Who Fails Where? LLM and Human Error Patterns in Endometriosis Ultrasound Report Extraction

Haiyi Li
haiyi.li@student.adelaide.edu.au
Adelaide University
Adelaide, Australia

Yutong Li
yutong.li@student.adelaide.edu.au
Adelaide University
Adelaide, Australia

Yiheng Chi
yiheng.chi@student.adelaide.edu.au
Adelaide University
Adelaide, Australia

Alison Deslandes
alison.deslandes@adelaide.edu.au
Robinson Inst., Adelaide University
Adelaide, Australia

Mathew Leonardi
leonam@mcmaster.ca
McMaster University
Hamilton, Canada

Shay Freger
Fregers@mcmaster.ca
McMaster University
Hamilton, Canada

Yuan Zhang
yuan.zhang01@adelaide.edu.au
Robinson Inst., Adelaide University
Adelaide, Australia

Jodie Avery
jodie.avery@adelaide.edu.au
Robinson Inst., Adelaide University
Adelaide, Australia

Mary Louise Hull
louise.hull@adelaide.edu.au
Robinson Inst., Adelaide University
Adelaide, Australia

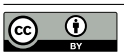
Hsiang-Ting Chen
tim.chen@adelaide.edu.au
Adelaide University
Adelaide, SA, Australia

Abstract

In this study, we evaluate locally deployed large language models (LLMs) for converting unstructured endometriosis transvaginal ultrasound (eTVUS) reports into structured data. Across 49 de-identified reports, we compared three on-premise LLMs (7B/8B and 20B parameters) against expert human extraction using a 185-field schema. The 20B model achieved the highest mean accuracy (86.02%), substantially outperforming the smaller models. Crucially, LLMs and humans exhibited complementary error patterns: the LLM excelled on structured fields (date formatting, measurement decomposition) where humans made protocol errors, while humans demonstrated superior performance on interpretive fields involving negation and clinical terminology. Targeted prompt engineering yielded only marginal gains, indicating that these errors reflect model limitations rather than instruction gaps. These findings support a human-in-the-loop workflow in which the LLM generates structured drafts, automated validation flags rule-verifiable errors, and human review focuses on fields requiring clinical interpretation.

CCS Concepts

• **Computing methodologies** → **Natural language processing**;
• **Applied computing** → **Health informatics**.



CHI EA '26, Barcelona, Spain

ACM ISBN 979-8-4007-2281-3/2026/04
https://doi.org/10.1145/3772363.3798872

Keywords

information extraction; large language models; human-in-the-loop; medical reporting

ACM Reference Format:
Haiyi Li, Yutong Li, Yiheng Chi, Alison Deslandes, Mathew Leonardi, Shay Freger, Yuan Zhang, Jodie Avery, Mary Louise Hull, and Hsiang-Ting Chen. 2026. Who Fails Where? LLM and Human Error Patterns in Endometriosis Ultrasound Report Extraction. In *Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems (CHI EA '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3772363.3798872

1 Introduction

Free-text ultrasound reports contain clinically valuable information, but key variables are embedded in heterogeneous narrative styles and local formatting conventions, limiting their use in analytics, model training, and auditing [14, 16, 18]. Across clinical domains, this structural barrier complicates secondary use and necessitates substantial manual abstraction [8, 9, 13, 16]. In settings where privacy requirements preclude cloud-based processing, extraction remains a manual, safety-critical task: abstractors must interpret clinical content while enforcing protocol constraints such as field decomposition, formatting standards, and missingness conventions [5, 18]. Our contextual inquiry identified recurring risk points, including terminology variation, inconsistent report detail, and verification-heavy routines that induce fatigue and increase silent transcription and field-alignment errors [5, 18]. These observations suggest that effective support tools should prioritise reviewability and accountability over full automation, enabling practitioners to calibrate their reliance on algorithmic assistance within real workflows [2, 7, 10].

Locally deployable LLMs offer a practical means of scaling abstraction without transmitting sensitive reports to external services [3, 12]. Recent studies demonstrate that LLMs can perform few-shot clinical extraction with substantial medical knowledge [1, 15], yet they also produce well-formed outputs that are semantically incorrect, failures that are difficult to detect from the output alone [5, 17]. In structured reporting, this problem is acute: schema-compliant responses may still mishandle negation, map terms to incorrect categories, or misinterpret context-dependent findings. Without visible indicators of error, users struggle to calibrate trust, leading to both over-reliance and under-reliance on model outputs [4, 6]. Research on clinical AI deployments reinforces this concern, showing that operational success depends less on standalone accuracy than on workflow integration and mechanisms that direct reviewer attention to high-risk outputs [2, 7]. Designing such mechanisms requires understanding where LLMs and humans each fail.

This study addresses that gap by investigating the error patterns of local LLMs and human abstractors to inform design guidelines for human-AI collaborative abstraction systems. Using 49 de-identified eTVUS reports and a 185-field extraction schema, we benchmark three on-premise models (7B to 20B parameters) against expert-verified human abstraction. All 49 reports come from a single clinic; however, given the clinical complexity of endometriosis and the unusually rich 185-field schema, we expect the observed reliability boundaries and reviewability-oriented controls to be informative for broader ultrasound report extraction and human-in-the-loop workflows. We nonetheless view multi-site validation as an important next step. Our analysis targets two dimensions with direct design implications: report-level variance, which governs review effort and suggests where batch processing is viable, and field-type error patterns, which reveal where human judgment remains essential and should be preserved. We find that LLMs and humans fail in complementary ways, motivating a division of labour in which automated validation catches mechanical errors and risk-based triage routes semantically sensitive fields to human review. This error pattern characterization not only highlights where models succeed or fail, but also directly informs the design of workflow controls that align reviewer effort with error risk. We also test whether targeted prompting can reduce errors on critical fields; the limited and inconsistent gains suggest that workflow-level safeguards, rather than prompt refinement, offer the more reliable mechanism for managing extraction risk.

2 Experiment

2.1 Data and Schema

The dataset consisted of unstructured, de-identified sonologists reports obtained from a specialized gynecology and obstetrics ultrasound clinic in Canada. These reports were heterogeneous, containing both structured data fields and free-text ultrasound narratives. To prepare the data for the pipeline, each report, originally in PDF format, was converted to plain text. We used a layout-preserving extraction process designed to retain semantic content while removing extraneous metadata and formatting artifacts.

Expert-verified reference labels (Verified Truth). To create the expert-verified reference labels (*Verified Truth*) for the 185-field schema, we used a two-stage human abstraction and verification process. First, a trained clinical research staff member with domain experience in endometriosis imaging studies performed the initial structured abstraction in a spreadsheet template (Excel) under the organization's standard operating conventions. During abstraction, the de-identified PDF report and the spreadsheet were viewed side-by-side (split-screen) and fields were entered directly into the corresponding columns, using predefined dropdowns/enumerations and formatting rules (e.g., date and numeric conventions). Missing or absent information was recorded using an explicit token to avoid implicit assumptions. Second, an independent domain expert in endometriosis ultrasound imaging research reviewed every extracted field against the source report and corrected discrepancies, producing the final *Verified Truth*. Quality control was conducted by (i) field-by-field cross-checking against the PDF, (ii) enforcing schema-level constraints (allowed values for categorical fields; canonical formats for dates/numbers), and (iii) resolving ambiguous cases through explicit adjudication against the report text rather than guesswork. To preserve double-blind review, we omit names here; the roles and qualifications are described in a verifiable, role-based manner.

The extraction target was defined by a structured endometriosis-centric schema. This schema was created by programmatically transforming the header row of a reference Excel data dictionary into a concise JSON schema. This file defined all key fields, their data types, and output format constraints, serving as the ground truth structure for both model prompting and final evaluation [1, 13]. The reference Excel file contained 185 fields in total. Each field was programmatically assigned a data type based on its values and intended use. The schema included five major data types: Numeric (6 fields), Date (2 fields), Text (19 fields), and Categorical (157 fields). The majority of fields were categorical, typically representing controlled vocabularies or discrete clinical options, while a smaller subset are free-text or numeric entries. This distribution reflected the highly structured nature of the target schema and the clinical emphasis on standardized reporting.

2.2 Extraction Pipeline

We designed an on-premise extraction pipeline to ensure full patient privacy and data sovereignty. The entire system operated offline, without reliance on external APIs, and is deployed on commodity hardware. The pipeline was built on the OLLAMA platform, using three different LLM models **gpt-oss:20b**, **llama3-8b** and **mistral-7b**. The workflow proceeded as follows:

(1) **Schema-Guided Prompting:** Each plain-text report was processed sequentially in batch mode. For every report, the model received the full textual content along with the JSON schema embedded within the prompt as an instructional template.
(2) **Inference:** The LLM generated a structured JSON object containing the extracted field-specific values. This output was saved as an intermediate file for validation.
(3) **Validation and Post-processing:** A rule-based validation layer was applied to the JSON outputs. This step normalized missing value indicators (e.g., `0`, `NA`, and empty strings), harmonized categorical variables to a controlled vocabulary,

and verified field completeness and data-type conformity [11].

The experiment ran on a personal computer equipped with an NVIDIA RTX 3090 GPU (24GB VRAM).

3 Preliminary Evaluation

Scoring and accuracy definition. We report *field-level accuracy* scored against an expert-verified reference (*Verified Truth*). For each report, we score each of the 185 schema fields as correct (1) if the prediction matches the reference *in meaning* after normalization, and incorrect (0) otherwise; report-level accuracy is the mean across fields, and we report the mean and standard deviation (SD) across 49 reports. For **numeric/date fields**, we convert outputs into a canonical representation (e.g., consistent units/format) and compare in that canonical form. For **text/categorical fields**, we treat minor spelling errors and formatting differences as correct when they preserve meaning; we apply lightweight normalization (case-folding plus whitespace/punctuation normalization) and accept semantically equivalent variants used in the reporting protocol (e.g., equivalent abbreviations or wording). Missing information is represented by an explicit token `NOT_MENTION`; a field is counted as correct when both prediction and reference indicate `NOT_MENTION`, and incorrect when one indicates missingness while the other provides a value. For protocol fields where `0` and `NA` both denote *not detected / not recorded*, we treat `0` and `NA` as the same missingness state during scoring. We did not perform significance testing; comparisons below are descriptive and intended to characterize performance level and variability in this setting.

Table 1: Aggregate LLM Backbone Comparison. Overall mean accuracy, and per-report standard deviation (Std) on the Sugo dataset, benchmarked against the verified ground truth.

Model Backbone	Mean Accuracy (%)	Std (%)
gpt-oss:20b	**86.02**	**6.87**
llama3-8b	80.53	4.58
mistral-7b	78.89	4.68
Clinical RA	98.40	2.13

Our quantitative results are summarized in Table 1. Using the same expert-sonographer annotated and double-checked reference labels (*Verified Truth*), we scored and compared three locally-deployed LLMs and a Clinical Research Assistant (Clinical RA) performing manual abstraction. The Clinical RA achieved a mean accuracy of **98.40%** with low variability (SD 2.13%). Among the three LLMs, `gpt-oss:20b` achieved the highest mean accuracy in our dataset (**86.02%**), while `llama3-8b` and `mistral-7b` achieved mean accuracies of 80.53% and 78.89%, respectively. Notably, while `gpt-oss:20b` achieved the highest mean accuracy among the LLMs, it also showed the largest per-report variance (SD 6.87%), indicating less consistent performance across reports.

Figure 2 further illustrates these distributional differences. The box-and-whisker plot shows that `gpt-oss:20b` attains a higher median accuracy but also a wider interquartile range (IQR), with a small number of outliers where accuracy drops markedly (e.g., below 65%). Overall, this suggests that the larger model performs better on average in our dataset but is more sensitive to a subset of challenging reports, whereas smaller models exhibit a narrower (but lower) performance range.

3.1 Error Analysis by Field Type

To investigate the sources of divergence across 185 fields in complex endometriosis ultrasound reports, we stratified performance using a small number of broad, protocol-aligned categories (Figure 3). While more fine-grained field groupings (e.g., by subjectivity or cognitive complexity) are desirable, in this domain many fields sit on blurry boundaries (e.g., categorical fields that still require semantic judgment), and overly specific groupings risk brittle, hard-to-generalize conclusions. We therefore report results at this coarse granularity. Unless noted otherwise, the LLM results in this subsection focus on `gpt-oss:20b` as the best-performing local backbone in our study, to highlight its typical error distribution.

The analysis reveals complementary failure modes between the LLM and the human extractor. The LLM performs best on more structured, protocol-constrained fields, achieving its highest accuracy on **Date Fields (97.3%)** and **Numeric Fields (92.7%)**. Remaining errors in these categories are primarily omissions and schema-alignment failures (e.g., incomplete decomposition of multi-dimensional measurements). In contrast, errors are more concentrated in semantically sensitive Text and Categorical fields, where the model often fails through omissions or inconsistent terminology/ontology mapping [14, 16, 17]. Given that errors on semantically nuanced fields remain challenging, we conducted a follow-up experiment to test whether targeted prompt strategies could mitigate these errors.

In contrast, the human extractor's errors were rarely clinical misinterpretations, but instead were predominantly data-entry protocol failures. A common error involved correctly reading a 3D nodule measurement from the report but failing to split it across three separate required database fields.

3.2 Error Analysis on Clinically Important Fields

To examine whether prompt engineering could improve performance on key items, we conducted a follow-up experiment using a critical-field prompt for `gpt-oss:20b`. This prompt explicitly identified the seven most clinically critical fields, with the goal of improving extraction consistency for these items. The critical-field prompt achieved marginally higher mean accuracy (88.8%, SD = 6.23) compared to the generic prompt (87.0%, SD = 6.44). The critical-field prompt produced only a small change in mean accuracy on these fields, and the effect was not stable relative to report-level variability. While prompt engineering showed marginal improvements, our analysis suggests that these improvements are not sufficient to resolve the deeper semantic challenges in clinical fields. We interpret this result narrowly: the specific *importance-emphasis* prompting strategy we tested yields limited and unstable gains in our setting, but it does not rule out other, more advanced prompting or tool-augmented strategies. In the near term, we expect auditable workflow safeguards (e.g., rule-based validation, risk-prioritized review, and structured error logging) to be the more dependable mitigation for semantic errors in schema-constrained extraction.[1, 5].

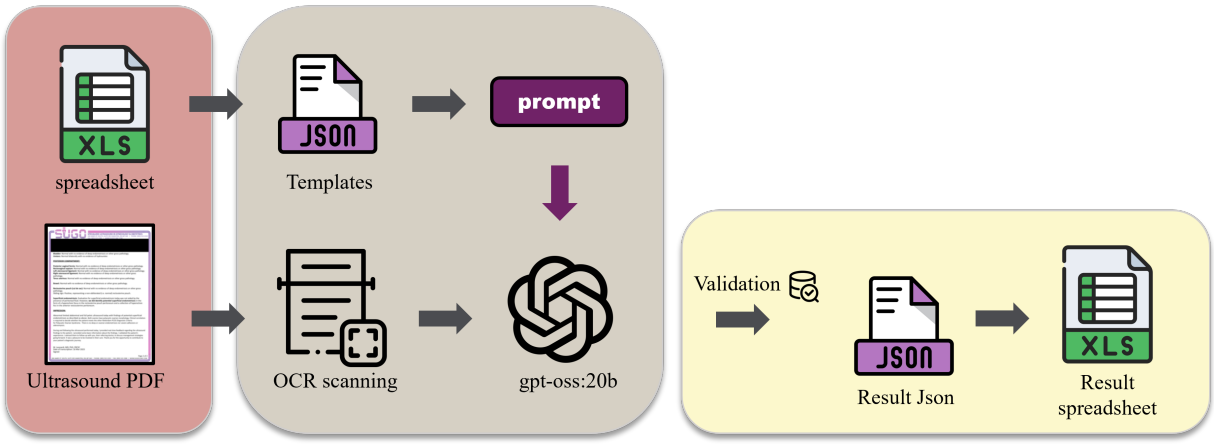


Figure 1: Overview of the on-premise workflow. De-identified ultrasound PDFs and the organization's spreadsheet template are processed locally by a layout-aware LLM backend to produce a schema-aligned structured JSON draft. A field stratification strategy and rule-based validation prioritize semantically sensitive fields for review. An interactive human-in-the-loop interface supports rapid trace-back from each field to relevant PDF text anchors, enabling verification and correction before exporting a verified structured dataset for clinical research use.

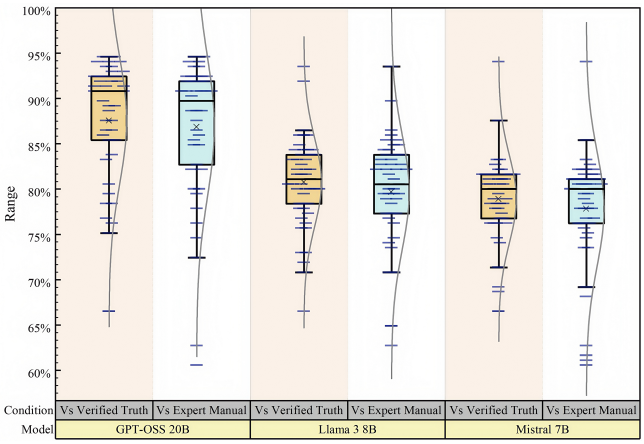


Figure 2: Report-level accuracy distributions for three locally-deployed LLMs. `gpt-oss:20b` attains a higher median accuracy but exhibits larger variability, including occasional low-accuracy outliers.

4 Discussion

Complementary Failure Modes Enable Task Allocation Our results reveal that LLMs and human abstractors fail on different field types, establishing a basis for differentiated task allocation. The LLM achieved near-human accuracy on Date (97.3%) and Numeric (92.7%) fields, where errors were predominantly mechanical (incomplete measurement decomposition, format mismatches) and detectable through rule-based validation. In contrast, the LLM struggled with Categorical and Text fields (77.3% and 80.0%), which require interpreting negation, mapping synonymous terms to controlled vocabularies, and inferring clinical intent. These semantic errors are syntactically well-formed, confirming that schema compliance alone cannot ensure extraction quality.

Actionable workflow controls. To make this implication operational rather than purely high-level, we propose three concrete controls: (C1) *rule-based validation with exception-based review* for protocol-checkable fields (format/range/enum checks for dates and measurements), with *targeted re-check* for relatively important columns and *random report-level audits* to maintain safety; (C2) *risk triage* semantically sensitive fields (negation/conditional findings and terminology/ontology mapping) to mandatory human verification; and (C3) require *evidence-oriented trace-back* for reviewed fields (one-click links from each extracted value to its supporting source span in the report) to enable fast spot-checking and accountability without rereading entire PDFs.

Safety and accountability boundary. Auto-validation does not mean auto-acceptance. In our workflow, LLM outputs remain human-in-the-loop artifacts and require a final human sign-off before becoming part of a verified dataset. We recommend explicit safety boundaries, including audit logs (traceable field-level edits), confidence gating and escalation rules for low-confidence or atypical cases, and an explicit assignment of responsibility for final verification. In high-stakes clinical applications, even low error rates can have unacceptable consequences if erroneous values are presented as verified. To address this, our proposed workflow imposes explicit **confidence gating thresholds** below which outputs are always escalated to human review, and requires documented **human sign-off** for every field before inclusion in any dataset used for clinical research. We further recommend maintaining **field-level audit logs** that record model outputs, reviewer edits, and timestamps, enabling traceability and accountability in post-hoc review.

Extraction Variance as a Deployment Criterion. Mean accuracy alone obscures operational risk. The 20B model achieved the highest average accuracy (86.02%) but also the widest variance (SD = 6.87%), with outliers below 65%. Inspection of these cases revealed

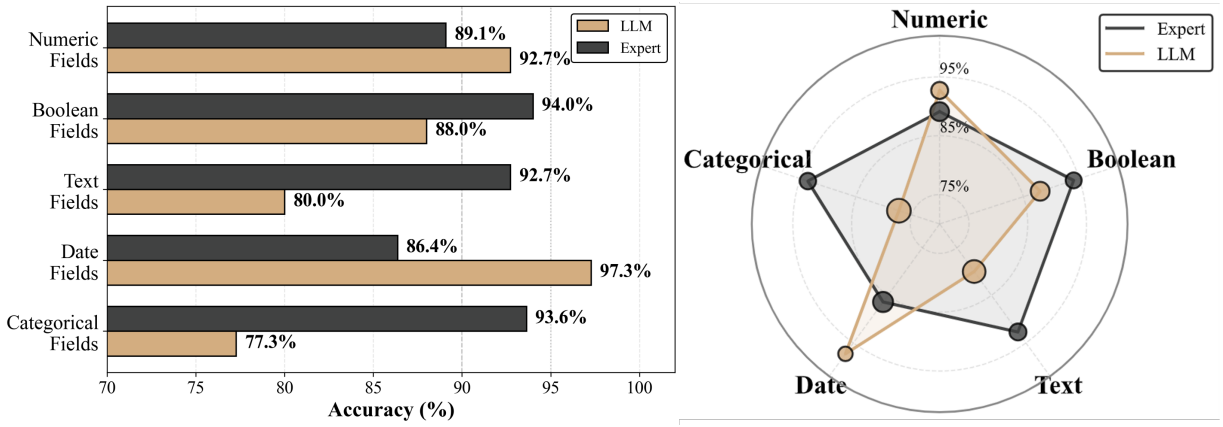


Figure 3: Performance breakdown by field type. In our setting, the LLM performs best on structured fields (Date, Numeric) and worse on semantically nuanced fields (Text, Categorical).

two patterns: reports with atypical formatting triggered cascading failures, and reports dense with negated or conditional findings led to accumulated errors across categorical fields. Smaller models showed lower variance but at a lower accuracy level, suggesting more consistent but conservative outputs.

High variance translates to unpredictable review burden. A reviewer calibrated for occasional errors may overlook reports where a third of extractions fail. This observation supports mechanisms such as confidence-based triage, routing structurally atypical or low-confidence reports to full review. More broadly, model selection should be framed as a variance-accuracy trade-off: in some operational contexts, a smaller model with predictable, recoverable errors may prove more practical than a larger model whose sporadic failures are harder to detect.

5 Limitation

Our evaluation is bounded by a modest sample (49 reports) from a single institution, which may not capture variation in reporting styles across sites. Given the sample size, we report mainly descriptive statistics; future work should apply paired significance tests and bootstrapping to quantify uncertainty on larger multi-site datasets, and validate the workflow on multi-site cohorts.

6 Conclusion

We present a systematic evaluation of locally deployed LLMs for structured extraction from endometriosis transvaginal ultrasound reports, where data-sovereignty requirements preclude cloud-based processing. Three findings inform human-AI abstraction workflow design. First, the 20B model achieved the highest field-level accuracy (86.02%) while remaining feasible for on-premise deployment. Second, LLMs and human abstractors exhibit complementary error patterns—LLMs excel on structured fields while humans outperform on interpretive ones—suggesting a division of labour rather than full automation. Third, critical-field emphasis prompting yielded no meaningful improvement on semantically sensitive fields, though alternative strategies (few-shot exemplars, chain-of-thought, constrained decoding) remain unexplored. These findings support a human-in-the-loop workflow in which local LLMs generate structured drafts at scale, automated validation flags mechanical errors, and targeted human review addresses fields requiring clinical judgement. Data-sovereignty constraints restricted evaluation to locally deployable models (7B–20B); cloud-hosted or domain-specialised medical backbones (e.g., MedGemma) may exhibit different error patterns and warrant future comparison. Future work should validate these patterns on multi-site datasets, explore a broader prompting space, and develop lightweight mechanisms for confidence-based triage.

Acknowledgments

This work was supported by the Australian Government through the Medical Research Futures Fund: Primary Health Care Research Data Infrastructure Grant 2020, the Australasian Society of Ultrasound in Medicine Research Grant 2022, Endometriosis Australia, and Australia's Economic Accelerator Ignite Grant 2025. We also acknowledge the contributions of Dr. Yuan Zhang, Associate Professor Jodie Avery, and all others who provided their support and guidance throughout this work. We would like to thank Endometriosis Australia for their additional support as paper funders.

References

[1] Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large language models are few-shot clinical information extractors. *arXiv preprint* (2022). https://arxiv.org/abs/2205.12689 arXiv:2205.12689.

[2] Emma Beede, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Ruamviboonsuk, and Laura M. Vardoulakis. 2020. A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3313831.3376718

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, et al. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 1877–1901. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[4] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. 5, CSCW1, Article 188 (April 2021), 21 pages. doi:10.1145/3449287

[5] Felix Busch, Lena Hoffmann, Daniel Pinto Dos Santos, Marcus R Makowski, Luca Saba, Philipp Prucker, et al. 2025. Large language models for structured reporting in radiology: past, present, and future. *European Radiology* 35, 5 (2025), 2589–2602.

[6] Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. 2015. The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems. In *2015 International Conference on Healthcare Informatics*. 160–169. doi:10.1109/ICHI.2015.26

[7] Carrie J. Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. "Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 104 (Nov. 2019), 24 pages. doi:10.1145/3359206

[8] Sergio M Castro, Eugene Tseytlin, Olga Medvedeva, Kevin Mitchell, Shyam Visweswaran, Tanja Bekhuis, and Rebecca S Jacobson. 2017. Automated annotation and classification of BI-RADS assessment from radiology reports. *Journal of Biomedical Informatics* 69 (2017), 177–187.

[9] Mary F Davis, Subramaniam Sriram, William S Bush, Joshua C Denny, and Jonathan L Haines. 2013. Automated extraction of clinical traits of multiple sclerosis in electronic medical records. *Journal of the American Medical Informatics Association* 20, e2 (2013), e334–e340.

[10] Geraldine Fitzpatrick and Gunnar Ellingsen. 2013. A Review of 25 Years of CSCW Research in Healthcare: Contributions, Challenges and Future Agendas. *Comput. Supported Coop. Work* 22, 4–6 (Aug. 2013), 609–665. doi:10.1007/s10606-012-9168-0

[11] Sami-Ramzi Leyh-Bannurah, Zhe Tian, Pierre I Karakiewicz, Ulrich Wolffgang, Guido Sauter, Margit Fisch, et al. 2018. Deep learning for natural language processing in urology: state-of-the-art automated extraction of detailed pathologic prostate cancer data from narratively written electronic health records. *JCO Clinical Cancer Informatics* 2 (2018), 1–9.

[12] Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus* 15, 6 (2023).

[13] Guergana K Savova, Eugene Tseytlin, Sean Finan, Melissa Castine, Timothy Miller, Olga Medvedeva, et al. 2017. DeepPhe: a natural language processing system for extracting cancer phenotypes from clinical records. *Cancer Research* 77, 21 (2017), e115–e118.

[14] Seyedmostafa Sheikhalishahi, Riccardo Miotto, Joel T Dudley, Alberto Lavelli, Fabio Rinaldi, and Venet Osmani. 2019. Natural Language Processing of Clinical Notes on Chronic Diseases: Systematic Review. *JMIR Medical Informatics* 7, 2 (27 Apr 2019), e12239. doi:10.2196/12239 PubMed: 31066697. Also available at: http://medinform.jmir.org/2019/2/e12239/.

[15] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, et al. 2023. Large language models encode clinical knowledge. *Nature* 620, 7972 (2023), 172–180.

[16] Yanshan Wang, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, Yuqun Zeng, Saeed Mehrabi, Sunghwan Sohn, and Hongfang Liu. 2018. Clinical information extraction applications: A literature review. *Journal of Biomedical Informatics* 77 (2018), 34–49. doi:10.1016/j.jbi.2017.11.011

[17] Yuqing Wang, Yun Zhao, and Linda Petzold. 2023. Are large language models ready for healthcare? a comparative study on clinical language understanding. In *Machine Learning for Healthcare Conference*. PMLR, 804–823.

[18] David L. Weiss and Curtis P. Langlotz. 2008. Structured Reporting: Patient Care Enhancement or Productivity Nightmare? *Radiology* 249, 3 (2008), 739–747. doi:10.1148/radiol.2493080988 PMID: 19011178.