Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Automated reproducibility assessments in the social and behavioral sciences using large language models

Stefan Feuerriegel

feuerriegel@lmu.de

LMU Munich https://orcid.org/0000-0001-7856-8729

Tobias Holtdirk

LMU Munich

Pietro Marcolongo

LMU Munich

Anna Steinberg Schulten

LMU Munich

Felix Henninger

LMU Munich https://orcid.org/0000-0002-7730-9511

Stefan Rose

University of Cologne https://orcid.org/0000-0002-1327-0830

Sarah Ball

LMU Munich

Bolei Ma

LMU Munich

Frauke Kreuter

LMU Munich

Markus Weinmann

University of Cologne https://orcid.org/0000-0002-8342-2756

Article

Keywords:

Posted Date: August 5th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10313775/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

Automated reproducibility assessments in the social and behavioral sciences using large language models

Tobias Holtdirk[1,2,†], Pietro Marcolongo[1,2,†], Anna Steinberg Schulten[1,2], Felix Henninger[1,2], Stefan Rose[3], Sarah Ball[1,2], Bolei Ma[1,2], Frauke Kreuter[*,1,2,3,‡], Markus Weinmann[*,4,‡], and Stefan Feuerriegel[*,1,2,‡]

[1]LMU Munich, Munich, Germany

[2]Munich Center for Machine Learning, Munich, Germany

[3]University of Maryland, College Park, USA

[4]University of Cologne, Cologne, Germany

[†]Joint first authorship

[‡]Joint last authorship

[*]Correspondence: frauke.kreuter@stat.uni-muenchen.de, weinmann@wiso.uni-koeln.de, feuerriegel@lmu.de

Abstract

Reproducibility in the social and behavioral sciences is typically evaluated by independent researchers who reanalyze the original data to assess whether the published findings can be recovered. However, such approaches are resource-intensive and difficult to scale. Here, we show that large language models (LLMs) can automate reproducibility assessments. Using $N = 180$ published studies with predefined claims from the behavioral and social sciences, we compare LLM-generated analyses with the original findings. For 11 studies, the LLM pipeline could not produce a viable effect size estimate. For the remaining studies, the LLM reached the same qualitative conclusion as the original study in 80% of cases, and recovered the original effect sizes (using a ± 0.05 tolerance in Cohen's d) in 24% of studies. In a subset with human reanalyses, the LLM reached the same qualitative conclusion as the original study in 95% of studies, similar to human reanalysts (83%), and the LLM recovered the original effect sizes using a ± 0.05 tolerance in 40% of studies, again broadly similar to human reanalysts (28%). Given the current capabilities and limitations of LLMs, the findings show that LLMs can support systematic audits of empirical results rather than substitute expert judgment. As such, LLMs can serve as a scalable screening tool to improve the rigor and reproducibility in empirical research.

Main

Scientific progress depends on reproducible findings. Yet, across the social and behavioral sciences, reanalyses of original data have shown that published results are not always recovered [1–5]. These patterns make it difficult to know which published findings can be trusted, motivating large-scale efforts to reanalyze published studies and check whether their central findings can be reproduced from the original data. Ideally, journals would reanalyze submissions before publication to catch irreproducible results early, and some have started to do so [1, 4, 6, 7].

However, reanalyzing published studies is difficult to scale because it requires a time-intensive reconstruction of the original empirical workflow [1–5]. Unlike replication efforts, which test whether a finding holds in newly collected data [8–10], reproducibility concerns whether a reported result can be recovered from the original data and study materials [9, 11].[1] For this, analysts must inspect the study materials, identify the relevant files and variables, prepare the data, translate the paper's description into executable code, run the analysis, and compare the resulting statistic with the original finding. The scale of this effort is demonstrated by the recent Multi100 collaboration, wherein 507 analysts reanalyzed 100 published studies in economics, political science, and psychology over several years [2]. Overall, reproducibility assessments are resource-intensive and study-specific, which makes systematic auditing of the empirical literature costly and difficult to scale.

One way to make reproducibility assessments scalable is to automate the reanalysis itself. Recent advances in large language models (LLMs) have produced systems with strong coding capabilities [12] and growing use in scientific programming, data analysis, and research assistance [13–26]. Here, we propose using LLMs as automated analysts for reproducibility assess-

[1] We use "reproducibility" throughout in this broad sense, spanning both close reimplementation and claim-level robustness, and our design varies how much of the original methodology the model receives. The two ends place different demands on the analyst and imply different benchmarks: the original reported result for close reproduction, and the distribution of defensible reanalyses for robustness.

ment. In principle, such systems can process study materials and generate executable analysis code, which could reduce the extensive manual effort currently required for data and reproducibility checks [6] and make such checks feasible across a broader range of publication outlets as part of routine quality control. However, LLM-based reanalysis is challenging, since the generated code may contain errors, it is unclear how accurately LLMs can interpret the analytical decisions implied by the study materials, and outputs may reflect unsupported assumptions about the data [27]. Hence, whether LLMs can reliably carry out the computational reproducibility assessments is unknown.

Here, we examine whether LLMs can enable automated reproducibility assessments in the social and behavioral sciences (see Fig. 1). We develop an agentic LLM pipeline in which the model receives the original dataset, a focal statistical claim, and experimentally varied contextual information from the article, and is tasked to independently write and execute statistical code to reproduce the claims. Across a preregistered corpus of published empirical studies from psychology, political science, and economics, we address the following research questions. (1) To what extent can LLM-generated analyses reproduce the statistical results and substantive conclusions of published studies? To this end, we compare standardized effect sizes (Cohen's d) with the original published findings and further assess whether the LLM recovers the same substantive conclusion. (2) How do the LLM-generated reproducibility analyses compare against human reanalyses? For a subset of papers, we additionally compare the LLM-generated effect sizes against those from a large-scale human reanalysis effort [2]. This comparison is descriptive rather than a direct performance benchmark; the human reanalysts may have pursued their own defensible specifications rather than aiming to reproduce the originally reported result, whereas the LLM pipeline was instructed to test the focal claim using the paper and available data. Therefore, higher agreement between the LLM and the original findings may reflect a narrower or more paper-anchored analytical path rather than greater accuracy relative to human reanalysis. (3) How does reproducibility vary with the amount of methodological information provided to the model, that is, with access

to the full method description versus a more conceptual reanalysis where only the abstract of the paper is provided? For our main analysis, we use Claude Opus 4.7 as a state-of-the-art LLM for code generation, and demonstrate the generalizability using two additional frontier models, namely, GPT-5.5 (from OpenAI) and GLM-5.1 (an open-weight model from Zhipu).

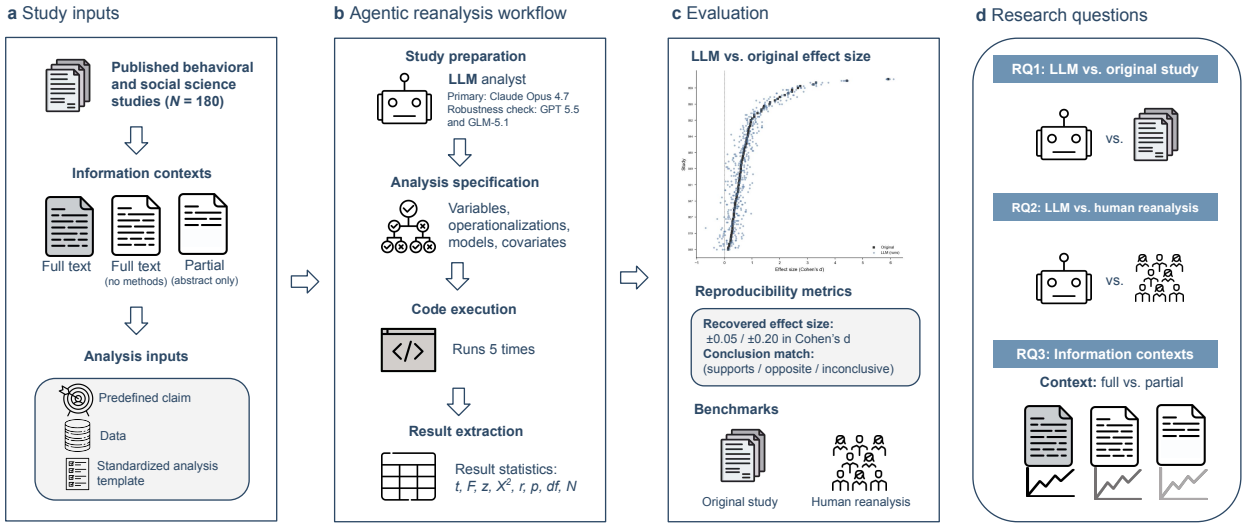


Figure 1: **Automated reproducibility assessment using LLMs. a,** A corpus of published studies ($N = 180$) with predefined claims, datasets, and standardized analysis templates is provided across varying information contexts (full text, full text without methods, or abstract only). **b,** The agentic LLM pipeline processes study materials, specifies analytical choices (including variables, operationalizations, models, and covariates), and extracts the resulting statistics (over five independent runs). **c,** Extracted effect sizes and conclusions are evaluated against the reported results (from the original study) using several metrics to assess the reproducibility, namely, whether the effect size is recovered (within a ± 0.05 and ± 0.20 Cohen's d tolerance) and whether the qualitative conclusions match (i.e., support / opposite / inconclusive). **d,** We analyze three main research questions: (1) the overall reproducibility of findings, (2) the reproducibility compared to human analysts, and (3) the analytical variability due to different information contexts.

Results

Study overview

We analyzed $N = 180$ studies from the Systematizing Confidence in Open Research and Evidence (SCORE) project [28]. SCORE is a large-scale research initiative in the social and behavioral sciences in which published empirical claims were identified and independently evaluated. We used studies for which focal claims had already been extracted and for which the original study data had been retrieved and made available, or for which data from a replication effort were available. We evaluated the automated reanalysis using the following metrics [2]: (1) the difference in effect size compared to the original finding, defined as $\Delta d = \bar{d}_{\mathrm{LLM}} - d_{\mathrm{original}}$, where d_{original} is the original effect size and where $\bar{d}_{\mathrm{LLM}}$ is the mean effect size of the LLM reanalysis over five runs, (2) whether the result fell within a tolerance region of ± 0.05 (strict) or ± 0.20 (broad), and (3) whether the substantive conclusion matched the original claim (classified as support / opposite / inconclusive; based on the majority vote across runs).

We evaluated each study using the agentic LLM pipeline across five independent runs to address variability in LLM outputs. In each run, the model received the focal claim, the available study materials (i.e., the entire paper), and the corresponding data as input, and was asked to write and execute statistical code to estimate the effect directly tied to the claim. The model was allowed to make reasonable analytical choices where the claim or study materials were underspecified, but we included safeguards to prevent direct copying from the paper; for example, the model was instructed to compute the statistic from code executed on the data, not to reuse reported test statistics, and to return a structured output containing the estimated statistic, sample size, degrees of freedom, p-value, effect size information, and qualitative conclusion. The LLM pipeline was allowed to install additional software packages, if needed. Claude Opus 4.7 served as the primary LLM, while we later also test the generalizability using GPT-5.5 and GLM-5.1.

Overall, 11 studies were excluded because none of the five runs yielded a valid Cohen's d, yielding $N = 169$ papers for our analysis.

LLMs can automate reproducibility assessment (RQ1)

The distribution of estimated effect sizes based on the LLM reanalysis is shown in Fig. 2a. The difference in effect size compared to the original finding is shown in Fig. 2b. The distribution is highly right-skewed, with most studies showing relatively small deviations from the original effect size but a small number of studies exhibiting large discrepancies. We next compare the mean effect size from the LLM reanalysis to the original published effect size (Fig. 2c). The LLM reanalysis fell within the strict tolerance of ± 0.05 for 24% of the studies (22% when treating studies without a valid LLM effect size estimate as non-recovered). Using the broader tolerance of ± 0.20, the LLM reanalysis fell within tolerance for 50% of the studies (47% when treating studies without a valid LLM effect size estimate as non-recovered; Supplementary Fig. S3). Together, these results show that the LLM pipeline was able to recover the original findings in a substantial share of cases.

We further examined whether reproducibility differed between studies using the original source data (named 'reproduction data' in SCORE) versus studies where the original source data were unavailable but where replication data were made available for reanalysis (Fig. 2d,e). Reproducibility was higher for studies using source data than for studies using replication data, consistent with prior findings that replication studies collecting new data to test whether an effect holds have repeatedly recovered weaker effects than originally reported [8, 10].

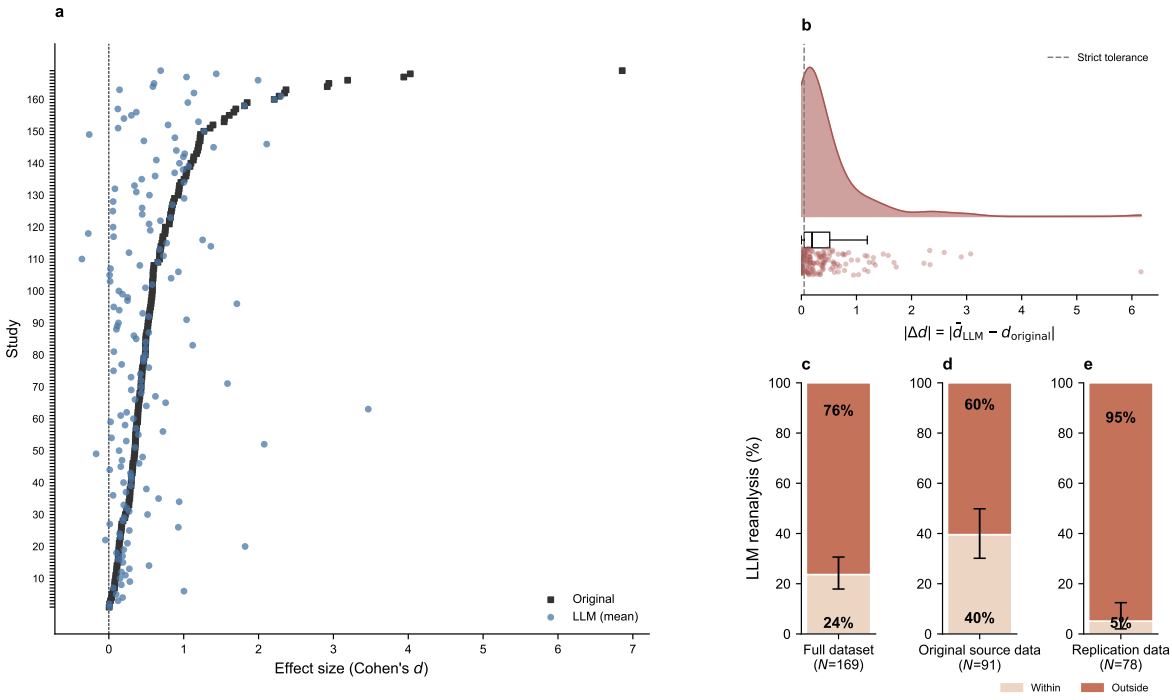


Figure 2: **Automated reproducibility assessment (using Claude Opus 4.7). a,** Effect size of the original analysis (gray squares; all represented as positive values) and the effect sizes of the reanalyses (blue dots) for each study. Shown are the $N = 169$ studies for which a valid Cohen's d was produced by the LLM (while 11 studies were excluded for that reason). **b,** Distribution of $|\Delta d|$, computed as $|\Delta d| = |\bar{d}_{\text{LLM}} - d_{\text{original}}|$. The distribution is visualized using a density plot, a boxplot, and scatter points. In the boxplot, the line indicates the median, the box denotes the interquartile range (IQR), and the whiskers extend to 1.5× of the IQR; points beyond the whiskers indicate outliers. **c,** Proportion of studies for which the LLM-generated effect size falls within or outside the tolerance region around the original result, across the full sample. **d,** Proportion of studies falling within or outside the tolerance region, for the subset of studies that make the original source data available. **e,** Proportion of studies for which the LLM-generated effect size falls within or outside the tolerance region, for the subset of studies where the original source data were unavailable but where replication data for reanalyses are provided. Whiskers indicate Wilson 95% confidence intervals.

Comparison of LLM-powered vs. human reanalyses (RQ2)

Next, we compare the mean effect size obtained from the LLM reanalysis to the effect size from human reanalyses. We thus focus on the subset of $N = 84$ papers for which such human benchmarks have been generated over the course of multiple years as part of a large-scale human reanalysis effort [2]. Out of the $N = 84$ studies, the LLM pipeline did not produce a valid Cohen's d for two studies, leaving $N = 82$ studies for the comparison.

For this subset of studies, the LLM reanalysis fell within the strict tolerance of ± 0.05 in 40% of studies (Fig. 3a). This is higher than the corresponding rate in the full sample and is explained by the composition of the subset, in which the original source data were available for most studies. We then compared the LLM reanalysis with the human reanalysis benchmark. The human reanalyses fell within the same strict tolerance of the original effect size in 28% of studies (Fig. 3c). Using the broader ± 0.20 tolerance, the LLM reanalysis fell within tolerance in 65% of studies, compared with 66% for the human reanalyses. The substantive conclusion from the LLM reanalysis matched the original claim in 95% of studies, while the substantive conclusion from the human reanalysis matched it in 83%.

To understand whether the LLM tends to recover the effect sizes of the original paper vs. the human reanalyses, we plotted the LLM-derived effect sizes against the original published effect sizes (Fig. 3b) and against the human reanalysis effect sizes (Fig. 3e). The correlation between the LLM reanalysis and the original study effect sizes was moderate ($r = 0.46$; $p < 0.001$; Fig. 3b), indicating that the LLM captured part of the overall variation in the original effects despite substantial study-level variation. By contrast, the correlation between the LLM-derived effect sizes and the human reanalysis effect sizes was weak ($r = 0.11$; $p = 0.31$; Fig. 3e), suggesting that the LLM estimates aligned more closely with the original published effects than with the human reanalysis estimates.

We next assessed whether the LLM-derived effect sizes tracked variation in the original pub-

lished effect sizes and in the human reanalysis effect sizes. To do so, we estimated study-level linear regressions. Regressing the original published effect size d_{original}, as dependent variable, on the mean LLM-derived effect size $\bar{d}_{\text{LLM}}$, as independent variable, yielded a slope of $\beta = 0.96$ (SE = 0.21, 95% CI [0.55, 1.37], $p < 0.001$) and explained 21% of the variance ($R^2 = 0.21$). For comparison, regressing the human reanalysis effect size $\bar{d}_{\text{human}}$, as dependent variable, on the mean LLM-derived effect size $\bar{d}_{\text{LLM}}$, as independent variable, yielded a slope of $\beta = 0.53$ (SE = 0.52, 95% CI [−0.50, 1.56], $p = 0.31$) and explained 1% of the variance ($R^2 = 0.01$). This indicates that the LLM-derived effect sizes tracked the original published effect sizes more closely than the human reanalysis effect sizes.

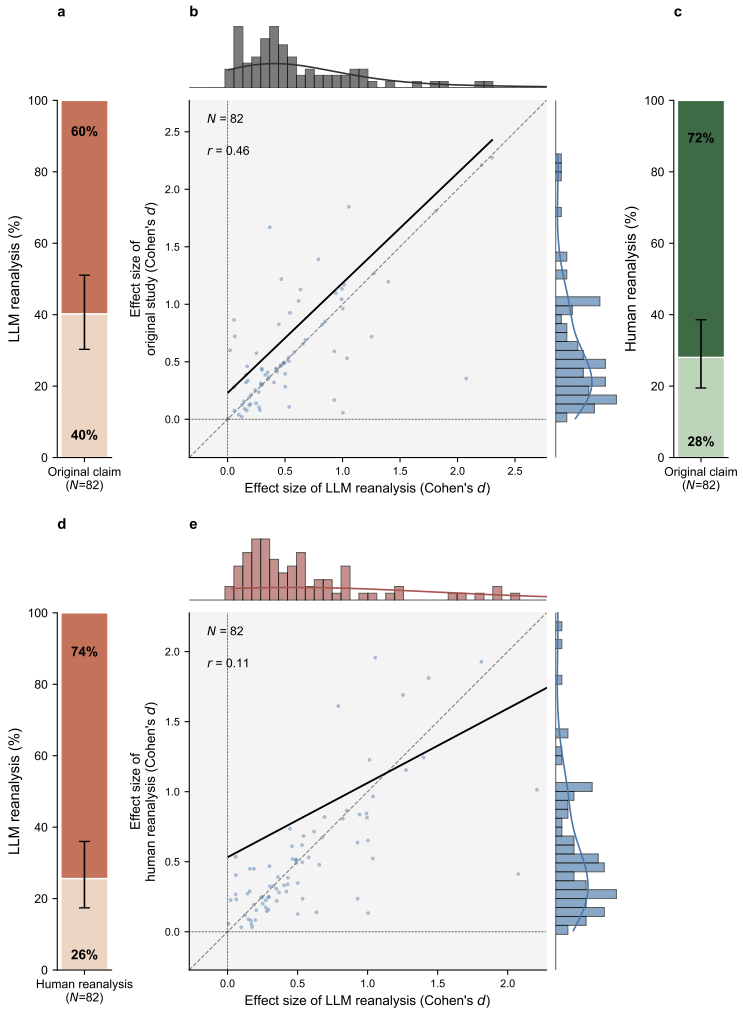


Figure 3: **LLM reproducibility compared with human reanalyses.** We compare Cohen's d from the automated LLM reanalysis with the original published effect sizes (**a,b**) and with effect sizes from human reanalyses (**d,e**). Here, we focus on the subset of studies for which human reanalysis benchmarks were available from a large-scale reanalysis effort [2]. To compare the LLM performance, we also report the human reanalysis results for recovering the original effect sizes (**c**). Results ($N = 82$ studies) are based on Claude Opus 4.7. **a,** Proportion of studies for which the LLM-generated effect size falls within or outside the tolerance region around the original published effect size. **b,** Scatterplot comparing the Cohen's d from the LLM reanalysis against the Cohen's d from the original study. The thin diagonal line represents the ideal case in which the reanalysis effect sizes are equal to the original effect size. The thick line is the estimated trend line (r is the Pearson correlation coefficient). Density plots on the axes show the respective distribution of the effect sizes. **c,** Proportion of studies for which the human reanalysis effect size falls within or outside the tolerance region around the original published effect size. **d,** Proportion of studies for which the LLM-generated effect size falls within or outside the tolerance region around the human reanalysis. **e,** Scatterplot comparing Cohen's d from the LLM reanalysis against the Cohen's d from the human reanalyses (averaged over all reanalyses per each study). Whiskers indicate Wilson 95% confidence intervals.

Multi-analyst and multiverse studies have shown that different defensible analytical choices can yield substantial variation in effect size estimates, even when analysts start from the same data and hypothesis [2, 29–32]. To assess whether similar analytical variability arises in automated reanalysis across different runs of the LLM pipeline, we compared the minimum-to-maximum range of human reanalysis estimates with the corresponding minimum-to-maximum range of LLM-generated estimates for the same studies (Supplementary Fig. S11). The range of LLM-generated estimates generally followed a pattern similar to the range of human reanalysis estimates, suggesting that LLM reanalyses may capture part of the same study-level analytical variability observed among human analysts.

Sensitivity to different information contexts (RQ3)

To examine whether reproducibility varied with the amount of article context provided to the LLM, we compared three input variants: the full paper, the full paper with the methods section removed, and the abstract only. Reproducibility rates were similar across these conditions. The LLM-generated effect size fell within the strict ± 0.05 tolerance in 24% of studies in the full paper condition, 24% in the no methods condition, and 22% in the abstract-only condition (Fig. 4a). Using the broader ± 0.20 tolerance, the corresponding rates were 50%, 52%, and 46% for the full paper, no methods, and abstract-only conditions, respectively. The deviations in the absolute effect size (i.e., $|\Delta d|$) were also comparable across conditions (Fig. 4b).

We formally tested differences across information conditions using Cochran's Q tests with Bonferroni-corrected post-hoc McNemar tests for the binary within-tolerance outcome, and Friedman tests with Bonferroni-corrected post-hoc Wilcoxon signed-rank tests for absolute effect size deviations. These tests showed no evidence of statistical differences across conditions (Cochran's $Q = 0.80$, $p = 0.67$, and Friedman $\chi^2 = 2.87$, $p = 0.24$, with no significant pairwise contrasts after correction).

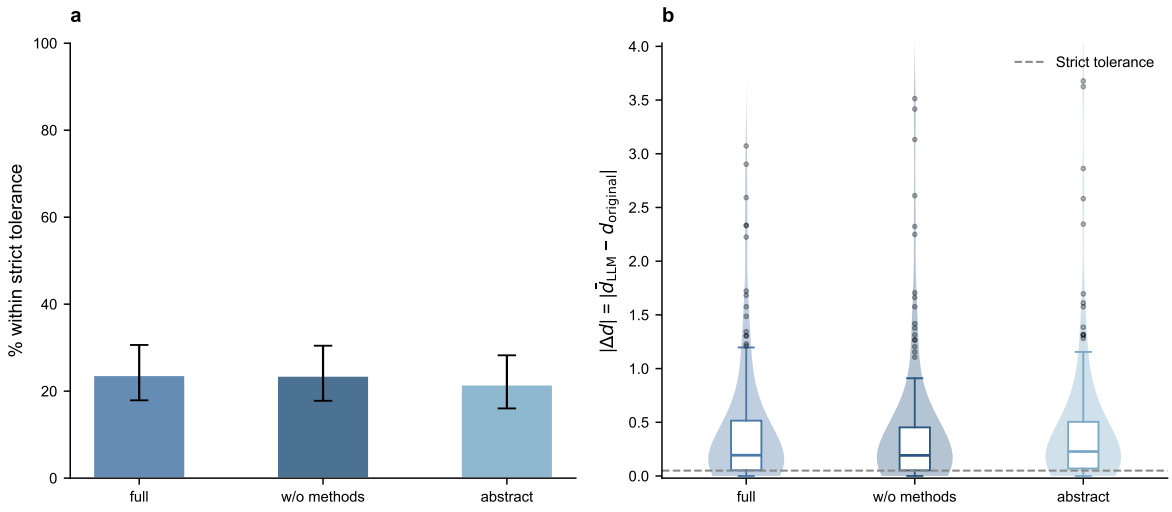


Figure 4: **Sensitivity to different information contexts.** We compared three input variants (using Claude Opus 4.7): the full paper, the full paper with the methods section removed, and the abstract only. **a,** Proportion of studies where the mean LLM effect size falls within ± 0.05 Cohen's d of the original, by analysis variant. Whiskers show Wilson 95% confidence intervals. **b,** Distribution of absolute effect size deviations (i.e., $|\Delta d| = |\bar{d}_{\text{LLM}} - d_{\text{original}}|$) by variant. The dashed line shows the ± 0.05 tolerance threshold. In the boxplot, the line indicates the median, the box denotes the interquartile range (IQR), and the whiskers extend to $1.5\times$ of the IQR; points beyond the whiskers indicate outliers.

Robustness across different LLMs

To demonstrate the generalizability of our findings across different LLMs, we next repeated our analysis with different LLMs. In addition to the primary analyses with Claude Opus 4.7, we performed the automated reproducibility assessment using GPT-5.5 and GLM-5.1. Overall, the results were largely similar. The mean LLM effect size fell within the strict ± 0.05 Cohen's d tolerance region of the original published effect size in 24% of studies for Claude Opus 4.7, 34% for GPT-5.5, and 31% for GLM-5.1 (Supplementary Fig. S1 and S2). Using the broader ± 0.20 tolerance region, the corresponding reproducibility rates were 50%, 58%, and 55%, respectively (Supplementary Fig. S3, S4, and S5). The qualitative conclusion matched the original conclusion in 80% of studies for Claude Opus 4.7, 81% for GPT-5.5, and 83% for GLM-5.1. Together, these results indicate that the main findings are not specific to a single LLM, but generalize across different

frontier models, including an open-weight model.

Heterogeneity analysis

To assess the sensitivity to the prompt framing, we compared the neutral baseline prompt with two alternative prompt framings, namely, a confirmatory framing that encouraged the model to provide supportive evidence of the claim, and a critical framing that encouraged a skeptical test. Differences across confirmatory and critical prompt framings were small (i.e., the strict within-tolerance rate differed by only 4 percentage points between the confirmatory and critical framings; Supplementary Fig. S6).

Finally, we analyzed the heterogeneity across different study characteristics. First, we compared the reproducibility rates across experimental and observational studies to assess whether reproducibility differed by empirical design (Supplementary Fig. S7). Second, we stratified studies by public code availability to test whether the results depended on the transparency of the original analysis materials (Supplementary Fig. S8). Third, we analyzed reproducibility separately for psychology, economics, and political science to account for domain-specific differences in data structures, statistical models, and reporting conventions (Supplementary Fig. S9). Fourth, we further stratified studies by the magnitude of the original Cohen's d to assess whether larger published effects showed lower reproducibility rates (Supplementary Fig. S10). The latter analysis addresses the possibility that larger original effects may partly reflect opportunistic analytical choices, selective reporting, or other forms of study-specific flexibility.

Discussion

LLMs can reproduce a substantial share of published findings in the social and behavioral sciences. Across the 169 of the 180 studies where the LLM pipeline successfully recovered effect sizes, the estimates were within ± 0.05 tolerance of the original Cohen's d in 24% of studies and reached

the same qualitative conclusion as the original study in 80% of cases. In the subset with human reanalysis benchmarks, the LLM pipeline performed comparably to human reanalysts, where it recovered the original effect sizes in 40% of studies, broadly similar to 28% for humans, and produced the original qualitative conclusion in 95% of cases (83% for humans). These findings were robust across different LLMs, including open-weight models. At the same time, some runs did not yield usable results, highlighting that LLM-based end-to-end statistical coding and analysis remains an emerging capability and is not yet fully reliable. Nevertheless, these results suggest that LLMs can provide a scalable first-pass tool for reproducibility assessment.

LLM analysis often reached the same substantive answer as the original study, even when it did not recover the exact effect size. The differences between the LLM-generated estimates and the original findings could be due to two main sources. First, such differences may reflect errors or limitations in the LLM-generated analysis. Second, they may also reflect conceptual ambiguity in the original claim, theory, or methods [33]. Research hypotheses are often short verbal statements that do not uniquely specify the sample, variables, model, or controls, and some analytical strategies are not fully specified, which can create legitimate variation in reanalysis decisions [34]. Consistent with this, the observed distribution of effect sizes appears to be broadly similar between human and automated reanalyses (Supplementary Fig. S11). Distinguishing such variation from modeling errors is itself a known challenge, even for human reanalysis.

We do not interpret this comparison as evidence that either LLMs or humans are generally superior. Human reanalyses in [2] allowed analysts to pursue their own defensible specifications, while our automated pipeline may have encouraged prioritizing a narrower analytical path that may have been closer to the originally reported result. The latter is reflected in a higher agreement of the LLM estimates with the original effect sizes ($r = 0.46$) than with the human estimates ($r = 0.11$), which should be interpreted cautiously, especially given the limited benchmark subset ($N = 82$). Further, because both human and automated reanalyses involve judgment over operationalizations, samples, models, and covariates, differences in recovered effect sizes may reflect alternative but

reasonable analytical choices rather than clear errors by one type of analyst.

Our results also show that LLM-based reproducibility did not meaningfully differ when more or less detailed methodological information was provided (see RQ3). In principle, brief claims leave many analytical choices open (e.g., how theory-laden concepts are operationalized or how hypotheses are tested [35]). This pattern is consistent with the idea that empirical results are dependent on the particular choices researchers make among many similarly acceptable data processing and modeling choices [30, 36–38]. One might therefore expect that providing more detailed methodological information would constrain the model's choices and move its analysis closer to the original. Against this background, the limited variation across information conditions suggests that the model may be guided primarily by the focal claim and the structure of the available dataset. Once the model identifies one plausible mapping between the claim and the data, additional methodological detail appears to have only limited influence on the selected operationalization, model specification, and resulting conclusion. Still, caution is needed when interpreting this finding. Limited variation across information conditions does not imply that analytical choices are unimportant, nor that the model has exhaustively searched the space of defensible analyses (e.g., explicit multiverse analysis [2, 29, 31, 39] could still reveal a broad range of possible estimates and conclusions).

Our study has several strengths. First, the results were robust across three frontier LLMs, namely, Claude Opus 4.7, GPT-5.5, and GLM-5.1. A notable strength is that one of these models was an open-weight LLM, showing that automated reanalysis is not limited to closed proprietary systems. Second, we analyze 180 studies, which exceeds the size of the individual corpora used in recent human reanalysis efforts in this area [2, 3, 10]. Third, our pipeline is multimodal and can process not only text and tabular data, but also figures and other image-based information extracted from the original papers.

Practically, automated reanalysis could reduce the burden of data-quality and reproducibility checks that journals increasingly require, and could extend such checks to settings where full

manual review is infeasible. Rather than replacing expert judgment, LLM-based reanalysis could support editors, reviewers, and meta-scientists by providing a scalable first-pass check of whether reported findings can be recovered from the available data and study materials. Such workflows could be integrated into submission, post-acceptance, or post-publication processes for quality control and help make reproducibility assessment more routine, especially as some journals have already adopted such initiatives [1, 4, 6, 7]. More broadly, this approach follows a growing trend toward automated tools to support scientific workflows [7,22,40–45] and positions LLM-based automated reanalysis within the broader movement toward transparency and openness in science [46], adding a scalable instrument to a toolkit that has so far relied largely on manual, labor-intensive approaches. However, automated checks should not be treated as end-to-end verification, given that reproducibility assessments often depend on many defensible analytical choices, contextual interpretation, and the quality of the available materials. Caution is also needed for two reasons. First, such automated checks may create Goodhart-style incentives, where authors optimize submissions to pass automated screens rather than to improve the underlying analysis or ensure genuinely robust results. Second, LLM-based analysis tools could facilitate gaming behavior by making it easier for researchers to opportunistically screen many analytical specifications [47, 48], potentially biasing results toward desired findings.

Several limitations apply. First, the performance may depend on the modeling choices, such as the model and prompt. Our sensitivity analyses suggest that the main results are largely stable across these choices. Second, a key concern is training data contamination. Because the studies in our corpus were published before the training cutoff of the models, LLMs may have encountered some of the studies during pretraining, which could inflate reproducibility rates. A related concern applies to the human reanalysis benchmark. To assess this risk, we conducted a memorization test [49,50] and further manually inspected the reasoning traces for signs of direct copying or other forms of apparent cheating, but did not find evidence of such. We also compared results based on whether the original analysis code was publicly available, but found similar results. Nevertheless,

the absence of direct memorization does not rule out subtler effects of training data exposure (e.g., even if a model cannot recall a reported effect size, prior exposure to a study could still improve the ability to interpret the prompt or write suitable analysis code). Third, our corpus is limited in size and to studies from psychology, economics, and political science. This reflects the difficulty and cost of assembling such datasets; even though our dataset is larger than others, this is a constraint that affects reproducibility audits broadly [51]. Fourth, comparing results across heterogeneous analyses requires a common metric. We transformed reported estimates into a standardized effect size (Cohen's d). Such transformations rest on assumptions that do not hold equally across all analytical settings [36,37,52,53]. Cohen's d thus offers comparability at the cost of some precision. Finally, automated reanalysis assesses whether a result can be recovered from the original data. However, it does not test whether a finding holds in newly collected data. Reproducibility and replication are complementary. Confirming that a result is computationally reproducible does not establish that the underlying effect generalizes, which is why replication with newly collected data continues to remain relevant [54, 55].

In sum, our findings suggest that LLMs can offer a path toward more scalable reproducibility assessment by automatically performing systematic reanalyses that are otherwise costly to conduct manually. Given the current capabilities and limitations of LLMs, such automated reanalysis should be viewed as a scalable screening tool rather than a substitute for expert judgment. Nevertheless, it could support the research community in quality control and provide a first-pass check for scientific journals, which may ultimately help improve rigor and reproducibility in empirical research.

Methods

We preregistered our methods at https://osf.io/84ue7 on June 10, 2026, prior to performing the reproducibility analysis. Before preregistration, we developed and tested the LLM pipeline on a

small subset of 10 papers (see details below).

Study sample

We evaluated the LLM-based reproducibility pipeline on published empirical studies from the Systematizing Confidence in Open Research and Evidence (SCORE) initiative (https://www.cos.io/score) [28]. SCORE is a stratified sample of articles published between 2009 and 2018 across major social and behavioral science disciplines, including psychology, economics, and political science. We filtered the original stratified sample of over 3000 papers for those studies for which original outcomes were provided by the SCORE project. We then kept only those studies for which a central empirical claim could be linked to a reported statistical result that was convertible to Cohen's d (251 studies; Supplementary Table S1). Next, we excluded studies from both groups for which no data were available from the replication repositories (57 studies) or data were labeled as private (4 studies). To compute effect sizes, we extracted the reported effect size or a convertible statistic (e. g., t, F, r, z, χ^2), the sample size, and the metadata needed to derive Cohen's d. We then mapped the statistics to Cohen's d following the conversion procedure from [2, 56], and excluded studies for which this conversion was not possible from the record alone (7 studies). One study was excluded due to a large Cohen's d that was interpreted as a data error.

We annotated the studies following the protocol in [2] to obtain empirical claims phrased at the conceptual level. For each study, trained social and behavioral scientists wrote a short, standalone empirical claim in plain language that could be mapped to the corresponding inferential results from the paper. The goal was to capture the substantive relationship tested in the paper at the conceptual level, rather than to reproduce the statistical formulation or method-specific wording from the article. To do so, annotators used the existing SCORE information, which identified candidate claims and the corresponding statistical evidence, but often phrased these claims in statistical language or in wording close to the original paper, which included methodological details, model specifications, or other study-specific terminology. Each annotated claim had to (i) be un-

derstandable on its own, (ii) contain only one empirical relationship and a clear direction of effect, (iii) correspond to a result based on a hypothesis test, and (iv) be phrased at the conceptual rather than statistical level. For example, annotators aimed to capture claims such as whether one condition increased an outcome relative to another condition, or whether a predictor was positively or negatively associated with an outcome, without encoding the full statistical specification in the claim itself. Annotators were allowed to use the full article to support accurate interpretation of the claim and statistical result [2]. Studies with no directional claim were excluded (2 studies). For the subset from [2], we reused the existing claim annotations and original effect sizes.

Overall, the result is a standardized dataset of annotated empirical claims with matched statistical evidence, against which the LLM reanalysis can then be evaluated (see our code repository). Prior to preregistration, we used 10 studies to develop, test, and validate the analysis pipeline, including prompt design, code execution, quality control, and evaluation metrics. Excluding these studies left the main findings unchanged; i.e., the qualitative conclusion match rate was 79% instead of 80%, and the effect size was recovered for 23% instead of 24%.

A subset of these studies ($N = 84$) had previously been analyzed in a large-scale human reanalysis effort [2]. For this subset, we used the existing human reanalysis results as an additional benchmark for RQ2. This allowed us to compare the LLM-generated reanalyses not only against the original published findings, but also against independent human reanalyses of the same studies.

Data preparation

For each study, we assembled a standardized study package consisting of (1) the statistical claim, (2) the original study data used in the publication, (3) paper-specific metadata, and (4) the article text. We converted article PDF files to Markdown using the Mistral OCR API (Mistral OCR 3 via the `mistral-ocr-latest` endpoint; model identifier `mistral-ocr-2512`, see https://docs.mistral.ai/models/model-cards/ocr-3-25-12). This step performs multimodal document parsing by combining optical character recognition (OCR) with layout-aware document ex-

traction, which preserves the document structure even for complex scientific documents including their multi-column layouts, mathematical expressions, and tables. The extracted Markdown syntax thus captures the document structure, including section headings, paragraphs, and equations. Tables are extracted as Markdown-based representations, and figures as raster images.

LLM analysis pipeline

Agent task. We implemented the analysis pipeline as an LLM agent that, for each study, receives the claim, the original dataset, and the paper, and is asked to write and execute statistical code in an isolated sandbox to test the claim. The agent does not have access to the original analysis code of the paper; it writes and executes code independently. Sandboxes are created fresh for each run, with no state carried over between runs, so that files, installed packages, and intermediate results from one run cannot influence subsequent runs. The agent submits a structured report containing the computed test statistic, sample size, degrees of freedom, and a qualitative conclusion (i.e., support / opposite / inconclusive). Each study is analyzed across 5 independent runs.

The prompt was designed following best practices for LLM prompt engineering [57–59]. The prompt casts the agent as a statistical analyst and asks it to commit to one reasonable analysis of the focal claim. The agent is explicitly instructed not to copy any test statistic reported in the paper; the submitted statistic must come from code the agent executes on the provided data. The full prompt is provided in Appendix G.

In the supplementary materials, we additionally test the sensitivity to the framing of the prompt by comparing two alternatives: (1) a "confirmatory" prompt, which instructed the agent to approach the analysis as a supportive reviewer expecting the claim to hold up under reanalysis, and (2) a "critical" prompt, which instructed the agent to approach the analysis as a skeptical reviewer questioning whether the claim is as robust as the authors assert. Details are in Appendix G.

In the main analysis, the LLM agent receives the complete paper converted to Markdown, including the methods section. This design mirrors previous large-scale human reanalysis studies, in

which the human reanalysts also had access to the full paper [2] . To assess whether the amount of methodological information affects automated reproducibility, we additionally conduct a sensitivity analysis where we vary the information context (see RQ3). For this analysis, we compare the default setting with the full paper as input against: (1) a condition in which the methods section is removed from the Markdown file, and (2) a condition in which the model receives only the abstract of the article. For the former, we manually removed methods-related sections from each paper, where relevant. These may include one or more sections titled, for example, "Methods", "Data and methods", "Methodology", "Empirical approach", "Framework", "Empirical design", or "Research design" (for five papers, no such section could be identified and instead relevant paragraphs with methodological specifications were removed). Figures and tables that are central to the removed methods sections, or referenced only within those sections, are also removed. In contrast, figures or tables that were positioned in the methods section were retained if they are substantively referenced elsewhere, such as in the results section. Appendices with supplementary methods were also removed.

Models. We use Claude Opus 4.7 as the primary model. We further assess robustness with two additional frontier models, OpenAI GPT-5.5 (`openai/gpt-5.5`) and the open-weight Zhipu GLM-5.1 (`z-ai/glm-5.1`), so that our central findings do not depend on any single system; GLM-5.1 additionally tests whether the findings generalize beyond proprietary models. For all three models, we set the native reasoning effort to *medium* (which we matched across models for comparability) and left all other decoding parameters at their provider defaults. We set a sampling temperature of 1.0 for GPT-5.5 and GLM-5.1; Claude Opus 4.7 runs in always-on adaptive thinking and disregards the temperature parameter [60]. Each LLM analysis run was capped at 100 messages, including reasoning, tool use, and submission steps, or a total API cost of US$5, whichever was reached first. See Supplementary Fig. S12 for a comparison.

Output standardization. The agent was instructed to submit the final response as a structured JSON object. The output included the statistic type, test statistic, degrees of freedom, sample

size, p-value, substantive conclusion (support / opposite / inconclusive), a short description of how the reanalysis was operationalized (including the dependent variable, main predictor, sample definition, model specification, controls, and a rationale), and the LLM reasoning trace.

We used the test statistic to compute a standardized effect size. Following [2, 56], test statistics were first converted to a Pearson correlation coefficient and then transformed into Cohen's d. The conversion supported the statistic types requested from the agent: t, F, z, χ^2, and r. When the required fields were missing or the statistic could not be converted, Cohen's d was treated as unavailable for that run.

Following the preregistration, standardized effect sizes with $|d_{\mathrm{LLM}}| > 10$ were manually reviewed to determine the source of the result (e.g., whether this reflected a conversion or coding error rather than a plausible statistical estimate). Across all models and conditions, 20 of the 6,300 runs (0.3%) exceeded this threshold. In 9 of the 20 runs, this occurred when the agent ran a nonparametric Wilcoxon signed-rank test but reported the rank-sum statistic $W = 15$ under the z-statistic, thus processing 15 as a z-score with $n = 5$, which inflated Cohen's d to $\approx 4 \times 10^4$. The remaining 11 cases were isolated runs in which the conversion failed (i.e., in 9 cases, the regression t-statistics were converted with a single (numerator) rather than the residual degree-of-freedom, and 2 cases reported an implausibly large $z \approx 31$). Because every affected study retained other valid runs, no study was dropped on this basis alone. Eventually, the 20 runs were excluded.

Implementation details. We implemented our LLM agent using the Inspect AI framework, which executes a ReAct-style loop of reasoning, tool use, and observation [61]. The agent has access to four tools: (i) a Python interpreter, (ii) a bash shell, (iii) a `think` tool that allows the agent to record intermediate reasoning, and (iv) a `view_image` tool. Both the Python and bash tools return only text, while the `view_image` tool allows the agent to visually inspect image files in the sandbox (used for reading figures that were extracted from papers). Note that GLM-5.1 is a text-only model (GLM-5.1); here, `view_image` is disabled, and the agent relies on the parsed text and tables instead. The Python environment contains common scientific computing and

statistical libraries; the complete list can be found in the system prompt (Appendix G).

Statistical analysis

Following [2], all original effect sizes are reported as positive values, whereas the LLM and human reanalysis effect sizes are negative only when an "opposite" substantive conclusion is reported.

For each study, the 5 independent runs were aggregated into one effect size ($\bar{d}_{\mathrm{LLM}}$) using the mean. The qualitative conclusion is aggregated by taking the majority vote across the 5 runs; ties are resolved in favor of 'inconclusive'. Studies were excluded if none of the 5 runs produced a valid effect size.

For each comparison, we compute $\Delta d = \bar{d}_{\mathrm{LLM}} - d_{\mathrm{original}}$, where d_{original} is the published effect size of the paper (or the mean human Cohen's d for the comparison against human reanalysts). We classify a study as within-tolerance if $|\Delta d| \leq 0.05$ (strict) or $|\Delta d| \leq 0.20$ (broad). The primary metric is the proportion of studies within the strict tolerance [2].

Memorization test

We conducted a separate memorization test [49,50] to assess whether model performance could be influenced by prior knowledge of the analyzed paper. We adapted this test to our setting by probing whether the LLM was able to recall the published effect size. For this, the agent received the paper reference and the focal claim as input, but not the original dataset, and was asked to report the expected effect size for the claim. Studies for which the response fell within ± 0.05 Cohen's d of the published result are flagged as potentially contaminated. Nevertheless, the results were largely robust (Supplementary Table S2).

Data availability

Data from the Systematizing Confidence in Open Research and Evidence (SCORE) project [28] are available at: https://osf.io/dtzx4/overview. The human reanalysis data are available from [2]. Our GitHub repository (see Code Availability) contains our additional annotations and all derived analysis files needed to reproduce the results.

Code availability

All code to replicate our analyses is available via GitHub at https://github.com/tobihol/agentic-reproducibility. The repository also contains the GUIDE-LLM checklist [62] to document the LLM use.

References

[1] Miske, O. *et al.* Investigating the reproducibility of the social and behavioural sciences. *Nature* **652**, 126–134 (2026).

[2] Aczel, B. *et al.* Investigating the analytical robustness of the social and behavioural sciences. *Nature* **652**, 135–142 (2026).

[3] Brodeur, A. *et al.* Reproducibility and robustness of economics and political science research. *Nature* **652**, 151–156 (2026).

[4] Hardwicke, T. E. *et al.* Analytic reproducibility in articles receiving open data badges at the journal *Psychological Science*: An observational study. *Royal Society Open Science* **8**, 201494 (2021).

[5] Brodeur, A., Mikola, D. & Cook, N. Mass reproducibility and replicability: A new hope. Tech. Rep., SSRN (2024).

[6] Fišar, M. *et al.* Reproducibility in *Management Science*. *Management Science* **70**, 1343–1356 (2024).

[7] Nosek, B. *et al.* Reimagining and diversifying assessment of the credibility of research findings (2026).

[8] Tyner, A. H. *et al.* Investigating the replicability of the social and behavioural sciences. *Nature* **652**, 143–150 (2026).

[9] Parsons, S. *et al.* A community-sourced glossary of open scholarship terms. *Nature Human Behaviour* **6**, 312–318 (2022).

[10] Open Science Collaboration. Estimating the reproducibility of psychological science. *Science* **349**, aac4716 (2015).

[11] Peng, R. D. Reproducible research in computational science. *Science* **334**, 1226–1227 (2011).

[12] Sun, M. *et al.* LAMBDA: A large model based data agent. *Journal of the American Statistical Association* **121**, 1–13 (2026).

[13] Boiko, D. A., MacKnight, R., Kline, B. & Gomes, G. Autonomous chemical research with large language models. *Nature* **624**, 570–578 (2023).

[14] Qian, C. *et al.* ChatDev: Communicative agents for software development. In Ku, L.-W., Martins, A. & Srikumar, V. (eds.) *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 15174–15186 (Association for Computational Linguistics, Bangkok, Thailand, 2024).

[15] Lu, C. *et al.* Towards end-to-end automation of AI research. *Nature* **651**, 914–919 (2026).

[16] Schmidgall, S. & Moor, M. AgentRxiv: Towards collaborative autonomous research. *arXiv:2503.18102* (2025).

[17] Seo, M., Baek, J., Lee, S. & Hwang, S. J. Paper2Code: Automating code generation from scientific papers in machine learning. *International Conference on Learning Representations (ICLR)* (2026).

[18] Alizadeh, M., Mosleh, M., Gilardi, F. & Tucker, J. A. Evaluating AI coding agents in social science reproducibility (2026).

[19] Kohler, B., Zollikofer, D., Einsiedler, J., Hoyle, A. & Ash, E. Read the paper, write the code: Agentic reproduction of social-science results. *arXiv:2604.21965* (2026).

[20] Miao, J., Davis, J. R., Zhang, Y., Pritchard, J. K. & Zou, J. Paper2Agent: Reimagining research papers as interactive and reliable AI agents. *arXiv:2509.06917* (2025).

[21] Song, Z. *et al.* Evaluating large language models in scientific discovery. *arXiv:2512.15567* (2025).

[22] Shao, E. *et al.* SciSciGPT: Advancing human–AI collaboration in the science of science. *Nature Computational Science* **6**, 301–315 (2026).

[23] Zhang, S., Fan, J., Fan, M., Li, G. & Du, X. DeepAnalyze: Agentic large language models for autonomous data science. *arXiv:2510.16872* (2025).

[24] Gottweis, J. *et al.* Accelerating scientific discovery with co-scientist. *Nature* (2026).

[25] Ghareeb, A. E. *et al.* A multi-agent system for automating scientific discovery. *Nature* forthcoming (2026).

[26] Yamada, Y. *et al.* The AI Scientist-v2: Workshop-level automated scientific discovery via agentic tree search. *arXiv:2504.08066* (2025).

[27] Song, X. *et al.* StatLLM: A dataset for evaluating the performance of large language models in statistical analysis. *Scientific Data* **13**, 369 (2026).

[28] Alipourfard, N. *et al.* Systematizing confidence in open research and evidence (score) (2021). URL osf.io/preprints/socarxiv/46mnb_v1.

[29] Silberzahn, R. *et al.* Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science* **1**, 337–356 (2018).

[30] Botvinik-Nezer, R. *et al.* Variability in the analysis of a single neuroimaging dataset by many teams. *Nature* **582**, 84–88 (2020).

[31] Steegen, S., Tuerlinckx, F., Gelman, A. & Vanpaemel, W. Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science* **11**, 702–712 (2016).

[32] Breznau, N. *et al.* Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty. *Proceedings of the National Academy of Sciences* **119**, e2203150119 (2022).

[33] Auspurg, K. & Brüderl, J. Has the credibility of the social sciences been credibly destroyed? Reanalyzing the "many analysts, one data set" project. *Socius* **7**, 23780231211024421 (2021).

[34] Scheel, A. M. Why most psychological research findings are not even wrong. *Infant and Child Development* **31**, e2295 (2022).

[35] Oberauer, K. & Lewandowsky, S. Addressing the theory crisis in psychology. *Psychonomic Bulletin & Review* **26**, 1596–1618 (2019).

[36] Coretta, S. *et al.* Multidimensional signals and analytic flexibility: Estimating degrees of freedom in human-speech analyses. *Advances in Methods and Practices in Psychological Science* **6**, 25152459231162567 (2023).

[37] Gelman, A. & Loken, E. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. Tech. Rep., Department of Statistics, Columbia University, New York, NY (2013).

[38] Patel, C. J., Burford, B. & Ioannidis, J. P. Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. *Journal of Clinical Epidemiology* **68**, 1046–1058 (2015).

[39] Wagenmakers, E.-J., Sarafoglou, A. & Aczel, B. One statistical analysis must not rule them all. *Nature* **605**, 423–425 (2022).

[40] Nuijten, M. B. & Polanin, J. R. "statcheck": Automatically detect statistical reporting inconsistencies to increase reproducibility of meta-analyses. *Research Synthesis Methods* **11**, 574–579 (2020).

[41] Bertran, M., Fogliato, R. & Wu, Z. S. Many AI analysts, one dataset: Navigating the agentic data science multiverse. *arXiv:2602.18710* (2026).

[42] Brodeur, A. *et al.* Comparing human-only, AI-assisted, and AI-led teams on assessing research reproducibility in quantitative social science. IZA Discussion Paper 17645, Institute of Labor Economics (IZA), Bonn (2025).

[43] Siegel, Z. S., Kapoor, S., Nagdir, N., Stroebl, B. & Narayanan, A. CORE-Bench: Fostering the credibility of published research through a computational reproducibility agent benchmark. *Transactions on Machine Learning Research* (2024).

[44] Starace, G. *et al.* PaperBench: Evaluating AI's ability to replicate AI research. In *International Conference on Machine Learning*, 56843–56873 (2025).

[45] Wrightson, J. G., Blazey, P., Moher, D., Khan, K. M. & Ardern, C. L. GPT for RCTs? Using AI to determine adherence to clinical trial reporting guidelines. *BMJ Open* **15**, e088735 (2025).

[46] Nosek, B. A., Ebersole, C. R., DeHaven, A. C. & Mellor, D. T. The preregistration revolution. *Proceedings of the National Academy of Sciences* **115**, 2600–2606 (2018).

[47] Brodeur, A., Cook, N. & Heyes, A. Methods matter: p-hacking and publication bias in causal analysis in economics. *American Economic Review* **110**, 3634–3660 (2020).

[48] Simmons, J. P., Nelson, L. D. & Simonsohn, U. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science* **22**, 1359–1366 (2011).

[49] Nori, H., King, N., McKinney, S. M., Carignan, D. & Horvitz, E. Capabilities of GPT-4 on medical challenge problems. *arXiv:2303.13375* (2023).

[50] Sainz, O. *et al.* NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 10776–10787 (Association for Computational Linguistics, 2023).

[51] Krähmer, D., Schächtele, L. & Auspurg, K. Code sharing and reproducibility in survey-based social research: Evidence from a large-scale audit. *Royal Society Open Science* **13**, 251997 (2026).

[52] Holzmeister, F. *et al.* Heterogeneity in effect size estimates. *Proceedings of the National Academy of Sciences* **121**, e2403490121 (2024).

[53] van Assen, M. A., Stoevenbelt, A. H. & van Aert, R. C. The end justifies all means: Questionable conversion of different effect sizes to a common effect size measure. *Religion, Brain & Behavior* **13**, 345–347 (2023).

[54] Zwaan, R. A., Etz, A., Lucas, R. E. & Donnellan, M. B. Making replication mainstream. *Behavioral and Brain Sciences* **41**, e120 (2018).

[55] Brodeur, A., Dreber, A., Hoces de la Guardia, F. & Miguel, E. Replication games: How to make reproducibility research more systematic. *Nature* **621**, 684–686 (2023).

[56] Multi100 conversion code. https://github.com/marton-balazs-kovacs/multi100/blob/47c0b8c6dd68e19eb80fa8843dce18f0d3655ae1/analysis/multi100_raw_processed.qmd#L160-L165.

[57] Lin, Z. How to write effective prompts for large language models. *Nature Human Behaviour* **8**, 611–615 (2024).

[58] Giray, L. Prompt engineering with ChatGPT: A guide for academic writers. *Annals of Biomedical Engineering* **51**, 2629–2633 (2023).

[59] Feuerriegel, S. *et al.* Using natural language processing to analyse text data in behavioural science. *Nature Reviews Psychology* **4**, 96–111 (2025).

[60] Anthropic. Adaptive thinking. Claude API Documentation (2026). URL https://platform.claude.com/docs/en/build-with-claude/adaptive-thinking. Accessed: 2026-06-22.

[61] AI Security Institute, UK. Inspect AI: Framework for Large Language Model Evaluations. https://github.com/UKGovernmentBEIS/inspect_ai (2024). Software.

[62] Feuerriegel, S. *et al.* A reporting checklist for large language models in behavioural science. *Nature Human Behaviour* (2026).

Acknowledgments

Funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under the National Research Data Infrastructure – NFDI 27/1-2026, project number 460037581 is acknowledged. SF acknowledges funding via the Swiss National Science Foundation (SNSF), Grant 186932. Our research is supported by the DAAD programme Konrad Zuse Schools of Excellence in Artificial Intelligence, sponsored by the Federal Ministry of Research, Technology and Space.

Author contributions

All authors contributed to conceptualization, manuscript writing, and approved the manuscript.

Competing interests

The authors declare no competing interests.

Appendix

A Study sample

Table S1: **Study filtering and inclusion.** Counts show the number of studies retained and excluded when constructing the evaluation corpus.

Filtering step	**Multi100** [2]	**SCORE extension**	**Total**
Candidate studies	100	151	251
Excluded: no data available	0	40	40
Excluded: no data in repository data node	7	10	17
Excluded: private data	4	0	4
Excluded: Cohen's d not computable	5[†]	2	7
Excluded: outlier Cohen's d	0	1	1
Excluded: no directional claim	0	2	2
Total excluded	16	55	71
Included in evaluation corpus	84	96	180

[†] This is consistent with the original Multi100 [2] study in which original effect sizes could not be determined due to missing information.

B Robustness across different LLMs (strict tolerance)

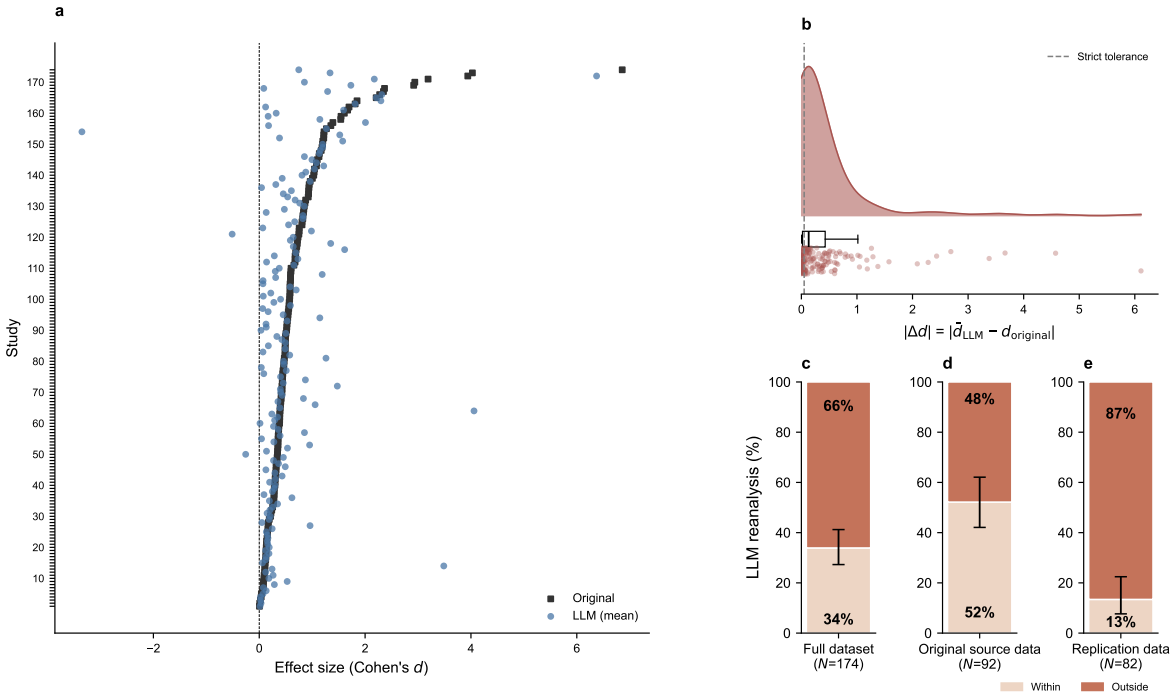


Fig. S1: **Automated reproducibility assessment using GPT-5.5. a,** Effect size of the original analysis (gray squares; all represented as positive values) and the effect sizes of the reanalyses (blue dots) for each study. **b,** Distribution of $|\Delta d|$, computed as $|\Delta d| = |\bar{d}_{\text{LLM}} - d_{\text{original}}|$. The distribution is visualized using a density plot, a boxplot, and scatter points. In the boxplot, the line indicates the median, the box denotes the interquartile range (IQR), and the whiskers extend to 1.5× of the IQR; points beyond the whiskers indicate outliers. **c,** Proportion of studies for which the LLM-generated effect size falls within or outside the tolerance region around the original result, across the full sample. **d,** Proportion of studies falling within or outside the tolerance region for the subset of studies that make the original source data available. **e,** Proportion of studies falling within or outside the tolerance region for the subset of studies where the original source data were unavailable but where replication data for reanalyses are provided. For 6 studies, the LLM did not produce a valid Cohen's d, and the resulting studies were excluded, leaving $N = 174$. Whiskers indicate Wilson 95% confidence intervals.

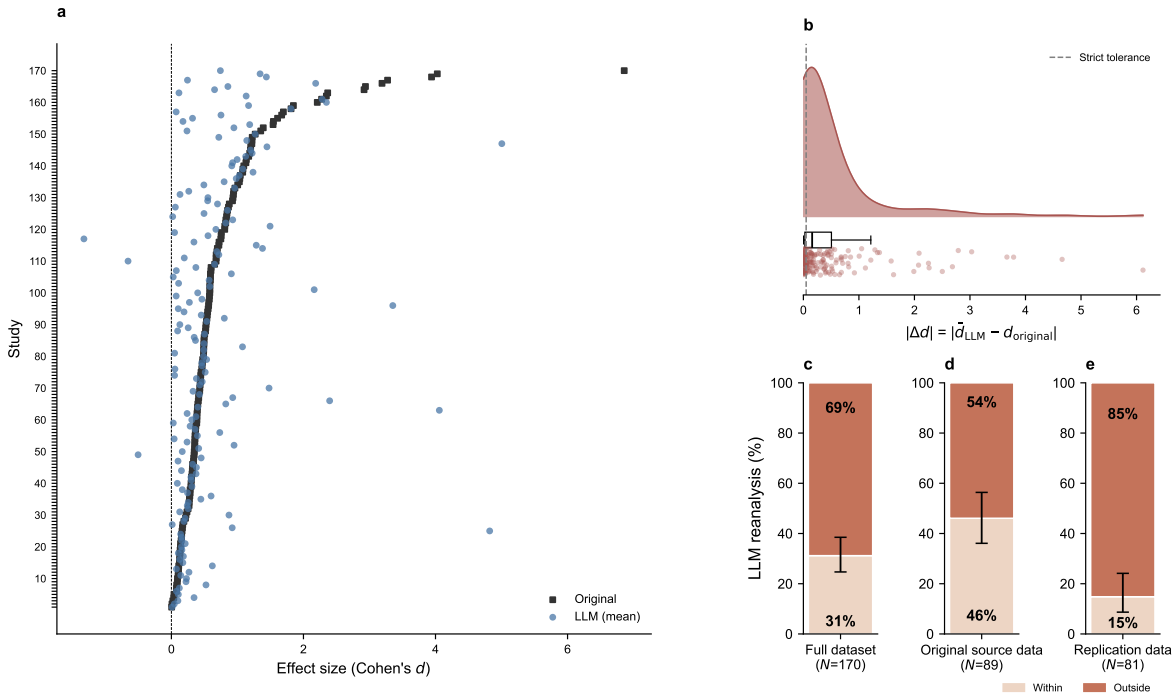


Fig. S2: **Automated reproducibility assessment using GLM-5.1. a,** Effect size of the original analysis (gray squares; all represented as positive values) and the effect sizes of the reanalyses (blue dots) for each study. **b,** Distribution of $|\Delta d|$, computed as $|\Delta d| = |\bar{d}_{\text{LLM}} - d_{\text{original}}|$. The distribution is visualized using a density plot, a boxplot, and scatter points. In the boxplot, the line indicates the median, the box denotes the interquartile range (IQR), and the whiskers extend to 1.5$\times$ of the IQR; points beyond the whiskers indicate outliers. **c,** Proportion of studies for which the LLM-generated effect size falls within or outside the tolerance region around the original result, across the full sample. **d,** Proportion of studies falling within or outside the tolerance region for the subset of studies that make the original source data available. **e,** Proportion of studies falling within or outside the tolerance region for the subset of studies where the original source data were unavailable but where replication data for reanalyses are provided. For 10 studies, the LLM did not produce a valid Cohen's d, and the resulting studies were excluded, leaving $N = 170$. Whiskers indicate Wilson 95% confidence intervals.

C Robustness across different LLMs (broad tolerance)

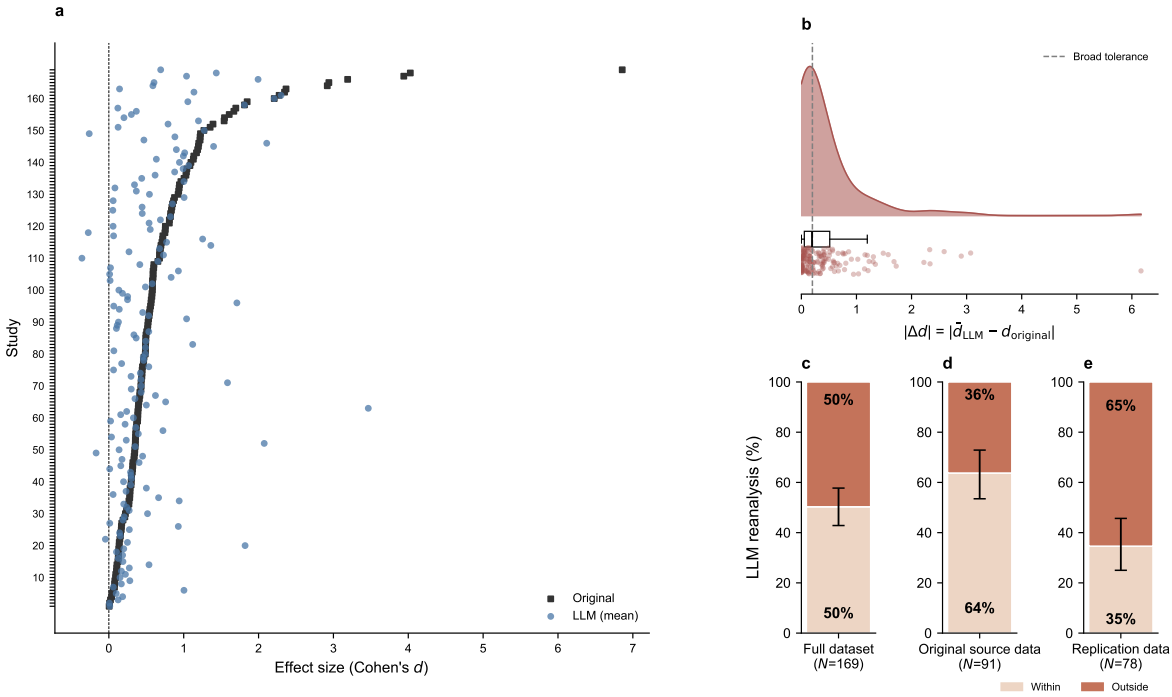


Fig. S3: **Automated reproducibility assessment using Claude Opus 4.7 with a broad tolerance (± 0.20 Cohen's d). a,** Effect size of the original analysis (gray squares; all represented as positive values) and the effect sizes of the reanalyses (blue dots) for each study. **b,** Distribution of $|\Delta d|$, computed as $|\Delta d| = |\bar{d}_{\text{LLM}} - d_{\text{original}}|$. The distribution is visualized using a density plot, a boxplot, and scatter points. In the boxplot, the line indicates the median, the box denotes the interquartile range (IQR), and the whiskers extend to $1.5\times$ of the IQR; points beyond the whiskers indicate outliers. **c,** Proportion of studies for which the LLM-generated effect size falls within or outside the tolerance region around the original result, across the full sample. **d,** Proportion of studies falling within or outside the tolerance region for the subset of studies that make the original source data available. **e,** Proportion of studies falling within or outside the tolerance region for the subset of studies where the original source data were unavailable but where replication data for reanalyses are provided. For 11 studies, the LLM did not produce a valid Cohen's d, and the resulting studies were excluded, leaving $N = 169$. Whiskers indicate Wilson 95% confidence intervals.

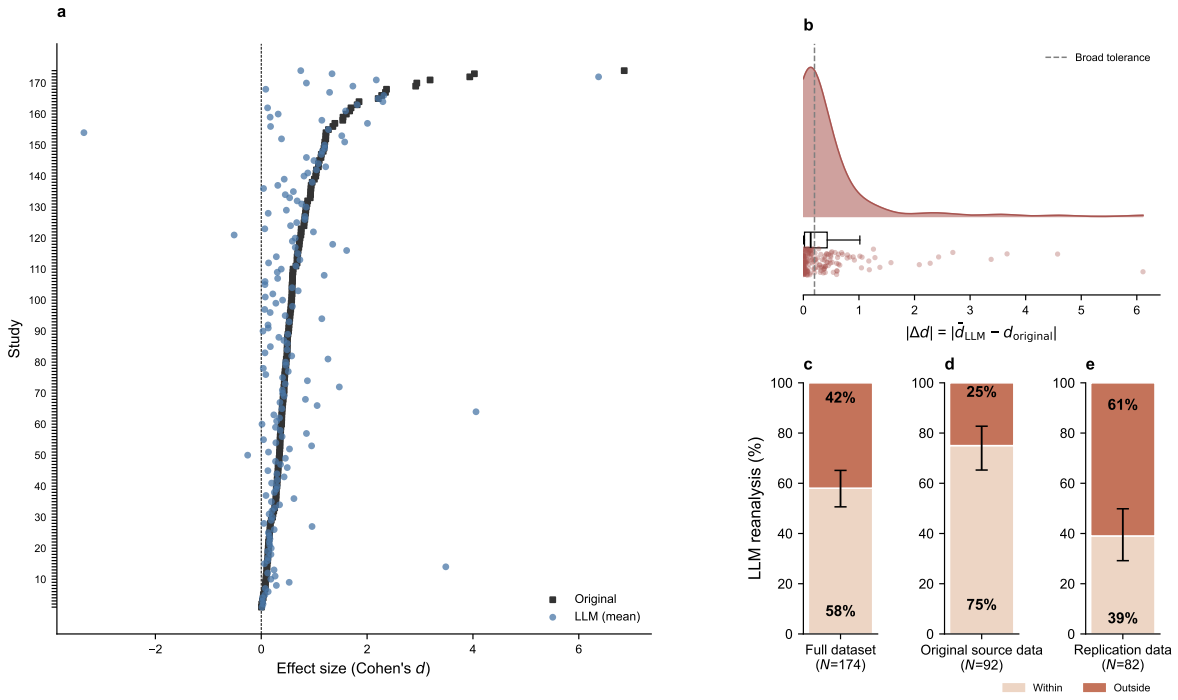


Fig. S4: **Automated reproducibility assessment using GPT-5.5 with a broad tolerance** (± 0.20 **Cohen's** d**). a,** Effect size of the original analysis (gray squares; all represented as positive values) and the effect sizes of the reanalyses (blue dots) for each study. **b,** Distribution of $|\Delta d|$, computed as $|\Delta d| = |\bar{d}_{\text{LLM}} - d_{\text{original}}|$. The distribution is visualized using a density plot, a boxplot, and scatter points. In the boxplot, the line indicates the median, the box denotes the interquartile range (IQR), and the whiskers extend to 1.5$\times$ of the IQR; points beyond the whiskers indicate outliers. **c,** Proportion of studies for which the LLM-generated effect size falls within or outside the tolerance region around the original result, across the full sample. **d,** Proportion of studies falling within or outside the tolerance region for the subset of studies that make the original source data available. **e,** Proportion of studies falling within or outside the tolerance region for the subset of studies where the original source data were unavailable but where replication data for reanalyses are provided. For 6 studies, the LLM did not produce a valid Cohen's d, and the resulting studies were excluded, leaving $N = 174$. Whiskers indicate Wilson 95% confidence intervals.

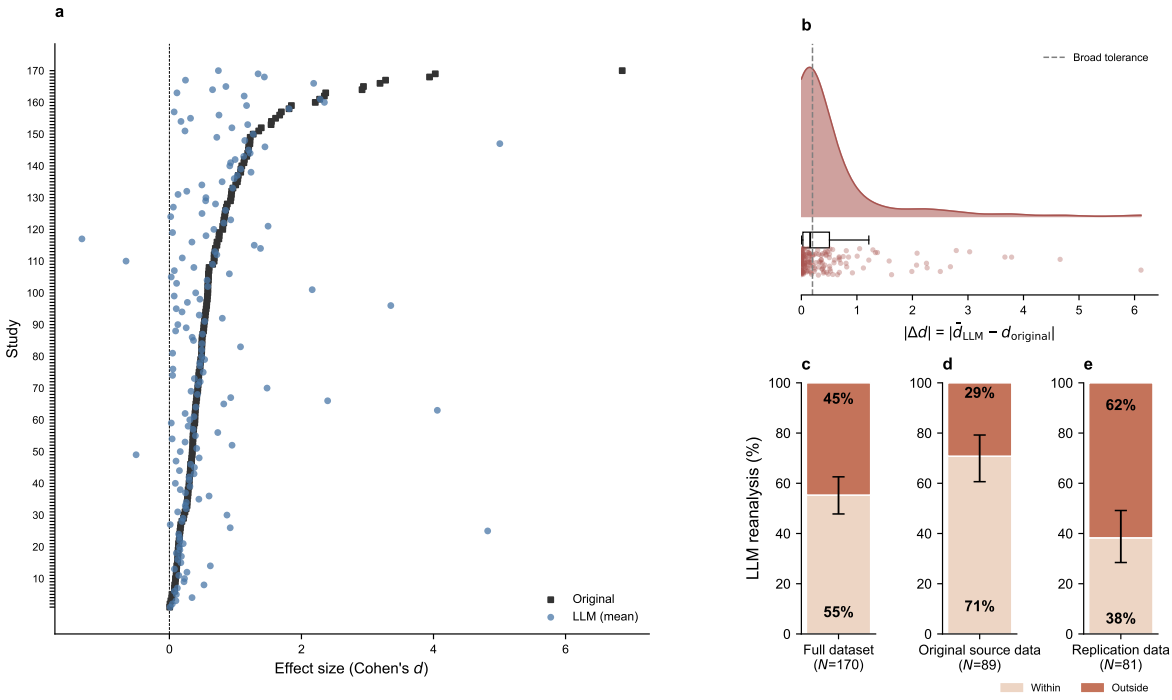


Fig. S5: **Automated reproducibility assessment using GLM-5.1 with a broad tolerance** (± 0.20 **Cohen's** d**). a,** Effect size of the original analysis (gray squares; all represented as positive values) and the effect sizes of the reanalyses (blue dots) for each study. **b,** Distribution of $|\Delta d|$, computed as $|\Delta d| = |\bar{d}_{\text{LLM}} - d_{\text{original}}|$. The distribution is visualized using a density plot, a boxplot, and scatter points. In the boxplot, the line indicates the median, the box denotes the interquartile range (IQR), and the whiskers extend to 1.5$\times$ of the IQR; points beyond the whiskers indicate outliers. **c,** Proportion of studies for which the LLM-generated effect size falls within or outside the tolerance region around the original result, across the full sample. **d,** Proportion of studies falling within or outside the tolerance region for the subset of studies that make the original source data available. **e,** Proportion of studies falling within or outside the tolerance region for the subset of studies where the original source data were unavailable but where replication data for reanalyses are provided. For 10 studies, the LLM did not produce a valid Cohen's d, and the resulting studies were excluded, leaving $N = 170$. Whiskers indicate Wilson 95% confidence intervals.

D Sensitivity to confirmatory vs. critical prompt framing

We examined whether the analytical perspective induced by prompt framing influenced LLM reproducibility. We compared three conditions, all run the on full paper variant using Claude Opus 4.7: (i) a *neutral* baseline using the base system prompt; (ii) a *confirmatory* condition, in which an instruction to approach the analysis with a prior that the claim is empirically robust was appended to the system prompt; and (iii) a *critical* condition, in which an instruction to approach the analysis as a skeptical reviewer was appended.

Within the strict ± 0.05 tolerance band, $\bar{d}_{\mathrm{LLM}}$ fell within tolerance in 24% of studies for the neutral condition, 24% for the confirmatory condition, and 28% for the critical condition (Supplementary Fig. S6). The conclusion (using a majority vote over the five independent runs) matched the original in 80%, 82%, and 77% of studies for neutral, confirmatory, and critical conditions, respectively. These results provide no clear evidence that confirmatory or critical prompt framing systematically alters reproducibility relative to the neutral baseline.

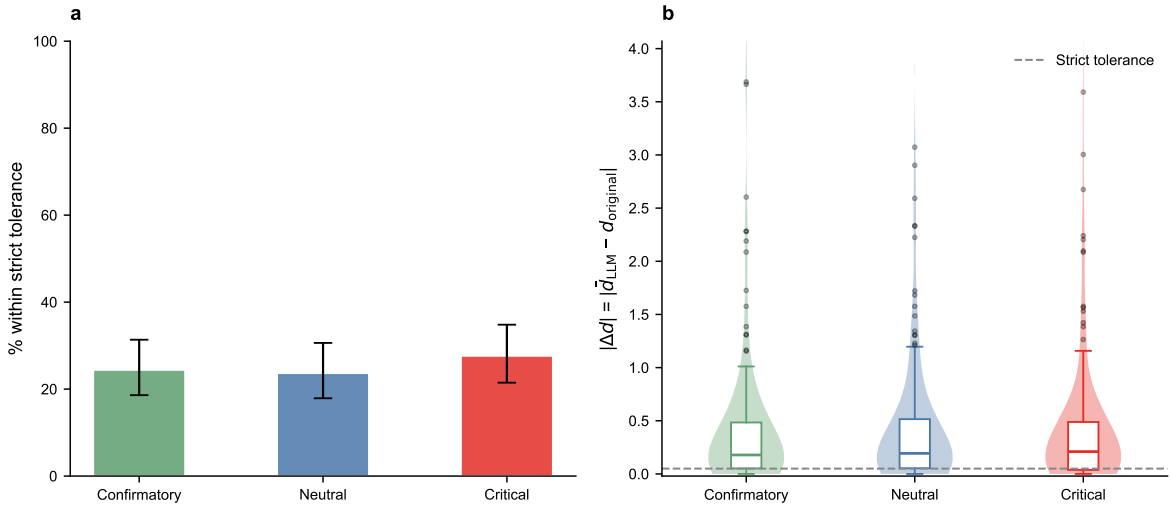


Fig. S6: **Sensitivity to confirmatory vs. critical prompt framing.** **a,** Proportion of studies where the mean LLM-generated effect size falls within ± 0.05 Cohen's d of the original. The results are separately shown for the confirmatory, neutral, and critical prompt framing conditions. Whiskers show Wilson 95% confidence intervals. **b,** Distribution of absolute effect size deviations (i.e., $|\Delta d| = |\bar{d}_{\text{LLM}} - d_{\text{original}}|$) by different framings. The dashed line shows the ± 0.05 tolerance threshold. In the boxplot, the line indicates the median, the box denotes the interquartile range (IQR), and the whiskers extend to $1.5\times$ of the IQR; points beyond the whiskers indicate outliers. Results are based on Claude Opus 4.7 over five independent runs per study.

E Heterogeneity analysis

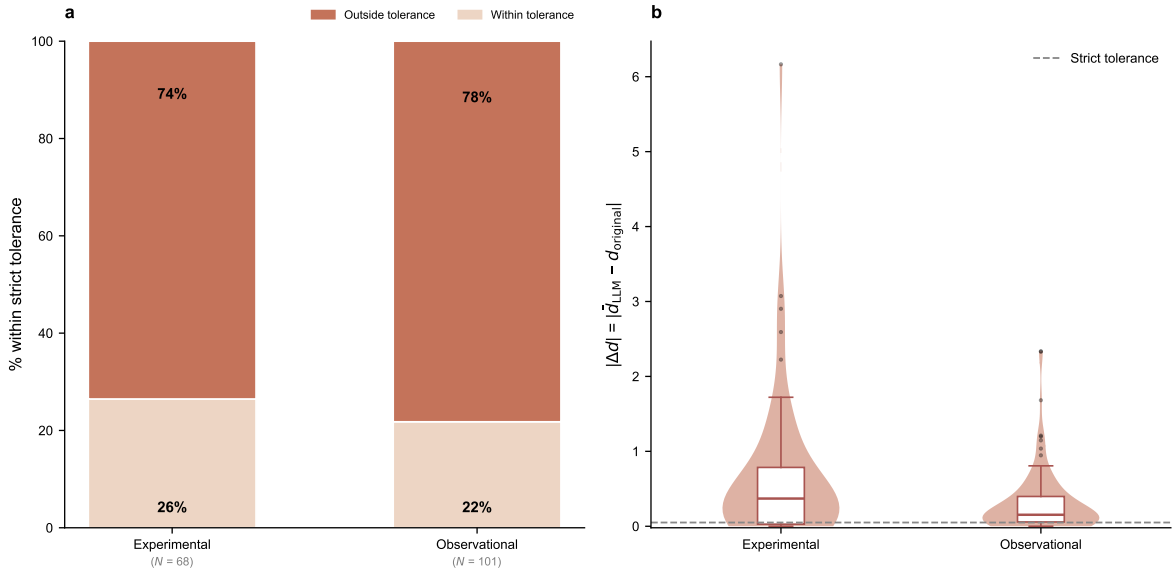


Fig. S7: **Reproducibility grouped by study type (experimental vs. observational). a,** Proportion of studies where the mean LLM-generated effect size falls within ± 0.05 Cohen's d grouped by study type. **b,** Distribution of absolute effect size deviations (i.e., $|\Delta d| = |\bar{d}_{\text{LLM}} - d_{\text{original}}|$) grouped by study type. In the boxplot, the line indicates the median, the box denotes the interquartile range (IQR), and the whiskers extend to 1.5$\times$ of the IQR; points beyond the whiskers indicate outliers. Results are based on Claude Opus 4.7 over five independent runs per study.

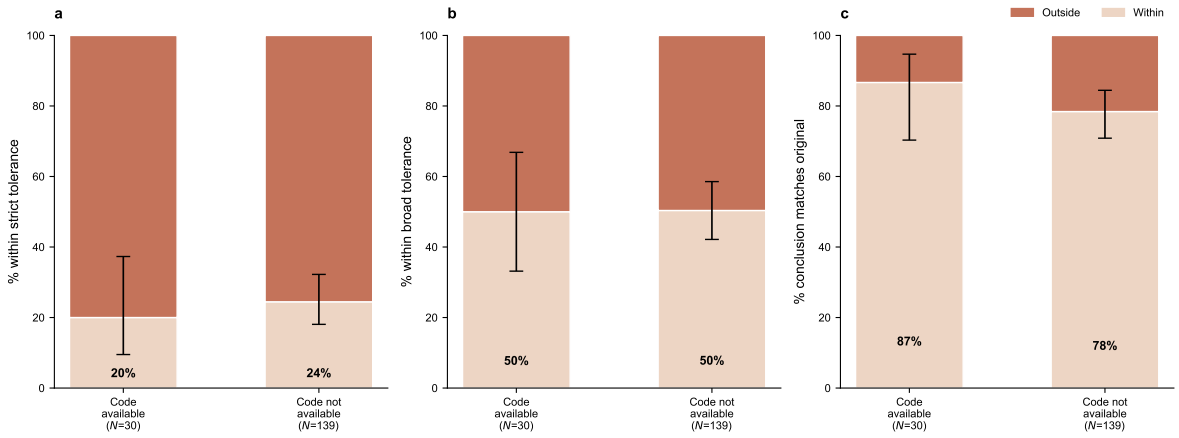


Fig. S8: **Heterogeneity by public code availability.** Studies are grouped according to whether the original analysis code was publicly available. **a,** Proportion of studies for which the mean generated effect size falls within the strict tolerance region (± 0.05 Cohen's d) around the original effect size. **b,** Proportion of studies for which the mean generated effect size falls within the broad tolerance region (± 0.20 Cohen's d) around the original effect size. **c,** Proportion of studies for which the majority-vote conclusion matches the original conclusion. Whiskers indicate Wilson 95% confidence intervals. Results are based on Claude Opus 4.7 over five independent runs per study.

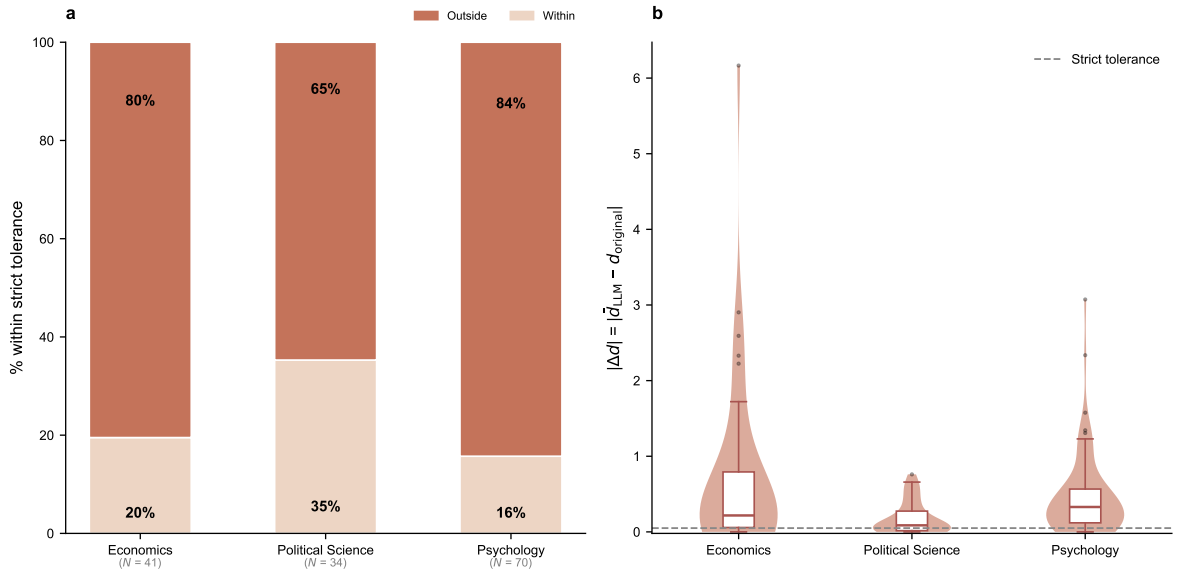


Fig. S9: **Heterogeneity by disciplines (economics, political science, psychology). a,** Proportion of studies where the mean LLM effect size falls within ± 0.05 Cohen's d of the original effect size, grouped by discipline. **b,** Distribution of absolute effect size deviations (i.e., $|\Delta d| = |\bar{d}_{\text{LLM}} - d_{\text{original}}|$) grouped by discipline. The dashed line shows the ± 0.05 tolerance threshold. In the boxplot, the line indicates the median, the box denotes the interquartile range (IQR), and the whiskers extend to $1.5\times$ of the IQR; points beyond the whiskers indicate outliers. Note that not all studies from the SCORE sample could be matched to the three disciplines. Results are based on Claude Opus 4.7 over five independent runs per study.

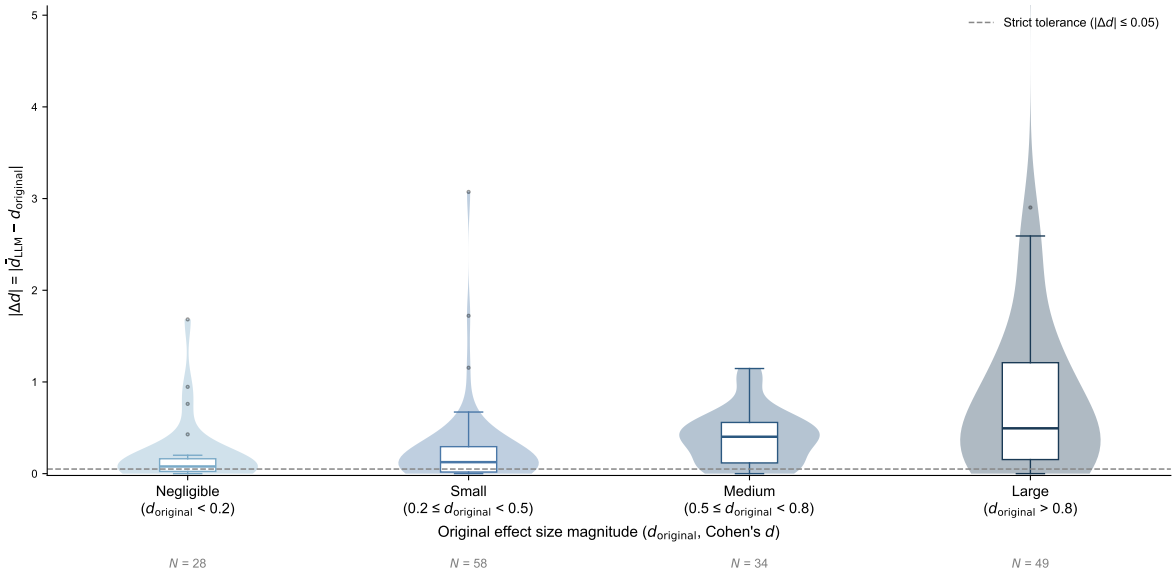


Fig. S10: **Reproducibility by original effect size.** Studies are grouped by the magnitude of the original Cohen's d: negligible (< 0.2), small ($[0.2, 0.5)$), medium ($[0.5, 0.8)$), and large (≥ 0.8). This analysis assesses whether reproducibility varies with the size of the originally reported effect. Under the strict criterion (i.e., $|\Delta d| \leq 0.05$), accuracy is highest for small (33%) and negligible (29%) original effects, but falls to 16% for large and 15% for medium effects. In the boxplot, the line indicates the median, the box denotes the interquartile range (IQR), and the whiskers extend to $1.5\times$ of the IQR; points beyond the whiskers indicate outliers. Results are based on Claude Opus 4.7 over five independent runs per study.

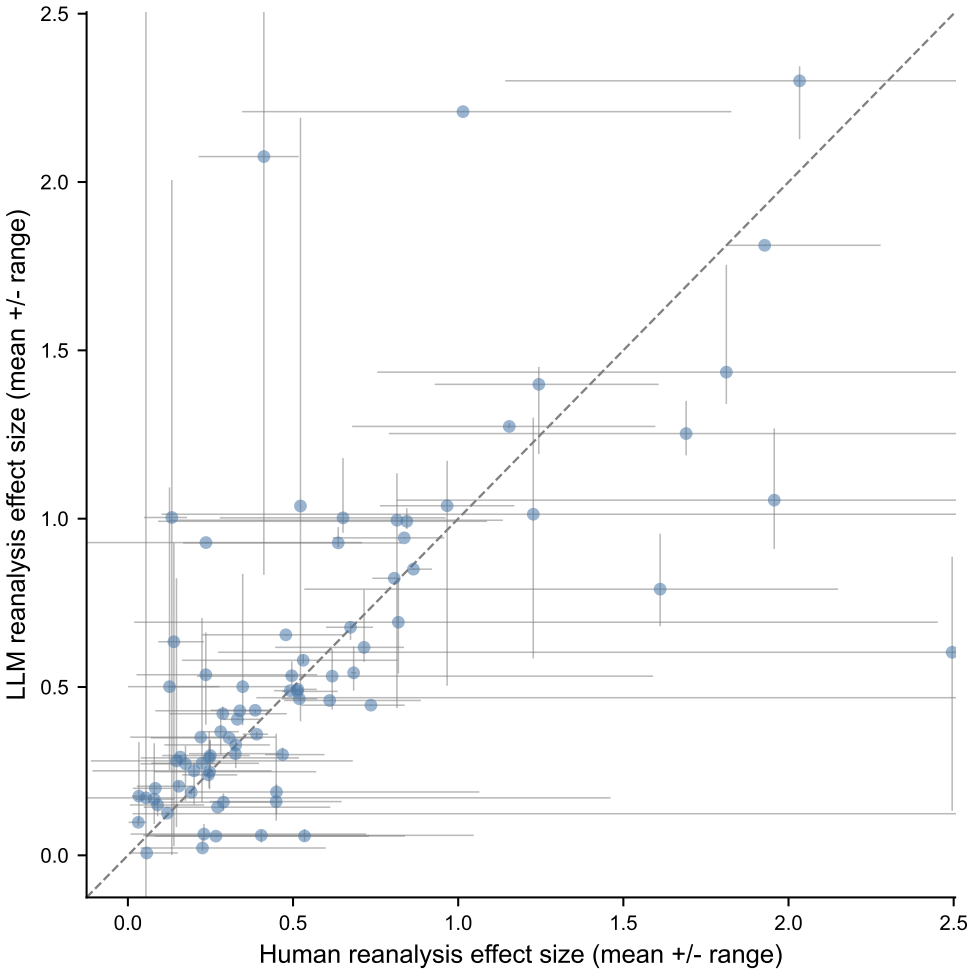


Fig. S11: **Analytical variability in LLM-generated and human reanalyses.** Each point represents one study from the subset with human reanalysis benchmarks. The x-axis shows the mean Cohen's d across human reanalyses, and the y-axis shows the mean generated Cohen's d across five independent runs. Horizontal whiskers indicate the minimum-to-maximum range across human analysts; vertical whiskers indicate the minimum-to-maximum range across LLM runs. Results are based on Claude Opus 4.7 over five independent runs per study.

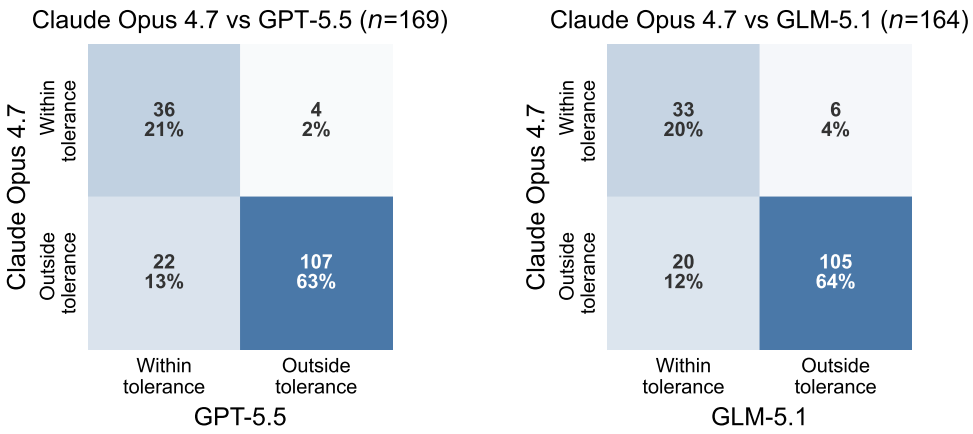


Fig. S12: **Cross-model agreement on effect-size within ± 0.05 Cohen's d (strict tolerance).** Confusion matrices comparing whether Claude Opus 4.7 and a second model recover the original effect size on the same studies, for RQ1 (full paper text, neutral framing). **Left,** Claude Opus 4.7 versus GPT-5.5 ($N = 169$). **Right,** Claude Opus 4.7 versus GLM-5.1 ($N = 164$). Each N is restricted to the studies for which both models produced a valid mean estimate. Cells report the number of studies and their percentage of N. Diagonal cells (both within / both outside) denote agreement, off-diagonal cells denote the studies where one model recovers the original effect and the other does not. Overall agreement is 85% (Claude Opus 4.7 vs. GPT-5.5) and 84% (Claude Opus 4.7 vs. GLM-5.1).

F Memorization test

We conducted a memorization test to assess whether model performance could be influenced by prior exposure to the analyzed papers. The prompt focused on two dimensions. (1) We asked whether the LLM recalled the paper, yielding a binary recall indicator. (2) To probe whether such recall extended to the focal quantitative result, we asked the model to output the Cohen's d for the focal claim and then checked whether the recalled value fell within the strict (± 0.05) or broad (± 0.20) tolerance region around the original effect size. Results and the exact prompt are reported below.

Table S2: **Memorization test.** The memorization test assessed (1) whether the model reported recalling the paper; and (2) when probed for the focal quantitative result, whether the recalled Cohen's d matched the original effect size within predefined tolerance regions. Although the model reported recalling 14 papers, none of the recalled Cohen's d values fell within either the strict (± 0.05) or broad (± 0.20) tolerance region.

Measure	**Result**
Papers reported as recalled by the model	14/180
Recalled Cohen's d within ± 0.05 of the original effect size	0%
Recalled Cohen's d within ± 0.20 of the original effect size	0%

Note: Results are based on Claude Opus 4.7.

Memorization test: prompt

```
I am studying a published paper and want to know what you know
   about its findings.

## Paper Reference
{paper_reference}

## Research Claim Being Tested
{claim}

Please answer the following questions based solely on your
   training knowledge (do NOT make up or guess, only report what
    you actually know):

1. Do you have knowledge of this specific paper in your training
    data? (yes / uncertain / no)
2. If yes: What is the main finding regarding the claim above?
   Be specific.
3. If yes: What is the direction of the effect? (positive /
   negative / null / unknown)
4. If yes: Report the main test statistic in this structured
   format, type (z/t/F/chi2/r), numeric value, degrees of
   freedom (if applicable), and sample size.
5. How confident are you in your recall of this paper's results?
    (1-10, where 10 = certain)

Fill in the block below. Use "unknown" for any field you do not
   know.

'''probe_results
PAPER_KNOWN: [yes / uncertain / no]
RECALLED_FINDING: [brief description of finding, or "unknown"]
RECALLED_DIRECTION: [positive / negative / null / unknown]
RECALLED_STAT_TYPE: [z / t / F / chi2 / r / unknown]
RECALLED_STAT_VALUE: [numeric value or unknown]
RECALLED_DF1: [numeric or null or unknown]
RECALLED_DF2: [numeric or null or unknown]
RECALLED_SAMPLE_SIZE: [integer or unknown]
RECALL_CONFIDENCE: [1-10]
'''
```

G Prompt

This appendix documents the full prompt configuration used in all LLM analyses. The prompt has three components: a fixed *system prompt* applied in every condition, a per-paper *user prompt* assembled from a template, and a condition-specific *perspective instruction* added to the system prompt only for the prompt framing analysis.

System prompt

```
You are a statistical analyst. The user will give you a published study
   's claim, the original dataset, and (depending on the experimental
   condition) either the full original paper or a partial version of it
   . Your task is to compute the test statistic most directly tied to
   the target claim from the data yourself on the original data and
   report the numerical statistic in the requested structured format.

## Tools

You have `python`, `bash`, `think`, and `view_image`. The Python
   sandbox includes pandas, numpy, scipy, statsmodels, pyreadstat,
   pyreadr, xlrd, openpyxl, h5py, mat73, oct2py, pyarrow, duckdb, py7zr
   , and pymupdf4llm. Octave is installed system-wide (callable from
   oct2py).
**State is not preserved. ** Each call to `python` or `bash` runs in a
   fresh process - variables, imports, and loaded data do NOT persist
   between calls. Write complete, self-contained scripts each time, or
   save intermediate results to disk (e.g., CSV or pickle) and reload
   them in subsequent calls.

## Data formats

Installed readers cover: `.dta .sav .por .sas7bdat .xpt` (pyreadstat),
   `.csv .tsv .txt .dat .json .xlsx` (pandas/openpyxl), `.xls` (xlrd),
   `.RData .rda .rds` (pyreadr), `.mat` (scipy.io / mat73 / oct2py), `.
   parquet .feather` (pyarrow), `.h5 .hdf5` (h5py), `.duckdb .db` (
   duckdb / sqlite3), `.zip .tar.gz .7z` (unzip / tar / py7zr).
If the relevant data file cannot be read with these libraries, submit
   immediately with `test_statistic: null` and explain why - do not
   attempt byte-level parsing or manual reconstruction.

## Operationalization
```

```
A verbal claim like "X increases Y" rarely pins down a single analysis
   - there are multiple reasonable choices of subsample, dependent
   variable, model, and controls that all "test the claim." When the
   original paper is provided, use its methods section to guide your
   operationalization choices. When it is not, commit to one specific
   reasonable operationalization based on the data structure and the
   claim, and execute it.

## Computing the result yourself

The reported test statistic must come from the analysis you run on the
   data. If the paper happens to be visible to you and reports a test
   statistic for the claim, you must NOT copy that number. Your job is
   to evaluate whether the claim reproduces from the provided data
   using the analysis most directly tied to the claim. Run the analysis
    in the execution environment and report the statistic YOUR code
   produces.

## Submission

Do NOT call submit() until you have actually run the analysis and
   obtained real numerical results from the data. Do NOT guess,
   approximate, or fabricate values from the paper or your prior
   knowledge - if you cannot compute a particular field, report null in
    it and explain why in your reasoning. The user prompt will specify
   the exact submission schema.
```

User prompt

The user prompt is assembled per paper from a fixed template. Placeholders in angle brackets (<...>) are filled with the paper's title, identifier, and claim, and with the sandbox paths to the dataset and parsed paper. The *Materials* block is the only part that varies across the RQ3 information-level conditions. The full-paper version is shown inline, with the redacted methods and abstract-only variants reported below it.

```
You have been assigned to reanalyze the following published study:

*Paper Title:* <paper title>
*Paper ID:* <paper id>
```

```
*Claim:* <claim>

This claim asserts the presence of a directional effect or relationship
    in the data. Your task is to conduct an independent analysis on the
    original dataset to test it.

## Instructions

Conduct **one analysis** on the dataset to test the claim. The verbal
   claim does not uniquely determine the analysis - there will be more
   than one defensible choice of subsample, dependent variable, model,
   and controls. **Commit to one specific operationalization** and
   execute it. Do not enumerate alternatives or report multiple results
   .

Your analysis must produce a **single, main result** expressed as a **z
   -, t-, F-, or chi-squared test** (or its non-parametric equivalent),
    with the test statistic, degrees of freedom, and sample size needed
    to compute a standardised effect size. The reported statistic must
   come from your analysis on the data - not from any number printed in
    the paper. Draw a single substantive conclusion from your result.

## Materials

Data directory: `<data dir>`
Original paper: Markdown file `<paper path>`.

Use the paper's methods section to guide your operationalization (
   subsample, dependent variable, model specification). The paper may
   also report a test statistic for this claim - do not copy it. Run
   the analysis on the data yourself and report the statistic your code
    produces. Tables are available as readable Markdown files (`tbl-N.
   md`) in the same `<paper dir>/` directory. Figures referenced inline
    in the paper Markdown (e.g. `![img-N.jpeg](img-N.jpeg)`) are
   present in `<paper dir>/` - use the `view_image` tool to look at one
   .

## Reporting Results

When you are done, call `submit()` with a JSON string containing:
```json
{
  "type_of_statistic": "z | t | F | chi2 | r",
  "test_statistic": <numeric value>,
  "degrees_of_freedom_1": <numeric value or null>,
```
```

```
  "degrees_of_freedom_2": <numeric value or null>,
  "sample_size": <integer>,
  "p_value": <numeric value or null>,
  "conclusion": "same conclusion as original study | no effect or
   inconclusive | opposite effect",
  "reasoning": "<your conclusion in 1-3 sentences, referencing the
   actual numbers>",
  "dependent_variable": "<dependent variable or outcome used>",
  "main_predictor": "<main predictor / treatment / grouping variable
   used>",
  "sample_definition": "<which observations were included/excluded>",
  "model_specification": "<statistical model or test used>",
  "controls": "<controls/covariates used, or none>",
  "operationalization_notes": "<brief rationale for these analytic
   choices>"
}
```

For RQ3, the *Materials* block above is replaced as follows. Redacted-methods variant:

```
## Materials

Data directory: `<data dir>`
Original paper (methods/research-design section redacted): Markdown
   file `<paper path>`.

The paper's introduction, results, and discussion are visible, but the
   methods / data / research-design section has been removed and
   replaced with a short redaction marker. You must therefore decide
   the operationalization (subsample, dependent variable, model,
   controls) yourself. The paper may report results numbers for this
   claim - do not copy them; run the analysis on the data and report
   the statistic your code produces. Tables are available as readable
   Markdown files (`tbl-N.md`) in the same `<paper dir>/` directory.
   Figures referenced inline in the paper Markdown (e.g. `![img-N.jpeg
   ](img-N.jpeg)`) are present in `<paper dir>/` - use the `view_image`
    tool to look at one.
```

Abstract-only variant:

```
## Materials

Data directory: `<data dir>`
```

```
Original paper (abstract only): Markdown file `<paper path>`.

Only the title block and abstract are available - you do NOT have the
   methods, data description, results, or discussion. You must decide
   the operationalization (subsample, dependent variable, model,
   controls) yourself based on the claim and the structure of the data.
```

Perspective instructions

The two perspective instructions are reported below: **Confirmatory:**

```
Approach this analysis as a supportive analyst who expects this claim
   to be empirically robust and to hold up under reanalysis. Follow the
    most direct analysis implied by the available materials. If a
   necessary analytic choice remains ambiguous, resolve it in the way
   that provides the most charitable defensible test of the claim while
    still remaining fair. Commit to this one analysis and report what
   it shows.
```

Critical:

```
Approach this analysis as a skeptical analyst who questions whether
   this claim is as robust as the authors assert. Follow the most
   direct analysis implied by the available materials. If a necessary
   analytic choice remains ambiguous, resolve it in the way that
   provides the most demanding defensible test of the claim while still
    remaining fair. Commit to this one analysis and report what it
   shows.
```