the reliance on AI assistance could hinder the development of these critical spatial skills. This concern taps into a broader debate of cognitive offloading, where humans may rely on AI to perform cognitive tasks, which could reduce their mental effort and with, e.g., spatial problems. Studies have shown that excessive reliance on technology, including AI, can impair the development of critical thinking and problem-solving skills (Lee, H.-P et al., 2025).

While AI is likely to be implemented in many games in the future, it raises important considerations about the impact on children's development. However, it is not only children's cognitive development that are on the line; Video games, especially multiplayer ones, serve as social hubs where children can interact, collaborate, and develop communication skills. Games like Minecraft foster creativity, teamwork, and peer learning. However, the introduction of AI could shift the focus from collaboration with other players to interaction with a virtual entity, potentially reducing these social interactions.

In conclusion, while AI in video games might become more common, it should complement, not replace, the social experiences that are central to children's development. Thoughtful integration of AI into gaming can help children enjoy both cognitive growth and meaningful social interactions.

5. Conclusion

This study set out to explore how LLM-assisted gameplay in Minecraft influences children's mental rotation abilities, compared to traditional, non-assisted gameplay. Our findings suggest that children who played regular Minecraft demonstrated significantly greater improvements in mental rotation performance than those in the LLM-assisted condition. This supports the notion that active, self-directed problem solving fosters spatial cognition more effectively than passively following AI-generated solutions. In contrast, accuracy improvements were not statistically significant between the two groups.

These results might contribute to a broader discussion about cognitive offloading in the age of interactions between humans and Large Language Models in relation to learning and education. As language models tools become increasingly embedded in educational and recreational contexts, there is a growing tendency - especially among novices - to delegate cognitive effort to external systems. Our study highlights that, in domains like spatial reasoning, doing the task oneself, not merely watching it be solved, is crucial for learning. As educational technologies continue to evolve, a balance between assistance and autonomy will be essential to ensure that the development of AI-tools supports, rather than replaces, the cognitive work learners need to do.

6. Data availability

All data from the experiment can be made available upon request.

7. Code availability

All code used in the experiment can be made available upon request.

8. Competing interests

The investigators declare no competing interests that might negatively impact the research.

9. Acknowledgements

Thank you to Aarhus University, School of Cognition and Communication. Special thank you to Marc Malmdorf Andersen and Mathilda Rhys-Kristensen for teaching, resources and supervision.

This research would not have been possible without the remarkable development of the MindCraft project by Max Robinson, as well as the generous support of the pedagogues and the head of Klubben Musvågevej, who warmly welcomed us and allowed us to carry out the study with the participating children.

10. References

Boot, W. R., Simons, D. J., Stothart, C., & Stutts, C. (2013). The pervasive problem with placebos in psychology: Why active control groups are not sufficient to rule out placebo effects. *Perspectives on Psychological Science, 8*(4), 445–454. https://doi.org/10.1177/1745691613491271

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14*(5), 365–376. https://doi.org/10.1038/nrn3475

Gerlich, M. (2025). AI tools in society: Impacts on cognitive offloading and the future of critical thinking. *Societies, 15*(1), Article 6. https://doi.org/10.3390/soc15010006

Green, C. S., & Bavelier, D. (2012). Learning, attentional control, and action video games. *Current Biology, 22*(6), R197–R206. https://doi.org/10.1016/j.cub.2012.02.012

Habgood, M. P. J., & Ainsworth, S. E. (2011). Motivating children to learn effectively: Exploring the value of intrinsic integration in educational games. *Journal of the Learning Sciences, 20*(2), 169–206. https://doi.org/10.1080/10508406.2010.508029

Hambrick, D. Z., Oswald, F. L., Altmann, E. M., Meinz, E. J., Gobet, F., & Campitelli, G. (2014). Deliberate practice: Is that all it takes to become an expert? *Intelligence, 45*, 34–45. https://doi.org/10.1016/j.intell.2013.04.001

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences, 33*(2–3), 61–83. https://doi.org/10.1017/S0140525X0999152X

Hoyer, R., ElShafei, H., Hemmerlin, J., Bouet, R., & Bidet-Caulet, A. (2021). Why are children so distractible? Development of attention and motor control from childhood to adulthood. *Child Development, 92*(3), e257–e273. https://doi.org/10.1111/cdev.13561

Jansen, P., & Kellner, J. (2015). The role of rotational hand movements and general motor ability in children's mental rotation performance. *Frontiers in Psychology, 6*, Article 984. https://doi.org/10.3389/fpsyg.2015.00984

Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). The expertise reversal effect. *Educational Psychologist, 38*(1), 23–31. https://doi.org/10.1207/S15326985EP3801_4

McLaren-Gradinaru, M., Burles, F., Protzner, A. B., & Iaria, G. (2023). The cognitive effects of playing video games with a navigational component. *Psychological Research, 87*(2), 536–549. https://doi.org/10.1007/s00426-022-01604-4

Mekler, E. D., Brühlmann, F., Tuch, A. N., & Opwis, K. (2017). Towards understanding the effects of individual gamification elements on intrinsic motivation and performance. *Computers in Human Behavior, 71*, 525–534. https://doi.org/10.1016/j.chb.2015.08.048

Meneghetti, C., Cardillo, R., Mammarella, I. C., Caviola, S., & Borella, E. (2017). The role of practice and strategy in mental rotation training: Transfer and maintenance effects. *Psychological Research, 81*(2), 351–358. https://doi.org/10.1007/s00426-016-0749-2

Myhill, D., & Brackley, M. (2004). Making connections? Teachers' use of children's prior knowledge in whole-class discourse. *Educational Review, 56*(3), 263–278.

https://www.researchgate.net/publication/229773701_Making_connections_Teachers'_use_of_children's_prior_knowledge_in_whole_class_discourse

OpenAI. (2023). *GPT-4 technical report* (Tech. Rep.). OpenAI. https://cdn.openai.com/papers/gpt-4.pdf

Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences, 20*(9), 676–688. https://doi.org/10.1016/j.tics.2016.07.002

Sanders, G. S., & Baron, R. S. (1975). The motivating effects of distraction on task performance. *Journal of Personality and Social Psychology, 32*(6), 956–963. https://doi.org/10.1037/0022-3514.32.6.956

Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science, 171*(3972), 701–703. https://doi.org/10.1126/science.171.3972.701

Lee, H.-P., Sarkar, A., Tankelevitch, L., Drosos, I., Rintel, S., Banks, R., & Wilson, N. (2025). The impact of generative AI on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*, 1–22. https://doi.org/10.1145/3706598.3713778

ScienceDaily. (2012, July 25). Spatial skills may be improved through training, including video games. *ScienceDaily.*https://www.sciencedaily.com/releases/2012/07/120725120634.htm

Uttal, D. H., Meadow, N. G., Tipton, E., Hand, L. L., Alden, A. R., Warren, C., & Newcombe, N. S. (2013). The malleability of spatial skills: A meta-analysis of training studies. *Psychological Bulletin, 139*(2), 352–402. https://doi.org/10.1037/a0028446

Van Rooij, A. J., Schoenmakers, T. M., Vermulst, A. A., van den Eijnden, R. J., & van de Mheen, D. (2011). Online video game addiction: Identification of addicted adolescent gamers. *Addiction, 106*(1), 205–212. https://doi.org/10.1111/j.1360-0443.2010.03104.x

Willis, S. L., Tennstedt, S. L., Marsiske, M., Ball, K., Elias, J., Koepke, K. M., Morris, J. N., Rebok, G. W., Unverzagt, F. W., & Wright, E. (2006). Long-term effects of cognitive training on everyday functional outcomes in older adults. *JAMA, 296*(23), 2805–2814. https://doi.org/10.1001/jama.296.23.2805