

# RAPTOR: Generative AI for Parsing Colorectal Cancer Referrals to Streamline Faster Diagnostic Standard Pathways

*Sofiat Abioye<sup>1</sup>, Shazad Ashraf<sup>2</sup>, Junaid Qadir<sup>4</sup>, Adam Byfield<sup>3</sup>, Anusha Jose<sup>3</sup>, William Poulett<sup>3</sup>, Ben Wallace<sup>3</sup>, Adil Butt<sup>2</sup>, Colm Forde<sup>2</sup>, Marcus Mottershead<sup>2</sup>, Simon Fallis<sup>2</sup>, Andrew Beggs<sup>2</sup>, Aneel Bhangu<sup>2</sup>, Lukman Akanbi<sup>1</sup>, Muhammad Bilal<sup>1</sup>*

<sup>1</sup> Faculty of Business and Law, Birmingham City University, United Kingdom

<sup>2</sup> University Hospital Birmingham, Birmingham, United Kingdom

<sup>3</sup> NHS England (National Health Service), United Kingdom

<sup>4</sup> College of Engineering, Qatar University, Doha, Qatar

**Abstract.** Delays in processing urgent cancer referrals hinder Faster Diagnostic Standards (FDS), with manual extraction of patient data (demographics, symptoms and test results) remaining a bottleneck in colorectal two-week wait (2WW) pathways. We evaluate generative AI (GenAI) for automating structured data extraction from colorectal cancer (CRC) 2WW referrals, comparing the reasoning capabilities of GPT-4o-Mini and DeepSeek-R1 against clinician-led extraction. Both models achieved near-human precision (GPT-4o-Mini: 94.83%, DeepSeek-R1: 93.72%) while reducing the processing time by 10-fold. Key challenges included non-deterministic output, OCR noise (e.g. handwritten annotations, overlapping text), and contextual ambiguity, notably misclassified checkboxes, symptom misattribution, and numerical inconsistencies (e.g. fecal immunochemical test (FIT) unit conversions). We also proposed an uncertainty quantification mechanism to flag uncertain extractions for human review. Despite residual limitations, GenAI shows the potential to improve efficiency, standardisation, and equity in cancer pathways by alleviating administrative burdens. Future work should prioritise hybrid AI-clinician workflows, domain-specific fine-tuning, and real-world validation to ensure reliable clinical integration.

**Keywords:** Generative AI · Colorectal Cancer · Urgent Referrals · Faster Diagnostic Standard · Large Language Models

## 1 Introduction

Rapid processing of urgent cancer referrals is critical for early diagnosis under the NHS Faster Diagnostic Standards (FDS), yet systemic inefficiencies persist in real-world practice. General practitioners (GPs) submit suspected cancer cases via the Electronic Referral Service (e-RS), but interoperability gaps between primary and secondary care systems require clinicians to manually extract data

from referrals—a process prone to delays, human error, and variability due to inconsistent form structures, missing fields, and ambiguous terminology [5]. These bottlenecks exacerbate the risks of delayed diagnoses [3]. Robust automated tools to parse referral data to streamline diagnostic pathways are vital [19].

The heterogeneity of CRC 2WW referrals complicates automated extraction. Critical data fields—such as fecal immunochemical test (FIT) results, hemoglobin values, and symptoms—are often missing, inconsistently labeled (e.g., “weight loss” vs. “unintentional weight drop”), or embedded in non-machine-readable formats like handwritten annotations or checkboxes [10]. Semantic ambiguities arise when key details (e.g., rectal bleeding) are inferred rather than explicitly entered, while redundant or conflicting FIT values risk misinterpretation. Unit inconsistencies (e.g., g/L vs. g/dL) and unstructured free-text narratives necessitate manual reconciliation, diverting clinician time. Embedded images in form headers or logos further degrade OCR performance [13]. These barriers underscore the urgent need for AI-driven solutions capable of contextual disambiguation to reliably and rapidly parse semi-structured referrals at scale.

Current approaches to automate referral processing—rule-driven heuristics or traditional machine learning—struggle with the variability of CRC 2WW referrals [18]. Large language models (LLMs) like GPT, DeepSeek, and BERT-based architectures demonstrate transformative potential, achieving clinician-level precision in structured data extraction (e.g., >98% accuracy in retrieval-augmented workflows [7]) and outperforming expert systems in tasks like negation detection from unstructured data [2]. While proprietary models marginally surpass open-source alternatives in domain-specific performance [20], both enable scalable digital technologies. Key challenges remain: LLM opacity complicates clinical validation, high computational costs hinder adoption, and biased or hallucinated outputs—clinically unsound responses lacking evidence—compromise reliability despite adaptive prompt tuning [11]. Robust evaluation frameworks are essential to address these concerns before clinical deployment.

This study evaluates the *in silico* efficacy of GPT-4o-Mini and DeepSeek-R1—state-of-the-art reasoning LLMs—in automating data extraction from semi-structured colorectal cancer (CRC) 2WW referrals. The objectives of our study include (1) comparing the performance of LLM in analysing complex referrals, (2) benchmarking AI accuracy and efficiency against manual extraction from clinicians, and (3) integrating risk-aware assurance by design to ensure robustness for clinical deployment. CRC serves as an example due to its diagnostic burden: Only 50% of 2WW cases meet the NHS 28-day FDS target, with 90% not having cancer, underscoring the need for efficient prioritisation [6,8]. Our pipeline simulates real-world triage by processing referrals via LLMs, iteratively resolving ambiguities (e.g., conflicting FIT values, implicit symptoms) through clinician feedback. Methodologically, we quantify time savings and error reduction against manual workflows, to reduce delays. The novelty lies in the comparison of LLM for CRC 2WW referrals and scalable human-AI collaboration, critical to clinical trust. Early results indicate a 10-fold reduction in processing

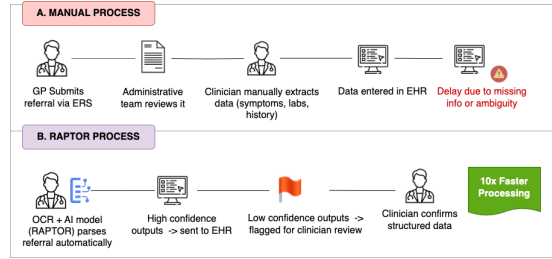


Fig. 1: **Clinical workflow comparison.** Top: manual CRC referral processing involves delays due to ambiguous documentation. Bottom: RAPTOR automates extraction via OCR and LLMs, flagging low-confidence fields for review.

time, demonstrating how LLM can augment (not replace) clinician judgment to accelerate pathways without compromising safety (Figure 1).

## 2 Material and Methods

We used a synthetic data set replicating real-world CRC 2WW referral workflows, carefully designed by clinical domain experts and without confidential patient data. Ethical approval is not necessary for synthetic data analysis, per institutional and local governance guidelines, and compliance was verified with the Health Research Authority decision tool.<sup>1</sup> The study was registered as a service improvement project with University Hospitals Birmingham (UHB) Cancer Services.

### 2.1 CRC Referrals Dataset Generation

A synthetic dataset was created to replicate CRC 2WW referral workflows, designed by clinicians with over five years of post-GMC registration experience. To capture real-world variability, the dataset was segmented into high-, intermediate-, and low-risk CRC likelihood categories. Referrals included diverse data entry scenarios such as typed and handwritten text, checkboxes, and overridden values—to capture documentation variability. A structured parsing approach then categorized referrals into the following categories:

1. *Patient Details:*
  - Demographics (age, sex, NHS number)
  - Red flag symptoms (rectal bleeding, bowel habit changes, weight loss)
  - Diagnostic tests (FIT, hemoglobin, MCV, ferritin, TTG, renal function)
  - Medical history (comorbidities, medications, smoking/alcohol status)
  - WHO performance status
2. *GP Details:*
  - Clinician/practice identifiers
  - Referral metadata (date, priority, declarations)

The referrals were converted to PDFs and stored in version-controlled repository.

<sup>1</sup> <https://www.hra-decisiontools.org.uk/research/>

## 2.2 LLM Selection, OCR Adaptation, and Prompt Tuning

We selected two models (i.e., GPT-4o-Mini and DeepSeek-R1) for automated extraction of structured JSON from semi-structured CRC 2WW referral PDFs. These models were chosen for their strong clinical document performance, fast inference, and accessible deployment. Both models, accessed in their public versions (January 2025), were tested in zero-shot mode to assess their generalisability without domain-specific fine-tuning. Initial OCR attempts with Tesseract and multimodal LLMs struggled with checkbox states, strikethroughs, and handwritten annotations. We fine-tuned Google Document AI on 200 synthetic CRC referrals simulating NHS artifacts (e.g., handwriting, occlusion). Fifteen fields were annotated on 300 DPI grayscale images. A 70:30 train:test split (stratified by CRC risk) was used to evaluate OCR performance. AutoML Vision OCR (6 epochs, batch size 4, lr 3e-5) achieved 90.5% F1 (93.5% precision, 87.8% recall). Full setup is detailed in the GitHub<sup>2</sup>. We employed a structured pipeline for data extraction as described below:

1. *OCR Processing*: Fine-tuned Document AI extracts text from referrals.
2. *Pre-processing*: Data cleaning, validation, and normalisation.
3. *Zero-shot LLM Reasoning*: GPT-4o-Mini or DeepSeek-R1 generate JSON using tailored prompts to handle complex data without hallucination.
4. *Output Structuring*: Final JSON data stored for clinical decision support.

The extraction prompt was iteratively refined to improve precision and adaptability. Key optimisations included hierarchical data grouping (symptoms, tests, medical history), dynamic unit-of-measure mapping for generalisability, mandatory field enforcement with null placeholders, error-resistant unit of measurement (UOM) normalisation, and hallucination controls to restrict outputs to referral-derived content. The final prompt was designed as a domain-agnostic framework compliant with clinical standards, adaptable to various oncology referrals.

## 2.3 Evaluation of LLMs and Risk Assessment

Model outputs (GPT-4o-Mini and DeepSeek-R1) were compared to the validated ground truth of clinicians. Accuracy was defined as the percentage of referrals in which LLM outputs matched expert-extracted data. Discrepancies were blindly reviewed by two CRC triage consultants to distinguish true errors from clinically acceptable variations, focusing on misinterpretations (e.g., symptom misattribution, numerical inconsistencies) and structured field misclassifications. The findings informed refinements to improve clinical relevance. The NHS AI Quality Community of Practice (AIQCoP) assessed risks per NHS standards, focusing on clinical safety and workflow integration. Key risks—OCR noise, misclassification, and model interpretability—were mitigated through safeguards like human oversight for ambiguities. Administrative impacts, including prioritisation biases, were also addressed. This ensured AI outputs met NHS reliability standards while balancing automation with diagnostic safety. Statistical analyses

<sup>2</sup> <https://github.com/bilalcodehub/swiftcare-ai>

were performed using Python. Descriptive statistics included counts, percentages, and ranges to summarise data distributions (e.g., symptom prevalence, test results) and model accuracy rates.

### 3 Results

Analysis of 111 synthetic CRC referrals (Table 1) revealed a median age of 64.5 years (range 21–125), near-equal gender distribution (52.25% male), and predominant rectal bleeding (53.15%), weight loss (25.23%), and iron deficiency anemia (IDA, 45.05%). Lab heterogeneity included widely varying FIT scores (median 9  $\mu\text{g/g}$ , range 2-400) and documentation errors (e.g., Hb: 119.0 g/dL). Most patients (77.4%) were triaged via urgent Straight-to-Test pathways.

Referrals were classified into low (41.4%), intermediate (6.3%), and high (52.3%) CRC risk groups.

Table 1: Clinical and Demographic Characteristics of CRC Referrals

<b>(A) Demographics/ WHO Status</b>		<b>(B) Presenting Symptoms</b>	
Characteristic	Value	Symptom	n (%)
Age, years (Range)	64.5 [21–125]	Rectal Bleeding	59 (53%)
Gender	Male 52%	IDA	50 (45%)
	Female 48%	Weight Loss	28 (25%)
Ethnicity	86% Caucasian	Abd/Rectal Mass	30 (combined)
	Others <8% each	Other Symptoms	<10% each
WHO Score 0–4	54%, 23%, 7%, 11%, 5%	<i>Abbreviations: Abd = Abdominal Pain; IDA = Iron Deficiency Anemia;</i>	
<b>(C) Medical History &amp; Lifestyle</b>		<b>(D) Labs and Risk Pathways</b>	
Variable	n (%)	Test Panel	Values (Range)
Consultation Recorded	111 (100%)	Hematology	Hb: 73 [55–150], MCV: 3 [50–67], Ferritin: 42 [4–6]
Smoking Status	50 (45%)	Renal Function	Urea: 16 [3–6], Creatinine: 16 [44–67], eGFR: 16 [90]
Allergies Reported	49 (44%)	Other Markers	FIT: 110 [2–400], TTG: 14 [1–100]
Medications Logged	17 (15%)	CRC Pathway	STT 77%, OPD 23%
Alcohol Intake Missing	101 (91%)	Risk Group	High 52%, Low 41%, Mid 6%
		<i>Abbreviations: FIT = fecal immunochemical test; STT = straight-to-test; OPD = outpatient clinic.</i>	

The Inter-Observer Variability (IOV) study (Table 2) revealed GPT-4oMini achieved near-human accuracy (94.83%) in parsing CRC referrals, surpassing DeepSeekR1 (93.72%) against expert annotation. Field-level analysis highlights complementary strengths: DeepSeek outperforms in eGFR (96.40% vs.88.29%) and last consultation (77.48% vs.74.77%), while GPT-4oMini leads in rectal bleeding (91.89% vs.61.26%), weight loss (78.38% vs.73.87%), and hemoglobin (81.98% vs.67.57%). Both excel at structured demographics (e.g., clinic address, telephone), yet diverge in clinical markers—DeepSeek excels at iron deficiency

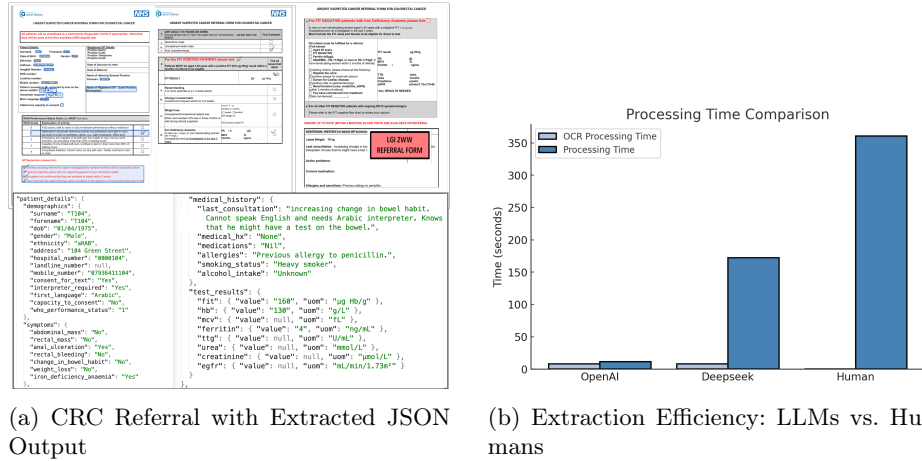


Fig. 2: Illustration of LLM-powered referral extraction and comparative efficiency.

anaemia (63.96% vs.59.46%) and WHO performance status (91.89% vs.90.09%). Each parser thus suits different tasks (e.g., DeepSeek for eGFR, GPT-4oMini for hemoglobin), guiding model selection or ensemble use. While neither replicates human expertise, GPT-4oMini’s aggregate performance suggests its utility as a parsing adjunct in resource-limited settings.

Confidence analysis (Table 2) reveals GPT-4oMini’s mean confidence (93.81%) aligns closely with its actual accuracy (94.83%), indicating generally well-calibrated self-assessment. Still, specific fields deviate: DeepSeekR1 reports 97% confidence on FIT tests yet attains 91.89% accuracy, whereas GPT-4oMini shows tighter confidence–accuracy coupling (88.29% vs.89.19%). DeepSeekR1 achieves near-perfect agreement in structured fields (e.g., 100% confidence and accuracy for addresses), but overstates confidence for hemoglobin (76% vs.67.57%) compared to GPT-4oMini (83.78% vs.81.98%). Both models slightly overestimate confidence in demographics (e.g., GPT-4oMini at 98.65% vs.98.20% for ethnicity), while last consultation differs sharply (GPT-4oMini: 74.86% vs.74.77%, DeepSeekR1: 54%). These patterns suggest GPT-4oMini offers more nuanced confidence estimates, whereas DeepSeekR1 exhibits a tighter but occasionally rigid margin. Figures 2a and 2b illustrate automated CRC referral extraction and comparative extraction times.

AIQCoP risk analysis found most system risks mitigated by technical safeguards, with residual challenges shaping future assurance. Semi-structured input constraints limited unexpected scenarios, while OCR fine-tuning addressed edge cases like strikethroughs or tick elements. Confidence-driven scoring rerouted ambiguous data to administrative review and flagged for clinician validation. Unmitigated risks mainly arose from clinical history free-text variability, where unstructured narratives risked omission or inconsistency despite schema-based

Table 2: Attribute-Level Performance and Agreement Analysis of GPT-4o-Mini and DeepSeek-R1 in Parsing CRC Referrals

Section	Field	GPT-4o-Mini DeepSeek-R1 Agreement (A/D) (%)				
		Acc.	Conf.	Acc.	Conf.	
<b>Patient Details</b>	Address	100.00	99.91	97.30	100.00	100 / 0
	Capacity to consent	88.29	88.74	95.50	100.00	93.5 / 6.5
	Consent for text	96.40	96.85	96.40	100.00	93.5 / 6.5
	Date of birth	99.10	99.55	99.10	100.00	100 / 0
	Ethnicity	98.20	98.65	98.20	100.00	100 / 0
	First language	98.20	98.65	99.10	100.00	100 / 0
	Forename	98.20	98.65	98.20	100.00	100 / 0
	Gender	99.10	60.99	100.00	100.00	100 / 0
	Hospital number	98.20	98.65	99.10	100.00	100 / 0
	Interpreter required	94.59	95.05	98.20	100.00	100 / 0
	Landline number	100.00	100.00	100.00	100.00	100 / 0
	Mobile number	100.00	99.91	99.10	100.00	100 / 0
	Surname	98.20	98.65	99.10	100.00	100 / 0
	WHO performance status	90.09	83.69	91.89	88.00	87.9 / 12.1
<b>Symptoms</b>	Abdominal Mass	99.10	99.55	100.00	100.00	100 / 0
	Anal Ulceration	97.30	97.12	100.00	100.00	94.3 / 5.7
	Change In Bowel Habit	88.29	84.33	61.26	63.00	63.6 / 36.4
	Iron Deficiency Anaemia	59.46	59.82	63.96	71.00	59.7 / 40.3
	Rectal Bleeding	91.89	89.91	61.26	63.00	75.4 / 24.6
	Rectal Mass	96.40	96.22	98.20	100.00	100 / 0
	Weight Loss	78.38	78.11	73.87	82.00	71.1 / 28.9
<b>Medical History</b>	Alcohol Intake	99.10	99.55	100.00	100.00	100 / 0
	Allergies	99.10	99.01	94.59	99.00	100 / 0
	Last Consultation	74.77	74.86	77.48	54.00	94.3 / 5.7
	Medical Hx	93.69	93.42	90.99	96.00	93.9 / 6.1
	Medications	94.59	94.32	93.69	88.00	100 / 0
	Smoking Status	99.10	99.55	98.20	100.00	100 / 0
<b>Test Results</b>	eGFR	88.29	87.93	96.40	97.00	84.7 / 15.3
	Ferritin	100.00	100.00	100.00	100.00	100 / 0
	FIT	88.29	89.19	91.89	97.00	89.6 / 10.4
	Hemoglobin (Hb)	81.98	83.78	67.57	76.00	82.3 / 17.7
	MCV	99.10	99.28	99.10	100.00	98.1 / 1.9
	TTG	99.10	98.56	99.10	100.00	98.1 / 1.9
	Urea	99.10	99.19	99.10	100.00	98.1 / 1.9
	Creatinine	99.10	99.19	99.10	100.00	98.1 / 1.9
<b>GP Details</b>	Clinic address	100.00	100.00	100.00	87.00	100 / 0
	Date of decision to refer	99.10	99.64	100.00	78.00	100 / 0
	Date of referral	99.10	99.64	99.10	78.00	99.1 / 0.9
	Email	100.00	100.00	100.00	100.00	100 / 0
	Fax number	100.00	100.00	100.00	100.00	100 / 0
	GP Name	100.00	100.00	100.00	100.00	100 / 0
	Telephone	100.00	100.00	100.00	100.00	100 / 0
<b>Overall</b>	–	<b>94.83</b>	<b>93.81</b>	<b>93.72</b>	<b>93.00</b>	<b>94.65 / 7.73</b>

Acc. = Accuracy; Conf. = Confidence; A/D = Agreement / Disagreement with ground truth.

mitigation. Full mitigation requires rigorous testing with synthetically engineered forms covering the whole range of patient scenarios.

## 4 Discussion

Our evaluation of LLMs versus human experts in the referrals parsing revealed that both models achieved near-human accuracy, but exhibited critical limitations in the interpretation of fields requiring vision capabilities in clinical PDFs.

These vulnerabilities, resulting from inconsistent data capture practices, require standardised structured input systems for reliable clinical deployment.

While LLMs advance structured data OCR,<sup>3</sup> they falter at interpreting handwritten annotations and implicit contextual cues, contrasting traditional OCR’s printed-text focus [12]. Our analysis reveals critical spatial reasoning failures: mislinking elevated FIT values without explicit statements, and irreversible error propagation from glyph-level misinterpretations (handwritten “Y” → “N”). These limitations mirror EHR extraction challenges [4], where hybrid systems outperform pure LLM approaches. Though multimodal architectures like *Med-PaLM M* [16] show promise, persistent weaknesses in distinguishing visual semantics (e.g., ✓ vs. ✕) underscore the need for integrated document understanding.

LLM-driven extraction can be deployed in three principal models. *Autonomous systems* operate independently but pose risks of critical errors in closed-loop workflows [17]. *Manual verification* ensures human oversight but undermines automation benefits, reducing efficiency. *Assistive systems* strike a balance between efficiency and safety by flagging discrepancies based on uncertainty scores, allowing clinicians to intervene when necessary.

The hybrid clinician-AI collaboration approach (assistive systems) enhances prior methods [15] by identifying conflicting annotations for rapid resolution, reducing automation bias while preserving human agency [9]. In urgent triage, this balances algorithmic assistance with clinical oversight [14]. However, as Agarwal et al. [1] caution, human-AI collaboration is not always optimal, as cognitive belief updating biases may prevent clinicians from fully leveraging AI assistance.

Collaboration with NHS England’s AIQCoP identified diverse clinical parsing risks (non-standard data entry, ambiguous contradictions, data drift), necessitating assurance-by-design integration. Proactive mitigations—embedding referral pathway logic for missing data flags, multilayered validation, and dynamic input monitoring—codified safeguards through fallback protocols and continuous auditing, shifting from reactive correction to preemptive risk reduction. While OCR fine-tuning addressed known limitations, its issue-specificity leaves vulnerability to unidentified live data challenges. Recommended adaptations combine generalised OCR solutions with semi-structured data advantages: systematic analysis of real-world inputs to preemptively catalog edge cases. This dual approach exemplifies how clinically anchored assurance protocols bridge algorithmic efficiency with healthcare’s human complexities, ensuring AI augmentation aligns with governance while retaining adaptability to emergent real-world variability.

This study evaluates general-purpose LLMs in a zero-shot setting using synthetic CRC 2WW referrals authored by NHS clinicians. While performance on real-world referrals remains untested, a prospective NHS validation is underway. Our current focus is colorectal cancer; future work will assess generalisability across other cancer pathways. Only two LLMs (GPT-4o-Mini, DeepSeek-R1) were evaluated; broader benchmarking with open-source models is planned. Although zero-shot inference avoids domain-specific supervision, incorporating few-shot or fine-tuned baselines may yield deeper insights. RAPTOR includes

<sup>3</sup> <https://github.com/getomni-ai/benchmark>

field-level confidence scoring to guide human oversight, with its real-world impact on triage accuracy and clinician workload to be assessed during deployment.

## 5 Conclusion and Future Work

This study demonstrates GPT-4o-Mini and DeepSeek-R1 have achieved near-human accuracy with 10-fold efficiency gains in colorectal cancer (CRC) referral processing. These models reveal complementary specialisations: GPT-4o-Mini excels in temporal reasoning while DeepSeek-R1 prioritises quantitative symptom data. Critical limitations in nonstandard input interpretation (handwritten annotations, checkboxes) necessitate our hybrid framework integrating real-time uncertainty quantification with clinician oversight—detecting low-confidence extractions (ambiguous fecal immunochemical test values) for prioritised review. In collaboration with the NHS England AI Quality Community of Practice (AIQ-CoP), engagement efforts codified assurance protocols to address 29 identified risks through dynamic monitoring and fallback mechanisms. Future progress requires multimodal architectures combining visual-semantic parsing with domain-adaptive fine-tuning, enabling AI to accelerate cancer pathways without compromising the indispensable role of clinical judgment in high-stakes decisions.

## References

1. Agarwal, N., Moehring, A., Rajpurkar, P., Salz, T.: Combining human expertise with artificial intelligence: Experimental evidence from radiology. Working Paper 31422, National Bureau of Economic Research (July 2023). <https://doi.org/10.3386/w31422>, <http://www.nber.org/papers/w31422>
2. Althari, G., Alsulmi, M.: Exploring transformer-based learning for negation detection in biomedical texts. *IEEE Access* **10**, 83813–83825 (2022)
3. Benary, M.: Leveraging large language models for decision support in personalized oncology. *Jama Network Open* **6**, e2343689 (2023). <https://doi.org/10.1001/jamanetworkopen.2023.43689>
4. Cui, H., Fang, X., Xu, R., Kan, X., Ho, J.C., Yang, C.: Multimodal fusion of ehr in structures and semantics: Integrating clinical records and notes with hypergraph and llm. arXiv preprint arXiv:2403.08818 (2024)
5. Hanauer, D., Mei, Q., Law, J., Khanna, R., Zheng, K.: Supporting information retrieval from electronic health records: a report of university of michigan’s nine-year experience in developing and using the electronic medical record search engine (emerse). *Journal of Biomedical Informatics* **55**, 290–300 (2015). <https://doi.org/10.1016/j.jbi.2015.05.003>
6. Hanna, T.P., King, W.D., Thibodeau, S., Jalink, M., Paulin, G.A., Harvey-Jones, E., O’Sullivan, D.E., Booth, C.M., Sullivan, R., Aggarwal, A.: Mortality due to cancer treatment delay: systematic review and meta-analysis. *bmj* **371** (2020)
7. Jabal, M.S., Warman, P., Zhang, J., Gupta, K., Jain, A., Mazurowski, M., Wiggins, W., Magudia, K., Calabrese, E.: Language models and retrieval augmented generation for automated structured data extraction from diagnostic reports. arXiv preprint arXiv:2409.10576 (2024)

8. K.J., M.: Faecal immunochemical testing (fit) in patients with signs or symptoms of suspected colorectal cancer (crc): a joint guideline from the association of coloproctology of great britain and ireland (acpgbi) and the british society of gastroenterology (bsg). *Gut* **71**, 1939–1962 (2022). <https://doi.org/10.1136/gutjnl-2022-327985>
9. Ou, Y.J., Barnett, A.J., Mitra, A., Schwartz, F.R., Chen, C., Grimm, L.J., Lo, J.Y., Rudin, C.: A user interface to communicate interpretable AI decisions to radiologists. In: Proc. SPIE 12467, Medical Imaging 2023: Imaging Informatics for Healthcare, Research, and Applications. p. 124670C (2023). <https://doi.org/10.1117/12.2654068>
10. Poyah, P., Quraishi, T.: The impact of a new triage and booking system on renal clinic wait times. *Canadian Journal of Kidney Health and Disease* **7** (2020). <https://doi.org/10.1177/2054358120924140>
11. Rani, M., Mishra, B.K., Thakker, D., Babar, M., Jones, W., Din, A.: Biases and trustworthiness challenges with mitigation strategies for large language models in healthcare. In: 2024 International Conference on IT and Industrial Technologies (ICIT). pp. 1–6. IEEE (2024)
12. Smith, R.: An overview of the tesseract ocr engine. In: Ninth international conference on document analysis and recognition (ICDAR 2007). vol. 2, pp. 629–633. IEEE (2007)
13. Suresh, B.: Empowering community clinical triage through innovative data-driven machine learning (2024). <https://doi.org/10.20944/preprints202403.1432.v1>
14. Topol, E.J.: High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine* **25**(1), 44–56 (2019). <https://doi.org/10.1038/s41591-018-0300-7>
15. Tschandl, P., Rinner, C., Apalla, Z., Argenziano, G., Codella, N., Halpern, A., Janda, M., Lallas, A., Longo, C., Malvey, J., Paoli, J., Puig, S., Rosendahl, C., Soyer, H.P., Zalaudek, I., Kittler, H.: Human–computer collaboration for skin cancer recognition. *Nature Medicine* **26**(8), 1229–1234 (2020). <https://doi.org/10.1038/s41591-020-0942-0>
16. Tu, T., Azizi, S., Driess, D., Schaeckermann, M., Amin, M., Chang, P.C., Carroll, A., Lau, C., Tanno, R., Ktena, I., et al.: Towards generalist biomedical ai. *Nejm Ai* **1**(3), AIoa2300138 (2024)
17. Tucker, C.: Algorithmic exclusion. the fragility of algorithms to sparse and missing data. Center on Regulation and Markets at Brookings, working paper (February), <https://www.brookings.edu/wp-content/uploads/2023/02/Algorithmicexclusion-FINAL.pdf> (2023)
18. Wang, Z., Guo, R., Sun, P., Qian, L., Hu, X.: Enhancing diagnostic accuracy and efficiency with gpt-4-generated structured reports: a comprehensive study. *Journal of Medical and Biological Engineering* **44**(1), 144–153 (2024)
19. Yang, X., Nejatian, N., Shin, H., Smith, K., Parisien, C., Compas, C., Martin, C., Flores, M., Zhang, Y., Magoč, T., Harle, C., Lipori, G., Mitchell, D., Hogan, W., Shenkman, L., Bian, J., Wu, Y.: Gatortron: a large language model for clinical natural language processing (2022). <https://doi.org/10.1101/2022.02.27.22271257>
20. Zhang, J., Zhang, Q., Wang, B., Ouyang, L., Wen, Z., Li, Y., Chow, K.H., He, C., Zhang, W.: Ocr hinders rag: Evaluating the cascading impact of ocr on retrieval-augmented generation. arXiv preprint arXiv:2412.02592 (2024)