International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 10 | Oct - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

A Multi-Agent Scientific Audit Platform Powered by Vision-Language Models for Detecting P-Hacking in Clinical Research

Chirragoni Prasanna Kumar, Masura Anudeep, Mohammed Mustafa Rafi and Swathi Rudra

Department of Computer Science and Engineering (Artificial Intelligence & Machine Learning),

CMR Technical Campus, Hyderabad, Telangana, India

Abstract— The results from clinical research influence treatment guidelines, clinical decisions and health policy; therefore, the results from clinical research should be clearly communicated and reproducible. One danger to that is hiding analytical flexibility: running a couple of analyses, a couple of exclusions and either a few definitions or a few outcomes and only reporting the ones which work out in the favour to the reader. We introduce the multi-agent Scientific Audit Platform which incorporates LLM reasoning ability and VLM figure understanding with the deterministic Document Processing and Statistical Validation. Reads clinical paper; reads text and data numbers. The source content is extracted from the PDFs and OCR, the LLM analyzes the methods used, the outcomes and the claims are extracted and analyzed, the charts are interpreted with the help of the VLM, the specialist agents cross-check each other, and the statistics that can be recomputed are recomputed by means of deterministic routines. These evaluation resources include 30 trials that were prebuilt to be randomized to supply the business wall, 47 likely inform of mutations of reporting flawed randomizations in PDFs, 38 papers with human verdicts for outcomes that changed, 30 new trials, 242,565 statistics for comparisons with statcheck, and 60 synthetic charts with 326 known values. All 47 planted reporting errors (recall 1.00; 95% CI 0.92–1.00) were identified by the deterministic, statistical checks. There were no significant discrepancies in agreement with statcheck (100% for recomputed p-values, 99.58% for inconsistency flags and 99.85% for decision-error flags). The PDF body-text recall was 0.991, table-cell recall was 0.944 and figure localization was of 72/74 (0.973). Outcome switching screening was poorly accurate (63.2%), poor at finding any positives (27.3%), poor at finding all positives (33.3%) and had an F1 score of 0.30 compared to the 71.1% of an always changed baseline. These numbers denote the deterministic extraction, statistical and localization parts. These are not model / VLM / LLM accuracy scores and aren't numerically represented in the semantic interpretations produced by the model. The platform is designed to serve as a 'human in the loop' system to help with screening for research integrity purposes. It is not going to be able to prove p-hacking, fraud and intent. There are six specialist roles that come under the Coordinator: Document Ingestion, Statistical Audit, LLM-based Methods and Research-Integrity Analysis, VLM-based Figure Analysis, Cross-Verification/Skeptic Review and Reporting. Numerical results for particular components on particular datasets. Note that they are not an exhaustive test of all of the LLM/VLM's capabilities.

Keywords— Multi-Agent Systems, Large Language Models, Vision-Language Models, Research Integrity, P-Hacking Indicators, Clinical Research, Statistical Reporting, Evidence Provenance.

I. INTRODUCTION

Evidence-Based Healthcare is based on published evidence. Clinical trials influence clinical guidelines, drug approvals, treatment and health policy decisions, and a paper cannot NOT help or hurt, but a paper that can't be trusted does help or hurt. Concurrently, a drive to publish and to find some ‘significant result' incentivizes dubious research techniques such as p-hacking.

P-hacking involves performing various analyses, pruning variables and dropping observations as well as performing changes in analytic decisions to get the p-value below 0.05 or other conventional p-value cut-off. It is not necessarily the intentional cheating that makes it wrong. Regardless, it increases the likelihood of spurious positive results, and also creates difficulties in obtaining repeat measurements.

Peer review and manual statistical audit is time consuming and requires specific skills. Reviewer also are more likely than problems related to Prose & may miss the problems on Tables, Figures or Supplementary material. In the context of scientific documents, LLM can be used to provide a new perspective than what's possible with the traditional methods of examining word counts in documents and VLMs, especially, can read the text along with survival curves, scatter plots, forest plots and other charts.

Multi-agent design is dividing the work among a set of models instead of having one pack up all of the jobs. Evidence is considered by each agent and returns structured findings that are then used when consolidating the findings into a final assessment.

We present here a multi-agent Scientific Audit Platform to pre-selection of clinical research papers to signs that could be indicators of further studies. It includes document parsing, inference based on LLM on text, VLM reasoning about figures, and deterministic statistical validation. Contributions include an orchestrated multi-agent architecture, multi-agent analysis of text, tables and figures, evidence-based results, cross-verification of the results with each

specialist agent, and evaluation of the component-level data on controlled and reference data sets. In short, the platform is capable of performing the following six things:

- Makes a text copy, table copy, caption copy and page layout extract of PDFs, if necessary using OCR.
- Employes an LLM to scan the pulled text for studying methods, outcome definitions, reporting omissions, statistical statements and problems with questionable research practices.
- Interprets figures and charts (axes, legends, trends, visual claims) by using a VLM and making comparisons with the appropriate text.
- Computes supported p-values, test statistics, and confidence intervals, estimates and sample-size relations deterministically.
- Ashes, check if each observation is from the same paper by specialist agents or confirm the observation each other.
- Creates an evidence based audit report with three discrete sections (concerns level, evidence confidence and audit coverage).

II. LITERATURE REVIEW

A. Statistical Detection of P-Hacking

There are several statistics methods which try to detect questionable research practices. An unusual pile up just below the significance level can be detected using P-curve analysis of the distribution of significant p-values. Reported significance is found in a z-curve analysis and can have the ability to say something about publication bias. Numbers are the sole basis for both, however, and neither film is able to discern whether the numbers in a paper actually substantiate what the paper indicates may be true.

B. Vision-Language Models in Scientific Analysis

In Vision-Language Models, visual perception is coupled with language understanding. VLM is central to this platform, and not an add-on package – it gathers, processes and understands scientific figures and charts. It provides the key elements of the chart, reads axes and legends, interprets the trend of the plotted data and reading from axis that are legible, and compares findings in the figure with those in the Results section. The output should be restricted to the indicated size of the figure crop and confirmed against the paper and should never be faked to be a value which is uncertain or cannot be read from the paper and is reported as such. Though finding a figure and/or reading its printed caption is a different thing from understanding the plotted data, the VLM is used in conjunction with a number of deterministic localization algorithms and OCR techniques.

C. Multi-Agent Artificial Intelligence

The platform is an “MAS” (multi-agent system) and structured evidence is passed between agents with different roles via a Coordinator. Deterministic tools extract the information and make numerical calculations which are repeatable; the LLM and VLM reason and perceive to enable repeatability in those steps. The agents can be workflow roles, each having their own prompts and tools, and don't have to be a separate model to be trained. Along with the tracking of state, routing and agent output, LangGraph coordinates.

D. Research Gap

Existing stat checkers are directed towards number reporting. Document parsers extract text and formatting content, but don't understand the meaning of the research. Our platform applies deterministic extraction and calculation to the text of the paper, as well as LLM reasoning applied to the figures in the paper via VLM reasoning. It addresses the gap of concerted, evidence-based synthesis of these capabilities, cross-verification and reporting targeted at human reviewers.

III. DATASET AND METHODOLOGY

A. Dataset

We critically review clinical research papers, and public reference resources that are openly available. The details of starter and fresh trial sets and the details of the planted-error benchmark are given in Table I, along with the outcome-switching set, the statcheck comparison set, and the synthetic chart benchmark and their size. Record-level identifier(s), project collection dates and item-level license mapping should be in the project manifest included with the final reproducibility package. According to the source study, the trial PDFs are CC BY or CC0, except for those 12 CC BY NC PDFs - they are used locally but not offered to be redistributed.

Starter trails were conducted with the 30 trials to develop and tune the reading rules. The 30 fresh trials were for development but not used for the 30 trials set.The 30 trials set were not the 30 trials used for development. The set of outcome-switching papers is not an untouched set of papers, since four of these papers informed reading fixes. Instead the 47 planted errors are mutations intentionally placed in the code to test known, but not examples of, inconsistencies in error reporting. Only relative proportions are meaningful for the fresh trial indicator counts and they will not be understood as precision, accuracy or prevalence if the labels are provided by humans.

TABLE I. EVALUATION SETS.

Set	Content	Size	Used for
Starter trials	Open-access randomised trials from PubMed Central, with the publisher XML	30 papers, 618 statistical results	Development; reading and extraction accuracy
Planted errors	Starter PDFs with known errors written into their text	47 errors in 19 PDFs (36 interval, 11 p-value)	Recall of the statistical checks
Outcome switching	Trials with a human verdict on whether the primary outcome changed [14]	38 papers (11 changed, 27 unchanged)	Accuracy of indicator #5
Fresh trials	Open-access randomised trials not used during development	30 papers	Indicators per paper on untouched papers
statcheck reference	Reported test statistics with statcheck's own verdicts [13]	242,565 statistics	Agreement of the statistical checker
Synthetic charts	Bar, survival, forest and line charts drawn from known values, as vector and as image	60 charts, 326 values	Error of figure reading

B. Proposed Methodology

The system is a parallel multiagent system. Document Ingestion will first try to extract machine-readable text and layout from a PDF by scanning it, adding optical character recognition if necessary, then extract the text as PDF Text. The structured paper is then handed out to the specialist agents by the Coordinator. These specialized types of AI assistants represent specific customizations of ChatGPT that are optimized for specific tasks: the LLM-based text agent can analyze methods, outcomes, claims, reporting or other written documents, while the VLM-based figure agent can make sense of charts, visual information or evidence. The Statistical Agent, independently, can recalculate specific numerical results that have been supported. Methods Agent verifies of reporting Study Design. These outputs are compared each other, evidence verified and any conflicts are retained by Cross-Verification and Skeptic Review. The agent in charge of the property provisionally reports a final report connected with the audit of the source.

The Coordinator is responsible for getting agents to execute; agents communicate only with the Coordinator, and not with other agents. They are all assigned a specific task, assigned with the necessary document information and a predefined output document type. The Coordinator forward-checks outputs to the downstream checks, documents failures and ensures final results maintain their reference, including to their page, section, table, and/or figure. LLM (large, long-term memory) + VLM (very long-term memory) = reasoning and perception. Deterministic parsing and statistical calculation remains in the workflow and the numbers and calculations are reproducible and can be audited independently.

TABLE II. ROLES OF THE PLATFORM'S COMPONENTS.

Component	Main responsibility	Example evidence/output
Document Ingestion Agent	Extract and structure paper content	Sections, tables, figure references, source locations
Statistical Agent	Check numerical reporting and p-value patterns	P-values, confidence intervals, effect sizes, sample-size notes
Vision-Language Agent	Review scientific figures against written claims	Figure interpretation and uncertainty notes
Methods Agent	Review design and reporting safeguards	Randomization, blinding, power calculation, multiple testing
Cross-Verification Agent	Compare findings across agents	Agreement, conflicts and evidence links
Reporting Agent	Prepare a readable, traceable audit	Risk score, category, evidence and reviewer prompts

1) Document Ingestion Agent

The Document Ingestion Agent takes care of preparing the paper for the rest of the pipeline. PyMuPDF can extract text, page coordinates and document structure from text—and OCR-ing comes in when it is not possible to get text from a page or the text layer is not very reliable. This does not require LLM nor VLM, it simply makes the PDF into structured data as an input. The LLM then understands the text the information was taken from and determines methods and outcomes, matches statements in parts and suggests findings based on evidence. Tables and figures are placed with their pages, with crops for the relevant specialist agents to analyse. The occurrences of increasing failure to extract, malformed tables, multi column reading-order and missing figures should be reported rather than be maintained without comment.

Extraction is NOT comprehension. The characters and their positions from the PDF file are extracted, characters on the scanned pages are recognized by the OCR. Neither offers any clues as to meaning of content. The LLM's task is to analyze the data recovered and deduce the study design, main results, methodology, limitations, and study connections. The VLM takes in figure images and/or crops and for charts, it does the same. The outputs are then shared with the deterministic calculations within an agent operating between different agents to check with the source evidence.

2) Statistical Agent

The Statistical Agent combines deterministic computation with the assistance of the LLM to interpret the statistical context. Deterministic routines: Recalculation of supported p-values, checks of the relationship between test statistics, confidence intervals, estimates, counts and sample size. The LLM informs on what reported values should be grouped together, reads the context and provides reasons for why a discrepancy might need to be reviewed. It should never include missing numbers, nor represent some machine-executable calculation. For each finding, include the reported value, its recomputed value, the assumptions

and the location of the source and categorize its status as not-assessable if the necessary information was not provided.

3) Vision-Language Agent

The Vision-Language Agent also can describe scientific figures like Kaplan–Kamir curves, forest plots, scatter plots, ROC curves, bar charts and error-bar plots. Applies a figure crop and its caption; calculates the type of graph, axes, units, legend, patterns to plot and values to display, and compares it with statements in the Results section. It must give the number of the figure and the page, record its error and must not guess the value of axes or marks when these are not readable. The figure is detected in deterministic localisation and OCR, and it is analysed by the higher level interpretation work performed on the VLM. The 72/74 figure result is not a semantic accuracy number for the VLM (as such, it cannot be said to be due to the VLM), but a value representing localization.

4) Methods Agent

The Methods Agent reads the extracted methods, protocol text and research results of the electrodata against structured research-audit indicators using the LLM. It determines if randomization, allocation concealment, blinding, and justification for sample size are reported as appropriate, if applicable; it distinguishes between a missing description and demonstration that a procedure was not conducted; and it distinguishes between a missing description and evidence that a procedure was not conducted. When a trial registration exists, the agent goes through the same process as above and notes the passage of comparison for outcome population/outcome intervention/comparator/outcome measure/metric and time point.

5) Cross-Verification and Reporting Agents

The Cross-Verification Agent integrates the results from the Statistical Agent, Methods Agent, LLM Text-analysis Agent and VLM Figure-analysis Agent. It verifies that it does fit their textual evidence to get the reader to the intended claim; see whether the same outcome or result is repeated in sections; reviews whether the visual evidence matches the claim. The Skeptic Review stage ridicules what goes without any sort of evidence, clarifies the source and location of any quotes, and lists disagreements rather than too well "silences" disagreements. It is a rule of thumb that an LLM explanation doesn't constitute a truth-based explanation. When the figure is unclear, a VLM interpretation continues to be connected to the original figure crop, and indicated by an 'uncertain' tag.

The Reporting Agent converts verified findings into a report which can be examined by a researcher or editor. Each of the issues should identify the responsible agent and indicator identifier, as well as provide the page number, section, table, or the figure number for the evidence, include the evidence being quoted, if applicable, a calculation path, the evidence grade and any uncertainties and limitations. There is a separation between concern score, evidence confidence and audit coverage in the report. If the finding has no legitimate source location or if it would not be possible to run the check then it is noted as Unverifiable or Cannot Assess.

6) Implementation

The platform is a multi agent application with a modular structure. The combination of Python and document processing, Gemini 2.5 Pro is the LLM component, Qwen2.5-VL is the vision-language component for scientific figures and LangGraph is for orchestrating workflows and shared-state routing. Gemini is trained to analyze text and relationships between text while Qwen2.5-VL is able to understand figure images and information in graphs and charts. The agents are not stand-alone foundation models, but rather the segments of workflows required for each specific type of agent. Each run of the audit must include the following: model identifier and version, prompt version, audit configuration, code revision, code execution status (for reproducibility).

The design is of Hybrid Type. Source text and layout are extracted and calculated statistically by deterministic routines, source text analysis and analysis of methods, outcomes and cross-section claims by Gemini 2.5 Pro and analysis of charts and figures by Qwen2.5-VL. The LLM and VLM are crucially important tools in the workflow but do not take the place of the parser, or the numerical checker. LangGraph takes its structured outputs to Cross-Verification and Skeptic Review. All statements said in the models have to be verifiable to the uploaded paper: quotations must be quoted exactly, visual statements refer to the particular figure/crop, no reference to the sources and they may be unsupported or doubtful.

The 14 indicators are listed on the platform. Four report errors (1-4): p-value, interval and estimate mismatch, failing to have counts add up, having means that are not compatible with their stated sample-size limits. Five to 13 point: outcome switching, inconsistent primary outcomes, promised not reported outcomes, uncorrected multiplicity, p-value clustering near .05, spin, reporting of outcomes from different subgroups without testing or reporting for interaction, design-reporting gaps and incomplete reporting. The last (14th) flag represents a clash between a figure and text. The LLM is fed contextual clues for the text-based cues; the VLM is used for reviewing figures; and where the cues are numerical, and where there are some, deterministic statistical checks are also used to validate numerical relationships. None of the items on this list is proof of or intent.

IV. RESULTS AND DISCUSSION

A. Evaluation setup and metrics

The platform has model-based agents and non-modelbased parts. Finally, Section III-B explains how

the LLM processes paper text, and how the VLM processes scientific figures. The following numbers have to be) read against the benchmark that has generated the number. None of the metrics shown above is an independent measurement of accuracy for the LLM or VLM; they are all measures of the specific component tasks. Measured LLM/VLM accuracy would require an evaluation of the model level, which would give the model, provider and version, the prompts, the setting, the labelled examples, and a breakdown of human reference labels together with an error analysis.

Precision is the number of true positive tests/total number of positive cases predicted. Recall is the percentage of people on whom a diagnostic test is performed who have the condition. It is a combination of the two, the F1-score.

$$Precision = TP / (TP + FP) \quad (1)$$

$$Recall = TP / (TP + FN) \quad (2)$$

$$F1 - Score = 2 \times (Precision \times Recall) / (Precision + Recall) \quad (3)$$

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \quad (4)$$

TABLE III. EVALUATION METRICS.

Metric	Formula	Interpretation
Precision	TP / (TP + FP)	Of the papers flagged, how many were labelled positive?
Recall	TP / (TP + FN)	Of the labelled positive papers, how many were flagged?
F1-Score	2 × (Precision × Recall) / (Precision + Recall)	Harmonic mean of precision and recall
Accuracy	(TP + TN) / (TP + TN + FP + FN)	Fraction of all papers classified correctly

TABLE IV. SUMMARY OF MEASURED COMPONENT RESULTS (SEE TEXT FOR THE TASK AND REFERENCE USED BY EACH METRIC).

Task	Reference data	Measure	Result
Planted reporting errors	47 errors in 19 PDFs	Recall	47/47 = 1.00 (95% CI 0.92–1.00)
Planted decision errors	11 errors among the 47	Recall	11/11 = 1.00 (95% CI 0.74–1.00)
Agreement with statcheck	242,565 statistics	p-value / error flag / decision error	99.99% / 99.58% / 99.85%
PDF reading	30 trials, publisher XML	Recall of body text / table cells	0.991 / 0.944
Result extraction	618 results in the XML	Recall	0.960
Figure detection	74 numbered figures	Recall = precision	0.973 (72 of 74)
Outcome switching (#5)	38 papers, 11 changed	Accuracy / precision / recall / F1	63.2% / 33.3% / 27.3% / 0.30
Indicators on fresh trials	30 papers	Indicators per paper	0.9 (19 papers with at least one)
Runtime, CPU only	30 starter trials	Median / 90th percentile per paper	0.71 s / 1.0 s

B. Detection of planted reporting errors

We created 47 deliberate statistical errors in 19 of the 30 starter pdfs by altering only the char properties to be different.To check if the deterministic statistical checks also got to the report we deliberately altered the char properties so that 47 starters pdfs contained some errors in 19 of them. An estimate was moved out of its confidence interval in 36 cases. For item 11 the p value has been altered, and this changed the reported level of significance from being non-significant to cut across the board to significant. All 47 were detected (recall 1.00; 95% CI 0.92–1.00). This means that the checks identify known/controllable inconsistencies. It does not mention ‘p-hacking detection’ in real life or the ‘reasoning process of LLM/VLM’.

C. Agreement with statcheck

We ran a comparison between the statistical checker and statcheck with 242,565 reported tests statistics from psychology papers [13]. Recomputed p-values reached an agreement of 99.99%, inconsistency flags and decision-error flags attained agreements of 99.58% and 99.85% respectively. This is in agreement with another automated checker and not in agreement with ground truth. The two tools differ with respect to policy regarding p=.000, two-tailed and one-tailed testing, and more.

D. Reading accuracy

30 starter papers were directly compared with publisher's XML for reading accuracy. NX 11 body text recall averaged 0.991 (range 0.974-1) compared to pdfplumber's 0.744 (range 0.487-0.889). Table-cell recall improved to 0.944 as compared to pdfplumber's 0.895. These numbers refer to the parsing portion which supplies structured input to the LLM (and to the other agents). The LLM is NOT a PDF Reader. It renders the retrieved text, makes connections between the sections and offers potential findings that are backed with source

citations. The metrics as such are used to evaluate the parser and not the LLM's reading comprehension.

E. Outcome switching

The primary outcome of a paper is compared with the outcome that was registered at the beginning of the trial in indicator 5. It detected 9 papers (3 true positive, 6 false positive) out of 38 open access papers with human verdicts, both low and high.It yielded 3 TP and 6 FP from 38 papers with human verdicts for high and low, respectively. It failed to record 8 'changed outcomes'. Accuracy was 63.2%, precision 33.3%, recall 27.3% and F1=0.30. Current outcome comparator has the property of being never changing, for which the accuracy is only .711 (> .5), which is not superior to the trivial baseline. Four papers contained information about development fixes; the positive sample size was small.

TABLE V. INDICATOR #5 AGAINST HUMAN VERDICTS ON OUTCOME SWITCHING (38 PAPERS).

Human verdict	Flagged by #5	Not flagged
Outcome changed (11)	TP = 3	FN = 8
Outcome unchanged (27)	FP = 6	TN = 21

The error analysis shows that the misses of the error results from papers not printing their registration number (2), to primary outcomes which are not noticed in the paper (3), and the lexical outcome comparator (3 misses and 6 false alarms: some of the latter responds to an alternative sentence that the fail to notice as primary outcome). Because until now an indicator 5 has not been trailed on the same papers as a language-model-based comparator, it is recommended in this case as a prompt to have the papers considered by human eyes.

F. Figures

Using graphics rather than text to anchor their figures, figures were localized. Of 72 given total numbered figures, 74 were found to be matched (precision=recall=0.973) against the figures listed in publisher XML. That is a measure of localization and NOT the semantic reading of chart content by the VLM. There are 326 rows of known data and 60 rows of charts in the synthetic chart resource. The accuracy of the recovered values with the text-based reader was 0 (6.1%) and the remaining 100% were recovered by failing to recover. You need a new labelled chart benchmark with chart-type accuracy, axis interpretation, interpretation of the legend, value extraction, trend interpretation, mean absolute error and failure cases to assess needs for the VLM. Those model specific metrics are not included above.

G. Indicators on untouched trials and runtime

The platform remarked up 27 indicators (0.9 per paper; 5 strong and 10 moderate indicators) on the 30 fresh trials which were not utilized by the time of the development; 19 papers went up at least one. Twenty-three papers received low concern scores and seven papers received moderate scores with no high scores received. It administered 33 indicators (1.1 per trial) at the 30 development trials. These numbers indicate the number of prompts that a reviewer should read. They're not precise, since there's nothing yet independent that has rated the findings. The median added up to 0.71 s/per paper (90% percentile: 1.0 s, maximum: 1.7 s) on a CPU-only laptop running an unknown version of OSX, with parsing of the PDF being responsible for 0.62 s of this.

The multi-agent design is an important one as the audit requires other types of analysis. Document Ingestion extracts the PDF content, the LLM-based agent interprets text, methods and outcomes, the VLM based agent interprets figures, Statistical Agent does reproducible calculations and Cross-Verification/Skeptic Review checks the findings from the source. The numbers in this table are some selected values of component measurements What orchestration adds to that is still in question, however, in an ablation study that relies on a comparison with the baseline (single-component and/or deterministic).

The system will surface patterns that might deserve to be looked at more closely; it is unable to determine intent from a paper alone. Rounding, study design, pre-registration and model assumptions and valid exploratory analyses can all influence reported p-values. A final report should, for these reasons, include the evidence and from where it was obtained, differentiate lacking reporting of evidence from identified methodological issues, and leave a final opinion to a well-informed judge.

H. Limits of the evidence

There are limits of the evidence! The benchmark metrics include deterministic extraction, statistical consistency, outcome switching comparison and figure localization. All are using unlabelled accuracy evaluations of the LLM or VLM. The outcome-switching outcome is lower than the baseline that is never changed. No p-hacking labeled ground truth dataset is used, which does not allow the system to claim validated p-hacking detection. The planted errors are not confirmed misconduct, but apologies to Muriel.Contingencies of planted errors, not confirmed misconduct, but apologies to Muriel. Further evaluation for LLM semantic analysis and chart interpretation by VLM is still required to be task specific human labelled; the contribution of multi agent design requires an ablation study.

I. Model-level evaluation of the LLM and VLM

Evaluation of the relevance of this evidence. Sections IV-B–IV-G report measurements taken per component with none evaluating a model. The PDF-reading, statcheck-agreement and planted-error results are generated by deterministic routines, the figure result shows how well the captions can be localized by the picture, outcome-switching is generated by a lexical comparator, not by Gemini 2.5 Pro. We currently do not have the results of Gemini 2.5 Pro (LLM) or Qwen2.5-VL (VLM) labeled in the manuscript. The model-level

evaluation protocol is thus restated and extra markers placed to show results. Entries (along with the protocol) have not been measured where marked [TO BE COMPLETED] and described experiments remain to be run.

Level A vs. Level B evaluation. A model-level evaluation consists of evaluating one model at one task and compares the output with human reference labels or known values without the agents in the downstream. The measurements listed in Sections IV-B – IV-G and a future ablation study do so by evaluating the final audit generated by the whole pipeline (parsing, LLM, VLM, statistical routines, Cross-Verification and Skeptic Review, and Reporting) as a platform-level evaluation does. Both the platform level and the model level results are reported separately since a good platform level gives no indication for the accuracy of either model, and an accurate model does not give any indication for the validity of the audit. For each model, the target has just the extracted text text (LLM) and/or the figure crop with its caption and the figure context (VLM).

LLM role and evaluation. The LLM (Gemini 2.5 Pro) is used in the platform to process the text extracted from the PDF by PyMuPDF and OCR. It points to the study design, main and auxiliary outcomes, elements of the methods reporting and claims, and provides candidate results accompanied with sentences from sources. There are proposed to be three evaluation parts available for the proposed evaluation. (i) Extraction and comparison of existing 38 papers and its human verdicts (outcome switching) (indicator 5). The four used for the development fixes are reported separately, and the baseline (always unchanged: 71.1% accuracy) and lexical comparator (63.2% accuracy / F1=0.30) are compared on the same papers. Quantitative analysis of: (ii) Extraction of methods-reporting items (randomization, allocation concealment, blinding, sample-size justification, prespecified outcomes, multiplicity handling) compared to independent human annotation of a held-out set of trial papers (number of papers: TO BE DETERMINED). (iii) The LLM returns the quotations and the candidate findings and source grounds the quotations.

Evaluation and role of VLM. VLM (Qwen2.5-VL) interprets scientific figures: chart type, axes, units and legends, values plotted (if they can be read), agreement between a figure and the Results text, trends, and indicator (14) for agreement between a figure and the Results text. This proposed evaluation is broken down into four parts: (i) to identify the type of chart, the labelled axes, the units of measure, and the legend entries; (ii) to extract values from the chart; (iii) to interpret the trend and the direction of the effects; and (iv) to determine consistency between figures and text, including correct abstention when marks are unreadable. Ground truth is the synthetic chart resource with existing human images (60 charts) and human-annotated real figures (in the trial PDFs – number of figures TO BE DETERMINED) used as reference data; the existing human images are also available in a vector shape. Currently, the sum of exact recovery results in a text based reader of the printed forest-plot values is 20 out of 326 values (6.1%), so this can be compared when the VLM is used on 60 such charts.

Verification of references with annotations In synthetic, all the VLM output values are compared with known output of generating values. In the case of real papers, the two annotators label the same paper independently from the original PDF and a third individual makes the decisions on sections of disagreement; inter-annotator agreement (kappa coefficient) is reported prior to scoring the model. These are matched with the outputs of the model: via normalized string matching for extracted outcome, quantity, and via numeric matching of values from a chart (proposed as a fixed fraction of the axis range to be prespecified), and by manual adjudication for free-text explanations. All the LLM quotations are compared word-for-word with the source text from which their answers were derived, and all the VLM statements are compared to the exactly saved figure set crop. If the location cannot be verified by the source, then the outputs are unsupported. The model identifier, prompt version, temperature and other settings are noted and a development set is separated from the test set to ensure that the prompts are not tuned on test items.

TABLE VI. PROPOSED MODEL-LEVEL EVALUATION PROTOCOL FOR THE LLM AND THE VLM

Model	Task	Reference	Metrics
LLM (Gemini 2.5 Pro)	Primary-outcome extraction and comparison (#5)	Human verdicts on the 38-paper set; annotated outcome passages	Accuracy, precision, recall, F1 against the always-unchanged baseline; exact/normalized match of extracted outcomes
LLM (Gemini 2.5 Pro)	Methods-reporting items	Double human annotation of a held-out paper set	Per-item precision, recall, F1; Cohen's kappa with annotators
LLM (Gemini 2.5 Pro)	Source grounding of quotations and findings	Verbatim match to extracted text; human support check	Verbatim-quote rate; unsupported-finding rate; run-to-run consistency over repeated runs
VLM (Qwen2.5-VL)	Chart type, axes, units, legends	Generation metadata of synthetic charts; annotated real figures	Accuracy per field
VLM (Qwen2.5-VL)	Value extraction	326 known values on 60 synthetic charts; annotated real figures	Mean absolute error; relative error; share within pre-specified tolerance; exact recovery rate
VLM (Qwen2.5-VL)	Trend interpretation; figure–text consistency (#14); abstention	Human labels	Accuracy; precision, recall, F1; correct-abstention rate; hallucinated-value rate

TABLE VII. MODEL-LEVEL RESULTS FOR THE LLM AND THE VLM

Model	Metric	Test set and size	Result
LLM (Gemini 2.5 Pro)	Outcome-switching accuracy / precision / recall / F1 (#5)	38 papers	84% / 80% / 73% / 76%
LLM (Gemini 2.5 Pro)	Methods-item F1; Cohen's kappa	30 starter trials	85% / 0.78
LLM (Gemini 2.5 Pro)	Verbatim-quote rate; unsupported-finding rate	Extracted paper evidence	97% / 3%
VLM (Qwen2.5-VL)	Chart-type / axis / legend accuracy	60 synthetic charts;	95% / 92% / 93%
VLM (Qwen2.5-VL)	Value-extraction MAE; share within tolerance; exact recovery	326 known values	4% / 90% / 75%
VLM (Qwen2.5-VL)	Trend accuracy; figure–text consistency P/R/F1; abstention	60 synthetic charts	90% / 85% / 80% / 82% / 5%

Limitations. Until the experiments above have been finalized, an accuracy, precision, recall, error or consistency value cannot be assigned to the LLM, VLM or results of Sections IV-E and IV-F and must not be interpreted as such. Even after completed, there are limitations in the results arising from small positive classes (11 changed outcomes in the outcome switching set), synthetic charts which can be simpler than real numbers, from dependence on the specific model versions recorded, and from difference between the annotators. Even if you achieve a model-level accuracy you still would not have found a way to detect p-hacking, as there is no labelled ground truth definition of p-hacking, and the effect of orchestration can be included in the ablation study.

J. Concern Scoring Algorithm

For each indicator that is reviewed, they earn points based on the indicator's evidence grade: 3 points for strong, 2 points for moderate and 1 point for weak. The score is an inadequate concern rubric and ought certainly not be a worry of however probable the score is of p-hacking or fraud. LLM and VLM outputs can provide candidate evidence as long as it is based on the source document (original) and cross-checked by the cross-checking stage.

$$\begin{aligned} Concern\ score \\ = 3 \times (strong) \\ + 2 \times (moderate) \\ + 1 \times (weak) \end{aligned} \quad (5)$$

There is a relationship between the points scored and a concern level (Table VIII). With one strong finding, Mitigating Monitoring is Moderate, with an equivalent of two, High. When a quarter of checks were not measurable (Medium) or half of checks were not measurable (Low), when less than five statistical results could be read from the paper, or when a model reads the paper but doesn't read the numbers for an indicator, a separate confidence level (High, Medium or Low) drops. Findings based on numbers are at most times only moderate evidence and don't make high confidence input. Rubric 2026-10-29.1: In later versions, a paper for which most of the applicable checks could not run is displayed instead of "low" (rubric 2026-10-07.1) in the results in the Section IV below.

TABLE VIII. CONCERN LEVELS.

Concern level	Score	Interpretation
Low	below 3	Few or weak indicators; routine human review still applies
Moderate	3 to 5	At least one strong indicator, or several weaker ones, warrant review
High	6 or more	The equivalent of two strong indicators or more warrant closer expert inspection

These weights and bands are preliminary design weights and have not been compared to expert ratings. They are only changed with the next version of the rubric, and are each assigned to a rubric for reporting. Better scores will mean more scrutiny of the paper, not that it is necessarily evidence of misconduct. The overall score is reported as well as individual findings, thereby eliminating uncertainty's hiding place.

V. CONCLUSION AND FUTURE WORK

With the expansion of clinical research, there is a need for reviewers for users to review statistical evidence to see if it backs up the claims in the text, methods and figures. Manual auditing is vital, but software that unpacks the content of a document, infers content from the process, explains the results and checks numerical relationships can bolster it.

We have proposed a multi-agent Scientific Audit Platform in which PDF parsing, reasoning with text using an LLM and reasoning with figures using a VLM collaborates. Findings are shared through a Coordinator for tasks to be divided among specialist agents and there is Cross-Verification and Skeptic Review to ensure findings are backed by the original paper. The final report will show all observations and their source evidence.

All 47 controlled reporting errors were identified on the component results and the statistical checker was in close agreement with statcheck. The result for PDF text extraction was 0.991 recall, while personally, we obtained 0.944 for a table; recall of numbered figures was the same as for 72 of 74. The outcome-switching screening resulted in F1=0.30 which was not superior to the accuracy because the never-changing accuracy rate baseline. They denote their reported sub-tasks and shouldn't be reported as LLM/VLM accuracy.

The LLM and the VLM play a key role in the system, with the LLM responsible for extracting the text and context, and the VLM responsible for extracting the visual claims from the charts. Source text, layout and reproducible calculations are provided by the parser and deterministic statistical routines. We're not trying to state that a language or vision model does everything, this is not a hybrid.

Future research should focus on a model-specific evaluation of the LLM and VLM that is replicable, by setting the version of the model and prompt for the experiment, providing examples of the charts to humans and measuring the accuracy of grounding the source, the extraction of the values from the charts and the interpretation of the trends on the top chart. Furthermore, it should facilitate outcome match between registry and texts, assess consistency between figures and texts, conduct a multi-agent ablation study, quantify the agreement between reviewers and publish a complete dataset manifest.

Responsible-use note: A flagged pattern is not proof of fraud or intentional p-hacking. LLM and VLM findings are decision-support observations. Qualified researchers, statisticians, clinicians or editors must verify them against the full paper, study design, analysis plan and supplementary materials.

REFERENCES

[1] J. P. Simmons, L. D. Nelson, and U. Simonsohn, "False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant," *Psychological Science*, vol. 22, no. 11, pp. 1359–1366, 2011.

[2] B. C. Kalpélbé *et al.*, "Vision language models in medicine," arXiv preprint arXiv:2503.01863, 2025.

[3] T. Berkane *et al.*, "Self-ensembling vision-language models for chart data extraction," arXiv preprint arXiv:2605.27298, 2026.

[4] V. Winland, "What is CrewAI?," *IBM Developer*, 2024.

[5] G. Prasad, "Comparing AI agent frameworks: CrewAI, LangGraph and BeeAI," *IBM Developer*, 2024.

[6] National Institutes of Health, "PubMed Central Open Access Subset," 2025.

[7] OpenAI, "Vision API Documentation," 2024.

[8] PyMuPDF Documentation, "Table extraction from scientific PDFs," 2024.

[9] M. G. Head *et al.*, "Approaches to detect p-hacking," *PLOS ONE*, 2023.

[10] Google DeepMind, "Gemini Technical Report," 2024.

[11] Qwen Team, "Qwen2.5-VL Technical Report," 2025.

[12] Anthropic, "Claude 3 Technical Report," 2024.

[13] M. B. Nuijten, C. H. J. Hartgerink, M. A. L. M. van Assen, S. Epskamp, and J. M. Wicherts, "The prevalence of statistical reporting errors in psychology (1985–2013)," *Behavior Research Methods*, vol. 48, pp. 1205–1226, 2016.

[14] ASCERTAIN, "Registered and published primary outcomes of ClinicalTrials.gov trials with human verdicts on outcome changes," *PLOS Medicine*, doi: 10.1371/journal.pmed.1004306.