Synthetic peer review:
Using large language models to automatically detect deviations from preregistrations and to evaluate their strictness

Master thesis by Lukas Sidler
Supervised and examined by Prof. Dr. Malte Elson
Psychology of Digitalisation, University of Bern

Bern, January 2025

Correspondence concerning this article should be addressed to Lukas Sidler, University of Bern, Institute of Psychology, Department of Psychology of Digitalisation, Fabrikstrasse 8 CH-3012 Bern. lukas.sidler@students.unibe.ch. Matriculation number: 18-106-724

Abstract

The replication crisis in psychology highlights the need for rigorous peer review and adherence to preregistrations. However, deviations from preregistrations often go unnoticed due to limitations in the peer-review process, including subjective judgments and high workloads for reviewers. Large language models (LLMs) offer a potential solution by providing consistent, efficient, and rapid evaluations, free from subjective biases. This study evaluates for the first time the performance of GPT-4o in detecting deviations between preregistrations and publications and assessing their strictness, which refers to their level of detail and rigidity. Using zero-shot prompting, GPT-4o analysed 300 preregistration-publication pairs (PSPs) with tasks divided into extraction (94 items) and comparison (46 items). Performance was benchmarked against human raters (van den Akker et al., 2024). Results revealed substantial test-retest reliability (κ = .74) but low interrater reliability (in the range of κ = .30) and low accuracy (of approximately 60%), insufficient for practical application. Surprisingly, a simplified prompt focusing on only nine key items yielded no improvement. Analyses indicated that task complexity, predefined response labels, and token limitations were key barriers. However, GPT-4o performed well in some aspects, namely extracting number of participants and providing justifications for classifications. Future studies should explore alternative methodologies, such as few-shot prompting and simplified question formats, to enhance performance. Despite its shortcomings, GPT-4o's efficiency in processing PSPs offers promise for selective integration into research evaluation practices.

Keywords: *Replication Crisis, Preregistration, Synthetic Peer Review, LLM, GPT-4o*

Contents

Introduction

In recent years, scientific fields, particularly psychology, have faced a significant challenge: the replication crisis. The findings of numerous influential studies cannot be reliably reproduced. For instance, a study by Stodden et al. (2018) found that only 26% of articles published in Science since 2011 could be accurately reproduced, even with available data. Furthermore, a large-scale replication project by the Open Science Collaboration (2015), which re-examined 100 experimental and correlational studies from leading psychology journals, revealed that the replicated effects were, on average, only about half as strong as originally reported, and that merely 36% of these replications produced statistically significant results — compared to 97% of the original studies.

Preregistration

To address this issue, various reforms have been proposed, including preregistration and enhanced peer review. The preregistration of an analysis plan serves methodological rigor by encouraging researchers to make analytical decisions — such as formulating hypotheses, determining sample sizes, or operationalising variables — independently of the data subsequently collected (Nosek et al., 2018). Typically, preregistrations are uploaded to independent repositories such as the Open Science Framework (https://osf.io/) or Clinical Trials (https://clinicaltrials.gov/). These platforms archive preregistrations and make them publicly accessible.

Preregistration establishes a clear distinction between a priori formulated predictions and post hoc adjustments that may arise from the results (HARKing). This approach supports the differentiation between confirmatory analyses, which test hypotheses, and exploratory analyses, which generate hypotheses (Nosek et al., 2018). It enables documentation of which analytical steps were originally planned, and which may have been added later. This is intended to prevent motivation, incentives, or cognitive biases from influencing decisions that could lead to methodologically questionable outcomes (Nosek et al., 2018).

Although preregistration offers numerous advantages, it is not sufficient on its own to ensure methodological rigor and transparency. Studies indicate that the actual effectiveness of preregistrations in reducing researcher degrees of freedom remains unclear (Ikeda et al., 2019;

Claesen et al., 2022; Heirene et al., 2024; van den Akker et al., 2024, as cited in Hahn et al., 2024). This is often attributed to the insufficient quality of preregistrations and a lack of consensus regarding appropriate criteria for assessing the quality of a preregistration (Hahn et al., 2024).

Apart from the quality or strictness of the preregistration — that is, whether it provides enough detail to limit researcher degrees of freedom — consistency also plays a significant role in the effectiveness of preregistration. This consistency pertains to the alignment between the preregistered plan and the actual conduct of the study. When preregistrations lack sufficient detail or researchers deviate from the preregistered plan, the purpose of preregistration is compromised, allowing for greater researcher flexibility and increased potential for p-hacking or biased decisions (van den Akker et al., 2024).

Empirical findings, however, suggest that this consistency is often low, with researchers in psychology and other scientific disciplines frequently deviating from their original preregistered plans without transparently disclosing these changes in their publications. It is estimated that up to 90% of studies published in psychology journals, such as *Psychological Science*, contain undisclosed deviations between the paper and the preregistration (Claesen et al., 2021, as cited in van den Akker et al., 2024). Further, a study by van den Akker et al. (2023) revealed that in a sample of psychological studies, approximately half of the preregistered hypotheses were no longer explicitly identifiable in the published articles.

Peer review

To ensure adherence to preregistrations and thereby guarantee their effectiveness, the peer-review process plays a critical role. Peer review is widely regarded as the gold standard in scientific publishing, functioning both as a selection mechanism and as a process for refining research reports (Elson et al., 2020). However, studies such as Mathieu et al. (2013) indicate that peer reviewers rarely access or systematically evaluate preregistrations. Contributing factors include a lack of awareness, that reviewers are already heavily burdened, and that the task of meticulously comparing two nearly identical documents for inconsistencies appears laborious and lacks appeal (Mathieu et al., 2013). More recent research, such as the study by

Syed (2023) further supports this, finding that only 18% of reviewers mention preregistrations and merely 5% actually examine them.

The current peer review process has also been fundamentally criticised. Concerns have been raised about the subjectivity inherent in the peer review process, which may compromise the fairness and accuracy of evaluations (Park et al., 2014; Lipworth et al., 2011; King et al., 2018; Lee et al., 2013; Abramowitz et al., 1975, as cited in Verharen, 2023). And also the considerable effort and resources required to provide timely and comprehensive feedback on scientific work is problematic (Horbach & Halffman, 2018, as cited in Liang et al., 2023). The rapid growth in scientific publications and the increasing specialization within fields further intensify these challenges (Price, 1963; Jones, 2009, as cited in Liang et al., 2023). Peer review alone is estimated to consume over 100 million researcher hours annually, costing billions, while finding enough qualified reviewers for the rising number of submissions is increasingly challenging (Kovanis et al., 2016; Lee et al., 2013, as cited in Liang et al., 2023). These limitations in providing high-quality feedback pose a foundational issue for the sustainable growth of science and amplify existing disparities. Researchers from marginalized institutions or resource-limited regions often struggle to access valuable feedback, which perpetuates systemic inequalities within the scientific community (Bourdieu, 2018; Merton, 1968, as cited in Liang et al., 2023).

Synthetic Peer review

In the context of these problems, the advancements of artificial intelligence (AI) and machine learning (ML) tools to enhance the peer-review process presents promising potential.

In particular, the Generative Pretrained Transformer (GPT) models developed by OpenAI have shown impressive abilities in producing human-like text, which raises the question of their applicability in an academic review setting (Brown et al., 2020, as cited in Robertson, 2023). Natural language processing (NLP) tools are usable for analysing large datasets, yet their application to scientific peer review remains a challenge due to the specialized language and structure of such reports (Chowdhary, 2020; Hirschberg & Manning, 2015; Yadav &

Vishwakarma, 2020, as cited in Verharen, 2023). Early studies have already indicated the potential of using AI in academic writing and even peer reviews: OpenAI's ChatGPT has been shown to effectively analyse language in peer review reports, providing consistent and accurate evaluations of sentiments (Verharen, 2023).

A proof-of-concept for using large language models (LLMs) like GPT to review research papers is presented by Tyser et al. (2023). The study demonstrates that LLMs can deliver consistently high-quality reviews almost instantly. Similarly, Robertson (2023) investigated GPT-4's effectiveness as a peer reviewer, finding that while its reviews were comparable to human feedback in helpfulness on average, they exhibited more variability, indicating occasional inconsistencies. Liang et al. (2023) extended this by developing an automated pipeline using GPT-4 to provide structured feedback on research manuscripts, addressing aspects like significance, novelty, and improvement suggestions. Their large-scale evaluations revealed substantial overlap between GPT-4's comments and human reviewers, especially for weaker papers. While GPT-4 functions for fast, general suggestions, limitations include a focus on specific feedback types and challenges in critiquing complex methods.

To date, no studies have investigated whether LLMs can automatically detect deviations from preregistrations and assess their strictness, which is an essential aspect of peer review, ensuring research transparency and accountability. The present study addresses this gap by evaluating the ability of LLM to identify these deviations and to assess the strictness of preregistrations themselves. By examining whether LLM can effectively evaluate the level of detail and rigidity in preregistration documents, this study aims to explore its potential to support and enhance research evaluation practices.

Criticism

Even though there are proposals to end the human-dependent peer-review system completely and to move to AI-based methods (Irfanullah, 2023) many scientists remain sceptical. For example Hosseini and Horbach (2023) critically evaluate the integration of LLMs into the peer review process. They emphasise that significant challenges remain. These include

reproducibility issues, opacity in model training and data usage, risks of amplifying existing biases, and threats to confidentiality when handling sensitive data. Furthermore, the social and epistemic functions of peer review, which involve dialogue on research norms and values, could be undermined if LLMs replace human reviewers. They also highlight ethical concerns, particularly the potential for misuse of LLMs to generate fraudulent or biased reviews and emphasize that LLMs should not independently replace human judgment. Instead, they recommend transparent disclosure of LLM use by reviewers and editors, rigorous training on responsible AI integration, and the development of policies to address the unique risks posed by LLMs (Hosseini & Horbach, 2023).

LLM and choice of model

The task of comparing papers and evaluating their strictness differs from generating feedback, as discussed in the papers mentioned above, and therefore presents a distinct challenge for LLMs. This raises the question of whether alternative approaches are required, and which model may best address the specific demands of this task.

LLMs are AI systems specialised in natural language processing (NLP). They are based on deep neural networks and are trained on extensive datasets to enable human-like text processing (Brown et al., 2020). The Transformer architecture, introduced by Vaswani et al. (2017), forms the foundation of these models. It uses mechanisms such as self-attention to analyse contextual relationships between words, regardless of their position within a text. Rather than memorising an entire corpus of text word by word, the attention mechanism assigns weights to the relevance of words in a context window for a given target word (Hommel et al., 2021). It is important to emphasise that LLMs do not possess actual understanding of the text. Instead, they calculate probabilities for the next tokens, words, and sentences without knowing the entire content of the text or its conclusion in advance.

The training of LLMs begins with pretraining on massive general datasets, during which patterns and linguistic structures are learned. Fine-tuning then adapts these models to specific

tasks by training them on smaller, domain-specific datasets (Howard & Ruder, 2018). LLMs vary significantly in functionality and applicability depending on their training method and fine-tuning.

Tasks like information extraction—relevant to this study—encompass a wide range of activities (Li et al., 2023). Extracting structured information from unstructured texts involves tasks such as Named Entity Recognition, Relation Extraction, Relation Classification, Event Detection, Event Argument Extraction, and Event Extraction (Li et al., 2023). These tasks each require specialised pretrained models. Thousands of such models are available, for example, through open-source libraries like Hugging Face (https://huggingface.co).

In recent years, however, a trend towards "task-agnostic" approaches has emerged (Brown et al., 2020). These approaches leverage pretrained models that are flexibly transferable to various tasks by using simple instructions or examples, without the need for extensive task-specific fine-tuning datasets, albeit with minor performance trade-offs (Navarro et al., 2022; Howell et al., 2023, as cited in Yuen et al., 2024). Techniques such as zero-shot prompting (where the model is prompted to perform a task without prior examples) and few-shot prompting (where the model is provided with a number of examples within the prompt) are instrumental in shaping answer structures and eliciting reasoning patterns resembling human thought processes (Yuen et al., 2024). Scaling up language models greatly improves task-agnostic, few-shot performance, sometimes even reaching competitiveness with prior state-of-the-art finetuning approaches (Brown et al., 2020).

As numerous LLMs are released week upon week claiming superior performance in specific tasks, it has made it increasingly difficult to discern true advancements and leading models. Closed-source LLMs like OpenAI's GPT and Anthropic's Claude generally outperform open-source alternatives, but the latter have made significant progress, occasionally claiming parity or superiority in certain tasks. This progress has notable implications for both research and business, prompting the need for comprehensive evaluations across a wide range of tasks to assess their general capabilities (Chen et al., 2023). While the proprietary nature of models

like ChatGPT limits transparency—details on its architecture, training data, and fine-tuning processes remain undisclosed—open-source LLMs present a promising alternative. They address issues such as high costs, transparency, and privacy concerns associated with closed-source models. However, as of late 2023, open-source LLMs like Llama-2 and Falcon still lag behind leaders such as GPT, which is widely considered the benchmark for performance (Chen et al., 2023). Encouragingly, the gap between open-source and closed-source models is narrowing, with some open-source LLMs now outperforming GPT-3.5-turbo on select benchmarks. The landscape, however, remains highly dynamic, with constant updates, new releases, and varied evaluation benchmarks complicating direct comparisons (Chen et al., 2023).

For the present study, these advantages and disadvantages of various LLMs were carefully weighed. While aspects such as transparency, costs, privacy concerns, and others favoured open-source models, GPT-4o by OpenAI, a proprietary model was selected primarily due to its leading performance benchmarks, mainstream recognition, and ease of implementation. According to OpenAI (2024), GPT-4o ('o' for 'omni') is their latest, fastest, highest intelligence model. It supports a context length of up to 128K tokens, enabling the analysis of extensive textual content, equivalent to the length of a longer novel. The model accepts both text and image inputs and generates outputs in both formats. Its knowledge cutoff extends to October 2023 (OpenAI, 2024).

Extraction and Comparison

Research with GPT-4o on the core components of the current tasks, extraction and comparison, presented a positive outlook.

Evaluating ChatGPT's Information Extraction Capabilities, findings of Li et al. (2023) reveal that ChatGPT performs poorly when it is provided with predefined labels to select from, but it excels when requested to generate predictions without predefined labels, relying only on task understanding and input, as confirmed by human evaluations (Li et al., 2023). Additionally, in their research ChatGPT delivers high-quality and reliable explanations for its decisions. However, a notable issue is its tendency toward overconfidence in predictions, leading to poor

calibration. Despite this, ChatGPT generally demonstrates a high degree of faithfulness to the original text in most cases (Li et al., 2023).

For “information comparison”, applying LLMs presents significant scalability challenges, primarily due to difficulties in managing extensive context within the constraints of model token limits (Yuen et al., 2024). Current state-of-the-art text comparison methods are constrained by their reliance on domain-specific training data, high computational demands, and limited scalability (Yuen et al., 2024). Minimizing information loss and managing token limits remain persistent challenges (Jaiswal and Milios, 2023, as cited in Yuen et al., 2024). Models with extended token limits often struggle with retaining relevant information, particularly from the middle of lengthy input contexts (Liu et al., 2023b, as cited in Yuen et al., 2024). Additionally, Liu et al. (2023b, as cited in Yuen et al., 2024) found that naive, untrained models without input context outperformed those provided with extended contexts on similar tasks. Yuen et al. (2024) introduced a system to automate large-scale text comparison with high accuracy (94%) and efficiency. The system works in four steps: First documents are condensed into summaries that retain key information, second it splits user-defined criteria into searchable pieces stored in a vector database, third it retrieves relevant passages from the database by conducting a semantic search and finally it compares the matches to generate analyses with confidence scores. The system’s comparisons were evaluated through human surveys, where participants assessed accuracy, justification, and confidence scores based on predefined criteria.

Method

Reference data set

For the present study, a dataset created by human raters from the study by van den Akker et al. (2024) was used as a benchmark to assess the capabilities of the LLM. For more detailed information on the data collection process of the reference dataset, readers may refer to this lead article. This dataset comprises coded responses to questions about preregistrations and their associated publications from 28 trained raters.

Van den Akker et al. (2024) selected 300 hypotheses from 459 preregistered psychological studies, either awarded the Preregistration Challenge Prize or marked with the Preregistration Badge prior to 2020. For each hypothesis—i.e., each preregistration-publication pair (PSP)—van den Akker et al. (2024) coded items using a protocol implemented via Qualtrics (available at https://osf.io/dpg3v). Each PSP was coded by two independent raters who subsequently resolved any discrepancies collaboratively.

The items included questions about the strictness of the preregistration and the paper—that is, whether they provided enough detail to limit researcher degrees of freedom—and consistency, meaning the alignment between the two. A preregistration and a study were considered strict if they were described in a specific (all steps were described) and precise (each described step allowed only one interpretation or implementation) manner. They were deemed consistent, when the researcher adhered to the preregistration description within the published paper (van den Akker et al., 2024). The items include questions about the operationalization of independent and dependent variables, including measurement methods, potential values, and the construction of variables from individual components (e.g., Likert scales). Further aspects included data collection (sample size, sampling frame), the statistical model (model type, specification, and use of variables), and the criteria for statistical inference. Questions also addressed the operationalisation of control variables, handling of missing data, treatment of violations of statistical assumptions, and the definition of exclusion criteria (van den Akker et al., 2024).

Test data set

For the test dataset, OpenAI's paid GPT-4o model was accessed through the chat interface available on their website (https://chatgpt.com). Data were collected between June 2024 and October 2024. Links to all interactions, results, and chats with GPT-4o are available on Github (https://github.com/Lukasior8/MA_Regcheck). All 300 PSPs from the study of van den Akker et al. (2024) were used, however 21 studies were excluded because either the preregistration or the paper were inaccessible. Additionally, 30 randomly selected PSPs were retested to assess test-retest reliability.

The studies and preregistrations were uploaded as separate PDF documents alongside the prompt text. GPT's responses were manually transferred into an excel sheet for further processing. Additionally, qualitative observations and notes regarding GPT's responses were recorded.

The same items as in the study by van den Akker et al. (2024) were used, albeit with adjusted phrasing. Certain items were excluded due to prompt length, such as questions about individual components of variables, or because they did not elicit categorical responses, such as items regarding authors' explanations for deviations. The two tasks—extracting information to assess strictness and comparing the document pairs—were split into two distinct prompts. One prompt contained 94 items for extraction, and the second prompt included 46 items for the comparison task.

Prompt engineering was developed iteratively, taking into account general tips for prompt optimization. Specifically, the empirically supported prompt principles outlined by Bsharat et al., (2023) were applied, including: omitting polite language; integrating the intended audience; using example-driven prompting; incorporating the phrases “Your task is” and “You MUST”; assigning a role to the LLM; employing delimiters; repeating specific words or phrases multiple times within the prompt; and clearly stating the requirements the model must follow to produce content, such as keywords, regulations, hints, or instructions. These principles and others were systematically incorporated.

Even though few-shot prompting is recommended, as it can increase the accuracy of the performance of LLMs (Brown et al., 2020), the length of the prompts was found to cause problems. Minimizing information loss and managing token limits remain persistent challenges for text comparison (Jaiswal & Milios, 2023, as cited in Yuen et al., 2024). Models with extended token limits often struggle to retain relevant information, particularly from the middle of lengthy input contexts (Liu et al., 2023b, as cited in Yuen et al., 2024). Additionally, naive, untrained models without input context outperform those provided with extended contexts on similar tasks (Liu et al., 2023b, as cited in Yuen et al., 2024).

Similarly, initial trials in this study demonstrated, that for this lengthy task with two uploaded PDF documents—which also contributed to the prompt length—further increasing the prompt size with additional shots or documents didn't seem sensible. Consequently, a zero-shot prompting strategy was employed. Zero-shot strategies can effectively incorporate chain-of-thought reasoning through prompts such as "Let's think step-by-step," which guide structured, logical responses (Kojima et al., 2022, as cited in Yuen et al., 2024).

To enable a direct comparison with the categories of van den Akker et al. (2024), the response options for GPT-4o were restricted to predefined labels Yes and No (Example*: "5.1. PreDV/PapDV: Is the DV's measure specified, e.g., the test, scale, etc.? (Y/N)."*). The responses from the reference dataset were transformed as well, to align with the same predefined categories. However, certain items or questions were not applicable to all studies (e.g., depending on experimental manipulations or questions regarding composite measures). To avoid creating individual prompts tailored to each PSP, dependent questions were incorporated. This meant that specific items were only to be answered if certain conditions were met (Example: *"2.2. PreMI1/PapMI1: If IV1 is experimentally manipulated, assess clarity in IV1's manipulation, e.g., the difference between conditions. (Y/N)."*). If the condition was not met, the response "*not applicable (X)*" had to be rated. Exceptions were items concerning the number of participants, where a numerical response was required *("7.3. PreDCP_Text/PapDCP_Text: Extract the exact number of participants (number).*") The only manual adjustment of the prompt was the instruction to focus solely on a specific sub-study in papers, as the PDF documents often contained multiple sub-studies.

Throughout the data collection process, refinements and modifications were applied to the prompts. The test data versions A to D represent iterative adjustments, encompassing changes such as rewordings, clarifications, and deletions. Version A corresponds to the initial version, while version D reflects the most refined and conceptually optimised version. All prompt versions, with variations highlighted in colour, are available on Github (https://github.com/Lukasior8/MA_Regcheck) for reference. Additionally, strictness ratings, comparison ratings and phrases including the main deviations were collected, although these were not included in the analysis.

Since the prompts were very long, and the task with its dependencies highly complex, a new minimal version was later developed with the expectation of achieving better results. This prompt version focused exclusively on the comparison. For this purpose, only the nine most important key items were specifically selected. To eliminate the need for dependencies, studies were carefully chosen, so that all items were applicable, reducing responses to only two categories: Y/N.

With this adjusted prompt, 52 PSPs were tested, including 16 that were tested twice for retest reliability.

Statistics

The R-code for the processing and analysis of the data is available on Github (https://github.com/Lukasior8/MA_Regcheck).

To evaluate GPT-4o's performance, confusion matrices were calculated. This tool is commonly used in machine learning to assess the performance of classification models, such as artificial neural networks, based on test data with known distinguishing features (Fahmy Amin, 2023). The principles of the confusion matrix, along with associated performance metrics like accuracy, sensitivity, and specificity for binary classifications, have been extended to three-class classifications according to Fahmy Amin (2023).

Additionally, Cohen's kappas were calculated to assess test-retest reliability as a measure of the stability and consistency of GPT-4o's responses under identical conditions.

Cohen's kappa was also calculated to evaluate interrater reliability, where the responses from van den Akker et al. (2024) and GPT-4o were each considered as a rater. The human raters were assumed to be correct in their evaluations, with no consideration given to potential error variance in their ratings. In this context, Cohen's kappa served as a chance-corrected agreement measure and a quality metric for GPT's performance. Calculations were conducted at the overall test level, the item level, the PSP level and across selected categories. Cohen's kappa is the most widely used measure of agreement in scientific literature (Zec et al., 2017); however, there are some controversies regarding its interpretation, which are addressed in the discussion.

Shapiro-Wilk tests and visual inspections, including histograms and Q-Q plots, were conducted to assess normality over items or papers. Furthermore, Pearson correlations with additional metrics, such as file size were calculated, to explore potential connections and explanatory patterns.

Results

This section presents the findings of the study. First, confusion matrices and quality criteria are reported, followed by the results of retest- and interrater kappas. Then results of different agreement measures per categories are reported, followed by agreement measures on item and PSP levels and finally results for the minimal key-item version.

From the 300 studies of van den Akker et al.'s (2024) dataset, 21 studies were excluded because either the preregistration or the paper were inaccessible. After these exclusions, 309 PSPs remained for the overall comparison including the 30 retest studies. Of the 140 items collected per PSP, only categorical responses that were also included in the dataset of van den Akker et al. (2024) were considered for the calculations. Non-categorical, open-ended questions about hypotheses or the content of variables were collected for completeness but were not evaluated. Additionally, different numerical strictness and consistency ratings were recorded for some prompt versions but were excluded from the analysis and subsequently omitted. In the final dataset, 118 items were analysed across all 309 PSPs, resulting in a total of 36’462 comparisons for the overall evaluation.

After entering the prompt and uploading the paper, formulating the output took an average of 65.07 seconds (SD = 6.09) for the extraction process and 68.57 seconds (SD = 8.30) for the comparison process.

Confusion matrix and quality criteria

Tables 1 and 2 display the classification prediction outcomes using a confusion matrix, presented both in absolute frequencies (Table 1) and percentages (Table 2). These findings illustrate the agreement between the “True” ratings of humans and “Predicted” ratings form GPT-4o for the categories "Yes," "No," and "X." Of the total 36’462 comparisons, only 19’956 comparisons, or 55%, were correctly classified by GPT-4o. The precision of GPT-4o—i.e., the proportion of correctly classified positive cases among all predicted positive cases —is particularly low at .16 for the "No" category (see Table 3). The model's sensitivity varies across the three classes. For class "Y," the sensitivity is .37, 95% CI [.36, .37], indicating that the model correctly identifies 37% of the actual "Y" cases. For class "N," sensitivity is .30, 95% CI [.29, .32].

The highest sensitivity was achieved for class "X," with a value of .89, 95% CI [.88, .89]. Specificity for class "X" is the lowest at .60, 95% CI [.59, .60], meaning that 60% of cases classified as "Y" or "N" in the reference data were correctly identified as not "X." The model's balanced accuracy calculated from the averages of sensitivity values for each class is .52.

Table 1

Confusion matrix total absolute

	Predicted Yes	**Predicted No**	**Predicted X**	**Sum**
True Yes	7177	5224	7256	19657
True No	382	1106	2172	3660
True X	938	534	11673	13145
Sum	8497	6864	21101	36462

Note. Table 1 compares the absolute frequencies of GPT-4o predictions (Predicted) to the judgments made by human raters (True) for 309 PSP and 118 items.

Table 2

Confusion matrix total in percent

	Predicted Yes	**Predicted No**	**Predicted X**	**Sum**
True Yes	19.68	14.33	19.90	53.91
True No	1.05	3.03	5.96	10.04
True X	2.57	1.46	32.01	36.04
Sum	23.30	18.82	57.87	100

Note. Table 2 compares the percentage frequencies of GPT-4o predictions (Predicted) to the judgments made by human raters (True) for 309 PSP and 118 items.

Table 3

Quality criteria

Class	Sensitivity	Specificity	Precision	NPV
Yes	.37 [.36, .37]	.92 [.92, .93]	.84 [.84, .85]	.55[.55, .56]
No	.30 [.29, .32]	.82 [.82, .83]	.16 [.15, .17]	.91[.91, .92]
X	.89 [.88, .89]	.60 [.59, .60]	.55 [.55, .56]	.90[.90, .91]

Note. Values are presented as the point estimate followed by the 95% CI in brackets. Sensitivity reflects the probability of correctly identifying positive cases, specificity the probability of correctly identifying negative cases, precision the proportion of predicted positives that are true positives, and NPV (Negative Predictive Value) the proportion of predicted negatives that are true negatives.

Test-Retest reliability total

Test-retest reliability was calculated using Cohen's kappa for 30 PSP and 118 items. Across a total of 3'540 classifications, the unweighted kappa coefficient was κ =.74 95% CI [.72, .75], indicating substantial agreement between test and retest ratings (according to Landis & Koch, 1977). However, the average test-retest kappa value across all items was .57 (SD = .32), ranging from -.09 to 1, which points to variability among the items. A significant, moderate positive correlation was found between the test-retest kappa and the percentage of prototypical or most frequent responses per item, $r(113) = 0.42$, $p < .001$, 95% CI [0.25, 0.56]. This indicates, that for items where GPT-4o frequently rated a particular category, the test-retest reliability was notably higher.

Interrater reliability total

Interrater reliability, as a measure of agreement between the reference data of human raters and the test data from GPT-4o, was calculated using Cohen's kappa (see Table 4). The kappa value for the overall test is κ = .31,95% CI [.31, .32], with an agreement rate of 54.92% across a total of 36'955 classifications. For the extraction items only, the kappa value is also κ =

.31, 95% CI [.31, .32], while for the comparison items only, it is κ =.28, 95% CI [.26, .29], each showing a similar level of agreement. The prompt versions A to D reflect incremental refinements, including minor changes such as rewordings, adjustments, or deletions; version A is the original version, while version D represents the latest version. The different versions of the test data show minimal variation, with values in a similar range between κ =.24 and κ =.33.

Table 4

Interrater Agreement for different parts and versions

Test	**Interrater kappa**	**Percent Agreement (%)**	***N***
Full Test	.31 [.31, .32]	54.92	36’955
Extraction	.31 [.31, .32]	53.31	25’831
Comparison	.28 [.26, .29]	58.65	11’124
Version A	.28 [.27, .30]	51.50	3’590
Version B	.29 [.28, .30]	51.84	5’374
Version C	.24 [.23, .26]	48.29	3’703
Version D	.33 [.33, .34]	57.11	24’288

Note. Kappa values are shown with their 95% CI in brackets. Percent Agreement (%) indicates the proportion of cases for which the test outcomes matched the reference. *N* represents the number of comparisons included per test scenario.

File Size

Furthermore, the results in Table 5 indicate similar interrater kappa values depending on file size. The kappa values across the terciles, based on the file size of the uploaded preregistrations (Extract Preregistration) and the papers (Extract Paper) for the extraction tasks, and both file sizes combined (Match) for comparison, are all in a comparable range.

Table 5

Interrater kappa depending on file size

Terciles	Extract Preregistration	Extract Paper	Match
Small	.31 [.28, .34]	.32 [.29, .35]	.30 [.27, .33]
Medium	.36 [.33, .39]	.30 [.27, .32]	.25 [.22, .28]
Large	.29 [.26, .32]	.30 [.27, .33]	.28 [.25, .31]

Note. Interrater kappa values with 95% CI are presented across three terciles (Small, Medium, Large), based on the file size of the uploaded preregistrations (Extract Preregistration) and the papers (Extract Paper) for the extraction tasks, and both file sizes combined (Match) for comparison.

Experiment type

Further analysis examined whether the type of hypothesis influenced interrater reliability (Table 6). The kappa values vary depending on the experiment type but are generally at a similar level. The “Effect” type demonstrated the highest agreement with a kappa value of κ =.34, 95% CI [.33, .35]. Lower kappa values were observed for the Association type (κ =.21, 95% CI [.18, .23]) and the Moderated Association type (κ =.23, 95% CI [.14, .31]).

Table 6

Interrater Agreement for different types of hypotheses

Hypotheses Type	Interrater kappa	Percent Agreement (%)	*N*
Association	.21 [.18, .23]	46.90	4908
Interaction / moderated effect	.31 [.29, .33]	55.09	8730
Effect	.34 [.33, .35]	56.97	20684
Mediated effect	.30 [.26, .34]	54.11	2155
Moderated association	.23 [.14, .31]	48.95	478

Note. Interrater agreement is presented for different types of hypotheses. Kappa values (with 95% CI in brackets) indicate the level of agreement beyond chance, while Agreement Percent reflects the proportion of exact matches. *N* represents the number of comparisons analysed per category.

Interrater reliability per PSP

Interrater reliability between GPT-4o and human raters was also examined for each PSP individually. The average kappa value across all PSPs is κ =.31 (SD = .15). The kappa values vary widely, ranging from -.09 to .79. A visual inspection of the distribution of kappa values across all PSPs using a histogram, reveals an approximately symmetrical distribution. Additionally, the Shapiro-Wilk test indicated that the distribution of kappa values across all PSPs does not significantly deviate from normality, W = .997, p = .89. This suggests that the kappa values across all PSPs are likely normally distributed, with no specific PSP items exhibiting notably high or low kappa values.

Interrater reliability per item

Interrater reliability was also analysed at the item level. The average kappa value across all items is very low at κ =.07 (SD = .12) but ranges from κ =-.04 to κ =.82. Similarly, the average

percent agreement across all items is low at 54.73% (SD = 22.13), comparable to the total balanced accuracy. A visual inspection of the kappa value distribution using a Q-Q plot and histogram, as well as the Shapiro-Wilk test, indicated that the kappa values across all items significantly deviate from normality, $W = .72$, $p < .001$. Notably, in the upper quantiles, there appear to be items with higher kappa values than would be expected under a normal distribution.

The four items with the highest kappa show values ranging from $\kappa = .31$ to $\kappa = .45$ and pertain to questions on the manipulation of the first independent variable (IV1) in the preregistration or the published paper. The next best seven items, with kappa values between $\kappa = .20$ and $\kappa = .27$, relate to methodological aspects such as conducting the power analysis to determine sample size, handling missing data, consistency in the number of participants, and the definition of the dependent variable in the preregistration and the published paper.

The analysis of the two numerical items concerning the extraction of the planned number of participants from the preregistration and the actual number from the paper shows, that in 59% of cases, the figures match the reference dataset. In 21% of cases, differing participant numbers were reported, and in 19% of cases, GPT-4o invented a participant number where none existed in the reference dataset.

At the item level, correlations between quality indicators were further examined. First, no significant correlation was found between kappa and the percentage of GPT's most frequent, prototypical responses per item, $r(116) = -.14$, $p = .123$,95% CI [−.32,.04].

Second, a significant positive correlation was observed between the percentage of GPT's most frequent, prototypical responses per item and the percent agreement per item, $r(116) = .28$, $p = .002$,95% CI [.11,.44].

Third, no significant correlation was detected between kappa Values per Item and the percent agreement per item, $r(116) = .12$, $p = .189$, 95% CI [−.06,.30].

Minimal Version

The following section presents the results of the new minimal prompt version, which focuses on the comparison of selected studies and 9 specifically chosen key items. After excluding PSPs that were erroneously selected, 36 studies, along with 16 retest pairs, were included in the overall comparison. The reference dataset from human raters shows balanced distribution of responses. As shown in Table 7, 240 responses or 51% fall into the "Yes" category. In comparison, GPT-4o classifies the comparison of studies as consistent more frequently, categorising 76% of items (354 out of 468) as "Yes."

The analysis of GPTs ability to identify discrepancies between papers, specifically items rated as "N," yielded the following results: sensitivity was 64%, specificity was 56%, precision was 32% and the negative predictive value was 83%. The accuracy or agreement rate between the two datasets is 58%.

The Interrater kappa value across all 468 comparisons is κ =.15, 95% CI [.06, .24]. The best-performing item in the analysis concerned the consistency of the number of participants. For this item the interrater kappa value was κ =.31, 95% CI [.05, .57]. Test-retest reliability for the new prompt was calculated with 16 retest pairs and a total of 140 comparisons. The unweighted Cohen's kappa retest-reliability is κ =.27, 95% CI [.14, .40].

Table 7

Confusion Matrix for the Minimal Version

	Predicted Yes	**Predicted No**	**Sum**
True YES	199	41	240
True No	155	73	228
Sum	354	114	468

Note. Table 7 compares the absolute frequencies of GPT-4o predictions (Predicted) to the judgments made by human raters (True) for the Minimal Version including 52 PSP and 9 items.

Discussion

The primary aim of this study was to investigate whether the large language model GPT-4o by OpenAI can automatically detect deviations between preregistrations and publications in a sample of psychological studies and assess their strictness, meaning their level of detail and rigidity.

Preregistrations are intended to promote methodological rigour in the current replication crisis by establishing a clear distinction between a priori formulated predictions and post hoc adjustments that may arise from the results (HARKing) (Nosek et al., 2018). Empirical findings, however, suggest that this consistency is often low, with researchers in psychology and other scientific disciplines frequently deviating from their original preregistered plans without transparently disclosing these changes in their publication or because they fail to preregister all relevant information (van den Akker et al., 2024). Peer reviewers are expected to oversee this process; however, preregistrations are rarely subjected to thorough review during the peer-review process due to an already high workload in this system (Mathieu et al., 2013; Syed, 2023). Previous studies have demonstrated that LLMs like GPT-4 can support the demanding peer-review process by providing qualitative feedback and analysing complex texts (Verharen, 2023; Robertson, 2023). However, their application in detecting discrepancies and assessing the strictness of preregistrations remains unexplored. This study aims to address this gap by investigating, for the first time, the ability of LLMs to automate these specific aspects of the peer-review process.

Using various zero-shot prompts, a total of 300 studies and preregistrations were analysed to determine whether the most widely used LLM, the web-based user interface of ChatGPT-4o, could provide comparable assessments to those of human raters in the study by van den Akker et al. (2024). The two tasks—extracting information to assess strictness and comparing the document pairs—were divided into two distinct zero-shot prompts. One prompt contained 94 items for extraction, while the second prompt included 46 items for comparing the respective document pairs. The total of 140 items corresponded to the questions posed to human raters in van den Akker et al.'s (2024) study and some additional questions.

Examination of agreements and discrepancies between the GPT-4o test dataset and the human benchmark reference dataset from van den Akker et al. (2024) were performed by statistical analysis using Cohen's kappa coefficients and other measures. The results indicate low interrater reliability, with Cohen's kappa in the range of κ = .30, low accuracy of approximately 60%, and significant differences in categorisation. This demonstrates that, under current conditions and in this experimental setup, the performance of GPT-4o cannot yet match that of human raters. While test-retest reliability was acceptable (κ = .74), the low sensitivity and precision values — particularly for the most critical category, detecting discrepancies "N" — are insufficient for practical application. In the course of data collection, iterative refinements and modifications were applied to the prompts, resulting in multiple test versions. The results showed that the different versions exhibited minimal variation, with only slight improvements observed in the later versions. An alternative minimal approach, which employed a simpler comparison prompt, fewer items without dependencies, and selected studies, surprisingly yielded similarly low interrater reliability and accuracy. This further confirmed that GPT-4o significantly deviates from human ratings even in this simplified setup.

The subsequent sections offer detailed discussion of the results, explanations for the observed performance, and recommendations for improvement.

Retest reliability

For a synthetic peer review to be effective, it is essential that results are replicable across multiple iterations, indicating high test-retest reliability. The calculation of test-retest reliability using Cohen's kappa revealed substantial reliability (according to Landis & Koch, 1977) across all retest items (κ = .74). However, test-retest reliability varied considerably by item, ranging from perfect reliability for some items to unacceptably low levels for others.

A potential explanation for this substantial variance lies in the observed significant, moderate positive correlation between the test-retest kappa per item and the percentage of the most frequent, or prototypical, response per item. This means that items where GPT-4o frequently provided the same response, had higher test-retest kappa values. This could suggest a bias in GPT's responses, where it tends to default to a commonly selected category, thereby achieving higher test-retest reliability. Conversely, for items with a lower percentage of the most frequent response—and thus greater variability in answers—test-retest reliability was poorer.

Interrater reliability total

In addition to being used as a measure of test-retest reliability, kappa was also calculated to evaluate interrater reliability, with the responses from van den Akker et al. (2024) and GPT-4o each considered as a rater. In this context, Cohen's kappa served as a chance-corrected agreement measure and a quality metric for GPT's performance. Beyond the low overall kappa value (κ = .31), the data were further analysed at the item and PSP levels to identify potential influencing factors on the evaluation results and to detect differences in outcomes.

At the item level, interrater kappa remained similarly low regardless of whether the overall test, only the extraction items, or exclusively the comparison items were considered. This suggests that the type of task associated with the two different prompts (extraction vs. comparison) did not have a substantial impact on agreement accuracy. This finding is surprising, given that the extraction prompt contained more than twice as many items as the comparison

prompt. However, the greater complexity of the comparison task, which required accessing two separate documents rather than merely extracting information, could be a contributing factor.

Furthermore, the specific prompt version made little difference overall. While the prompt versions “D” exhibited higher kappa values compared to the initial version “A”, the difference was relatively minor and, despite adjustments, did not fall within an acceptable range.

Tests for normality of interrater kappa values across all items revealed significant deviations from a normal distribution, with certain items systematically exhibiting higher kappa values. The items with the highest interrater kappa values related to questions on the manipulation of the first independent variable (IV1) in the preregistration or the published paper, methodological aspects such as conducting power analyses to determine sample size, handling missing data, consistency in the number of participants, and the definition of the dependent variable in the preregistration and the published paper. Additionally, numerical items concerning the extraction of the planned number of participants from the preregistration and the actual number from the paper also demonstrated comparatively high kappa values.

Correlations and Interpreting magnitude

One factor that may partially explain the variation in agreement values across items is the association with patterns of the response frequency. A significant positive correlation was observed between the percentage of the most frequent, or prototypical, response per item in GPT’s ratings and the interrater percent agreement per item. This indicates that for categories where GPT-4o frequently provides a specific response (e.g., "Y"), there is also higher interrater agreement.

This could be interpreted as evidence that, in addition to the previously mentioned bias—where GPT-4o almost exclusively provides one response for certain items—there are indeed simpler items. These items are correctly and consistently assigned to a category by GPT-4o and human raters, resulting in higher percent agreement.

For example, the item "pap_dcp_1," which asks about the data collection procedure—specifically, whether the sample size is described in a reproducible manner in the paper—applies to almost all papers and is therefore straightforward to answer.

In contrast, no significant correlation was observed between the percentage of the most frequent, or prototypical, response ratings by GPT-4o per item and the interrater kappa per item. This result is surprising, but may be explained by the statistical properties of the kappa statistic, specifically its chance-correction mechanism. Items with one-sided response patterns from GPT-4o (e.g., predominantly "Y" responses) exhibit low variance and therefore high agreement by default, which can lead to lower kappa values.

For the same reason probably, no significant correlation was observed between percent agreement per item and interrater kappa per item. This indicates that high agreement between human raters and GPT-4o does not automatically correspond to high interrater kappa values. This calls into question the interpretative suitability of the kappa coefficient at the item level, as a significant relationship between the two measures would have been expected based on interrater reliability principles.

A possible explanation is the kappa paradox. This paradox arises when kappa, despite high agreement between raters, produces low values, potentially leading to incorrect conclusions about the absence of agreement (Zec et al., 2017). The kappa paradox challenges the assumption that higher agreement is necessarily associated with higher kappa values. This issue becomes particularly evident when there are substantial differences in category prevalence. Sensitivity analyses have shown that the paradox arises when one category is overwhelmingly dominant—either due to its inherent nature or because one rater consistently assigns more cases to that category (Zec et al., 2017).

Similar explanations are offered by Sim and Wright (2005) who demonstrate that factors beyond agreement itself can influence the magnitude of kappa, complicating its interpretation. As they note, two key factors are prevalence and bias. Prevalence refers to whether the codes are equally likely or vary significantly in probability. When codes are equiprobable, kappa values tend to be higher. Conversely, bias occurs when the marginal probabilities for two raters differ.

Its impact is more pronounced for lower kappa values than for higher ones (Sim & Wright, 2005).

The number of codes also affects kappa's behaviour. Simulation studies by Bakeman et al. (1997) showed that fewer codes result in lower kappa values for fallible raters, whereas a larger number of codes increases kappa. These findings align with observations that kappa values are higher when codes are roughly equiprobable, underscoring the influence of distribution characteristics on kappa's magnitude.

These issues are particularly relevant to the current study. The items exhibit highly diverse distributions of ratings, complicating item comparisons. It is therefore difficult to determine which items perform particularly well or poorly, as the results depend heavily on whether percent agreement or interrater kappa is used as the comparison metric. Moreover, in some items, both human raters and GPT-4o have shown a tendency to heavily favour one category over others. As demonstrated, this likely results in a paradoxical kappa. Additionally, the analysis relies on only a few codes (Y/N/X or just Y/N), which likely contributes to lower kappa values.

Despite these limitations, Cohen's kappa remains one of the most widely used measures of interrater agreement (Zec et al., 2017). Even though its magnitude guidelines (e.g., those proposed by Landis & Koch, 1977) have been frequently criticised as subjective and arbitrary. As an alternative to Cohen's kappa, Gwet's AC1 has been suggested as a more robust measure, particularly in addressing the kappa paradox (Zec et al., 2017) and should be considered for implementation in future studies.

Interrater reliability per PSP

Another important aspect of a usable synthetic peer review is, that it does not matter which preregistration-paper pair (PSP) is compared, or that any differences are at least well understood. Tests for normality showed, that interrater kappa values across all PSPs tend to be normally distributed, with no systematic outlier papers.

The examined PSPs exhibited notable differences in the length and size of the preregistrations and papers, as well as in the type of hypothesis. Although the uploaded documents contribute to prompt length — and prompt length strongly influences output quality (e.g., Yuen, 2024)—neither the size of the preregistration files nor that of the papers appeared to impact interrater kappa values, and thus agreement accuracy. However, the analysis did not account for images and graphics within the documents, which might artificially inflate file sizes.

The type of hypothesis, however, did influence interrater reliability, although overall agreement levels remained similar. Effect hypotheses demonstrated the highest agreement, while lower kappa values were observed for association hypotheses. However, these differences were small and may be explained by the tendency of effect hypotheses to be simpler in structure. For example, questions related to statistical procedures for effect hypotheses are often more straightforward, involving clear group comparisons (e.g., t-tests, ANOVA). In contrast, questions about association hypotheses are more prone to varying interpretations, leading to greater challenges in achieving agreement.

Minimal version

Due to the length of the prompts and the high complexity of the task, particularly with respect to dependencies, a single reduced minimal version, was ultimately developed. This new version aimed to achieve better results by focusing exclusively on comparisons. To this end, nine key items were carefully selected. Additionally, specific studies were chosen to ensure all items were applicable, eliminating dependencies and thereby reducing responses to two categories (Y/N).

Surprisingly, the accuracy of the new prompt (58%) was only slightly higher than that of the old, longer prompt (52%). This indicates that for both versions, nearly half of the predictions were incorrect. Interrater reliability for the new prompt was even lower (κ = .15) compared to the old prompt (κ = .31). While comparing kappa values can be misleading due to the

aforementioned limitations of the measure, the results are nevertheless surprising, as the modifications were expected to lead to significantly improved performance.

Bias

With three categories in the longer, full version, there was a clear tendency to overclassify the category "X," which was associated with high sensitivity (89%) but low specificity (60%), resulting in a high false positive rate for category "X." The most important category, "No," exhibited the weakest performance, with sensitivity at 30% and precision at 16%, indicating frequent misclassifications. Similarly, the "Yes" category also showed weaknesses in sensitivity (37%), though it was recognised more accurately overall (with a precision of 84%) compared to "No."

With two categories in the minimal version, analyses of the confusion matrix and related metrics revealed a strong tendency of the model to favour the "Y" class. Actual cases belonging to the "N" class were frequently misclassified as "Y," as reflected in a low specificity (56%) and poor precision (32%) for the "N" category. The sensitivity for "N" was 64%, further highlighting deficiencies in recognition.

Given that the results reveal different biases depending on the prompt, it cannot be concluded that the model consistently favours one particular category. However, the model tends to be less strict than human raters, often overestimating agreement and being overly lenient in its assessments.

Causes and improvements

Overall, the results indicate inadequate performance of GPT-4o as a synthetic peer reviewer in automatically detecting discrepancies and assessing strictness. Several reasons may account for this:

First, a possible explanation lies in the fundamental architecture of the model, which may not be capable of meeting the requirements of scientific peer review. LLMs such as GPT-4o operate probabilistically and are designed to recognise and predict text patterns rather than conduct logical analyses or precise comparisons. They lack scientific understanding and face

challenges in selectively extracting relevant information from lengthy texts. These limitations may render GPT-4o fundamentally unsuitable for complex tasks such as the meticulous comparison of documents required in scientific peer reviews.

Second, although document size did not correlate significantly with model performance in this study, the length of the documents under review may still play a role, as suggested by research from Yuen et al. (2024). Scientific articles are often lengthy and contain extensive details, tables, and figures. Human reviewers navigate documents strategically, skipping irrelevant sections and focusing on essential content—a strategy that GPT-4o cannot systematically employ. A potential improvement to GPT's performance might involve pre-processing the articles by shortening them or filtering out irrelevant sections, enabling GPT-4o to focus on key content. Specifically, only the sub-study relevant to the specific question could be included in the analysed documents.

Third, the set of 140 questions that GPT-4o is required to answer, presents a highly complex task. Some items required additional verification for applicability and dependencies, necessitating further processing steps. Language models like GPT-4o may still encounter limitations in handling such complexity. Although this difficulty was partially addressed with the adjusted minimal version, achieving success in future studies may require an even more precise focus on specifically selected items. Research by Li et al. (2023) aligns with these suggestions, indicating that ChatGPT performs well on relatively simple information extraction tasks but struggles with more complex and challenging ones.

Fourth, in this context, the chosen zero-shot prompting approach may not be the most suitable option, because language models like GPT-4o as few-shot learners benefit from examples (Brown et al., 2020). Although an example was provided to illustrate the output format, initial attempts with few-shot prompting—including example texts and responses to the questions—proved impractical. The additional documents and responses would have significantly extended the context window, making the approach unfeasible. However, with a

reduced number of questions and shortened texts, few-shot prompting could potentially be more effective.

Fifth, the temperature setting plays a crucial role in controlling randomness during text generation by LLMs. Lowering the temperature favours words with higher probabilities while reducing the selection of less likely words (Zhao et al., 2024). This adjustment can increase the consistency and, consequently, the test-retest reliability of the generated results. Additionally, lower temperatures result in more rigid and accurate token choices. While this functionality is available in the API version of OpenAI, it is absent in the browser version, making it difficult to modulate randomness and improve reproducibility. Consequently, the output format was not always uniform or identical, sometimes exhibiting slight variations in formatting, annotations, or explanations.

A sixth reason for the observed accuracy may relate to the predefined categories. While such an approach was necessary for quantitative comparison with the human dataset, initial trials with prompt engineering already revealed that GPT-4o struggles to adhere to rigidly predefined response formats. This aligns with research by Li et al. (2023) on ChatGPT's information extraction capabilities, which indicates that GPT performs poorly when provided with predefined labels but performs better when required to generate predictions without predefined labels, relying only on task understanding and input.

This was also evident in qualitative annotations, which were sometimes offered as justifications for specific category assignments by GPT-4o. These annotations occasionally provided meaningful explanations that could be helpful in the peer review process. For example: "*Addition of Bayesian inference not mentioned in preregistration. Assessment: N (Procedure detailed only in the published study).*"

Moreover, GPT-4o sometimes struggled with consistent categorisation, even when the underlying information appeared reasonable. For instance, small deviations in the number of participants were sometimes categorised as discrepancies ("N") and at other times as agreements ("Y"). Examples include:

"*PREREG: The planned sample size is N = 395 participants, as determined by a Power analysis for sufficient power. PUBSTUD: Final sample size is N = 422 participants, exceeding the preregistered sample size. Assessment: Y (The published study includes a larger sample size, which is acceptable and not contradictory.)*"

"PREREG: Minimum of 100 participants per condition. PUBSTUD: 324 participants in total. Assessment: N (different total sample size)."

Such qualitative justifications for assessments were not systematically considered or evaluated for correctness in this study. According to Li et al. (2023), however, ChatGPT's explanations for its decisions are generally of high quality and reliable.

Strengths

Beyond these limitations, the current study has several strengths:

First, the use of GPT-4o, as one of the most commonly employed models, enhances the generalisability of the findings for a broader user base. Potentially, GPT-4o is already being used by researchers for synthetic peer reviews. By demonstrating its poor performance, this study highlights the risks associated with its use and the potential for harmful misinformation (compare also Ateia & Kruschwitz, 2024).

Second, the results were analysed and compared across multiple levels, including item level, PSP level, different prompt versions, retest reliability, associations with file size, and other dimensions. This multi-faceted approach provides a nuanced understanding of the model's performance for this task.

Third, the chosen human benchmark is a highly respected source for comparison. And as a large dataset with numerous PSPs and items, it offers a reliable comparison for evaluating actual performance.

Fourth, although the generalisability is temporally limited to the current GPT-4o model and methodologically to the chosen approach, this study is the first to investigate whether LLMs can automatically detect deviations from preregistrations and assess their strictness, an essential aspect of peer review. It provides a comprehensive comparison for future research with other models and serves as an example of a methodological framework.

Conclusion

The LLM GPT-4o was unable to match the evaluations of human raters from the study by van den Akker et al. (2024) in its task of automatically assessing strictness and identifying discrepancies. The methodology used in the present investigation, which involved answering a wide range of items, accounting for dependencies, and selecting answers from predefined categories, was intended to ensure comparability with the reference dataset but went beyond the primary core task. These requirements posed significant challenges for the automatic assessment by LLM and were not fully adapted to GPT-4o's strengths. It is therefore likely that following alternative approaches such as few-shot prompting, open-ended questions without predefined categories, and curated or shortened papers, improved results might have be achievable.

Despite its insufficient over all accuracy and frequent occurrence of errors or hallucinations, it can be noted that GPT-4o achieved good results for certain items or aspects of the task. Specifically, its ability to extract participant numbers and provide detailed text-based justifications for ratings could be valuable in a peer review process, in particular in terms of time efficiency. While human raters typically required 20 to 80 minutes to code a single PSP, with complex pairs taking several hours (van den Akker et al., 2024), GPT-4o completed each PSP in just over two minutes in this study.

However, the results demonstrate that direct application of this approach is not advisable. It remains unclear which items are answered correctly, and there is a lack of transparency regarding the sources of information—for example, which parts of the study were used or where hallucinations occurred. This raises the risk of automation bias, the tendency to follow suggestions from automated decision-making systems and place undue trust in technology while ignoring contradictory but correct information provided independently of automation.

Further arguments and problematic aspects, as highlighted by Hosseini and Horbach (2023), should also be considered when introducing synthetic peer reviewers. These include issues with reproducibility, opacity in model training and data usage, the risk of amplifying existing biases, and threats to confidentiality when handling sensitive data. Additionally, the social and epistemic functions of peer review—such as dialogue about research norms and values—could be undermined if LLMs were to replace human reviewers (Hosseini & Horbach, 2023). Other ethical concerns, such as the substantial energy and water consumption of LLMs, also provide arguments against their widespread use.

Regarding the generalisability of the present evaluation, it has to be noted, that the findings reflect only this specific model, at this time, under this particular study setup, and with this prompt version. As shown in prior research, GPT's performance changes over time, hindering reproducible results (Chen et al., 2023). Furthermore, modifying the input context or adjusting the sequence of presented information within a prompt, significantly influences the structure and quality of the output (Lu et al., 2021, as cited in Yuen et al., 2024). Nevertheless, the results seem promising enough to suggest that with further investigation and developing an optimised approach - such as alternative prompts, low temperature with the API version, or improved and larger models - a viable solution for automatically assessing deviations from preregistrations in synthetic peer review may be achieved in the future.

References

Ateia, S., & Kruschwitz, U. (2024). Can open-source llms compete with commercial models? Exploring the few-shot performance of current gpt models in biomedical tasks (arXiv:2407.13511). arXiv. https://doi.org/10.48550/arXiv.2407.13511

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners (arXiv:2005.14165). arXiv. https://doi.org/10.48550/arXiv.2005.14165

Bsharat, S. M., Myrzakhan, A., & Shen, Z. (2023). Principled instructions are all you need for questioning LLAMA-1/2, GPT-3.5/4. *arXiv*. https://doi.org/10.48550/arxiv.2312.16171

Chen, H., Jiao, F., Li, X., Qin, C., Ravaut, M., Zhao, R., Xiong, C., & Joty, S. (2023). Chatgpt's one-year anniversary: are open-source large language models catching up?. arXiv. https://doi.org/10.48550/arxiv.2311.16989

Elson, M., Huff, M., & Utz, S. (2020). Metascience on peer review: Testing the effects of a study's originality and statistical significance in a field experiment. *Advances in Methods and Practices in Psychological Science*, *3*(1). https://doi.org/10.1177/2515245919895419

Fahmy Amin, M. (2023). Confusion matrix in three-class classification problems: A step-by-step tutorial. *Journal of Engineering Research, 7*(1). https://doi.org/10.21608/erjeng.2023.296718

Hahn, L., Glöckner, A., Gollwitzer, M., Hellmann, J. H., Lange, J., Schindler, S., & Sassenberg, K. (2024). More than box-ticking? Assessing preregistration quality in psychological research. https://doi.org/10.31219/osf.io/wc7qr

Hommel, B. E., Wollang, F. M., Kotova, V., Zacher, H., & Schmukle, S. C. (2021). Transformer-based deep neural language modeling for construct-specific automatic item generation. *Psychometrika*, *87*(2), 749–772. https://doi.org/10.1007/s11336-021-09823-9

Hosseini, M., & Horbach, S. P. J. M. (2023). Fighting reviewer fatigue or amplifying bias? Considerations and recommendations for use of chatgpt and other large language

models in scholarly peer review. *Research Integrity and Peer Review, 8*(1), 4. https://doi.org/10.1186/s41073-023-00133-5

Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification (arXiv:1801.06146). arXiv. https://doi.org/10.48550/arXiv.1801.06146

Irfanullah, H. (2023, September 29). *Ending Human-Dependent Peer Review*. The Scholarly Kitchen. https://scholarlykitchen.sspnet.org/2023/09/29/ending-human-dependent-peer-review/

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310

Li, B., Fang, G., Yang, Y., Wang, Q., Ye, W., Zhao, W., & Zhang, S. (2023). Evaluating chatgpt's information extraction capabilities: An assessment of performance, explainability, calibration, and faithfulness (arXiv:2304.11633). arXiv. https://doi.org/10.48550/arXiv.2304.11633

Liang, W., Zhang, Y., Cao, H., Wang, B., Ding, D., Yang, X., Vodrahalli, K., He, S., Smith, D., Yin, Y., McFarland, D., & Zou, J. (2023). Can large language models provide useful feedback on research papers? A large-scale empirical analysis (arXiv:2310.01783). arXiv. https://doi.org/10.48550/arXiv.2310.01783

Mathieu, S., Chan, A.-W., & Ravaud, P. (2013). Use of trial register information during the peer review process. *PLoS ONE, 8*(4), e59910. https://doi.org/10.1371/journal.pone.0059910

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences, 115*(11), 2600–2606. https://doi.org/10.1073/pnas.1708274114

OpenAI. (2024, May 13). *Hello, GPT-4o*. OpenAI. https://openai.com/index/hello-gpt-4o/

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Robertson, Z. (2023). Gpt4 is slightly helpful for peer-review assistance: A pilot study (arXiv:2307.05492). arXiv. https://doi.org/10.48550/arXiv.2307.05492

Sim, J., & Wright, C. C. (2005). The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Physical Therapy, 85*(3), 257–268. https://doi.org/10.1093/ptj/85.3.257

Stodden, V., Seiler, J., & Ma, Z. (2018). An empirical analysis of journal policy effectiveness for computational reproducibility. *Proceedings of the National Academy of Sciences, 115*(11), 2584–2589. https://doi.org/10.1073/pnas.1708290115

Syed, M. (2023). Some data indicating that editors and reviewers do not check preregistrations during the review process. https://doi.org/10.31234/osf.io/nh7qw

van den Akker, O. R., Bakker, M., Van Assen, M. A. L. M., Pennington, C. R., Verweij, L., Elsherif, M. M., Claesen, A., Gaillard, S. D. M., Yeung, S. K., Frankenberger, J.-L., Krautter, K., Cockcroft, J. P., Kreuer, K. S., Evans, T. R., Heppel, F. M., Schoch, S. F., Korbmacher, M., Yamada, Y., Albayrak-Aydemir, N., … Wicherts, J. M. (2024). The potential of preregistration in psychology: Assessing preregistration producibility and preregistration-study consistency. Psychological Methods. https://doi.org/10.1037/met0000687

van den Akker, O. R., Van Assen, M. a. L. M., Enting, M., De Jonge, M., Ong, H. H., Rüffer, F., Schoenmakers, M., Stoevenbelt, A. H., Wicherts, J. M., & Bakker, M. (2023). Selective hypothesis reporting in psychology: Comparing preregistrations and corresponding publications. *Advances in Methods and Practices in Psychological Science*, *6*(3). https://doi.org/10.1177/25152459231187988

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. arXiv, 30, 5998–6008. https://arxiv.org/pdf/1706.03762v5

Verharen, J. P. (2023). ChatGPT identifies gender disparities in scientific peer review. *eLife, 12*, RP90230. https://doi.org/10.7554/eLife.90230.3.sa0

Yuen, T., Watt, G. A., & Lawryshyn, Y. (2024). Assisting humans in complex comparisons: Automated information comparison at scale (arXiv:2404.04351). arXiv. https://doi.org/10.48550/arXiv.2404.04351

Zec, S., Soriani, N., Comoretto, R., & Baldi, I. (2017). High agreement and high prevalence: The paradox of cohen's kappa. *The Open Nursing Journal, 11*(1), 211–218. https://doi.org/10.2174/1874434601711010211

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., … Wen, J.-R. (2024). A survey of large language models (arXiv:2303.18223). arXiv. https://doi.org/10.48550/arXiv.2303.18223

Selbstständigkeitserklärung:

„Ich erkläre hiermit, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen benutzt habe. Alle Stellen, die wörtlich oder sinngemäss aus Quellen entnommen wurden, habe ich als solche gekennzeichnet. Als Hilfsmittel habe ich Künstliche Intelligenz verwendet. Sämtliche Elemente, die ich von einer Künstlichen Intelligenz übernommen habe, werden als solche deklariert, und es finden sich die genaue Bezeichnung der verwendeten Technologie sowie die Angabe der «Prompts», die ich dafür eingesetzt habe. Mir ist bekannt, dass andernfalls die Arbeit mit der Note 1 bewertet, wird bzw. der Senat gemäss Artikel 36 Absatz 1 Buchstabe r des Gesetzes vom 5. September 1996 über die Universität zum Entzug des auf Grund dieser Arbeit verliehenen Titels berechtigt ist. Für die Zwecke der Begutachtung und der Überprüfung der Einhaltung der Selbständigkeitserklärung bzw. der Reglemente betreffend Plagiate erteile ich der Universität Bern das Recht, die dazu erforderlichen Personendaten zu bearbeiten und Nutzungshandlungen vorzunehmen, insbesondere die schriftliche Arbeit zu vervielfältigen und dauerhaft in einer Datenbank zu speichern sowie diese zur Überprüfung von Arbeiten Dritter zu verwenden oder hierzu zur Verfügung zu stellen."

Verwendete KI-Hilfsmittel:

- ChatGPT Version 4o wurde insbesondere für die Unterstützung bei der Erstellung des R-Skripts verwendet. Genauere Angaben zu diesen Prompts finden sich in den betreffenden R-Files.
- ChatGPT Version 4o wurde ausserdem für die Unterstützung bei Übersetzungen und Umformulierung verwendet. Verwendete Prompts waren beispielsweise: *“Translate the following phrase into scientific English.”* oder *“What is an alternative formulation for the following expression?”*

Ort / Datum: Bern, 06.01.2025

Unterschrift: