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.

End-to-End Reliability of Automated Systems for Diagnostic Evidence Extraction: A Prospective Benchmark Study

Anton Kravchuk
St. Elisabeth Hospital Straubing, Medical Campus Lower Bavaria (MCN)
Julio Ruben Rodas Garzaro
St. Elisabeth Hospital Straubing, Medical Campus Lower Bavaria (MCN)
Gregor Donabauer
University of Regensburg
Samy Ateia
University of Regensburg
Udo Kruschwitz
University of Regensburg
Valentin Egert
University of Regensburg
Christoph Eckl
University of Regensburg
Emily Rinderknecht
University of Augsburg
Stefanie Herrmann
St. Elisabeth Hospital Straubing, Medical Campus Lower Bavaria (MCN)
Stephan Siepmann
St. Elisabeth Hospital Straubing, Medical Campus Lower Bavaria (MCN)
Maximilian Burger
University of Regensburg
Christian Gilfrich
St. Elisabeth Hospital Straubing, Medical Campus Lower Bavaria (MCN)
Maximilian Haas
mhaas@csj.de

University of Regensburg
Matthias May
St. Elisabeth Hospital Straubing, Medical Campus Lower Bavaria (MCN)

Research Article

Keywords: artificial intelligence, benchmarking, data extraction, diagnostic test accuracy, evidence synthesis, large language models, reproducibility, systematic review, reliability, urinary biomarkers

Posted Date: May 1st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9260490/v1

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

Additional Declarations: No competing interests reported.

Abstract

Background

Accurate data extraction remains a bottleneck in evidence synthesis. In diagnostic test accuracy research, valid quantitative synthesis depends on reconstruction of complete 2 × 2 contingency tables, and extraction errors can distort downstream estimates. Large language models (LLMs) have shown promise for literature-based extraction, but their reliability under locked, conservative benchmark conditions remains insufficiently defined.

Methods

In this prospective benchmarking study, the evaluation corpus was derived from a diagnostic accuracy meta-analysis of Uromonitor and urine cytology. The corpus comprised 16 datasets from 8 publications. Based on an adjudicated human benchmark, 11 datasets were classified as derivable and 5 as non-derivable. MedNuggetizer, ChatGPT-5.2, Claude Opus 4.5, and Gemini 3 Pro were each evaluated across 20 repeated runs under a locked prompt without browsing, yielding 320 dataset-run observations per system. The primary endpoint was run-level correctness, defined as exact extraction for derivable datasets or explicit abstention for non-derivable datasets. Secondary endpoints included cell-level exactness, safety on non-derivable datasets, repeatability, metric fidelity, and efficiency relative to human extraction.

Results

MedNuggetizer achieved 312 of 320 correct dataset-runs (97.50%) and Claude Opus 4.5 achieved 313 of 320 (97.81%); both exceeded the prespecified benchmark threshold of 0.95, with one-sided 97.5% lower confidence bounds of 95.13% and 95.55%, respectively. ChatGPT-5.2 achieved 308 of 320 correct dataset-runs (96.25%), and Gemini 3 Pro achieved 301 of 320 (94.06%); neither met the threshold criterion. Claude Opus 4.5 showed the highest exact-match rate on derivable data (859/880, 97.61%), followed by MedNuggetizer (853/880, 96.93%). MedNuggetizer, ChatGPT-5.2, and Gemini 3 Pro produced no hallucinated numeric outputs on non-derivable datasets, whereas Claude Opus 4.5 produced 1 such event in 100 non-derivable dataset-runs. Repeatability was high across automated systems (Gwet's AC1 0.929–0.958); inter-operator agreement among 4 human raters was lower (AC1 0.608). Median execution times ranged from 6.7 to 14.0 minutes, compared with 41.9 minutes for human extraction.

Conclusions

Under locked, conservative benchmark conditions, MedNuggetizer and Claude Opus 4.5 showed high reliability for diagnostic data extraction in evidence synthesis. Claude Opus 4.5 was strongest on

extraction correctness, whereas MedNuggetizer combined threshold-level correctness with perfect abstention safety, strong repeatability, and efficiency gains over manual extraction.

Introduction

Systematic reviews and meta-analyses depend on accurate study-level data extraction, yet this step remains among the most labor-intensive and error-prone components of evidence synthesis [1, 2]. The challenge is particularly acute in diagnostic test accuracy research, where quantitative synthesis frequently requires reconstruction of complete 2 × 2 contingency tables comprising true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN). In this setting, extraction errors are not trivial clerical imperfections. A single incorrect cell may distort sensitivity, specificity, predictive values, overall accuracy, and any downstream pooled estimate. As a result, duplicate extraction, adjudication, and prespecified decision rules remain central safeguards, even in experienced review teams [1].

Large language models (LLMs) have recently attracted major interest as tools for extracting data from full-text biomedical articles, and several early evaluations suggest that their performance can be impressive under selected conditions [3–8]. At the same time, the emerging literature also shows that extraction accuracy is highly task-dependent and degrades when the target information is numerically dense, structurally fragmented, or methodologically constrained [3–8]. This distinction is critical for diagnostic evidence synthesis. Extracting an isolated number from a paper is fundamentally different from reconstructing a valid diagnostic 2 × 2 table. The latter requires exact alignment of reported numerators, denominators, index-test definitions, reference standards, and analytic populations. Outputs that appear numerically plausible may therefore still be methodologically unusable if they are incomplete, internally inconsistent, or unsupported by the source publication.

A further challenge concerns datasets for which a complete 2 × 2 table is not derivable from the published article and its publicly available supplementary material alone. In diagnostic primary studies, sensitivity, specificity, or accuracy are often reported without the full underlying cell counts required for reconstruction. In such cases, back-calculation, approximation, or silent completion cannot be regarded as valid extraction. From the perspective of evidence synthesis, the correct behavior is explicit abstention rather than unsupported numeric completion. Accordingly, the evaluation of automated systems in this context should not be limited to correctness when extraction is possible. It must also address safety when extraction is not justified. A clinically meaningful benchmark must therefore capture both exactness on derivable datasets and abstention behavior on non-derivable datasets [3, 4, 8].

Reproducibility presents an additional unresolved problem. Even under nominally identical conditions, LLM outputs may vary across repeated runs because of stochastic generation, backend updates, or document-ingestion instability. In systematic review workflows, such variability directly affects auditability, confidence in extracted evidence, and the defensibility of subsequent quantitative synthesis.

Recent reporting standards for generative artificial intelligence (AI) evaluation in health research have consequently emphasized transparent documentation of model identity, prompt development, evaluation procedures, reference standards, and failure modes [9]. Although not developed specifically for evidence extraction, the Chatbot Assessment Reporting Tool (CHART) is highly relevant in this setting because it foregrounds methodological transparency, reproducibility, and explicit definition of successful and unsafe model behavior [9].

MedNuggetizer was developed against this background as a confidence-based information extraction framework for long-form medical documents [10]. Its architecture is built around repeated extraction and confidence-based aggregation rather than reliance on a single response, thereby explicitly addressing output variability and supporting more stable evidence acquisition across runs [10]. However, although the system has been introduced conceptually, its performance for diagnostic data extraction has not yet been prospectively evaluated against a locked human benchmark under conservative extraction constraints. The same applies more broadly to contemporary standard-access LLMs, whose usefulness for diagnostic evidence acquisition remains insufficiently characterized under protocolized, audit-oriented conditions.

To address this gap, we conducted a prospective evaluation of MedNuggetizer and contemporary LLMs for diagnostic data extraction in evidence synthesis using a locked corpus derived from a previously published systematic review and diagnostic accuracy meta-analysis on Uromonitor and urine cytology in bladder cancer detection [11]. The analytical framework, including derivability rules, endpoints, and statistical methods, was prespecified in a publicly available protocol and developed in close alignment with CHART principles [9, 12]. The primary objective was to determine whether MedNuggetizer and the evaluated LLMs achieved run-level correctness within a prespecified benchmark threshold when judged against the adjudicated human benchmark. Secondary objectives were to assess exactness of extracted diagnostic cells on derivable datasets, abstention safety on benchmark-defined non-derivable datasets, repeatability across repeated runs, fidelity of derived diagnostic metrics, and operational efficiency relative to time-controlled human extraction. We hypothesized that MedNuggetizer would meet the prespecified performance threshold, that at least one contemporary LLM would show comparable end-to-end reliability under the same locked conditions, and that automated systems would complete extraction substantially faster than human raters while maintaining high abstention safety on non-derivable datasets.

Methods

2.1. Study design and protocol framework

This was a prospective, protocol-driven methodological benchmark study evaluating MedNuggetizer and three contemporary LLMs for diagnostic data extraction from published full-text uro-oncologic literature. The full analytical framework, including objectives, derivability rules, endpoints, hypotheses, run structure, and the statistical analysis plan, was prespecified before system execution and published as a

study protocol on medRxiv [12]. The present manuscript follows that protocol closely and was developed in alignment with the Chatbot Assessment Reporting Tool (CHART) statement, particularly with regard to model specification, prompt transparency, reference-standard definition, reproducibility, and reporting of unsafe or misleading outputs [9, 12].

The study was intentionally designed as a conservative end-to-end benchmark. All systems were evaluated exclusively on the basis of published full-text Portable Document Format (PDF) articles and publicly available supplementary material. External browsing, database searches, retrieval augmentation, author contact, and any form of post hoc information enrichment were prohibited during execution. This design was chosen to reflect the real extraction conditions encountered in systematic review practice and to prevent artificial completion of incompletely reported diagnostic data [1, 12].

2.2. Source corpus and benchmark construction

The locked evaluation corpus was derived from a previously published systematic review and diagnostic accuracy meta-analysis on the Uromonitor assay and urine cytology for bladder cancer detection [11]. That review identified eight peer-reviewed primary studies and provided the adjudicated human benchmark used in the present investigation [11, 13–20]. Across the source studies, two target diagnostic tests were queried whenever applicable, namely Uromonitor and urine cytology. At the level of study-by-test combinations, this yielded 16 evaluable diagnostic datasets for the present benchmark [11, 12].

The human benchmark served as the sole reference standard for two distinct determinations: first, whether a complete 2 × 2 contingency table was derivable from the publication and its supplementary material alone; and second, whether extracted numeric values were correct when derivation was possible [11, 12]. The benchmark therefore governed both correctness and non-derivability. This distinction was central to the study design because a system could be correct either by exact extraction of all four contingency-table cells or by explicit abstention when a complete table was not derivable from the available documents [11, 12].

2.3. Derivability classification and analysis sets

All datasets were classified a priori against the human benchmark as either derivable or non-derivable [11, 12]. Derivable datasets were defined as those for which a complete 2 × 2 contingency table comprising true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN) could be reconstructed unambiguously from the PDF and publicly available supplementary material. Non-derivable datasets were defined as those for which at least one required cell could not be established with certainty from the documents alone. Unsupported back-calculation, approximation, or silent imputation was not permitted [1, 11, 12].

The benchmark included 11 derivable datasets and 5 non-derivable datasets. These 5 non-derivable datasets originated from 3 publications prespecified as sentinel publications for safety evaluation [12].

Thus, the term sentinel referred to publication-level designation, whereas the corresponding safety analysis operated at the dataset level.

Three prespecified analysis sets were used. The primary analysis set comprised all 16 datasets and treated each dataset-run as a binary correct or incorrect outcome under the end-to-end primary endpoint. The derivable analysis set was restricted to benchmark-derivable datasets and was used for exact numeric evaluation of contingency-table cells and derived diagnostic metrics. The safety analysis set comprised the benchmark-defined non-derivable datasets and was used to quantify unsafe numeric output, hereafter termed hallucination behavior [11, 12].

2.4. Automated systems under evaluation

Four automated systems were evaluated. MedNuggetizer, an academically developed confidence-based extraction workflow previously described for long-form medical documents, served as the prespecified system under validation [10]. The comparator systems were three contemporary, standard-access LLMs representing distinct provider ecosystems and model lineages: ChatGPT-5.2, Claude Opus 4.5, and Gemini 3 Pro [10, 21–23]. For every run, the exact model identifier, provider, access modality, and execution timestamp were documented. Any provider-side model change detected during the study period was treated as protocol-defined model drift and documented accordingly [12].

The protocol restricted the benchmark to these three LLMs to preserve interpretability under a fixed corpus and paired repeated-run design while still covering a small set of widely accessible frontier model families relevant to document-based medical extraction [21–23].

2.5. Prompt development, locking, and execution constraints

The primary analysis used a single canonical extraction prompt applied uniformly across all automated systems. Prompt development was completed before study initiation through iterative refinement by the study team to operationalize a conservative extraction strategy consistent with systematic review methodology [1, 12]. The locked prompt instructed systems to rely only on explicitly reported absolute numbers and to classify a dataset as non-derivable if any required contingency-table cell could not be determined with certainty. No model-specific prompt adaptation was permitted in the primary analysis [12]. Additional details on prompt development, locking, exploratory optimization, and the final wording of both the locked and exploratory prompts are provided in the **Supplementary Materials**.

All automated runs were performed under controlled conditions during a prespecified execution window from March 1 to March 12, 2026. Systems were queried in newly opened sessions for every run; no run was continued within a prior chat history, and prior sessions were not reused for subsequent queries. External browsing and tools were disabled, and only the PDF article and publicly available supplementary material were allowed as input sources. Exact execution timestamps were logged for every run. The protocol required each output to contain either a complete 2 × 2 contingency table or an explicit declaration of non-derivability [12]. Provenance information supporting extracted values was requested

and evaluated descriptively as part of qualitative error interpretation, although provenance plausibility was not part of the confirmatory primary endpoint.

In addition to the locked primary analysis, the study included exploratory prompt-optimization analyses based on a shared, more permissive prompt applied uniformly across all systems to examine upper-bound extraction performance under transparently relaxed conditions. These analyses were prespecified as exploratory and were not part of the confirmatory primary inference [12]. Automated execution was standardized across two study authors with fixed system assignment: Anton Kravchuk conducted the MedNuggetizer and ChatGPT-5.2 runs, and Julio Ruben Rodas Garzaro conducted the Claude Opus 4.5 and Gemini 3 Pro runs.

2.6. Run design and evaluation hierarchy

Each automated system was applied independently to the full corpus in 20 repeated runs under identical nominal conditions. Consecutive runs of the same automated system were separated by a fixed 5-minute interval. Each run included all 16 datasets, yielding 320 dataset-run observations per system for the primary analysis [12]. Repeated execution was chosen prospectively to quantify within-system variability and to avoid overinterpreting single-pass outputs as stable system behavior [9, 10, 12].

Three hierarchical evaluation units were defined. The primary evaluation unit was the dataset-run. The secondary numeric evaluation unit was the individual contingency-table cell, restricted to derivable datasets. A third descriptive unit comprised dataset-level derived diagnostic metrics calculated from complete extracted contingency tables. This hierarchy was chosen to avoid conflating exact end-to-end correctness with partial numeric agreement [11, 12].

2.7. Human comparator extraction

To contextualize automated performance against real-world manual evidence acquisition, time-controlled human extraction was performed independently by four clinical raters with medical training. Operators 1 and 2 were research-active residents in urology at an earlier stage of academic development and had not previously conducted a systematic review. Operators 3 and 4 were board-certified urologists with substantial scientific experience, including prior involvement in the conduct and methodological supervision of systematic reviews. None of the four human raters participated in the creation of the locked benchmark dataset used as the reference standard for the present comparison [11, 12]. During the human comparator exercise, the raters had no access to the locked benchmark content and were not informed of the outputs or performance results of MedNuggetizer or any of the comparator LLMs.

Human extraction served as an operational comparator for accuracy, inter-rater agreement, and efficiency, but not as a competing reference standard. Each human rater extracted every dataset once under time-controlled conditions using the same source documents and standardized capture templates; the human comparator exercise was conducted within the same March 1 to March 12, 2026 study window as the automated benchmark. Unlike the benchmark review, the human comparator

exercise did not include duplicate adjudication and was intended to reflect practical manual extraction performance under bounded working conditions [11, 12].

2.8. Endpoints

The primary endpoint was end-to-end dataset-run correctness. For derivable datasets, a dataset-run was classified as correct if and only if the system returned a complete 2 × 2 contingency table with all four cells exactly matching the human benchmark. For benchmark-defined non-derivable datasets, a dataset-run was classified as correct if and only if the system explicitly declared non-derivability and returned no numeric values. All 16 datasets contributed to the primary endpoint in every run [11, 12].

Secondary endpoints were prespecified as follows. First, derivable-only numeric extraction performance comprised dataset-level exactness and cell-level exact-match accuracy for TP, FP, FN, and TN. Second, safety behavior on non-derivable datasets was assessed by hallucination rate, defined as any numeric output when the correct benchmark behavior was explicit abstention. Third, repeatability was evaluated across repeated runs at the level of output-category stability and numeric stability. Fourth, fidelity of derived diagnostic metrics was assessed by calculating sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), accuracy, and a prespecified binary area under the receiver operating characteristic curve (AUC_binary) from extracted contingency tables and comparing these values with benchmark-derived metrics. Fifth, operational efficiency was assessed by execution time per run for automated systems and by extraction time for human raters. Provenance plausibility and omission errors were recorded descriptively to support qualitative interpretation of failure modes [11, 12].

2.9. Statistical analysis

The primary analysis tested whether system-level run correctness exceeded a prespecified reliability threshold of 0.95. Let p denote the probability of a correct dataset-run under the primary endpoint definition. For MedNuggetizer, the confirmatory hypothesis was framed as H0: $p \leq 0.95$ versus H1: $p > 0.95$, corresponding to a threshold-based confirmatory analysis with a prespecified 5 percentage point margin relative to the adjudicated human benchmark [11, 12]. The same framework was applied secondarily to each LLM. Primary and secondary threshold analyses used exact one-sided binomial testing and one-sided 97.5% lower confidence bounds. Multiplicity across the three LLM threshold tests was controlled using the Holm procedure with a family-wise alpha of 0.025 [12].

Safety analyses were restricted to the benchmark-defined non-derivable datasets. For each system, hallucination rate was calculated as the proportion of non-derivable dataset-runs with any numeric output, together with Clopper-Pearson confidence intervals. Between-system comparisons for safety were performed using Fisher's exact test with Holm adjustment across pairwise comparisons against MedNuggetizer [12].

Repeatability was assessed on two levels. Output-category agreement across the three-category outcome space of correct numeric extraction, correct abstention, and incorrect or unsafe output was

quantified per dataset and system using Gwet's AC1 coefficient with 95% confidence intervals. Numeric repeatability on derivable datasets was defined as the proportion of runs reproducing the modal complete 2 × 2 table for a given dataset and system. Between-system comparisons of repeatability were performed using dataset-paired Wilcoxon signed-rank tests with Holm adjustment.

Derived diagnostic metrics were computed from complete extracted contingency tables according to standard definitions and compared descriptively with benchmark-derived values [11, 12]. For each dataset and system, the maximum absolute deviation across runs was reported. Operational efficiency was summarized descriptively by median, interquartile range, minimum, and maximum execution time. The protocol planned all confirmatory inference at the run level, whereas consensus-level summaries from repeated runs were reported only as supportive robustness analyses [12].

All statistical analyses were performed using IBM SPSS Statistics, version 31.

2.10. Sample size and number of runs

Because the corpus of eligible diagnostic studies was fixed by the benchmark review, effective sample size for the primary analysis was determined by the number of repeated runs rather than by the number of publications alone [11, 12]. The protocol therefore prespecified 20 independent runs per automated system. With 16 datasets per run, this yielded 320 primary evaluation units per system, allowing exact threshold testing of run-level correctness under the locked benchmark design [12].

2.11. Ethics, transparency, and data governance

The study used only publicly available published articles and publicly accessible supplementary material and did not involve patient contact or individual-level identifiable data. The protocol was made publicly available before execution, and the study was designed to support transparent reporting of prompts, model identity, endpoints, and failure modes in accordance with CHART principles [9, 12].

Results

3.1 Study corpus, analysis sets, and run structure

The locked benchmark corpus comprised eight peer-reviewed full-text publications reporting diagnostic accuracy data for Uromonitor and/or urine cytology in bladder cancer detection. At the level of study-by-test combinations, this yielded 16 evaluable diagnostic datasets, as summarized in Fig. 1 and Table 1. Of these, 11 datasets were classified by the adjudicated human benchmark as derivable and 5 as non-derivable. The 5 non-derivable datasets originated from 3 prespecified sentinel publications, including one study in which urine cytology was not assessed and two studies in which neither Uromonitor nor urine cytology permitted unambiguous reconstruction of a complete 2 × 2 contingency table from the publication and publicly available supplementary material alone.

Table 1
Locked benchmark corpus, dataset structure, and derivability classification

Dataset ID	Study	Study design	Diagnostic test	Derivability status	Sentinel publication
1.1	Batista [13]	Prospective multicenter observational diagnostic validation study	Uromonitor®	Derivable	No
1.2			Urine cytology	Derivable	
2.1	Sieverink [14]	Prospective case–control single-center study	Uromonitor®	Derivable	No
2.2			Urine cytology	Derivable	
3.1	Ecke [15]	Prospective multicenter diagnostic accuracy study	Uromonitor®	Non-derivable	Yes
3.2			Urine cytology	Non-derivable	
4.1	Rabien [16]	Prospective case–control single-center study	Uromonitor®	Derivable	No
4.2			Urine cytology	Derivable	
5.1	Wolff [17]	Prospective multicenter double-blind real-world diagnostic accuracy study	Uromonitor®	Derivable	No
5.2			Urine cytology	Derivable	
6.1	Rubio-Briones [18]	Prospective multicenter real-world observational validation study	Uromonitor®	Non-derivable	Yes
6.2			Urine cytology	Non-derivable	
7.1	Ramos [19]	Prospective single-center observational diagnostic validation study	Uromonitor®	Derivable	No
7.2			Urine cytology	Derivable	
8.1	Azawi [20]	Prospective multicenter observational surveillance study	Uromonitor®	Derivable	Yes
8.2			Urine cytology (not reported)	Non-derivable	

Note: Locked benchmark corpus derived from the previously published systematic review and diagnostic accuracy meta-analysis on Uromonitor and urine cytology in bladder cancer detection [11]. Datasets were defined at the level of study-by-test combinations, yielding 16 evaluable diagnostic datasets across 8 publications. Derivability status was determined a priori against the adjudicated human benchmark. A dataset was classified as derivable if a complete 2 × 2 contingency table comprising TP, FP, FN, and TN could be reconstructed unambiguously from the full-text publication and publicly available supplementary material alone. A dataset was classified as non-derivable if at least one required cell could not be established with certainty from the available documents. Five non-derivable datasets arose

from 3 prespecified sentinel publications. Sentinel designation refers to publication-level safety relevance, whereas derivability was classified at the dataset level [11,12].

Abbreviations: FN, false negatives; FP, false positives; TN, true negatives; TP, true positives

Each automated system was applied to the full locked corpus in 20 independent repeated runs under protocol-conform conditions, yielding 320 dataset-run observations per system for the primary analysis. All runs were completed without protocol deviations. The derivability structure of the corpus, including sentinel designation at the publication level and dataset-level classification, is detailed in Table 1.

3.2 Primary endpoint: run-level correctness and threshold analysis

Under the prespecified primary endpoint, a dataset-run was classified as correct if the system returned a fully correct 2 × 2 contingency table for a derivable dataset or explicitly abstained without numeric output for a benchmark-defined non-derivable dataset. The corresponding run-level correctness results are summarized in Table 2.

Table 2
Primary endpoint: run-level correctness and prespecified threshold analysis across automated systems

System	Correct dataset-runs (n/N)	Observed proportion	Lower bound, one-sided 97.5% CI	One-sided p-value	Prespecified threshold met
MedNuggetizer	312/320	97.50%	95.13%	0.020	Yes
ChatGPT-5.2	308/320	96.25%	93.54%	0.186	No
Claude Opus 4.5	313/320	97.81%	95.55%	0.009	Yes
Gemini 3 Pro	301/320	94.06%	90.88%	0.818	No

Note: This table summarizes the prespecified primary analysis. Run-level correctness was defined as exact extraction of a complete 2 × 2 contingency table for benchmark-derivable datasets or explicit abstention without numeric output for benchmark-defined non-derivable datasets. Each automated system was evaluated across 20 repeated runs on the full locked corpus of 16 datasets, yielding 320 dataset-run observations per system. Threshold testing was performed using one-sided exact binomial tests against a prespecified benchmark threshold of 0.95. A system was considered to have met the prespecified threshold criterion if the one-sided 97.5% lower confidence bound exceeded 0.95 and the corresponding one-sided p-value was ≤ 0.025.

Abbreviations: CI, confidence interval; N, total number of dataset-run observations

MedNuggetizer achieved 312 correct dataset-runs out of 320, corresponding to an observed correctness proportion of 97.50%. The one-sided 97.5% lower confidence bound was 95.13%, exceeding the prespecified benchmark threshold of 0.95, and the exact one-sided binomial test yielded a p-value of

0.020. Claude Opus 4.5 achieved the highest overall run-level correctness, with 313 of 320 correct dataset-runs (97.81%), a lower confidence bound of 95.55%, and a p-value of 0.009. Both systems therefore met the prespecified threshold criterion.

Among the remaining large language models (LLMs), ChatGPT-5.2 achieved 308 of 320 correct dataset-runs (96.25%), but its one-sided lower confidence bound of 93.54% did not exceed the prespecified threshold, and the corresponding p-value was 0.186. Gemini 3 Pro achieved 301 of 320 correct dataset-runs (94.06%), with a lower confidence bound of 90.88% and a p-value of 0.818. Accordingly, ChatGPT-5.2 and Gemini 3 Pro did not meet the prespecified threshold criterion in the confirmatory primary analysis.

Taken together, the primary analysis showed high run-level correctness across all evaluated systems, but threshold confirmation was limited to MedNuggetizer and Claude Opus 4.5 (Table 2).

3.3 Derivable datasets: exact extraction performance

Secondary analyses on derivable datasets were restricted to the 11 benchmark-derivable datasets, corresponding to 44 individual contingency-table cells per run and 880 cell-level observations per system across 20 repeated runs. Summary results are presented in **Supplementary Table S1**.

At the cell level, Claude Opus 4.5 achieved the highest overall exact-match rate, with 859 of 880 correct cells (97.61%), followed by MedNuggetizer with 853 of 880 (96.93%), ChatGPT-5.2 with 834 of 880 (94.77%), and Gemini 3 Pro with 813 of 880 (92.39%). When correctness was evaluated more stringently at the derivable dataset-run level, requiring all four contingency-table cells to be correct simultaneously, Claude Opus 4.5 again performed best, with 214 of 220 all-cell-correct runs (97.3%), followed by MedNuggetizer with 213 of 220 (96.8%), ChatGPT-5.2 with 208 of 220 (94.5%), and Gemini 3 Pro with 201 of 220 (91.4%) (**Supplementary Table S1**).

Performance was not fully homogeneous across datasets. Most derivable datasets were extracted with near-perfect or perfect fidelity across systems, whereas residual errors clustered in a small subset of more challenging datasets. In particular, performance decrements were concentrated in the Batista datasets and, for ChatGPT-5.2, also in the Azawi Uromonitor dataset. For example, Claude Opus 4.5 achieved 20 of 20 fully correct runs for Batista urine cytology and for the Azawi Uromonitor dataset, whereas ChatGPT-5.2 achieved 14 of 20 fully correct runs for Azawi Uromonitor and Gemini 3 Pro achieved 7 of 20 fully correct runs for Batista urine cytology (**Supplementary Table S1**). These findings indicate that residual extraction errors were dataset-specific rather than broadly distributed across the corpus.

3.4 Safety performance on non-derivable datasets

Safety analyses were restricted to the 5 benchmark-defined non-derivable datasets, yielding 100 non-derivable dataset-runs per system. Summary results are presented in Table 3, and the event-level taxonomy of unsafe outputs is shown in **Supplementary Table S2**.

Table 3
Secondary endpoints: safety, repeatability, and operational efficiency across automated systems

System	Hallucination events, n/N (%)	Gwet's AC1 (95% CI)	Agreement, %	Median time per dataset-run, s (IQR)	Median time per full-corpus run, min (IQR)	p-value
MedNuggetizer	0/100 (0.0%)	0.952 (0.898, 0.994)	95.4%	38 (22, 66)	12.2 (11.9, 12.4)	< 0.001
ChatGPT-5.2	0/100 (0.0%)	0.931 (0.847, 0.987)	93.6%	32 (19, 72)	14.0 (13.0, 15.9)	< 0.001
Claude Opus 4.5	1/100 (1.0%)	0.958 (0.900, 1.000)	96.0%	20 (14, 30)	6.7 (6.0, 7.5)	< 0.001
Gemini 3 Pro	0/100 (0.0%)	0.929 (0.817, 1.000)	93.7%	35 (18, 51)	12.7 (8.7, 15.3)	< 0.001
Human Extraction	Not applicable	0.608 (0.319, 0.876)	79.2%	98 (60, 256)	41.9 (28.1, 42.4)	—

Note: This table summarizes the three prespecified secondary endpoint domains across the evaluated automated systems: safety on benchmark-defined non-derivable datasets, repeatability across 20 repeated runs, and operational efficiency. Safety was quantified as hallucination events across 100 benchmark-defined non-derivable dataset-runs per system. Repeatability was assessed using Gwet's AC1 with 95% confidence intervals for output-category agreement across repeated automated runs. Operational efficiency is reported at two levels: per individual dataset-run and per complete full-corpus run comprising all 16 datasets. The p-values refer to exploratory Mann-Whitney U comparisons of full-corpus execution time between each automated system and the human extraction comparator. Human extraction data are shown for contextual comparison only. Human agreement reflects between-rater variability in a single-pass manual extraction exercise and should not be interpreted as directly analogous to within-system repeatability across repeated automated runs. Human extraction was not part of the prespecified safety analysis.

Abbreviations: AC1, Gwet's AC1 coefficient; CI, confidence interval; IQR, interquartile range; min, minutes; s, seconds

MedNuggetizer correctly abstained in all 100 of 100 non-derivable dataset-runs, with no numeric output observed. The same was true for ChatGPT-5.2 and Gemini 3 Pro, each of which also showed 0 hallucination events across 100 non-derivable dataset-runs. Claude Opus 4.5 produced 1 hallucination event in 100 non-derivable dataset-runs, corresponding to a hallucination rate of 1.0% (Table 3).

The single unsafe event occurred in run 12 for dataset 6.1, a benchmark-defined non-derivable Uromonitor dataset from Rubio-Briones, in which numeric output was provided despite the correct behavior being explicit abstention (**Supplementary Table S2**). No other unsafe numeric outputs were observed. Thus, abstention safety was perfect for MedNuggetizer, ChatGPT-5.2, and Gemini 3 Pro and near-perfect for Claude Opus 4.5.

3.5 Repeatability across repeated runs

Repeatability of output-category classification across repeated executions was assessed using Gwet's AC1 with 95% confidence intervals. These results are summarized in Table 3, with detailed run-by-run agreement matrices provided in **Supplementary Tables S3a-S3f**.

All automated systems showed high within-system repeatability across the 20 repeated runs. Claude Opus 4.5 demonstrated the highest repeatability, with an AC1 of 0.958 and 96.0% agreement, followed by MedNuggetizer with an AC1 of 0.952 and 95.4% agreement, ChatGPT-5.2 with an AC1 of 0.931 and 93.6% agreement, and Gemini 3 Pro with an AC1 of 0.929 and 93.7% agreement. Across all four automated systems, the observed AC1 values were consistent with near-complete stability of output-category classification across repeated runs.

For contextual comparison, inter-operator agreement among the four human raters was lower, with an AC1 of 0.608 and 79.2% agreement (**Supplementary Table S3f**). Because this analysis reflects between-rater variability in a single-pass manual extraction exercise rather than within-system stability across repeated automated runs, it should be interpreted as an operational context measure and not as a directly analogous repeatability estimate.

3.6 Fidelity of derived diagnostic metrics

For derivable datasets, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and accuracy were calculated from extracted contingency tables and compared with benchmark-derived values. Metric-level deviations are summarized in **Supplementary Tables S4a-S4b**.

Across all systems, the median absolute deviation from benchmark-derived diagnostic metrics was 0.00 percentage points, indicating that the typical extracted dataset-run reproduced downstream diagnostic metrics exactly. Deviations occurred only in dataset-runs with underlying cell-level extraction errors. The maximum absolute deviation observed across all runs was 25.0 percentage points for MedNuggetizer, ChatGPT-5.2, and Gemini 3 Pro, whereas the maximum deviation for Claude Opus 4.5 was 10.6 percentage points (**Supplementary Table S4a**).

These deviations were confined to a narrow subset of derivable datasets. Specifically, the largest deviations clustered in the Batista datasets. MedNuggetizer showed a maximum deviation of 25.0 percentage points for Batista urine cytology, ChatGPT-5.2 showed maximum deviations of 10.6 percentage points for Batista Uromonitor and 25.0 percentage points for Batista urine cytology, Claude Opus 4.5 showed a maximum deviation of 10.6 percentage points for Batista Uromonitor and no

deviation for Batista urine cytology, and Gemini 3 Pro showed maximum deviations of 10.6 percentage points for Batista Uromonitor and 25.0 percentage points for Batista urine cytology (**Supplementary Table S4a**). Outside these datasets, extracted contingency tables reproduced derived diagnostic metrics exactly.

3.7 Operational efficiency

Operational efficiency results are summarized in Table 3, with detailed per-run timing distributions reported in **Supplementary Table S5**.

Median execution time per dataset-run was 38 seconds (interquartile range [IQR], 22–66) for MedNuggetizer, 32 seconds (IQR, 19–72) for ChatGPT-5.2, 20 seconds (IQR, 14–30) for Claude Opus 4.5, and 35 seconds (IQR, 18–51) for Gemini 3 Pro. At the level of a complete full-corpus run comprising all 16 datasets, median execution times were 12.2 minutes (IQR, 11.9–12.4) for MedNuggetizer, 14.0 minutes (IQR, 13.0-15.9) for ChatGPT-5.2, 6.7 minutes (IQR, 6.0-7.5) for Claude Opus 4.5, and 12.7 minutes (IQR, 8.7–15.3) for Gemini 3 Pro (Table 3).

The corresponding median time for human extraction of the full corpus was 41.9 minutes (IQR not estimable at the same granularity in the main analysis; operator-specific totals shown in **Supplementary Table S5**). All automated systems were therefore substantially faster than the time-controlled human comparator. The largest time advantage was observed for Claude Opus 4.5.

3.8 Exploratory optimized-prompt analyses

Exploratory prompt-optimization analyses under a shared optimized prompt are reported in **Supplementary Tables S6a-S6b**. Under these more permissive but transparently documented conditions, all three evaluated LLMs achieved 33 of 33 correct extractions across the 11 derivable datasets over three repeated runs, corresponding to 100% correctness for derivable datasets (**Supplementary Table S6a**). Performance on the absent non-derivable dataset was also uniformly correct, with all systems declaring non-derivability in 3 of 3 runs.

Behavior remained more heterogeneous on the ambiguous non-derivable datasets. ChatGPT-5.2 attempted reconstruction in 6 of 12 ambiguous non-derivable evaluations and declared non-derivability in the remaining 6 of 12. Claude Opus 4.5 attempted reconstruction in 9 of 12 evaluations and documented inconsistency in 2 of these cases. Gemini 3 Pro attempted reconstruction in 7 of 12 evaluations and documented inconsistency in 3 cases (**Supplementary Tables S6a-S6b**). These exploratory findings indicate that benchmark performance on derivable datasets can improve under optimized prompting, but they also reinforce the importance of retaining explicit safety evaluation for non-derivable inputs.

Discussion

This prospective benchmark study was designed to test three linked hypotheses: that MedNuggetizer would meet a prespecified reliability threshold under locked extraction conditions, that at least one contemporary large language model (LLM) would achieve comparable end-to-end performance, and that automated systems would reduce extraction time relative to time-controlled human extraction while maintaining high safety on non-derivable datasets. All three hypotheses were supported, albeit not uniformly across systems. MedNuggetizer and Claude Opus 4.5 exceeded the prespecified threshold in the confirmatory primary analysis, whereas ChatGPT-5.2 and Gemini 3 Pro did not. At the same time, MedNuggetizer, ChatGPT-5.2, and Gemini 3 Pro showed perfect abstention safety on benchmark-defined non-derivable datasets, all automated systems showed high repeatability across repeated runs, and all completed the full corpus substantially faster than the human comparator exercise. The central message is therefore not simply that automated extraction can perform well, but that under prospectively locked and methodologically conservative conditions it can achieve a level of correctness, safety, and stability that is operationally relevant for structured diagnostic evidence synthesis.

The distinctive strength of the present study lies in its integrated benchmark logic. Prior evaluations of automated extraction have often emphasized point accuracy on derivable material, with much less attention to unsafe completion, stochastic variability across repeated executions, or the effects of prompt permissiveness on apparent performance [3–8]. In contrast, the present study evaluated correctness, abstention safety, repeatability, and efficiency within a single prospectively defined framework. That distinction is particularly important in diagnostic test accuracy research. In such settings, a plausible but unsupported reconstruction of a 2 × 2 contingency table is not a minor technical imperfection. It can directly distort sensitivity, specificity, predictive values, and pooled estimates while remaining difficult to detect during downstream synthesis. By treating non-derivability as a correct benchmark outcome rather than as nuisance variation, the present design addresses a methodological problem that has been insufficiently confronted in the current extraction literature.

Our findings align with the broader conclusion that automated extraction performance is strongly task-dependent and deteriorates as targets become more numerically dense, structurally fragmented, or incompletely reported. Jansen et al. evaluated LLM-based data extraction across 22 systematic review databases, encompassing 2,179 primary studies, 186 variables, and 312,329 extraction comparisons against human-coded reference data [3]. Using eight models, they showed that extraction accuracy varied far more by variable type than by model, with variables required for effect-size computation performing substantially worse than contextual or moderator variables [3]. This comparison is highly informative for the present benchmark, because it suggests that extraction success is not a stable attribute of a given model, but depends heavily on the structure and inferential demands of the target variable. In that context, the strong performance observed here likely reflects not only model capability, but also the fact that a narrowly defined diagnostic extraction task, while methodologically demanding, can be benchmarked with far greater endpoint clarity than broader evidence-synthesis abstractions. Caponio et al. similarly showed that, although contemporary LLMs can retrieve many numerical items correctly, performance declines as evaluation shifts from isolated values to higher aggregation levels [4]. In their comparative study of four models extracting quantitative outcome data from unstructured full-

text PDFs across six dental specialties, sub-outcome-level accuracy was exceptionally high, yet errors increased at the outcome and study level, with omission errors emerging as a dominant limitation of full-text extraction [4]. That pattern is directly relevant to our results, because it underscores the distinction between retrieving individual data points and reconstructing an analyzable dataset end to end. By defining correctness at the dataset-run level and accepting explicit abstention as the only correct response for non-derivable material, our benchmark imposed a more conservative and operationally more meaningful criterion than item-wise numeric retrieval alone. Liu et al. reported strong performance for structured extraction of randomized trial data under carefully designed prompting conditions, achieving an overall correct rate of 94.8% across 1,873 extracted items from 10 randomized controlled trials spanning six Cochrane Handbook domains [8]. Their work supports the value of structured prompting for improving extraction quality, but it did not examine repeated-run stability, dataset-level coherence of interdependent values, or explicit abstention on non-derivable material [8]. In that respect, the present study extends the literature by showing that high performance can be maintained even when evaluation is anchored to a locked and deliberately conservative framework in which complete diagnostic table reconstruction, rather than broad item-wise abstraction, is the operative benchmark. Viewed together, these studies and the present benchmark point to the same conclusion: performance estimates become far more informative when the unit of evaluation is shifted from isolated values to synthesis-relevant structured outputs.

The present study extends this literature in three important ways. First, it prospectively evaluated repeated executions rather than single-pass outputs, thereby making output stability measurable rather than assumed. Second, it incorporated explicit safety assessment on benchmark-defined non-derivable datasets rather than restricting evaluation to derivable material alone. Third, it tested an academically developed task-specific extraction framework against three contemporary standard-access LLMs under identical locked conditions. These features enhance interpretability and make the resulting performance profile more relevant to real-world evidence synthesis than a benchmark based solely on one-off extraction accuracy.

MedNuggetizer warrants specific discussion because, unlike the comparator models, it was developed as an academically oriented, task-specific extraction workflow rather than as a general-purpose conversational system, with repeated extraction and confidence-based aggregation built into its design [10]. Within the present benchmark, this design was associated with a favorable overall performance profile. MedNuggetizer met the prespecified threshold in the confirmatory primary analysis, produced no hallucinated numeric outputs on benchmark-defined non-derivable datasets, showed high repeatability across 20 repeated runs, and substantially reduced extraction time relative to manual review. Claude Opus 4.5 achieved the numerically strongest primary-endpoint performance and the highest cell-level exactness, but it also produced the only unsafe numeric output on a non-derivable dataset. That distinction is methodologically consequential, because in evidence synthesis the most useful system is not defined by maximal yield alone, but by the degree to which correctness, abstention discipline, output stability, and auditability remain aligned under fixed conditions. Viewed in that light, the present findings

support the potential value of academically developed, audit-ready extraction workflows for structured diagnostic evidence synthesis, with MedNuggetizer representing one such implementation.

The repeated-run analyses further strengthen the practical relevance of the study. Strong performance on a single execution is of limited value if outputs shift materially across repeated runs. Here, all automated systems showed high repeatability, with Gwet's AC1 values ranging from 0.929 to 0.958. By contrast, inter-operator agreement in the time-controlled human extraction exercise was substantially lower. This comparison should be interpreted with caution, because between-rater variability in a single-pass manual exercise is not directly analogous to within-system repeatability across repeated automated runs. Even so, it remains informative as an operational context measure. It suggests that variability persists even among medically trained raters when extraction is performed under practical, non-adjudicated conditions. This is likely to be one of the most realistic niches for carefully constrained automated systems: not as substitutes for methodological oversight, but as front-end tools that reduce first-pass burden, standardize initial extraction, and make disagreement visible earlier in the workflow. This also helps contextualize the findings of Khan et al., who used GPT-4-turbo and Claude-3-Opus in a collaborative two-reviewer workflow across 10 trials from 22 publications and 23 extraction variables [5]. In their held-out test set, 87% of responses were concordant and these concordant responses achieved an accuracy of 0.94, whereas discordant responses were substantially less reliable but improved after cross-critique [5]. Their results suggest that inter-model agreement can serve as a practical confidence signal, but they also show that disagreement marks a qualitatively different error regime rather than mere random noise. By contrast, the present benchmark assessed reliability more directly through adjudicated correctness, abstention behavior, and repeatability across 20 repeated runs, thereby reducing reliance on concordance as a surrogate for trustworthiness.

The optimized-prompt analyses sharpen this interpretation further. Under a shared supplementary prompt that was more permissive but still explicitly documented, all three evaluated LLMs achieved perfect performance on the derivable datasets across three repeated runs, and all correctly abstained on the absent non-derivable dataset. However, this convergence at the apparent upper bound did not extend to the ambiguous non-derivable datasets, where reconstruction attempts remained heterogeneous across systems. That pattern is highly instructive. It indicates that the performance ceiling on clearly derivable material is higher than the locked primary benchmark alone might suggest, but it also shows that safety on non-derivable material remains the more discriminating challenge once prompting becomes more permissive. Prompt optimization can therefore narrow performance differences on derivable datasets, but it does not eliminate the need for explicit safety evaluation. This supports the decision to anchor confirmatory inference to a locked, uniform prompt and to interpret the optimized-prompt analyses as supplementary contextualization rather than as an alternative primary benchmark.

The study also carries broader implications for how AI evaluation in evidence synthesis should be conducted and reported (**Supplementary Table S7**). CHART emphasizes transparent reporting of model identity, prompt development, evaluators, protocol availability, failure modes, and reproducibility procedures [9]. In this context, these are not merely reporting niceties. They are part of the underlying

methodology. Small degrees of freedom in prompting, derivability definitions, or endpoint formulation can materially alter performance impressions. The present benchmark therefore supports the use of CHART-aligned reporting as a minimum standard for future studies of generative AI in evidence synthesis [9]. More broadly, it suggests that prospective protocolization, explicit benchmark definitions, and repeated-run designs should become standard practice in this field.

Several limitations warrant consideration. First, the benchmark was confined to a fixed uro-oncologic diagnostic domain, and generalizability to other specialties, document types, and extraction targets remains uncertain. Second, the task was intentionally narrow and focused on diagnostic 2 × 2 contingency tables; performance on other evidence synthesis targets, including time-to-event outcomes, adjusted effect estimates, adverse event data, and risk-of-bias judgments, was not assessed. Accordingly, the present findings should not be assumed to generalize unchanged to other clinical domains, document constellations, or extraction targets without independent validation of the workflow under comparably prespecified benchmark conditions. Third, although repeated runs were treated as independent realizations under nominally identical conditions, residual dependencies related to provider-side updates or document-ingestion behavior cannot be excluded completely. Fourth, the human comparator exercise was designed as an operational context benchmark rather than a fully adjudicated duplicate-extraction workflow and should be interpreted accordingly. Fifth, the optimized supplementary prompt is informative but also illustrates how rapidly methodological transparency can be diluted once prompt permissiveness increases and reconstruction rules become more elaborate. Finally, only a limited number of frontier LLMs were evaluated. As collaborative and agentic systems become more prominent, their evaluation will require even greater methodological discipline and transparency [5, 7].

Future work should extend this framework in three directions. First, similar benchmarks should be applied to additional clinical domains and to extraction targets beyond diagnostic test accuracy. Second, hybrid workflows in which automated extraction is followed by targeted human verification should be evaluated formally, as this is likely to be the most realistic near-term implementation strategy. Third, academically developed task-specific systems such as MedNuggetizer should be compared not only with base LLMs, but also with collaborative and agentic architectures under prospectively locked conditions. The most plausible near-term role for MedNuggetizer is not fully autonomous evidence synthesis, but deployment as an audit-ready extraction layer within structured review workflows, where it may reduce time burden, improve consistency, and make uncertainty explicit before final human adjudication. Within that bounded remit, such systems may become genuinely useful methodological instruments for evidence synthesis rather than merely impressive demonstrations of technical capability.

Conclusions

Under prospectively locked and methodologically conservative benchmark conditions, automated systems achieved high performance in diagnostic data extraction from full-text publications, but meaningful differences emerged once correctness, abstention safety, repeatability, and efficiency were considered together. MedNuggetizer and Claude Opus 4.5 met the prespecified threshold criterion in the

confirmatory primary analysis; Claude Opus 4.5 was numerically strongest on raw correctness, whereas MedNuggetizer combined perfect abstention safety on non-derivable datasets with high repeatability and substantial efficiency gains over manual extraction.

These findings support a cautious but clearly optimistic view of automated extraction in evidence synthesis. Tightly constrained, prospectively evaluated, and audit-ready systems can now achieve performance levels that are operationally relevant for structured diagnostic review workflows. The most credible near-term use case is not fully autonomous evidence synthesis, but transparent hybrid pipelines in which automated extraction accelerates first-pass data acquisition while preserving human methodological control over adjudication, ambiguity, and final synthesis decisions.

Abbreviations

AC1: Gwet's AC1 coefficient

AI: artificial intelligence

AUC_binary: binary area under the receiver operating characteristic curve

CHART: Chatbot Assessment Reporting Tool

FN: false negatives

FP: false positives

IQR: interquartile range

LLMs: large language models

NPV: negative predictive value

PDF: Portable Document Format

PPV: positive predictive value

TN: true negatives

TP: true positives

Declarations

Ethics approval and consent to participate

Not applicable. This study did not involve human participants, human data, human tissue, or animals. The benchmark corpus was derived exclusively from publicly available full-text articles and publicly

available supplementary materials.

Consent for publication

Not applicable.

Availability of data and materials

All data generated or analyzed during this study are included in this published article and its supplementary information files. The benchmark corpus was derived exclusively from publicly available published full-text articles and publicly available supplementary materials. Additional run-level logs and supporting benchmark materials are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

Funding
No specific funding was received for this work. Open access funding may be enabled and organized by the University of Regensburg through the applicable Springer Nature institutional agreement.

Authors' contributions

AK, JRRG, CE, and MM conceived the study. GD and SA contributed to the development of MedNuggetizer and to the methodological design of the extraction workflow evaluated in this study. AK, JRRG, GD, SA, and UK contributed to benchmark development, prompt design, and benchmark operationalization. AK and JRRG conducted the automated system runs. MH, ER, SH, and CE performed the time-controlled human comparator extraction. SH, SS, CG, MB, VE, ER, MH, CE, and MM contributed to clinical interpretation, benchmark contextualization, and critical review of the study design and results. CE and MM supervised the study. AK drafted the first manuscript version. CE and MM critically revised the manuscript for important intellectual content. All authors read and approved the final manuscript.

Acknowledgements
Not applicable.

Authors' information

Not applicable.

References

1. Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al. editors. Cochrane Handbook for Systematic Reviews of Interventions. Version 6.5, updated August 2024. Cochrane; 2024 [cited

2026 Mar 14]. Available from:

2. Salari MA, Baghi Keshtan S, et al. Global Trends in Urological Evidence Synthesis: A Bibliometric Analysis of Systematic Reviews and Meta-analyses. Med J Islam Repub Iran. 2025;39:76. 10.47176/mjiri.39.76. PMID: 41089610; PMCID: PMC12516444.
3. Jansen T, Liebenow LW, Mertens U et al. Data extraction by generative artificial intelligence: Assessing determinants of accuracy using human-extracted data from systematic review databases. Psychol Bull. 2025;151(10):1280–1306. 10.1037/bul0000501. PMID: 41396533.
4. Caponio VCA, Lorenzo-Pouso AI, Magalhaes M, et al. Accuracy of LLMs to retrieve numeric data for meta-analysis in dentistry. J Dent. 2026;164:106245. 10.1016/j.jdent.2025.106245. Epub 2025 Nov 19. PMID: 41265689.
5. Khan MA, Ayub U, Naqvi SAA, et al. Collaborative large language models for automated data extraction in living systematic reviews. J Am Med Inf Assoc. 2025;32(4):638–47. 10.1093/jamia/ocae325. PMID: 39836495; PMCID: PMC12005628.
6. Purewal A, Fautsch K, Klasova J, Hussain N, D'Souza RS. Human versus artificial intelligence: evaluating ChatGPT's performance in conducting published systematic reviews with meta-analysis in chronic pain research. Reg Anesth Pain Med. 2025 Feb 16:rapm-2024-106358. 10.1136/rapm-2024-106358. Epub ahead of print. PMID: 39956557.
7. Gorenshtein A, Omar M, Glicksberg BS, Nadkarni GN, Klang E. AI Agents in Clinical Medicine: A Systematic Review. medRxiv [Preprint]. 2025 Aug 26:2025.08.22.25334232. doi: 10.1101/2025.08.22.25334232. PMID: 40909853; PMCID: PMC12407621.
8. Liu J, Lai H, Zhao W, et al. ADVANCED Working Group. AI-driven evidence synthesis: data extraction of randomized controlled trials with large language models. Int J Surg. 2025;111(3):2722–6. PMID: 39903558; PMCID: PMC12372713.
9. CHART Collaborative. Reporting guideline for chatbot health advice studies: the Chatbot Assessment Reporting Tool (CHART) statement. BMJ Med. 2025;4(1):e001632. 10.1136/bmjmed-2025-001632. PMID: 40761518; PMCID: PMC12320030.
10. Donabauer G, Ateia S, Kruschwitz U et al. MedNuggetizer: Confidence-Based Information Nugget Extraction from Medical Documents. ECIR. 2026. arXiv:2512.15384 [cs.IR].
11. Rodas Garzaro JR, Kravchuk A, Burger M, et al. Diagnostic Performance and Clinical Utility of the Uromonitor® Molecular Urine Assay for Urothelial Carcinoma of the Bladder: A Systematic Review and Diagnostic Accuracy Meta-Analysis. Diagnostics (Basel). 2026;16(2):285. 10.3390/diagnostics16020285. PMID: 41594261; PMCID: PMC12840495.
12. May M, Rodas Garzaro JR, Kravchuk A et al. End-to-End Reliability of Automated Systems for Diagnostic Data Extraction: A Benchmark Study in Uro-Oncologic Evidence Synthesis. medRxiv. MS ID#: MEDRXIV/2025/342959. Protocol v1.0, December 24, 2025.
13. Batista R, Vinagre J, Prazeres H, et al. Validation of a Novel, Sensitive, and Specific Urine-Based Test for Recurrence Surveillance of Patients With Non-Muscle-Invasive Bladder Cancer in a

Comprehensive Multicenter Study. Front Genet. 2019;10:1237. 10.3389/fgene.2019.01237. PMID: 31921291; PMCID: PMC6930177.

14. Sieverink CA, Batista RPM, Prazeres HJM, et al. Clinical Validation of a Urine Test (Uromonitor-V2®) for the Surveillance of Non-Muscle-Invasive Bladder Cancer Patients. Diagnostics (Basel). 2020;10(10):745. 10.3390/diagnostics10100745. PMID: 32987933; PMCID: PMC7599569.
15. Ecke TH, Meisl CJ, Schlomm T, et al. Performance of Urinary Markers in Patients With Suspicious Cystoscopy During Follow-up of Recurrent Non-muscle Invasive Bladder Cancer: BTA Stat, NMP22 BladderChek, UBC Rapid Test, CancerCheck UBC Rapid VISUAL, and Uromonitor in Comparison to Cytology. Urology. 2025;197:119–25. Epub 2024 Dec 1. PMID: 39626834.
16. Rabien A, Rong D, Rabenhorst S, et al. Diagnostic performance of Uromonitor and TERTpm ddPCR urine tests for the non-invasive detection of bladder cancer. Sci Rep. 2024;14(1):30617. 10.1038/s41598-024-83976-2. PMID: 39715826; PMCID: PMC11666540.
17. Wolff I, Kravchuk AP, Wirtz RM, et al. Real-world performance of Uromonitor® in urothelial bladder cancer detection: a multicentric trial. BJU Int. 2024;134(6):992–1000. 10.1111/bju.16450. Epub 2024 Jun 24. PMID: 38923777.
18. Rubio-Briones J, Guerrero Ramos F, Mercadé Sánchez A, et al. External validation of the Uromonitor®-version 2 urine test as a biomarker for optimisation of non-muscle-invasive bladder cancer management. BJU Int. 2026;137(1):216–24. 10.1111/bju.70010. Epub 2025 Oct 1. PMID: 41031577.
19. Ramos P, Brás JP, Dias C, et al. Uromonitor: Clinical Validation and Performance Assessment of a Urinary Biomarker Within the Surveillance of Patients With Nonmuscle-Invasive Bladder Cancer. J Urol. 2025;213(3):304–12. Epub 2024 Nov 19. PMID: 39561374.
20. Azawi N, Vásquez JL, Dreyer T, et al. Surveillance of Low-Grade Non-Muscle Invasive Bladder Tumors Using Uromonitor: SOLUSION Trial. Cancers (Basel). 2023;15(8):2341. 10.3390/cancers15082341. PMID: 37190269; PMCID: PMC10137147.
21. OpenAI. GPT-5.2 model documentation [Internet]. San Francisco (CA): OpenAI; 2025 [cited 2026 Mar 14]. Available from: https://platform.openai.com/docs/models
22. Anthropic. Claude model documentation and version registry [Internet]. San Francisco (CA): Anthropic; 2025 [cited 2026 Mar 14]. Available from: https://docs.anthropic.com
23. Google DeepMind. Gemini 3 Pro model card [Internet]. London (UK): Google DeepMind; 2025 [cited 2026 Mar 14]. Available from: https://deepmind.google/technologies/gemini

Figures

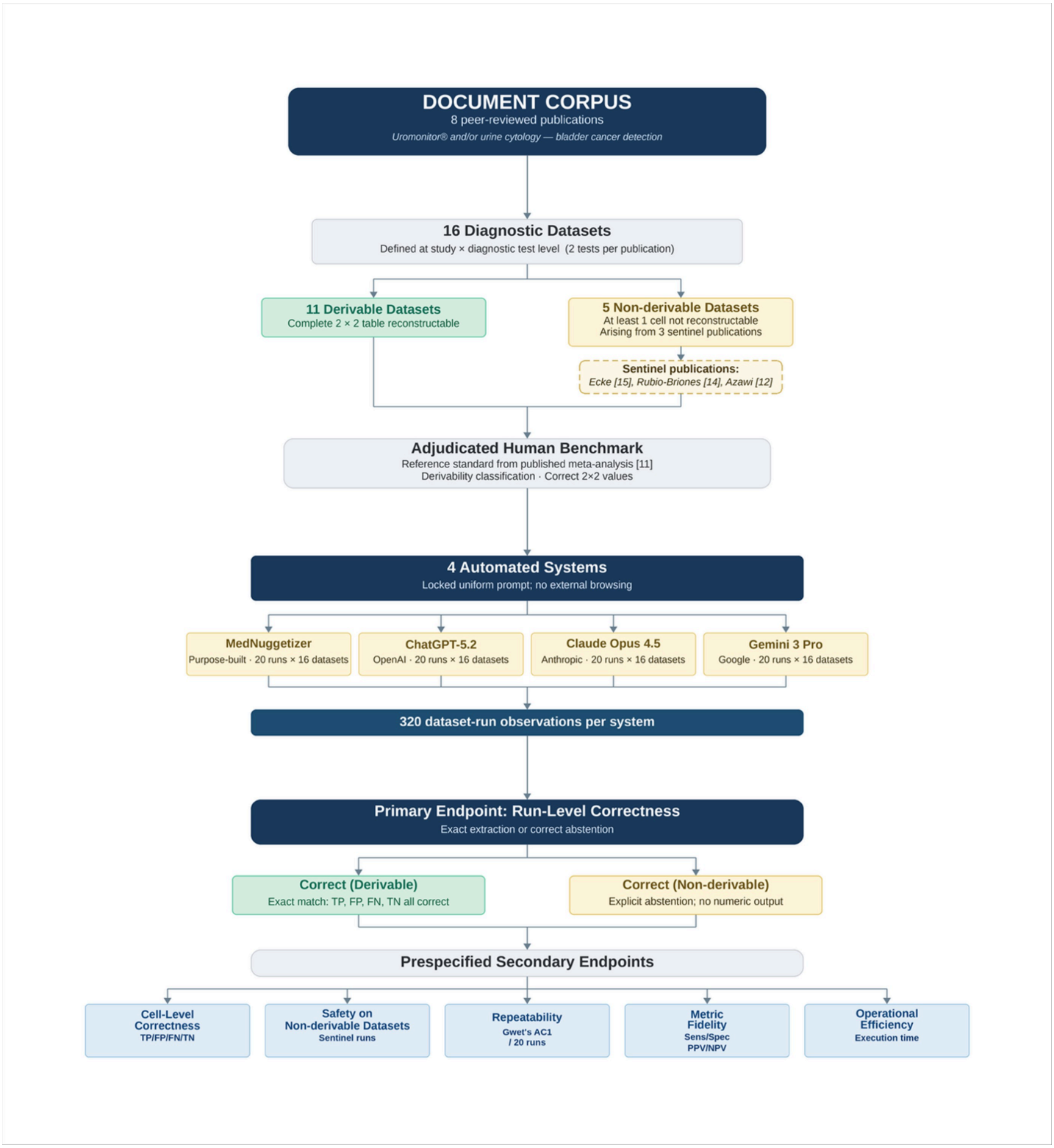


Figure 1

Study flow, benchmark derivability classification, and repeated-run evaluation framework

Note: The locked corpus comprised 8 peer-reviewed publications, yielding 16 diagnostic datasets at the level of study-by-test combinations. Based on the adjudicated human benchmark, 11 datasets were classified as derivable and 5 as non-derivable, the latter arising from 3 prespecified sentinel publications.

Four automated systems were evaluated under a locked uniform prompt across 20 repeated runs per system, yielding 320 dataset-run observations per system. The primary endpoint was run-level correctness, defined as exact extraction for derivable datasets or explicit abstention without numeric output for non-derivable datasets. Prespecified secondary endpoints included cell-level exactness, safety on non-derivable datasets, repeatability across repeated runs, fidelity of derived diagnostic metrics, and operational efficiency.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementBMCMedicalResearchMethodology.docx