

Perspective

Sovereign Large Language Models for Structured Data Extraction from Pathology Reports: A Perspective for the Clinical Laboratory

Ravi Shankar ^{1,2}

¹ Clinical Research and Innovation Office, Tan Tock Seng Hospital, National Healthcare Group, Singapore 308433, Singapore; ravisr.srivastava@gmail.com

² Rehabilitation Research Institute of Singapore (RRIS), Nanyang Technological University, Singapore 308232, Singapore

Abstract

The surgical pathology report remains one of the richest yet least computable artefacts in the clinical record. Diagnostic, prognostic, and treatment-relevant information is recorded predominantly as a free-text narrative that resists aggregation for research, quality monitoring, and cancer registration, while manual abstraction is slow, costly, and difficult to scale. Large language models (LLMs) have rapidly emerged as a means of converting unstructured pathology narrative into structured, analysis-ready data. This perspective examines the current state of the evidence, with particular reference to breast pathology, and foregrounds the distinction between proprietary cloud-hosted models and locally deployed open-weight models. Recent comparative studies indicate that open-weight models can approach the accuracy of proprietary systems for structured extraction, offering a privacy-preserving and cost-controlled alternative that keeps protected health information inside the institutional firewall—a decisive advantage under data-protection regimes such as Singapore’s Personal Data Protection Act (PDPA) and Human Biomedical Research Act (HBRA). We argue that hybrid architectures—pairing deterministic rule-based extraction for unambiguous fields with local LLMs for narrative reasoning—currently offer the most defensible route to laboratory deployment. We also highlight the “reality gap” between synthetic benchmark performance and real-world clinical accuracy, and the need to align studies with emerging reporting and appraisal frameworks (TRIPOD-LLM, PROBAST + AI). Structured extraction is compatible with the quality and traceability expectations of accredited laboratories only when it is verified before use, monitored over time, and kept under human oversight.

Keywords: large language models; natural language processing; pathology reports; structured data extraction; breast cancer; data sovereignty; open-weight models; cancer registry; TRIPOD-LLM



Academic Editor: Gassan Hodaifa

Received: 10 June 2026

Revised: 25 June 2026

Accepted: 26 June 2026

Published: 29 June 2026

Copyright: © 2026 by the author.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. The Pathology Report as Uncomputable Data

The pathology report carries the definitive information used for diagnosis, staging, treatment selection, and outcomes research. Despite decades of effort to encourage synoptic reporting, much of this content continues to be recorded as narrative prose rather than as discrete, machine-readable fields, and large legacy archives remain entirely free text [1]. The practical consequence is that the data most relevant to clinical research and to cancer registration must be re-derived after the fact, one report at a time.

Manual abstraction by trained annotators is the de facto reference standard, but it is labour-intensive, expensive, and inherently difficult to scale to the volumes required for registry operations or large retrospective studies [2,3]. This bottleneck has motivated successive generations of natural language processing (NLP) tooling, and it is the problem against which any automated extraction method must ultimately be judged: not whether it can produce a plausible structured record, but whether it can do so accurately, reproducibly, and at a cost that the laboratory can sustain.

2. From Rule-Based Pipelines to Large Language Models

Early automated extraction relied on rule-based systems and regular expressions, later supplemented by supervised machine-learning models. Domain-adapted transformer encoders such as the BERT family have performed extremely well on well-circumscribed breast pathology fields; a recent single-institution study reported near-perfect parsing accuracy for several fields of interest using a BioBERT-based pipeline [4]. These approaches, however, depend on labelled training data and careful feature engineering, and they require re-training whenever templates, vocabularies, or reporting conventions change—an operational liability in a working laboratory.

LLMs reframe the task. Through zero- or few-shot prompting, they extract structured information from natural-language instructions without bespoke labelled datasets [2]. Truhn and colleagues demonstrated that GPT-4 could convert unstructured histopathology narrative into structured fields with high accuracy, including cross-language validation [5]. In breast pathology specifically, zero-shot GPT-4 has been shown to outperform conventional supervised baselines for classification tasks [6], and related work has automated TNM stage assignment from pathology text with external validation [7], and a multilingual study confirmed that GPT-4 enables post-hoc structuring of free-text reports across languages with high feasibility [8]. The appeal is clear: a single, instruction-driven model can be redirected to new fields or new tumour types with a change in prompt rather than a new training cycle.

3. The Sovereignty Question: Why Local Deployment Matters

Proprietary, cloud-hosted models are the easiest to adopt, but using them for clinical extraction requires transmitting report text—frequently identifiable—to external infrastructure. This raises concerns about data protection, vendor lock-in, recurring cost, and reproducibility, since the underlying model can change without notice [9,10]. For institutions operating under strict data-governance regimes, these concerns are not merely theoretical. In Singapore, the PDPA and the HBRA constrain the movement and secondary use of patient and biomedical data; sending identifiable pathology narrative to an external application programming interface is, in many workflows, either impermissible or operationally restricted.

Locally deployed open-weight models offer a direct answer. Run on institutional graphics-processing hardware through lightweight inference runtimes, they keep all report text inside the institutional firewall, so that no patient data leaves the network [11]. Local deployment additionally supports auditability, version pinning for reproducible analysis, and optional fine-tuning on institutional data—for example, via low-rank adaptation, which has been shown to lift smaller open models to expert-level performance while preserving data sovereignty [11]. The trade-off is a shift in burden from recurring API cost to in-house hardware, engineering, and maintenance [9]; the governance benefit, for many laboratories, justifies it.

To make the deployment envelope concrete for laboratory stakeholders, the hardware requirements are more modest than is often assumed. With current four-bit post-

training quantisation, a 7-billion-parameter model occupies on the order of 6 to 8 gigabytes of graphics-processing-unit (GPU) memory and runs on a single consumer-grade card, whereas a 70-billion-parameter model occupies approximately 40 to 48 gigabytes and fits on one professional accelerator of 48 gigabytes, or on two 24-gigabyte consumer cards in a tensor-parallel configuration [12]. Working memory scales with parameter count, quantisation level, and the key-value cache required for the chosen context length, so report length and batch size, not the model alone, determine the practical footprint. Throughput is a property of the inference runtime rather than of the model: a paged-attention serving stack removes most key-value cache fragmentation and raises sustained throughput by roughly two to four times relative to naive batching, which allows a single mid-range server to process on the order of thousands of pathology reports per day in batched mode [13]. Local generation will not match the per-request latency of a heavily optimised proprietary endpoint; registry and research extraction is, however, performed in scheduled batches rather than interactively, so aggregate throughput, not single-report latency, is the operative metric, and it is comfortably sufficient for routine laboratory volumes.

4. Evidence on Open-Weight Versus Proprietary Models

The central empirical question is whether models that can be self-hosted are accurate enough to be useful. The evidence is increasingly encouraging. A comprehensive comparison across commercial and open-source models found that open-source systems matched the precision of GPT-4 for structuring pathology data, with performance varying chiefly according to prompt-engineering strategy and quantisation method rather than according to whether the model was open or closed [14]. In a breast cancer histopathology benchmark, large open-weight models reached accuracy statistically comparable to a trained human annotator (approximately 94.7%) and to GPT-4o (approximately 96.1%), while a smaller 70-billion-parameter model, though slightly below human accuracy, was noted as a practical self-hosting option given its reduced computational requirements [2]. Broader multi-use-case, multilingual evaluations reach a similar conclusion: small-to-medium general-purpose open models, with optimised prompting, can approach expert-annotator performance across diseases, languages, and institutions [15]. Table 1 summarises the principal approaches and their trade-offs.

Table 1. Approaches to structured data extraction from pathology reports and their characteristics. * Reported accuracy figures are indicative and drawn from heterogeneous studies and datasets; they are not directly comparable and should not be interpreted as head-to-head benchmarks.

Approach	Data Sovereignty	Reported Accuracy *	Principal Limitation
Rule-based/regex	Full (local)	High on deterministic fields	Brittle to template and wording changes; poor on narrative reasoning
Supervised models (e.g., BERT family)	Full (local)	≈0.99 on selected breast fields	Requires labelled training data and re-training when schemas change
Proprietary cloud LLM (e.g., GPT-4o, Gemini)	Limited (data leaves firewall)	≈96% on curated breast reports	Privacy, vendor lock-in, reproducibility, recurring cost
Open-weight LLM, local deployment	Full (local)	≈92–95%; approaches proprietary	Hardware and engineering burden; prompt- and quantisation-sensitive
Hybrid (rules + local LLM)	Full (local)	Field-dependent; best of both	Pipeline complexity; needs per-field validation

These figures should, however, be read with corresponding caution, because they are drawn from heterogeneous studies rather than from a common benchmark. The investigations summarised above differ in almost every dimension that influences a reported accuracy: they draw on different institutions, tumour types, and languages; they extract different field sets against locally defined gold standards; and they report performance using non-equivalent metrics, ranging from exact-match field accuracy to token-level F1 and aggregate inter-rater agreement. Quantisation level, prompt design, and post-processing also vary between studies and, as the comparative work itself observes, frequently account for more of the variance than the open-versus-proprietary distinction [14]. Few studies undertake a controlled head-to-head comparison on a single shared dataset, so the apparent convergence between open-weight and proprietary performance is better interpreted as broad directional evidence than as demonstrated equivalence; differences in dataset curation and the possibility of selective reporting may further inflate the published estimates. For a clinical laboratory audience, the defensible conclusion is therefore narrower than a headline of parity: open-weight models are now credible candidates for local extraction, but their adequacy for a given task must be established on the laboratory's own data rather than inferred from cross-study comparison.

Read together, the rows of Table 1 describe a single trade-off rather than a ranking. Rule-based extraction and supervised encoder models are fully local and highly accurate on the narrow fields for which they are built, but they are brittle: rules fail when wording or templates change, and supervised models require labelled data and re-training whenever the reporting schema changes. Proprietary cloud models remove that engineering burden and perform strongly on curated reports, yet they forfeit data sovereignty and introduce recurring cost, vendor lock-in, and version instability. Locally deployed open-weight models recover sovereignty and approach proprietary accuracy, at the price of in-house hardware and sensitivity to prompt and quantisation choices. The hybrid row is not a fifth competitor but a composition of the others: deterministic methods are assigned to the unambiguous fields they handle best, and a local model is reserved for the narrative fields that require interpretation, so that the principal limitation becomes pipeline complexity and the need for per-field validation rather than any single failure mode. The accuracy figures in the table are indicative only and, for the reasons set out above, should not be read as head-to-head benchmarks.

5. The Reality Gap: Synthetic Benchmarks Versus Clinical Text

Headline accuracy figures must be read with caution, because performance measured on clean or synthetic test material consistently overstates deployable accuracy. A dual-validation study of an agent-based extraction system for breast cancer synoptic reports illustrates the point starkly: models that achieved 93.8–99.0% accuracy on synthetic data fell to field-level recall of 61.8–87.7% on real clinical reports, a drop of 11–32 percentage points, and model size did not predict real-world performance [16]. For registry and research use, the relevant quantities are field-level recall and the false-negative rate, because a missed field is a silent gap in the structured record rather than a visible error. Commentators have similarly cautioned that error propagation during extraction is a primary risk that must be actively managed rather than assumed away [17].

The practical implication is unambiguous. A model's accuracy on a public benchmark, or even on a curated local sample, is not a substitute for validation on the laboratory's own reports, with their idiosyncratic templates, abbreviations, and dictation artefacts. Accuracy is a property of the model-and-data pair, not of the model alone.

6. Hybrid Architectures and the Case for Human Oversight

Different fields suit different methods. Deterministic fields—specimen type, laterality, accession and collection dates, identifiers, and coded values—are captured reliably and transparently by rules and regular expressions, and there is little reason to delegate them to a probabilistic model. Fields that demand interpretation of narrative—histologic subtype, grade, margin status, lymphovascular invasion, or the synthesis of receptor results—are where LLMs add the most value. A hybrid pipeline assigns each field to the most reliable available method, reserving the model’s reasoning capacity for the fields that genuinely require it (Figure 1).

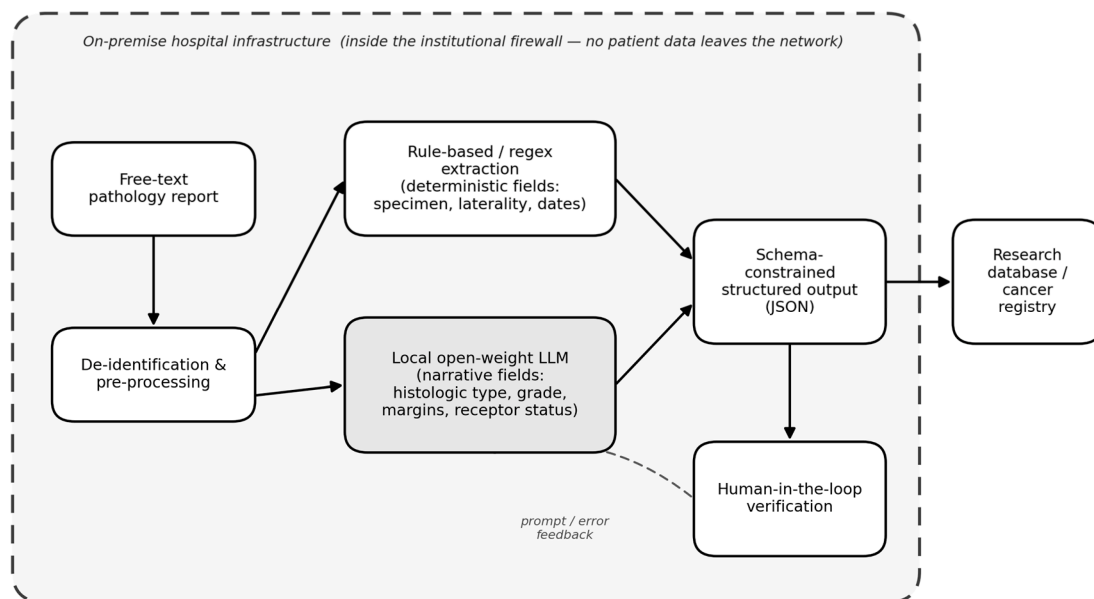


Figure 1. A sovereign hybrid pipeline for structured data extraction from pathology reports. Report text is de-identified and pre-processed inside the institutional firewall; deterministic fields are captured by rule-based extraction while narrative fields are handled by a locally deployed open-weight model. Outputs are constrained to a fixed schema and pass through human verification before reaching the research database or cancer registry. No patient data leaves the on-premise network.

In practical terms, a hybrid pipeline can be specified field by field. Identifiers, accession and collection dates, specimen type, and laterality are routed to deterministic rules and controlled dictionaries, whereas histologic subtype, grade, margin status, lymphovascular invasion, and the synthesis of receptor and biomarker results are routed to the local model under a fixed output schema. Error handling is layered: rule outputs are checked against controlled vocabularies and plausibility ranges; model outputs that fail schema validation or fall below a confidence threshold are flagged rather than written; and any field on which the rule and the model disagree is escalated for human adjudication. Integration with the laboratory information system is most robust through the messaging and resource layers already used for results reporting, namely HL7 version 2 messages or HL7 FHIR resources, so that structured output reaches the research database or cancer registry through an interface the laboratory already governs rather than through an ad hoc export. The operational burden of maintaining such a pipeline should not be understated. It requires a maintained validation set, documented version control of prompts, model weights, quantisation settings, and rule dictionaries, and a defined re-validation step whenever any of these components change. These maintenance obligations are the practical counterpart to the governance benefits described earlier, and they are the reason a hybrid pipeline must be treated as a managed laboratory process rather than as a single deployment event.

Two design choices materially improve robustness. First, schema-constrained decoding—forcing the model to emit output conforming to a predefined structure, typically JSON—sharply reduces malformed and hallucinated output. Second, and indispensably, human-in-the-loop verification is retained for fields that feed clinical or registry decisions. The objective of automation here is not to remove the expert but to direct expert attention efficiently, surfacing low-confidence extractions for review while routine, high-confidence fields pass through. The error feedback that verification generates can, in turn, refine prompts and inform any local fine-tuning.

Two further considerations determine whether human oversight actually reduces the labour it is intended to economise. The first is cognitive ergonomics. Verification yields a net saving over manual abstraction only when reviewers are shown a structured draft alongside the source span that justifies each field, so that the task becomes confirmation against highlighted evidence rather than re-reading the entire report. Presenting an unsourced, confident-looking extraction instead invites automation bias, the well-documented tendency to accept a plausible machine output without adequate scrutiny [18]. Oversight should therefore be designed to surface uncertainty rather than to conceal it: low-confidence and disagreement-flagged fields are presented prominently for review, routine high-confidence fields are sampled rather than exhaustively re-checked, and the interface records reviewer corrections so that error rates can be monitored over time. The second consideration is the silent gap created by a missed field, which, unlike a wrong value, produces no visible error. Human review alone is an incomplete safeguard against omission, and two model-side strategies complement it. Confidence thresholding, derived from token-level probabilities or from self-consistency across repeated generations, converts a silent omission into an explicit low-confidence flag that is routed for review. Ensemble or multi-sample extraction, in which the same field is generated several times or by more than one model and disagreement triggers human adjudication, further reduces the chance that a field is silently dropped. Together these measures shift the residual risk from undetected omission towards detected uncertainty, which is precisely the failure mode that a verification workflow is designed to handle.

7. Reporting, Appraisal, and Laboratory Quality

As extraction studies proliferate, comparability and trustworthiness depend on standardised reporting and appraisal. The TRIPOD-LLM statement provides a targeted reporting guideline for studies that develop, tune, or evaluate LLMs in healthcare, emphasising transparency, human oversight, and task-specific performance reporting [10]; it extends the TRIPOD + AI framework for prediction models more generally [19]. For critical appraisal and for systematic reviews of extraction studies, PROBAST + AI offers an updated instrument to assess quality, risk of bias, and applicability across both regression- and AI-based models [20]. Authors and reviewers in this space should treat these as the default scaffolding for design and reporting, and the World Health Organization's guidance on the ethics and governance of large multi-modal models situates the technical work within a broader accountability framework [21].

For the accredited laboratory, the analogy to existing quality expectations is direct. Under ISO 15189:2022 [22], examination methods must be verified before use, their validity monitored over time, and their performance traceable [23]. An extraction model deployed against clinical reports is, in this light, an examination process: it cannot simply be switched on. It must be validated against a local gold standard, monitored for drift as report styles and model versions change, and governed by documented procedures—precisely the discipline that organisational-preparedness analyses for LLMs in pathology informatics have called for [22].

8. Final Thoughts

LLM-based extraction of structured data from pathology reports is now feasible, and—critically for jurisdictions with strict data governance—it can be sovereign. Open-weight models deployed locally can approach the accuracy of proprietary cloud services while keeping patient data inside the institutional firewall. The most defensible path to deployment is neither a wholesale embrace of a single large model nor a retreat to brittle rules, but a hybrid pipeline: deterministic methods for unambiguous fields, a local model for narrative reasoning, schema-constrained output, validation against the laboratory's own reports, human oversight where decisions depend on it, and reporting aligned to TRIPOD-LLM and PROBAST + AI. The reality gap between benchmark and bedside means accuracy must be measured locally before reliance. Above all, automation does not transfer responsibility: the laboratory remains accountable for the validity of every result it releases, whether a human or a model first put it into structured form.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created.

Conflicts of Interest: The author declares no conflicts of interest.

Abbreviations

LLM	Large Language Model
NLP	Natural Language Processing
PHI	Protected Health Information
PDPA	Personal Data Protection Act (Singapore)
HBRA	Human Biomedical Research Act (Singapore)
GPU	Graphics-Processing Unit
JSON	JavaScript Object Notation
LoRA	Low-Rank Adaptation
ISO	International Organization for Standardization

References

1. Srigley, J.R.; McGowan, T.; MacLean, A.; Raby, M.; Ross, J.; Kramer, S.; Sawka, C. Standardized synoptic cancer pathology reporting: A population-based approach. *J. Surg. Oncol.* **2009**, *99*, 517–524. [[CrossRef](#)] [[PubMed](#)]
2. Balasubramanian, J.B.; Adams, D.; Roxanis, I.; de Gonzalez, A.B.; Coulson, P.; Almeida, J.S.; García-Closas, M. Medical Report Information Extractor: Leveraging large language models for structured information extraction from pathology reports. *arXiv* **2025**, arXiv:2502.12183.
3. Lee, D.; Vaid, A.; Menon, K.M.; Freeman, R.; Matteson, D.S.; Marin, M.L.; Nadkarni, G.N. Using large language models to automate data extraction from surgical pathology reports: Retrospective cohort study. *JMIR Form. Res.* **2025**, *9*, e64544. [[CrossRef](#)] [[PubMed](#)]
4. Park, P.; Choi, Y.; Han, N.; Park, Y.-L.; Hwang, J.; Chae, H.; Yoo, C.W.; Choi, K.S.; Kim, H.-J. Leveraging natural language processing for efficient information extraction from breast cancer pathology reports: Single-institution study. *PLoS ONE* **2025**, *20*, e0318726. [[CrossRef](#)] [[PubMed](#)]
5. Truhn, D.; Loeffler, C.M.; Müller-Franzes, G.; Nebelung, S.; Hewitt, K.J.; Brandner, S.; Bressem, K.K.; Foersch, S.; Kather, J.N. Extracting structured information from unstructured histopathology reports using generative pre-trained transformer 4 (GPT-4). *J. Pathol.* **2024**, *262*, 310–319. [[PubMed](#)]
6. Sushil, M.; Zack, T.; Mandair, D.; Zheng, Z.; Wali, A.; Yu, Y.N.; Quan, Y.; Butte, A.J. A comparative study of zero-shot inference with large language models and supervised modeling in breast cancer pathology classification. *arXiv* **2024**, arXiv:2401.13887.
7. Kefeli, J.; Berkowitz, J.; Acitores Cortina, J.M.; Tsang, K.K.; Tatonetti, N.P. Generalizable and automated classification of TNM stage from pathology reports with external validation. *Nat. Commun.* **2024**, *15*, 8916. [[CrossRef](#)] [[PubMed](#)]

8. Adams, L.C.; Truhn, D.; Busch, F.; Kader, A.; Niehues, S.M.; Makowski, M.R.; Bressem, K.K. Leveraging GPT-4 for post hoc transformation of free-text radiology reports into structured reporting: A multilingual feasibility study. *Radiology* **2023**, *307*, e230725. [[CrossRef](#)] [[PubMed](#)]
9. Dennstädt, F.; Hastings, J.; Putora, P.M.; Schmerder, M.; Cihoric, N. Implementing large language models in healthcare while balancing control, collaboration, costs and security. *npj Digit. Med.* **2025**, *8*, 143. [[CrossRef](#)] [[PubMed](#)]
10. Gallifant, J.; Afshar, M.; Ameen, S.; Aphinyanaphongs, Y.; Chen, S.; Cacciamani, G.; Demner-Fushman, D.; Dligach, D.; Daneshjou, R.; Fernandes, C.; et al. The TRIPOD-LLM reporting guideline for studies using large language models. *Nat. Med.* **2025**, *31*, 60–69. [[CrossRef](#)] [[PubMed](#)]
11. Gao, S.; Zhao, X.; Chen, L.; Yu, J.; Tian, S.; Zhou, H.; Chen, J.; Long, S.; He, Q.; Feng, X. Enhancing privacy-preserving deployable large language models for perioperative complication detection: A targeted strategy with LoRA fine-tuning. *npj Digit. Med.* **2025**, *8*, 773. [[CrossRef](#)] [[PubMed](#)]
12. Dettmers, T.; Pagnoni, A.; Holtzman, A.; Zettlemoyer, L. QLoRA: Efficient finetuning of quantized LLMs. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 10088–10115. [[CrossRef](#)]
13. Kwon, W.; Li, Z.; Zhuang, S.; Sheng, Y.; Zheng, L.; Yu, C.H.; Gonzalez, J.E.; Zhang, H.; Stoica, I. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th ACM SIGOPS Symposium on Operating Systems Principles (SOSP '23)*; ACM: New York, NY, USA, 2023; pp. 611–626.
14. Grothey, B.; Odenkirchen, J.; Brkic, A.; Schömig-Markiefka, B.; Quaas, A.; Büttner, R.; Tolkach, Y. Comprehensive testing of large language models for extraction of structured data in pathology. *Commun. Med.* **2025**, *5*, 96. [[CrossRef](#)] [[PubMed](#)]
15. Spaanderman, D.J.; Prathaban, K.; Zelina, P.; Mouheb, K.; Hejtmánek, L.; Marzetti, M.; Schurink, A.W.; Chan, D.; Niemantsverdriet, R.; Hartmann, F.; et al. Evaluating open-weight large language models for structured data extraction from narrative medical reports across multiple use cases and languages. *arXiv* **2025**, arXiv:2511.10658.
16. Hart, S.N.; Bergamaschi, T.S. Agent-based large language model system for extracting structured data from breast cancer synoptic reports: A dual-validation study. *JAMIA Open*, 2026; *in press*.
17. Daungsupawong, H.; Wiwanitkit, V. Large language models as tool for pathology report data extraction: Comment. *Turk. J. Pathol.* **2025**, *41*, 70. [[CrossRef](#)] [[PubMed](#)]
18. Goddard, K.; Roudsari, A.; Wyatt, J.C. Automation bias: A systematic review of frequency, effect mediators, and mitigators. *J. Am. Med. Inform. Assoc.* **2012**, *19*, 121–127. [[CrossRef](#)] [[PubMed](#)]
19. Collins, G.S.; Moons, K.G.M.; Dhiman, P.; Riley, R.D.; Beam, A.L.; Van Calster, B.; Ghassemi, M.; Liu, X.; Reitsma, J.B.; van Smeden, M.; et al. TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* **2024**, *385*, e078378. [[CrossRef](#)] [[PubMed](#)]
20. Moons, K.G.M.; Damen, J.A.A.; Kaul, T.; Hooft, L.; Navarro, C.A.; Dhiman, P.; Beam, A.L.; Van Calster, B.; Celi, L.A.; Denaxas, S.; et al. PROBAST+AI: An updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *BMJ* **2025**, *388*, e082505. [[CrossRef](#)] [[PubMed](#)]
21. World Health Organization. *Ethics and Governance of Artificial Intelligence for Health: Guidance on Large Multi-Modal Models*; WHO: Geneva, Switzerland, 2024.
22. ISO 15189:2022; Medical Laboratories—Requirements for Quality and Competence. ISO: Geneva, Switzerland, 2022.
23. Hart, S.N.; Hoffman, N.G.; Gershkovich, P.; Christenson, C.; McClintock, D.S.; Miller, L.J.; Jackups, R.; Azimi, V.; Spies, N.; Brodsky, V. Organizational preparedness for the use of large language models in pathology informatics. *J. Pathol. Inform.* **2023**, *14*, 100338. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.