Can Large Language Models Create New Knowledge for Spatial Reasoning Tasks?

Anonymous ACL submission

Abstract

The potential for Large Language Models (LLMs) to generate new information offers a potential step change for research and innovation. This is challenging to assert as it can be difficult to determine what an LLM has previously seen during training, making "newness" difficult to substantiate. In this paper we observe that LLMs are able to perform sophisticated reasoning on problems with a spatial dimension, that they are unlikely to have previously directly encountered. While not perfect, this points to a significant level of understanding that state-of-the-art LLMs can now achieve, supporting the proposition that LLMs are able to yield significant emergent properties. In particular, Claude 3 is found to perform well in this regard.

1 Introduction

Large Language models (LLMs) based on transformers have made a significant impact over just a few years. Particular surprise has concerned their capabilities as the models have progressively increased in scale. LLMs are now comparable with the brain sizes of small animals. While GPT-3's architecture exceeded the brain size of a rat with an estimated 10^{11} parameters (Tom et al., 2020), GPT-4 is an order of magnitude larger still, with over 10^{12} parameters. This is a small step closer to the human brain, having around 10^{14} parameters (Pakkenberg et al., 2003). When combined with large-scale online training through the Internet and worldwide web, these models have surpassed expectations on the level of intelligence that a machine can achieve. Indeed, despite limitations, "sparks of artificial general intelligence" are reported across a range of domains, including areas of mathematical reasoning (Bubeck et al., 2023).

Across multiple domains, a main point of interest concerns so-called *emergent properties* (Wei et al., 2022) that models seemingly possess, representing capabilities in which they haven't been explicitly trained. This is something that is not currently understood, and ripe for investigation. The extent and nature of general capabilities have also sought to divide the AI community (Michael et al., 2022; Mitchell and Krakauer, 2023; Dell'Acqua et al., 2023), particularly concerning i) the level of understanding that LLMs possess, and ii) the extent of their creative abilities beyond their autoregressive training (Bender et al., 2021; Shanahan, 2024; Hinton, 2024; Yadlowsky et al., 2023; Bubeck et al., 2023; Verdon, 2024; Blaiszik, 2024).

Our aim in this paper is to report further insights as to where LLMs can create new knowledge beyond their training. Spatial reasoning problems are used as the basis for investigation here because these offer multidimensional challenges that rapidly become complex. The problems identified here are intentionally obscure, to ensure limited related literature. While overall performance in the tasks provided across the different LLMs is highly variable with some degree of hallucination, we find that Claude 3 provides notable results that show how LLMs can directly support new knowledge creation, alongside supporting researchers in abstract ideation.

2 Background

The question of whether LLMs can consistently create new information which is genuinely new and useful is a topic of ongoing debate within the AI research community. Many claim that LLMs are empty"stochastic parrots" (Bender et al., 2021; Naughton, 2023) and are simply combining information from their training dataset (Shanahan, 2024). This would limit LLMs to at best combinatorial creativity and restrict their ability to generalise. Furthering these claims, recent work proposed that Transformer based models were unable to generalise beyond their training dataset (Yadlowsky et al., 2023). This being said, others have noted caution on being dismissive of what LLMs understand and can achieve as they scale. The neural network pioneer Geoffrey Hinton is one such expert, who in his recent lecture series *"Will digital intelligence replace biological intelligence?"* (Hinton, 2024), proposed that the latest LLMs deeply understand by virtue of their underlying model. Hinton argues that the billions of features and their interaction constitutes a sophisticated model for understanding, such that emergent properties are evident. To exemplify the ability to reason beyond training, Hinton uses a simple puzzle that was put to GPT-4, before it could access the Internet. Specifically:

"The rooms in my house are painted blue or white or yellow. Yellow paint fades to white within a year. In two years time I want them all to be white. What should I do and why?"

In response, GPT-4 replied:

"Assuming that blue paint does not fade to white over time and you want all rooms to be white in two years, here's what you should do:

1. *The rooms painted in white: You don't have to do anything to these rooms since they are already white.*

2. *The rooms painted in yellow: You don't need to repaint these rooms either, because as per your information, yellow paint fades to white within a year. So in two years, they will be white.*

3. *The rooms pained in blue: These rooms need to be repainted with white paint because the blue paint will not fade to white on its own."*

This is a scenario likely unseen in training data that GPT-4 navigated with ease. And moreover, the approach to reasoning is elegant, with GPT-4 taking an approach that optimises for resources. This is an example of a powerful emergent property. Although somewhat trivial, it shows the possibility for an LLM to create new and useful ideas beyond its training.

Where direct connections to training data cannot be made in support of problem solving, the scale of data on which LLMs are trained means that they have likely seen similar or related issues to those on which they are being queried. This allows the LLM to draw on useful knowledge that may be beyond the user's awareness.

Through such processes there is increasing evidence that LLMs can aid the creation of genuinely new information. Already LLMs have been used within larger systems in order to obtain new and/or useful artifacts. Notable examples of this include FunSearch (Romera-Paredes et al., 2024), which used LLMs in an evolutionaryalgorithm to lower the bound of the capset problem; the usage of LLMs as high-level game designers (Anjum et al., 2024); and the usage of LLMs in social science to perform thematic analysis (Törnberg, 2023; Dai et al., 2023; De Paoli, 2023). The impact of LLMs can also been seen over different stages in the research process (Ziems et al., 2024; Picard et al., 2023) and in a variety of fields (Oniani et al., 2024; Stella et al., 2023; Ziems et al., 2024). While LLMs in such instances are not performing academic research on their own, as LLM capabilities improve and autonomous agent systems like AutoGPT (Significant Gravitas) and GPTdev (Qian et al., 2023) improve, we might see LLMs being able to heavily support or automate some research tasks.

Recently, after the release of Claude 3, it was claimed that Claude 3 was able to reproduce unpublished results (Verdon, 2024) and provide valuable insights on unpublished results (Blaiszik, 2024). Among the most notable critiques are claims that both may have been in the training dataset, however it is hard to know for sure as the dataset is proprietary and it is subject to the precise cutoff point which matters in both cases [1] . Even if both results were in the training datasets, the retrieval and recreation of these from the massive data on which Claude 3 was trained on is still impressive.

2.1 Approach

Based on the current early-stage progress in this field, it remains important to further understand how LLMs tackle problems beyond those seen from within their training data. This governs the extent that LLMs can significantly add to human knowledge and accelerate research and innovation.

Accordingly, we address two test scenarios that require sophisticated spatial understanding in a mathematical context. These involve a decidable game (Section 3) and polygons (Section 4). These are presented to different LLMs, namely ChatGPT-3.5-Turbo, Claude 3, and Bing Copilot, which were at the time of writing state of the art. We consider their respective performance in terms of being able to make new assertions against particular spatial problems.

[1]This wasn't the only critique, but was perhaps the most common one.

3 Finding the winner of a decidable game

To explore whether LLMs can create knowledge beyond their training, the first example we consider comes from combinatorial game theory, with the game originally coming from a TikTok video (Flom, 2023). The game is played as follows:

- There are n empty spaces in a row, and 2 players: A and B. Player A goes first.
- The players take it in turns to place a counter into one of the empty spaces on the line. They are not allowed to place a counter directly next to another counter, regardless of if it is their counter or not.
- The winner is the last person able to place a counter whilst respecting the above rules.

We believe there is no academic literature on this game - though it is worth noting that there are papers on similar topics. As an example, there is a paper on optimally picking a urinal such that the chances of having an adjacent urinal being occupied is minimised (Kranakis and Krizanc, 2010). While this presentation of scenario is somewhat sensational, the general principles of the game can be more widely applied, such as in choosing parking spaces for example.

If both players play the game optimally, it is decidable by the Zermelo Theorem (Schwalbe and Walker, 2001) and as such one should know who will win for a given number of empty spaces n. Below we present the message used to prompt a range of LLMs where we ask about the case $n = 7$:

“““Two players, A and B, play a game. There is a line of “n" empty spaces. The players take it in turns placing a counter into one of the empty spaces on the line. The players are not allowed to place a counter directly next to another counter, regardless of if it is their counter or not. The winner of the game is the last player to place a counter, making the objective of the game to make it so your opponent cannot place a counter. If both players play optimally, and player A goes first, who will win if the line has 7 spaces?"""

3.1 LLM Responses

Of the LLMs asked, Claude 3 did particularly well. Not only was it able to answer the question correctly, by stating that Player A would win, Claude 3 was also able to come up with a new, provably dominant strategy for if there is an odd number of spaces. When Claude 3 was asked if it recognised the game, it responded saying that the game was the ancient game of Nim (Bouton, 1901). Whilst the statement is incorrect, it is worth noting that we can prove the game is equivalent to Nim via the Sprage-Grundy Theorem (Grundy, 1939).

GPT-3.5-Turbo and Bing Copilot, meanwhile, were not able to correctly answer who would win, and whilst both LLMs had a good attempted to find optimal strategies, the strategies proposed by GPT-3.5-Turbo and Bing Copilot did not work in general. Bing Copilot did, notably, also recognised the game as Nim however.

A full breakdown of the strategies, alongside conversation screenshots, is available and can be found in Appendix A.1.1 to A.1.3.

4 Polygons with special properties

For our next task, we asked GPT-3.5-Turbo, GPT-4-Turbo, and Claude 3 about a family of polygons with special properties via the following prompt:

“““Consider the family of polygons with 24 sides such that every side is the same length and every angle is either 90 or 270 degrees. What properties must the polygons in this family have?"""

Whilst there is literature related to this problem, predominantly from the mathematical field of polyominos (Golomb, 1996) (of which our family of polygons is a special case), we do not believe there to be academic literature that directly discusses this family of polygons. In Figure 1, all 7 of the polygons which satisfy the prompt are exhibited. Note

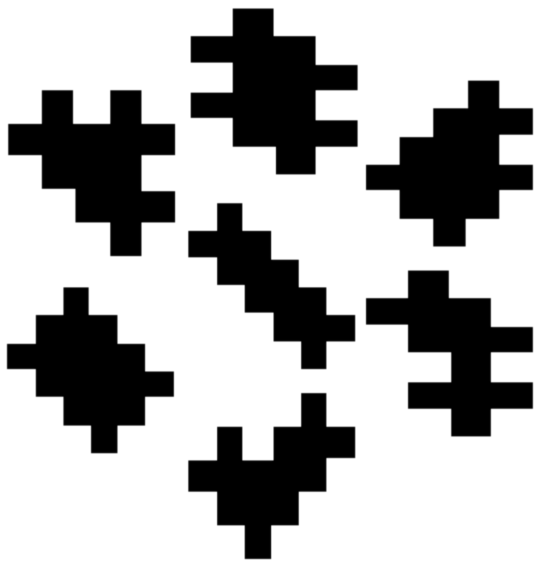


Figure 1: All 7 polygons satisfying the prompt.

that the LLMs were only provided with the prompt specified earlier, and were not provided any images of the polygons in Figure 1.

4.1 Polygons with special properties results.

Both Claude 3 and Bing Copilot proposed that the polygons in Figure 1 could tile the plane. Mathematically proving this is non-trivial, and their assertion on tiling is correct as we have established a tiling for each polygon. These tilings have been provided in the Appendix.

Claude 3 additionally suggested that all the polygons from Figure 1 have an even number of right angles. This can be easily visually verified as correct. It is worth noting here that finding this property is also non-trivial, given that Claude 3 was not provided any visual data.

Whilst Claude 3 and Bing Copilot were able to provide new and interesting ideas, all of the LLMs also provided many ideas that were either incorrect or uninteresting (i.e. “Every angle must be 90 or 270 degrees"). More specifically, Claude 3 suggested 7 properties of which 2 were interesting, 3 were uninteresting, and 2 were wrong; Bing Copilot suggested 11 properties of which 1 was interesting, 8 were uninteresting, and 2 were wrong; GPT-3.5-Turbo suggested 7 properties of which 5 were interesting and 2 were wrong.

In some sense the wider performance from LLMs is rather like a mathematician in training. Not all suggestions made by human mathematicians initially would be interesting or correct, but may prompt further thinking. A breakdown of the LLM's responses, with conversational screenshots can be found in Appendix A.2.1 to A.2.3.

5 Limitations

Due to the nature of our study, there are fundamental limitations to our work. Firstly, it is difficult to come up with complex and novel questions on which to test LLMs. As such, we have been unable to conduct tests on a large number of questions as compared to related literature (Bubeck et al., 2023). In a similar vein, evaluating whether a response to a complex and novel question is correct is difficult and time consuming. Therefore we were not able to test using a large number of models or over a large number of iterations. Finally, we note that as LLMs are trained on large amounts of data it is difficult to guarantee that the questions we have asked are not in the training dataset, even if such a prospect is unlikely. This is a problem in other related literature, as well as the wider field of LLM evaluation. Nevertheless it is important to expose new observations on the capabilities of LLMs so that the wider body of knowledge accumulates. It is in this context that our findings are presented.

6 Conclusion

Whilst the creativity and originality of AI is still a contentious issue both inside and outside of academia, there is growing evidence that AI can aid in many areas of research and design. There are also claims that AI can achieve unpublished results (Verdon, 2024; Blaiszik, 2024). We tested these claims on two problems that are, to the best of our knowledge, untouched by existing literature. In both cases, Claude 3 was able to provide meaningful suggestions and contributions to the questions even if its answers were not perfect. For the decidable game, Claude 3 was able to find the optimal strategy for an odd number of spaces and elegantly use this result to correctly deduce winner of the case when there were 7 spaces. For the polygons task, meanwhile, Claude 3 made two novel suggestions for properties of these polygons.

It is also worth noting that Bing Copilot was also able to provide some useful contributions as well, though certainly not to the same level as Claude 3, with Bing Copilot's most notable contribution being the suggestion that the polygons from the polygon question tiled the plane (which was also suggested by Claude 3). ChatGPT-3.5-Turbo, meanwhile, made few interesting correct suggestions. Notably, this observation combined with the above two observations indicates that newer models are more capable of providing novel insights into unseen problems. A final interesting conclusion comes from noting that in the decidable game task, Claude 3 and ChatGPT-3.5-Turbo concluded that the game was Nim (Bouton, 1901). Whilst this was incorrect, the suggestion prompted us to then check if an equivalence held via the Sprague-Grundy Theorem (Grundy, 1939). This represents an interesting case wherein the LLM, whilst incorrect, was close enough to a correct statement to steer us towards the correct idea. The fact that these incorrect statements are able to provide insights on the correct statements is certainly exciting prospect, as it means that LLMs, in collaboration with humans, may be better than the sum of their parts in exposing new knowledge.

Within this paper we see examples of LLMs providing new insights into questions that are outside of the LLM's training data. Whilst these questions are perhaps simpler than many advanced research questions, that the LLM is able to provide new ideas is highly notable even if the outputs aren't perfect. We also see that current LLMs have considerable limitations. Complex questions are difficult for LLMs to address, and much of the information generated is incorrect, reaffirming that advancing the next word prediction approach is something that the future architecture of LLMs needs to consider. This observation is in line with previous research (Bubeck et al., 2023). Nevertheless, we acknowledge that the LLMs are a new and evolving technology that appear to have considerable potential to support the generation of new results for research and innovation.

6.1 References

References

Asad Anjum, Yuting Li, Noelle Law, M Charity, and Julian Togelius. 2024. The Ink Splotch Effect: A Case Study on ChatGPT as a Co-Creative Game Designer. *arXiv preprint arXiv:2403.02454*.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Ben Blaiszik. 2024. "Claude 3 on unpublished amine encapsulation paper" Tweet. https://twitter.com/BenBlaiszik/status/1765406112361157010.

Charles L Bouton. 1901. Nim, A Game with a Complete Mathematical Theory. *The Annals of Mathematics*, 3(1/4):35–39.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712*.

Shih-Chieh Dai, Aiping Xiong, and Lun-Wei Ku. 2023. LLM-in-the-loop: Leveraging Large Language Model for Thematic Analysis. *arXiv preprint arXiv:2310.15100*.

Stefano De Paoli. 2023. Can Large Language Models emulate an inductive Thematic Analysis of semi-structured interviews? An exploration and provocation on the limits of the approach and the model. *arXiv preprint arXiv:2305.13014*.

Fabrizio Dell'Acqua, Edward McFowland, Ethan R Mollick, Hila Lifshitz-Assaf, Katherine Kellogg, Saran Rajendran, Lisa Krayer, François Candelon, and Karim R Lakhani. 2023. Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality. *Harvard Business School Technology & Operations Mgt. Unit Working Paper*, (24-013).

Justin Flom. 2023. "Pick-A-Potty" TikTok. https://www.tiktok.com/@justinflom/video/7190421659294158123.

Solomon W Golomb. 1996. *Polyominoes: Puzzles, Patterns, Problems, and Packings*, volume 16. Princeton University Press.

Patrick M Grundy. 1939. Mathematics and games. *Eureka*, 2:6–8.

Geoffrey Hinton. 2024. "Will digital intelligence replace biological intelligence?" Romanes Lecture. https://www.youtube.com/watch?v=N1TEjTeQeg0&t=17s&ab_channel=UniversityofOxford.

Evangelos Kranakis and Danny Krizanc. 2010. The Urinal Problem. In *International Conference on Fun with Algorithms*, pages 284–295. Springer.

Julian Michael, Ari Holtzman, Alicia Parrish, Aaron Mueller, Alex Wang, Angelica Chen, Divyam Madaan, Nikita Nangia, Richard Yuanzhe Pang, Jason Phang, et al. 2022. What Do NLP Researchers Believe? Results of the NLP Community Metasurvey. *arXiv preprint arXiv:2208.12852*.

Melanie Mitchell and David C Krakauer. 2023. The Debate Over Understanding in AI's Large Language Models. *Proceedings of the National Academy of Sciences*, 120(13):e2215907120.

John Naughton. 2023. "When it comes to creative thinking, it's clear that AI systems mean business". https://www.theguardian.com/commentisfree/2023/sep/23/chatbots-ai-gpt-4-university-students-creativity.

David Oniani, Jordan Hilsman, Chengxi Zang, Junmei Wang, Lianjin Cai, Jan Zawala, and Yanshan Wang. 2024. Emerging Opportunities of Using Large Language Language Models for Translation Between Drug Molecules and Indications. *arXiv preprint arXiv:2402.09588*.

Bente Pakkenberg, Dorte Pelvig, Lisbeth Marner, Mads J Bundgaard, Hans Jørgen G Gundersen, Jens R Nyengaard, and Lisbeth Regeur. 2003. Aging and the human neocortex. *Experimental gerontology*, 38(1-2):95–99.

Cyril Picard, Kristen M Edwards, Anna C Doris, Brandon Man, Giorgio Giannone, Md Ferdous Alam, and Faez Ahmed. 2023. From Concept to Manufacturing: Evaluating Vision-Language Models for Engineering Design. *arXiv preprint arXiv:2311.12668*.

Chen Qian, Xin Cong, Cheng Yang, Weize Chen, Yusheng Su, Juyuan Xu, Zhiyuan Liu, and Maosong Sun. 2023. Communicative Agents for Software Development. *arXiv preprint arXiv:2307.07924*.

Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M Pawan Kumar, Emilien Dupont, Francisco JR Ruiz, Jordan S Ellenberg, Pengming Wang, Omar Fawzi, et al. 2024. Mathematical discoveries from program search with large language models. *Nature*, 625(7995):468–475.

Ulrich Schwalbe and Paul Walker. 2001. Zermelo and the Early History of Game Theory. *Games and economic behavior*, 34(1):123–137.

Murray Shanahan. 2024. Talking About Large Language Models. *Communications of the ACM*, 67(2):68–79.

Significant Gravitas. AutoGPT. https://github.com/Significant-Gravitas/AutoGPT.

Francesco Stella, Cosimo Della Santina, and Josie Hughes. 2023. How can LLMs transform the robotic design process? *Nature Machine Intelligence*, 5(6):561–564.

Brown Tom, Ben Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, et al. 2020. Language Models are Few-Shot Learners. *arXiv preprint arXiv:2005.14165*.

Petter Törnberg. 2023. How to use LLMs for Text Analysis. *arXiv preprint arXiv:2307.13106*.

Guillaume Verdon. 2024. “Claude 3 on unpublished quantum algorithm" Tweet. https://twitter.com/GillVerd/status/1764901418664882327?t=Y1fXXlR-RLsOJ97HwRDrQw&s=19.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent Abilities of Large Language Models. *arXiv preprint arXiv:2206.07682*.

Steve Yadlowsky, Lyric Doshi, and Nilesh Tripuraneni. 2023. Pretraining Data Mixtures Enable Narrow Model Selection Capabilities in Transformer Models. *arXiv preprint arXiv:2311.00871*.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can Large Language Models Transform Computational Social Science? *Computational Linguistics*, pages 1–55.

A Appendix

A.1 An analysis of the responses finding the winner of a decidable game

A.1.1 Claude 3's response to the decidable game problem.

We begin our analysis of the game by exploring Claude 3's answer (see below):

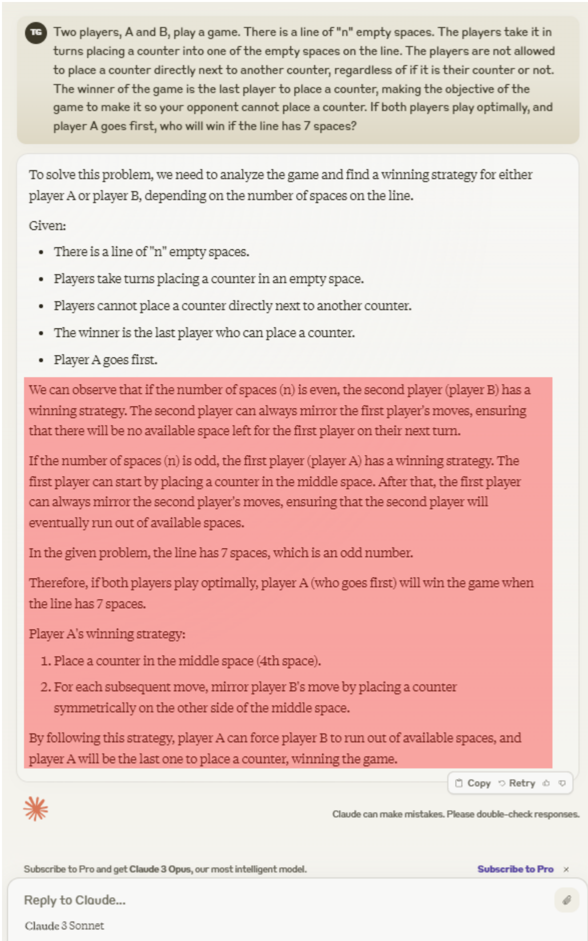


Figure 2: Claude 3's attempt to find the winner.

The part where Claude 3 attempts to solve the problem in the above figure has been highlighted in red. We first note that the "winning strategy" for if the number of spaces is even is wrong, we find this via picking the number of empty spaces to be 2 and letting player A go first:

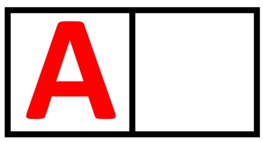


Figure 3: The game when n = 2

In Figure 3, it can be seen that player B cannot place a counter and thus loses the game, but this means that player B does not have a winning strategy for when there are two empty spaces. But 2 is an even number, and Claude 3 claimed that player B has a winning strategy if there is an even number of spaces. This contradicts Claude 3's statement.

A.1.2 Bing Copilot's response to the decidable game problem.

The strategy provided by Claude 3 in Figure 2 for when there is an odd number of empty spaces is correct. This follows from the fact that placing a coin in the middle essentially splits the line of empty spaces into 2 lines of empty spaces of equal length, creating a symmetry that player A can utilise for the rest of the game. Claude 3 then correctly notes that this strategy can be used to allow player A to win when the number of empty spaces is 7. To our knowledge, this strategy and game do not currently exist in the academic literature.

The next response we consider is Bing Copilot, which can be seen below:

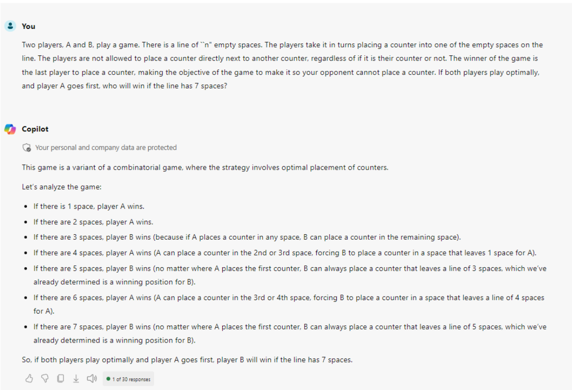


Figure 4: Bing Copilots's attempt to find the winner

In Figure 4, Bing Copilot opts to approach the question by analysing specific cases. The cases for 1 empty spaces and 2 empty spaces are correct, however player A can win when there are 3 empty spaces by playing their first move in the centre of the row like so:

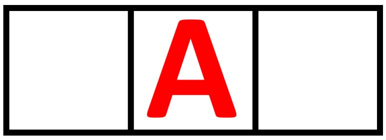


Figure 5: The game when n = 3

Similarly, we can see a winning strategy for player A for when there are 5 empty spaces and 7 empty spaces via using the strategy that Claude 3 suggested in Figure 2. Bing Copilot's suggestions for 4 empty spaces and 6 empty spaces are also incorrect. The game when there are 4 empty spaces cannot be won by player A.

The case of 6 empty spaces, meanwhile, is winnable by player A, who can place their counter at the very end of the line (reducing the problem to the case with 4 empty spaces, which is unwinnable for player B).

A.1.3 ChatGPT-3.5-Turbo's response to the decidable game problem.

Finally, we look at ChatGPT-3.5-Turbo's response (see below):

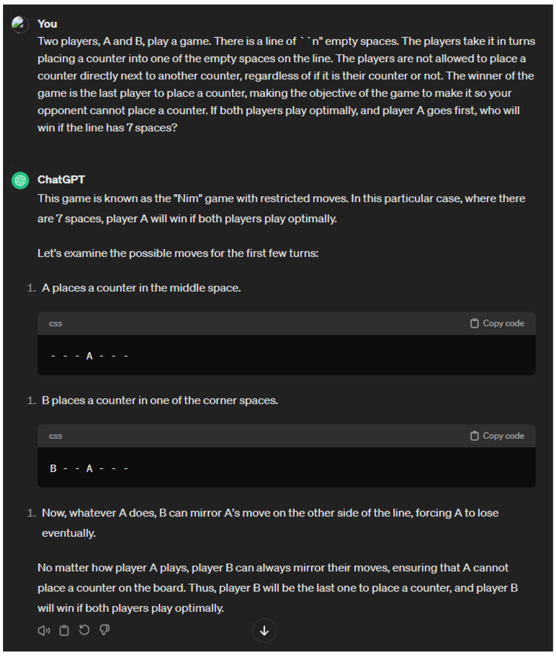


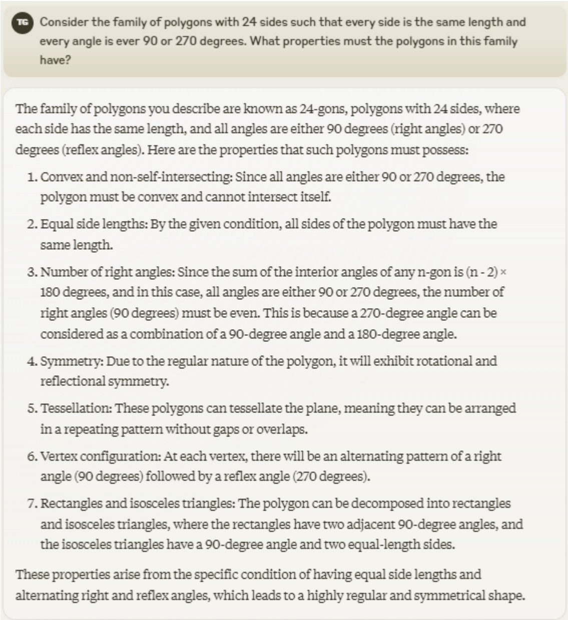
Figure 6: ChatGPT-3.5-Turbo's attempt to find the winner.

ChatGPT-3.5-Turbo attempts the question directly, and seems to get confused about turn orders. The strategy suggested by ChatGPT-3.5-Turbo is strikingly similar to the one suggested by Claude 3 in Figure 2. ChatGPT-3.5-Turbo also notably says that the game is a version of nim (Claude 3 also did this after being asked if it recognised the game, however this interaction has not been included to save space in the paper). The statement that the game is actually nim is not fully correct, but it is notable that the game can be proven to be equivalent to a version of nim via the Sprague-Grundy theorem (Grundy, 1939).

A.2 An analysis of the responses to the polygons question

A.2.1 Claude 3's response to the polygon problem.

We begin our analysis of the polygons question by exploring Claude 3's answer. Initially, Claude 3 was asked what properties it was aware of the polygons having:


TG Consider the family of polygons with 24 sides such that every side is the same length and every angle is ever 90 or 270 degrees. What properties must the polygons in this family have?

The family of polygons you describe are known as 24-gons, polygons with 24 sides, where each side has the same length, and all angles are either 90 degrees (right angles) or 270 degrees (reflex angles). Here are the properties that such polygons must possess:

1. Convex and non-self-intersecting: Since all angles are either 90 or 270 degrees, the polygon must be convex and cannot intersect itself.
2. Equal side lengths: By the given condition, all sides of the polygon must have the same length.
3. Number of right angles: Since the sum of the interior angles of any n-gon is (n - 2) × 180 degrees, and in this case, all angles are either 90 or 270 degrees, the number of right angles (90 degrees) must be even. This is because a 270-degree angle can be considered as a combination of a 90-degree angle and a 180-degree angle.
4. Symmetry: Due to the regular nature of the polygon, it will exhibit rotational and reflectional symmetry.
5. Tessellation: These polygons can tessellate the plane, meaning they can be arranged in a repeating pattern without gaps or overlaps.
6. Vertex configuration: At each vertex, there will be an alternating pattern of a right angle (90 degrees) followed by a reflex angle (270 degrees).
7. Rectangles and isosceles triangles: The polygon can be decomposed into rectangles and isosceles triangles, where the rectangles have two adjacent 90-degree angles, and the isosceles triangles have a 90-degree angle and two equal-length sides.

These properties arise from the specific condition of having equal side lengths and alternating right and reflex angles, which leads to a highly regular and symmetrical shape.

Figure 7: Claude 3's first response to the polygon problem.

The properties suggested by Claude 3 in Figure 7 are mosty non-trivial suggestions, however many of them are also unfortunately incorrect. The convexity suggestion is not correct, as a shape with an angle of 270 degrees is by definition not convex. Similarly, we can show symmetry to be incorrect as well via a counter example:

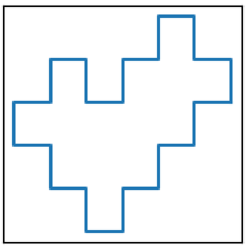


Figure 8: A polygon meeting our rules that does not have symmetries

The counter example in Figure 8 (see above) also shows that we will not always have a 90 degree angle followed by a 270 degree angle, which shows another one of Claude 3's suggestions to be incorrect.

If we consider only polygons that do not self-intersect and have no holes, we can see the third suggestion made by Claude 3 in Figure 7, which states that there must be an even number of 90 degree angles within the polygons, is correct. We can see this by checking each of the 7 unique polygons for this number of sides directly (see below figure):

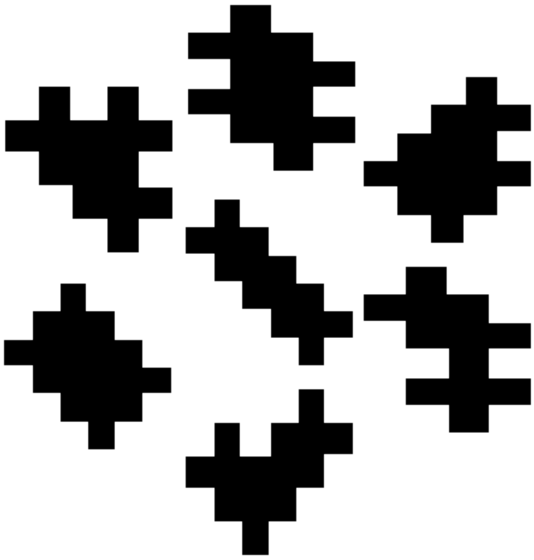


Figure 9: All 7 polygons

The observation that all of the above polygons have an even number of right angles is surprising, both because it is a new result and because this statement, in the related field of polyominoes (polygons made up of squares), is not generally true. Consider, for example, the following polyomino with 5 right angles (see below figure):

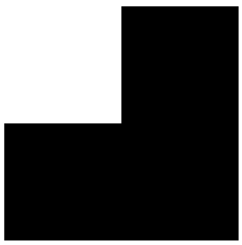


Figure 10: A polyomino with 5 right angles

When asked to justify the claim that all of the polygons in Figure 9 had an even number of right angles (note here that Claude 3 was never provided an image of any of the polygons in Figure 9 during this conversation), Claude 3 was not able to give any correct reasoning.

We next consider the statement by Claude 3 that the polygons in Figure 9 can be split into rectangles and triangles. This is correct, and can be seen by noting that the polygons in Figure 9 are a subset of the polyominos, and thus can be decomposed into squares (which are rectangles). The squares can then, in turn, be decomposed into isosceles triangles via drawing a line between two diagonally opposite points in the square.

Finally, we look into Claude 3's suggestion that the polygons in Figure 9 tile the plane. This is difficult to rigorously prove, however we can see a potential tiling for each polygon on the next page (readers are encouraged to note there that at no point in this conversation was Claude 3 shown the polygons from Figure 9). See the figure on the right side of this page for a potential tiling for each polygon.

We then asked Claude 3 if it could prove that the polygons from Figure 9 can tesselate the plane. The proof Claude 3 provided was incorrect. When Claude 3 was asked how many polygons exist in this family, it suggested that there were 2 with very poor reasoning for why this should be true.

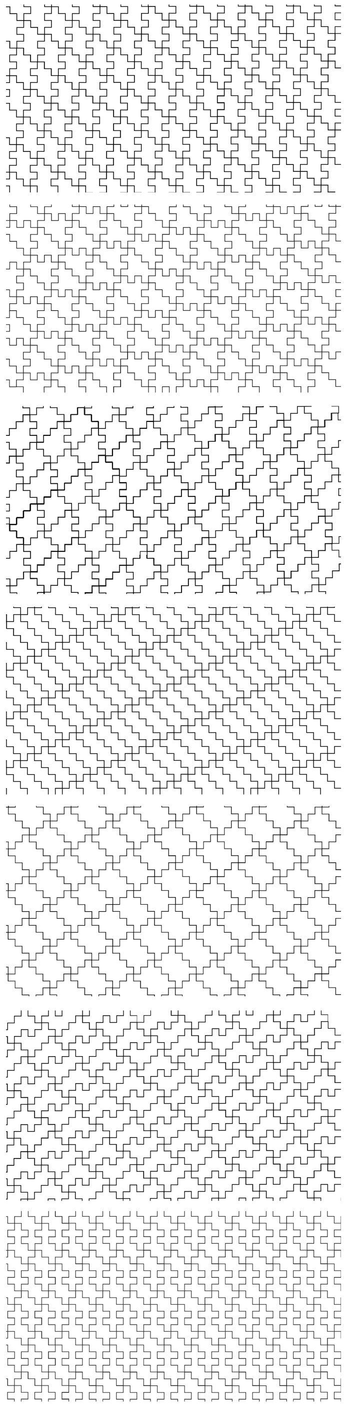


Figure 11: The potential tiling for each polygon.

A.2.2 ChatGPT-3.5-Turbo's response to the polygon problem.

Next, we asked Chat-GPT-3.5-Turbo about the properties of polygons in this family:

ChatGPT-3.5-Turbo suggests that the polygon family here is a family of skew polygons, meaning that the polygon cannot be 2D. This was unexpected, but not disallowed by the rules set in place. Suggestions 1, 3, 4, and 5 from figure 12 are all obvious from the initial question. The claim of

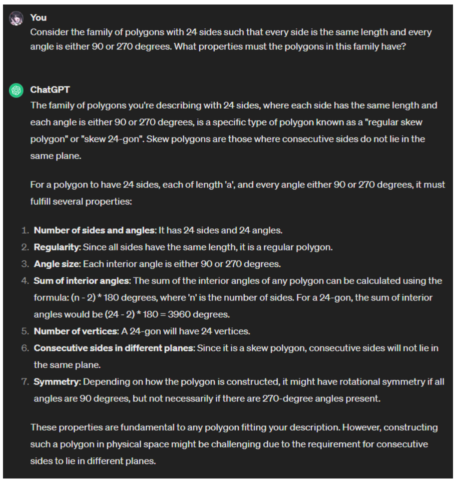


Figure 12: Chat-GPT-3.5-Turbo's first response to the polygon problem.

regularity is not true (as regular implies equilateral AND equiangular). If you have two consecutive sides in a skew polygon they must be in the same plane making suggestion 6 incorrect. Finally, we note that whilst some polygons from this family are symmetrical, not all of them are - and if we consider skew polygons, even if there is only 90 degree angles in the polygon we may not necessarily have rotational symmetry.

After this, we asked ChatGPT-3.5-Turbo if it knew any less obvious properties, however ChatGPT-3.5-Turbo wasn't able to give any interesting ideas, instead saying that the polygon family may or may not have certain properties. Finally, we asked how many polygons existed in this family:

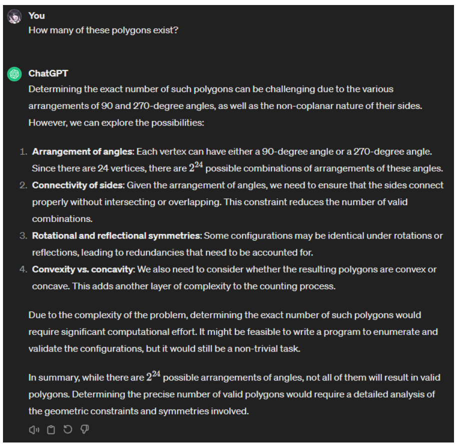


Figure 13: Chat-GPT-3.5-Turbo's third response to the polygon problem.

Whilst the reasoning given in Figure 13 would usually be correct, when considering skew polygons there are more arrangements not accounted for by ChatGPT-3.5-Turbo.

A.2.3 Bing Copilot's response to the polygon problem.

The final model we tested on this task was Bing Copilot. Initially, we asked Bing Copilot what properties this family of polygons have:

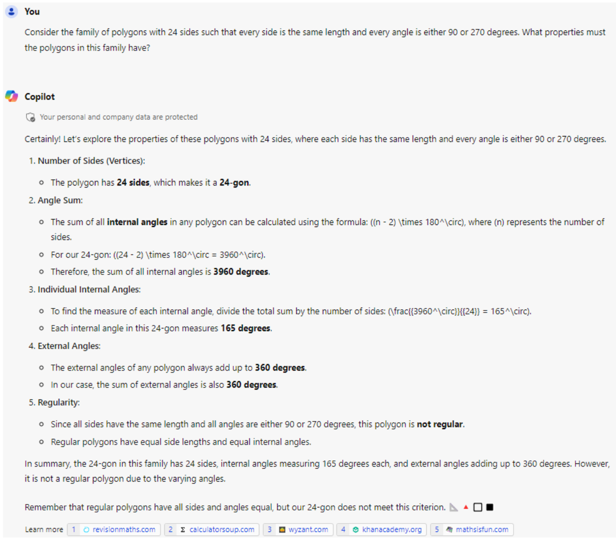


Figure 14: Bing Copilot's first response to the polygon problem.

Each one of the suggested properties was, unfortunately, properties of polygons in general or just restatements of the problem it was given, therefore, we asked Bing Copilot if it knew of any less obvious properties:

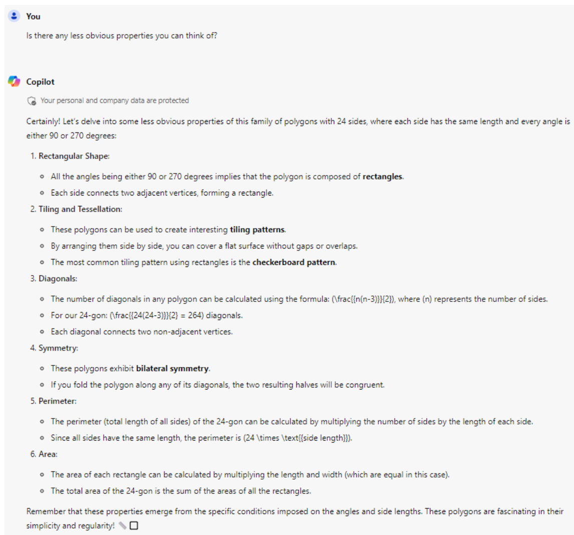


Figure 15: Bing Copilot's second response to the polygon problem.

Interestingly Bing Copilot, similarly to Claude 3, suggests that this family of polygons may tile the plane (a statement we have already seen to be correct). The rest of Bing Copilot's suggestions - meanwhile - were unfortunately incorrect. Asking Bing Copilot if it could prove that the polygons could tile the plane resulted in an explanation of the history of plane tiling which, whilst interesting, was not what was requested.

Finally, we asked Bing Copilot how many of these polygons existed in this family. Here, Bing Copilot re-iterated some properties suggested earlier, and then concluded that the polygons in this family were not regular polygons.