Laboratorija za eksperimentalnu psihologiju,
Odeljenje za psihologiju,
Filozofski fakultet Univerziteta u Beogradu

Understanding Space by Large Language Models

Dražen Domijan[1]

Enio Ibrišagić[1]

[1] Faculty of Humanities and Social Sciences,

University of Rijeka, Rijeka, Croatia

drazen.domijan@uniri.hr

Large language models (LLMs) have recently attracted significant interest and controversy due to their potential to perform various tasks at a level comparable to humans, leading some authors to claim that they are achieving artificial general intelligence (Sharkar, 2023). However, several recent studies have found that LLMs tend to produce non-human errors in language comprehension tests (Dentella et al., 2024; Murphy et al., 2025). In this study, we tested the ability of LLMs to reason about space and spatial relations using a maze-solving task as a benchmark. The task required finding a path from point X to point Y without crossing the walls of the maze and drawing it on a supplied image. We tested Grok, Gemini, ChatGPT-4o, and ChatGPT-5.2 on 20 problems (10 with a path from X to Y and 10 without one). Each problem was presented five times in random order, resulting in a total of 100 trials. We found that only ChatGPT-5.2 was able to solve the task and draw the path, if it existed. The others made errors such as crossing the walls, providing a verbal rather than a visual response, or inventing a new maze (a form of visual hallucination). We also asked the LLMs to provide metacognitive judgements about their performance. Interestingly, ChatGPT-5.2 typically gave lower confidence estimates (around 95%) than the others, which were 100% confident they had solved the task. We further tested ChatGPT-5.2's spatial abilities by devising new tests involving contour tracing and inside-out relations. Again, ChatGPT-5.2 was able to find a solution without making errors. We interpret this finding as suggesting that ChatGPT-5.2 is not a pure LLM but rather integrates an LLM with other techniques from classical AI, such as breadth-first search and the A* algorithm, which have been developed to find a path on abstract representations such as graphs.