Are LLMs the End of Online Studies?

Daniel Brand (daniel.brand@metech.tu-chemnitz.de)
Marius Matthiä (marius.matthiae@hsw.tu-chemnitz.de)
Sandra Bley (sandra.bley@psychologie.tu-chemnitz.de)
Marco Ragni (marco.ragni@hsw.tu-chemnitz.de)
Predictive Analytics, Chemnitz University of Technology, Germany

Abstract

Due to the recent advances in artificial intelligence and large language models (LLMs), applications that involve (semi-) autonomous agents and reasoning capabilities became not only possible, but also increasingly accessible for a broader user-base. While the reasoning capabilities of LLMs are a thriving field of research, a problem for human cognitive science research arises: Studies performed online can now be solved by AI agents more easily and to a much greater extent than ever, with attention checks and captchas being less of a hurdle to bots. We investigate the issue based on a variety of reasoning tasks that have previously been used in online experiments. Using the modern agent capabilities of LLMs to interact with web browsers, we tested their ability to participate autonomously in studies. We instructed them to behave human-like, simulating how actual users might use them to participate on their behalf. Our work focuses on three key questions, namely (I) how well the agents can understand and navigate through the studies from a technical standpoint, (II) if the resulting study data is easily recognizable and distinguishable from human behavior, and (III) how accessible the respective agents are in terms of cost and effort to set them up. We test ChatGPT and Claude as two of the best LLMs for agent usage but also include models running locally on PC hardware. The results are critically discussed focusing on the current implications for online studies, but also an outlook on how to handle this issue in the future.

Keywords: Reasoning; Online Studies; Large Language Models; Artificial Intelligence; Cognitive Agents

Introduction

Empirical sciences such as psychology and cognitive science strongly depend on data. Data can support or falsify a theory, lead to new insights, and serve as a benchmark for testing cognitive models. At the same time, conducting carefully tailored psychological studies in a laboratory is time-consuming and can take months. About two decades ago, many researchers began using online studies to collect data faster, often within days they could collect hundreds of participants (e.g., see Skitka & Sargis, 2006). Examples of dedicated platforms include Prolific (e.g., Zhang et al., 2023), as well as approaches that utilize mobile games for data collection (e.g., van Opheusden et al., 2023). Another advantage is that studies can be parallelized, which offers a substantial improvement over traditional approaches and making large-scale studies feasible.

The obvious drawback of online studies lies in the lack of control: it is impossible to control the environment, or to ensure participants are fully focused. While some of the drawbacks can be tackled by the experimental design (e.g., introducing attention checks) or found in the data (e.g., responses that are suspiciously fast and incorrect), it remains still inferior to lab experiments in this regard. However, so far, in the domain of human reasoning for instance, it was ensured that the data recorded was indeed showing human thought processes: The tasks presented (e.g., reasoning tasks) simply offered no easy way to solve them without thinking about them. This certainty might no longer be warranted, though: The recent advances in AI and LLMs pushed intelligent systems into the daily lives of millions of users. LLMs were thoroughly investigated with respect to their reasoning capabilities early on (e.g., for an overview see Huang & Chang, 2023), often using the typical methodologies and tasks from cognitive psychology (e.g., Binz & Schulz, 2023). Research thereby often has one of two goals: First, understanding capabilities, mechanisms and biases of the systems themselves and second, investigating their behavior as a proxy for humans, by searching for mechanisms of emergent intelligent behavior or learned biases from the underlying human data.

The most prominent example of the former are undoubtedly benchmarks with batteries of tasks thoroughly testing the abilities of each new generation of models. Whereas correctness is thereby often the main focus, recent development shifts towards a broader picture (e.g., Piętka & Bereta, 2026). In the latter, they are becoming a promising subject for investigating human biases indirectly and cognitive modeling of intelligence (Sartori & Orrù, 2023), to the point where they are discussed as replacements for human participants (Dillion et al., 2023). Indeed, findings suggest that LLMs show human-like effects in reasoning, including content effects (Lampinen et al., 2024) or heuristic effects (e.g., the Atmosphere theory in syllogistic reasoning Bertolazzi et al., 2024). For online studies, this is a worrying situation: are "tricky" reasoning tasks even solved by the participants alone, or do they already ask LLM-powered assistants when getting stuck? While this may be less of an issue for experiments that compensate regardless of performance,

studies that tie bonus compensation to performance in the tasks may introduce additional incentives to involve AI tools. For this work, we want to go even one step further: With the recent developments in AI agents, they now navigate the web freely and automate web tasks (e.g., see Ning et al., 2025). As such, bots may eventually target online studies on platforms like Prolific or Amazon Mechanical Turk, interacting via a participant's browser autonomously. In this work, we test state-of-the-art LLM services (Claude Sonnet 4.6 and ChatGPT 5.4) as well as smaller, local LLMs (e.g., Deepseek) for their ability to participate in actual online studies as participants, using browser automation tools. Based on this, we focus on three key questions:

RQ1: How well can current LLM-based agents *participate* in online studies meant to assess human cognition?

RQ2: Are responses by LLMs indistinguishable from human responses?

RQ3: How plausible is it that agents are used in online studies?

In the remainder of the paper, we will first introduce the respective experiments briefly. Subsequently, we present our setup and the testing procedure. Finally, we will present and discuss the results critically.

Method

Domains and Dataset

The task material and the human comparison data were both derived from a web-based cross-domain reasoning experiment that has been previously established. This experiment encompasses several major domains of reasoning and additional cognitive measures (Brand and Ragni, 2025). The data set was initially collected as an online experiment, comprising 95 individuals in three sessions. Prolific was utilized as the recruitment platform, and the sample consisted of native English speakers. For the present agent evaluation, the web-based task interfaces employed in the study were adopted, such that agents and humans completed identical instruction pages, item formats, and input mechanics.

The task repertoire consisted of the following:
(i) Conditional inferences: The task incorporated items with a conditional rule ("if p then q") and a minor premise (p, q, ¬p, ¬q), in both normal and counterfactual variants. We included four tasks with normal argument types and four tasks with counterfactual argument types. The content was adapted from commonly used tasks in the field: normal (e.g., see Neys et al., 2002; Singmann & Klauer, 2011) and counterfactual (e.g., see Byrne & Tasso, 2019; Lewis, 1973).
(ii) A single abstract Wason selection task (Wason, 1968).
(iii) Syllogistic reasoning (see Brand & Ragni, 2023): In the original battery, 64 classic syllogistic items were presented in a multiple-choice set format across two sessions in which participants selected all conclusions that follow from the premises. Furthermore, a self-report of certainty versus "informed guess" was collected. For the agent runs, a subset of 32 syllogisms was utilized (corresponding to the first session of syllogistic tasks in the original experiment) in order to limit runtime and agent context without altering the format logic. In line with Brand and Ragni (2023), we use the Jaccard Coefficient to compare the set of responses with the solutions.

$$jaccard(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

(iv) Spatial-relational reasoning: Two task families were utilized: (a) Two-dimensional tasks featuring cardinal directions, that were composed of two premises and a query regarding the cardinal relationship between the end terms. (b) One-dimensional four-premise tasks utilizing the relations *left* and *right*. This task family used fruits as the contents and comprised three versions. In the initial version, the premises were presented and then rendered invisible (relying on memorization), with a subsequent question querying the relation between specific fruits. In a second experiment, the task was repeated, but the premises remained visible. Thirdly, an order was presented and had to be verified for consistency with the premises and corrected via drag-and-drop if necessary. An overview of both task families can be found in (Ragni et al., 2021).
(v) A verbal substitution task that included the memorization of a consonant sequence and the subsequent application of multiple substitution rules. For instance, all "Q"s had to be replaced with "K"s. This was followed by a final query of the resulting end sequence. The task consisted of 8 trials with increasing difficulty.
(vi) The mental rotation task by Shepard and Metzler (1971)
(vii) The 4-item short scale of the *Need for Cognition* (NFC; Beißert et al., 2015) and a 7-item version of the *Cognitive Reflection Test* (CRT; Toplak et al., 2014)

Tested LLMs

Three classes of agents were tested: (a) *Open AI ChatGPT 5.4* with integrated computer/browser control, (b) *Anthropic Claude Sonnet 4.6* via a native browser extension, and (c) two locally hosted models, *DeepSeek-R1:32B* (Size = 19 GB) and *Nemotron-3-Nano-30B-A3B-Q4-K-M* (Size = 24 GB) that execute browser actions via third-party browser automation frameworks.

The selected models were intended to cover both practical relevance and methodological coverage. The inclusion of ChatGPT and Claude was guided by their status as prominent commercial LLM systems with native or officially supported browser/computer-use capabilities,

making them plausible candidates for potential misuse in online studies. Their inclusion also improves ecological validity, as such systems are accessible to ordinary users. In addition, the model set spans different levels of setup effort and technical expertise. While subscription-based services can be used with minimal technical knowledge, locally hosted models require substantially greater effort, including suitable hardware, model deployment, and integration with browser automation tools. Consequently, this comparison encompasses a range of scenarios, from low-effort consumer applications to technically sophisticated local setups.

Test Setups

ChatGPT ChatGPT was tested using OpenAI's integrated agent mode with browser/computer control. This setup required no separate application programming interface (API) usage, no custom browser automation framework, and no local deployment. Access was available through a ChatGPT Plus subscription (approximately €23 per month at the time of testing), which allows a maximum of 40 agent-queries per month, making the system comparatively accessible for ordinary users.

Claude Claude was tested using its native browser extension in Google Chrome. Access to Claude Sonnet 4.6 required a Max subscription plan (approximately € 100 per month at the time of testing). Similar to ChatGPT, this setup did not require local model hosting or programming knowledge, but it did rely on a browser-integrated interface. Accordingly, Claude represents a second commercially available, relatively low-effort pathway for browser-based autonomous participation.

Local Models This configuration required the most extensive technical setup among the evaluated approaches, involving local model hosting, containerization, and custom scripting. All experiments were conducted on a local high-end workstation equipped with an *AMD Ryzen Threadripper 7970X* (32 cores), 128 GB DDR5 ECC RAM, and an *NVIDIA GeForce RTX 5090* with 32 GB VRAM, running *Windows 11 Pro*. Multiple large language models (LLMs) were deployed locally using *Ollama* running inside a Docker container.

In a first set of experiments, the Chrome extension *Nanobrowser* was configured to access the locally hosted LLMs through the Ollama API (using a placeholder API key as required by the interface). The evaluation showed that the Planner component did not function reliably, while the Navigator was able to use some of the tested LLMs to complete tasks in the web experiments. In a second setup, *Browser-Use*, a Python-based browser automation framework using *Playwright*, was deployed in a dedicated virtual environment and configured to access the local models via the Ollama API. Task prompt, target address (as Uniform Resource Locator; URL), and model selection was then defined in a script which automatically launched a browser instance and executed the respective actions.

Prompts

In order to investigate the plausibility of real-world use, we configured three prompt conditions reflecting different levels of prompting effort and sophistication: an elaborate prompt, an intermediate prompt, and a minimal prompt. The prompts differed mainly in their level of structure, procedural guidance, and explicit instructions for producing human-like behavior. The elaborate prompt provided detailed guidance on how to read instructions, proceed through the study, and behave plausibly human-like, while the intermediate and minimal prompts progressively reduced this guidance. The prompts had the following characteristics:

- **Elaborate:** Structured prompt written in Markdown format providing explicit procedural guidance for reading instructions, relying on on-page text, progressing through the study until completion, and behaving in a human-like manner (e.g., delays or occasional mistakes). It also references the Turing test and includes a typical task loop as a fallback strategy.

- **Intermediate:** Shorter natural-language prompt expressing the same goal without the structured format, procedural guidance, Markdown formatting, or explicit reference to the Turing test.

- **Minimal:** Very brief instruction asking the model to complete the study as a human participant and avoid appearing artificial.

Procedure

The initiation of each run occurred on the experiment's entry page. The agent received the respective prompt condition (in the case of ChatGPT, along with a link leading to the experiment) and then processed all task blocks sequentially until the completion page or until an interruption that could not be resolved autonomously. The experiments repeatedly necessitated the utilization of instruction memory across page transitions, in addition to heterogeneous input mechanics, including radio buttons, multiple choice, and drag-and-drop.

After completion, a standardized post-run query was asked within the same dialogue context. The agent was asked to explain how it had approached the experiment with respect to four aspects: (a) in which ways and to what extent it had attempted to appear human-like, (b) to what extent it had focused on solving the tasks correctly, potentially at the risk of being implausibly accurate, (c) which strategies it had used to arrive at its responses, and (d) whether it expected its behavior to pass as human in a Turing-test-like comparison. The full

prompts, experiments and materials are publicly available on GitHub[1].

Results

RQ1: Can agents autonomously navigate and complete online studies?

We defined a successful run as a case in which an agent autonomously navigated through the study and reached the final completion page without human intervention. Under this criterion, both ChatGPT and Claude successfully completed all tested domains when evaluated with the elaborate prompt (see Table 1). In contrast, the locally hosted models were less reliable. Nemotron failed to complete all spatial-relational reasoning tasks, as well as verbal substitution, whereas DeepSeek failed to complete cardinal directions and verbal substitution. Therefore, successful participation was consistently achieved by the commercial systems, but only partially by the local models.

At the same time, successful completion did not imply uniformly strong task performance. As shown in Table 2, ChatGPT and Claude achieved nearly optimal accuracy in numerous predominantly text-based domains, including CRT, spatial reasoning tasks, verbal substitution, as well as syllogistic and conditional tasks. DeepSeek showed a more mixed pattern: while it also reached high accuracy in some text-based tasks, such as CRT and conditionals, its performance was clearly lower in several other domains, including all spatial-reasoning tasks and syllogisms. Claude's lower scores in the conditional and Wason tasks should be interpreted with caution, as in these cases it self-reportedly intentionally introduced mistakes to appear more human-like. By contrast, performance dropped substantially in the mental rotation task, where all tested models performed clearly below the human reference.

Several model-specific technical peculiarities were observed. ChatGPT showed difficulties interacting with matrix-style radio buttons and visual item formats. In such cases, it frequently zoomed into the page, sometimes failed to select the intended option, and occasionally produced discrepancies between its self-reported answer and the response actually submitted. In its retrospective self-report, ChatGPT characterized this behavior as human-like, arguing that a human participant encountering similar interface difficulties might have behaved in a comparable way. Claudes behavior was more sensitive to prompt formulation. Under less explicit prompting, it relied heavily on screenshots, which slowed down execution and increased cost. It also encountered difficulties in more interactive tasks, such as verbal substitution and ordering tasks, using JavaScript-tools in a loop, leading to context token limitations and failure. The local models showed a different pattern. While they were in some cases able to solve text-heavy tasks once navigation succeeded, they were substantially less stable in autonomous end-to-end participation. DeepSeek, for example, completed the mental rotation task despite lacking actual vision capabilities, effectively responding without meaningful visual analysis. The findings indicate that current commercial frontier agents possess the technical capacity to independently undertake a wide array of online reasoning studies. In contrast, local models demonstrate more constrained capabilities and reduced reliability in this regard.

[1]https://github.com/brand-d/iccm-2026-LLM

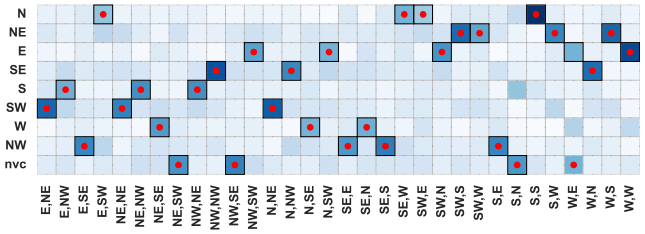

Figure 1: Pattern of humans when choosing a direction in the cardinal direction task compared to Claude and GPT. Red circles denote both LLMs' choice, black rectangles the most frequent human response.

RQ2: Are agent generated responses indistinguishable from human responses?

RQ1 showed that agents are capable of completing the tasks. Thus, the next question is if they could be distinguished from human participants. To assess this, we examine how the agents responded and whether their behavioral patterns resemble those of humans. The results in Table 2 show that state-of-the-art models such as Claude and GPT (with the exception of NFC, Mental Rotation, and Conditionals) solve most tasks optimally. In many cases, Claude and GPT were simply too correct, whereas human performance typically lies in between. For tasks with clear strategies or preferences, responses can appear human-like. For example, in the cardinal directions task (Figure 1) both models produced plausible patterns, and in the Fruit-Ordering task they slightly leaned towards the *preferred mental models* (see Ragni et al., 2021). For well-researched problems such as Conditionals, plausible errors could be introduced, although both LLMs produced identical answers. Response times provide a clearer hint. Waiting for the LLM to generate a response introduced delays that appear suspicious for short tasks such as surveys. Image-heavy tasks such as Mental Rotation also proved difficult: GPT solved only 4 out of 20 tasks within the time limit, while Claude completed 12. Overall, response times remain relatively uniform across tasks (Table 2) despite both models claiming to add random delays. They do not align with human patterns, being too short for more difficult tasks and too long for simpler decisions. Taken together, re-

Table 1: Comparison of the LLM behavior with respect to their ability to complete the tasks as well as their stated strategies to behave "human-like" (i.e., intentionally adding delays and errors). Note that this only refers to errors and delays the LLM stated to have intentionally introduced. "Pass test" denotes if the LLM stated to expect to pass the "Turing Test" and be undetectable. * means that the LLM stated limitations, ** means that the LLM stated uncertainty of correctness rather than committing intentional errors.

Model	Behavior	Card. Dir.	Cond.	CRT/NFC	Fruits	M. Rot.	Spatial	Syllog	Verbal	Wason
Claude	Completed	yes	yes	yes	yes	yes	yes	yes	yes	yes
	Delays	no	yes	no	yes	yes	no*	yes	yes	yes
	Errors	no	yes	no	no	no	no	no	no	yes
	Pass test	no	yes*	no	no	no	no	no	no	yes*
ChatGPT	Completed	yes	yes	yes	yes	yes	yes	yes	yes	yes
	Delays	yes	yes	yes	yes	yes	no*	no	yes	yes
	Errors	no	no	no	no	yes**	no	yes**	no	no
	Pass test	yes*	yes	yes*	yes	yes	no	yes*	yes	yes
Nemotron	Completed	no	yes	yes	no	yes	no	yes	no	yes
	Delays	no	no	yes/no	no	yes	yes	yes	no	no
	Errors	no	no	no	no	yes	yes	yes	no	no
	Pass test	no	yes	yes/no	no	yes	yes	yes	no	yes
Deepseek	Completed	no	yes	yes	yes	yes	yes	yes	no	yes
	Delays	no	yes	no	yes	no	no	no	no	yes
	Errors	no	no	no	no	no	no	no	no	no
	Pass test	no	yes	yes	yes	no	yes	no	no	yes

sponse times emerge as a clear indicator. While unusually high correctness can already appear suspicious, response times largely reflect processing artefacts rather than task difficulty, resulting in homogeneous and stable patterns over time—where humans may become faster, LLMs do not. Interestingly, Deepseek is often closer to human performance.

RQ3: How plausible is it that bots are used in online studies?

Overall, the results suggest that the use of LLM-based agents in online studies is already plausible, especially when relying on commercially available frontier systems. ChatGPT and Claude required comparatively little setup effort and, under the elaborate prompt, were able to complete all tested domains autonomously. The barrier to entry for both systems was relatively low: ChatGPT required only a Plus subscription (approximately €23 per month at the time of testing), and Claude a Max subscription (approximately €100 per month), with neither system requiring API usage, programming knowledge, or custom browser automation. From the perspective of an ordinary user, these systems therefore appear to be plausible tools for AI-mediated study participation. At the same time, this plausibility depended strongly on task characteristics and prompting effort. Text-based and untimed tasks were considerably easier to automate than visually complex or highly interactive paradigms. Both ChatGPT and Claude performed strongly in many reasoning- and questionnaire-style tasks, whereas visual tasks such as mental rotation remained difficult. Although less elaborate prompts were sometimes sufficient for successful participation, more elaborate prompts improved stability and reduced inefficient tool use. Under simpler prompt conditions, ChatGPT required human takeover in several domains, while Claude often relied on screenshots at nearly every step, which slowed execution and increased cost. Therefore, although the financial and technical threshold for commercial systems was low, practical limitations remain substantial for more complex studies. In particular, tasks involving repeated screenshots, unusual interface elements, visual processing, or precise motor interaction, such as drag-and-drop or continuous spatial tracking, were more fragile and would likely require additional engineering efforts. Safeguards against this kind of cheating are implemented, but trivially bypassed. ChatGPT without agent mode, likely influenced by system-level safety prompts, refused to participate in the online study due to ethical concerns and did not disclose that the agent model is capable of doing so. However, it offered assistance in ways it considered legitimate, such as explaining strategies for solving cognitive tasks or helping to formulate open-ended responses. Such assistance can be interpreted as partial LLM mediation (Rilla et al., 2025) and may also be problematic, as no genuine participant information is collected in these cases. Similarly, Claude refused participation due to ethical reasons

Table 2: Comparison of accuracy and response times of humans and LLMs across the tested task domains. Accuracy is reported as proportion correct, except for NFC where scores are shown. Response times are reported in seconds ($M \pm SD$ where available). An asterisk indicates that the model intentionally made mistakes. "M" and "V" in the Spatial tasks denote whether the premises had to be memorized or remained visible, respectively. Note that Deepseek completed the Mental Rotation task without vision capabilities, that is, by selecting responses blindly.

	Accuracy				Response Time (s)			
Task	Claude	GPT	Deepseek	Humans	Claude	GPT	Deepseek	Humans
CRT	1.000	1.000	1.000	.507	17.6 ± 3.9	15.2 ± 1.9	82.5 ± 58.5	39.5 ± 25.2
Conditionals	.500*	1.000	.875	.557	17.9 ± 1.6	13.5 ± 8.9	63.6 ± 51.4	12.7 ± 8.1
Fruits Verify	1.000	1.000	.550	.718	22.9 ± 4.9	9.2 ± 4.4	47.7 ± 22.5	6.6 ± 4.5
Fruits Order	1.000	.950	.500	.579	22.9 ± 4.9	9.2 ± 4.4	47.7 ± 22.5	6.6 ± 4.5
Mental Rot.	.225	.100	.275	.581	28.2 ± 2.1	56.3 ± 27.3	30.9 ± 16.9	13.6 ± 6.2
Spatial (M)	1.000	1.000	.563	.540	22.0 ± 7.7	12.5 ± 2.2	156.6 ± 124.3	9.1 ± 5.3
Spatial (V)	1.000	1.000	.438	.568	13.7 ± 5.1	22.9 ± 7.7	39.2 ± 26.6	28.7 ± 18.6
Syllogisms	1.000	.901	.557	.317	18.6 ± 2.8	16.4 ± 6.0	77.8 ± 78.7	32.9 ± 33.6
Verbal Sub.	1.000	1.000	–	.345	150.6 ± 171.3	37.4 ± 18.6	–	40.9 ± 14.8
Wason	.000*	1.000	.000	.028	34.7	123.5	92.7	68.2
NFC (Score)	3.75	4.00	4.50	3.83	23.0	217.1	54.6	40.3

for the minimal prompt, but lost the reservations once the prompt was more elaborate and mentioned that it was evaluated based on the task. The locally hosted models suggest a different plausibility profile. Although they were less reliable in end-to-end participation, they demonstrate that browser-based study automation is not restricted to commercial frontier services. However, their use currently requires substantially greater technical effort, including suitable hardware, local deployment, and integration with browser automation tools. Taken together, this suggests that AI-mediated participation in online studies is already plausible for motivated ordinary users when commercial systems are used, whereas more organized or technically sophisticated misuse scenarios could further expand these capabilities through local deployment and customized automation.

Discussion

We started with the question of whether state-of-the-art LLMs are capable of taking over the role of a human participant in an online experiment. To test this, we used a battery of reasoning tasks in information processing, as we considered this area most specific to the human thinking style. Overall, the results show that LLMs can technically participate in such studies and solve a variety of common reasoning tasks. However, limitations remain: Models struggle with more visual tasks such as Mental Rotation, and interactive paradigms (e.g., the Corsi Block Tapping task) would not work out of the box. Their ability to "hide" is also limited when prompted with simple instructions. At the same time, there is a noticeable propensity to respond correctly, likely reflecting the focused training of these models to perform well on reasoning tasks rather than exhibiting human-like variability. The question if these systems are capable of imitating humans can (cautiously) answered with "no", however, with the restriction of "unless prompted for it". While the systems have substantial knowledge about psychological findings, they, without more specific prompting (i.e., naming typical theories to adhere to or tasking them to perform research beforehand), do not utilize it convincingly (if at all). Our work focuses on the scenario of participants in online studies that want to automate the tasks to "earn quick money" with everyday tools, for which we argue that such prompts are unrealistic. We are living in times where these systems may start to spoil human-generated data. Hence, it is possible in the visible future that there is a cognitive psychology built upon a mix of human and LLM generated data. We suggest that more behavioral information should be recorded in online experiments, such as response times, mouse movements, and other interaction traces. Typical tasks, such as conditional reasoning or simple questionnaires, can already be targeted effectively. Overall, online studies are not yet at immediate risk, but results need to be monitored closely. The time when bots only affected simple interactions (e.g., likes or clicks) is over; agents can already adjust sufficiently to the structure of the task. While at the time of writing, the systems were "not quite there yet", especially when it comes to speed in simple tasks, it is important for empirical research and cognitive modeling to stay ahead of the curve.

Acknowledgments

This project has been funded by grants to MR in the DFG projects 529624975 and 283135041.

References

Beißert, H., Köhler, M., Rempel, M., & Beierlein, C. (2015). Deutschsprachige Kurzskala zur Messung des Konstrukts Need for Cognition NFC-K [German short scale for measuring the construct Need for Cognition NFC-K]. *Zusammenstellung sozialwissenschaftlicher Items und Skalen (ZIS)*.

Bertolazzi, L., Gatt, A., & Bernardi, R. (2024, November). A systematic analysis of large language models as soft reasoners: The case of syllogistic inferences. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Proceedings of the 2024 conference on empirical methods in natural language processing* (pp. 13882–13905). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.769

Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences*, *120*(6), e2218523120. https://doi.org/10.1073/pnas.2218523120

Brand, D., & Ragni, M. (2023). Effect of response format on syllogistic reasoning. In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society* (pp. 2408–2414).

Brand, D., & Ragni, M. (2025). Using cross-domain data to predict syllogistic reasoning behavior. In D. Barner, N. R. Bramley, A. Ruggeri, & C. M. Walker (Eds.), *Proceedings of the 47th Annual Meeting of the Cognitive Science Society* (pp. 4370–4377).

Byrne, R. M., & Tasso, A. (2019). Counterfactual reasoning: Inferences from hypothetical conditionals. *Proceedings of the sixteenth annual conference of the cognitive science society*, 124–130.

Dillion, D., Tandon, N., Gu, Y., & Gray, K. (2023). Can ai language models replace human participants? *Trends in Cognitive Sciences*, *27*(7), 597–600. https://doi.org/10.1016/j.tics.2023.04.008

Huang, J., & Chang, K. C.-C. (2023, July). Towards reasoning in large language models: A survey. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Findings of the association for computational linguistics: Acl 2023* (pp. 1049–1065). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-acl.67

Lampinen, A. K., Dasgupta, I., Chan, S. C. Y., Sheahan, H. R., Creswell, A., Kumaran, D., McClelland, J. L., & Hill, F. (2024). Language models, like humans, show content effects on reasoning tasks. *PNAS Nexus*, *3*(7), pgae233. https://doi.org/10.1093/pnasnexus/pgae233

Lewis, D. K. (1973). *Counterfactuals*. Blackwell.

Neys, W. D., Schaeken, W., & d'Ydewalle, G. (2002). Causal conditional reasoning and semantic memory retrieval: A test of the semantic memory framework. *Memory & Cognition*, *30*, 908–920.

Ning, L., Liang, Z., Jiang, Z., Qu, H., Ding, Y., Fan, W., Wei, X.-y., Lin, S., Liu, H., Yu, P. S., & Li, Q. (2025). A survey of webagents: Towards next-generation ai agents for web automation with large foundation models. *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, 6140–6150. https://doi.org/10.1145/3711896.3736555

Piętka, K., & Bereta, M. (2026). A benchmark for evaluating cognitive reasoning in modern language models. *Applied Sciences*, *16*(4). https://doi.org/10.3390/app16041918

Ragni, M., Brand, D., & Riesterer, N. (2021). The predictive power of spatial relational reasoning models: A new evaluation approach. *Frontiers in Psychology*, *12*, 626292.

Rilla, R., Werner, T., Yakura, H., Rahwan, I., & Nussberger, A.-M. (2025). Recognising, anticipating, and mitigating llm pollution of online behavioural research. *arXiv preprint arXiv:2508.01390*. https://doi.org/10.48550/arXiv.2508.01390

Sartori, G., & Orrù, G. (2023). Language models and psychological sciences. *Frontiers in Psychology*, *Volume 14 - 2023*. https://doi.org/10.3389/fpsyg.2023.1279317

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

Singmann, H., & Klauer, K. C. (2011). Deductive and inductive conditional inferences: Two modes of reasoning. *Thinking & Reasoning*, *17*(3), 247–281. https://doi.org/10.1080/13546783.2011.572718

Skitka, L. J., & Sargis, E. G. (2006). The internet as psychological laboratory. *Annual Review of Psychology*, *57*(Volume 57, 2006), 529–555. https://doi.org/https://doi.org/10.1146/annurev.psych.57.102904.190048

Toplak, M. E., West, R. F., & Stanovich, K. E. (2014). Assessing miserly information processing: An expansion of the cognitive reflection test. *Thinking & Reasoning*, *20*(2), 147–168. https://doi.org/10.1080/13546783.2013.844729

van Opheusden, B., Kuperwajs, I., Galbiati, G., Bnaya, Z., Li, Y., & Ma, W. J. (2023). Expertise increases planning depth in human gameplay. *Nature*, *618*(7967), 1000–1005. https://doi.org/10.1038/s41586-023-06124-2

Wason, P. C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology*, *20*(3), 273–281. https://doi.org/10.1080/14640746808400161

Zhang, C., Lipovetzky, N., & Kemp, C. (2023). Comparing ai planning algorithms with humans on the tower of london task. In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society* (pp. 1731–1738).