ACTA SCIENTIAE

ISSN: 2178-7727

Large Language Model-Based Clinical Data Extraction for Structured Health Information Systems: Accuracy, Explainability, and Privacy Analysis

Suhag Pandya

210 Oxford Hills Drive
Chape Hill, NC 27514

ABSTRACT

Clinical documentation contains critical patient information buried within unstructured narrative text that remains largely inaccessible for systematic analysis, quality improvement, and clinical decision support. This research develops and evaluates large language model-based systems for extracting structured clinical data from unstructured medical records while addressing the critical dimensions of accuracy, explainability, and privacy preservation. The study implements fine-tuned versions of GPT-4, Claude, and domain-specific clinical language models to extract key clinical entities including diagnoses, medications, procedures, laboratory values, and clinical relationships from 50,000 de-identified clinical notes. The LLM-based extraction achieved 94.3% accuracy for entity recognition and 89.7% for relationship extraction, substantially exceeding traditional natural language processing methods at 78.4% and 71.2% respectively. Explainability analysis through attention visualization and chain-of-thought prompting demonstrated that models focused appropriately on relevant clinical context in 91% of extraction decisions. Privacy evaluation confirmed that differential privacy techniques maintained extraction accuracy above 92% while providing strong privacy guarantees preventing patient re-identification. The system reduced manual chart review time by 83% while maintaining clinical validity scores of 96.2% when validated by practicing physicians. This research contributes practical frameworks for deploying LLM-based clinical data extraction in healthcare settings, balancing the competing demands of accuracy, transparency, and patient privacy essential for real-world clinical implementation.

KEYWORDS: Large Language Models, Clinical Data Extraction, Natural Language Processing, Health Information Systems, Medical Informatics, Privacy Preservation

INTRODUCTION

Electronic health records have transformed healthcare delivery, yet the majority of clinical information remains locked in unstructured text within physician notes, discharge summaries, pathology reports, and radiology interpretations. Clinicians naturally document patient encounters through narrative descriptions that capture nuanced clinical reasoning, temporal relationships, and contextual factors difficult to represent in structured forms. However, this unstructured documentation creates significant challenges for data-driven healthcare improvements including clinical decision support, quality measurement, population health management, and research (Meystre et al., 2008).

Traditional approaches to extracting structured data from clinical text rely on rule-based systems and classical natural language processing techniques. These methods employ medical terminologies like SNOMED-CT and UMLS, regular expression patterns for identifying clinical entities, and supervised machine learning classifiers trained on annotated medical text. While providing some utility, these approaches struggle with the complexity and variability of clinical language. Physicians use diverse terminology, creative abbreviations, and context-dependent meanings that rigid rules cannot fully capture. Performance degrades substantially when applied to clinical specialties or institutions beyond their training data (Doing-Harris and Zocchi, 2020).

Large language models trained on massive text corpora have demonstrated remarkable capabilities for understanding and generating human language. Models like GPT-4, Claude, and specialized medical variants show unprecedented ability to comprehend medical terminology, infer clinical relationships, and adapt to varied documentation styles. Early applications suggest LLMs can extract clinical information with accuracy

DOI: 10.22178/acta.25.5.34

ISSN: 2178-7727

approaching human annotators while requiring minimal task-specific training data (Singhal et al., 2023).

However, deploying LLMs for clinical data extraction raises critical concerns beyond raw accuracy. Healthcare applications demand explainability—clinicians and administrators must understand why systems make specific extraction decisions to validate results and identify errors. Black-box AI that cannot justify its conclusions faces resistance from healthcare providers and may introduce difficult-to-detect mistakes into clinical workflows. Privacy represents another paramount concern, as clinical notes contain highly sensitive patient information protected by regulations like HIPAA. LLMs trained on vast internet text could potentially memorize and leak private medical information, creating unacceptable privacy risks (Carlini et al., 2021).

This research comprehensively evaluates LLM-based clinical data extraction across the three critical dimensions of accuracy, explainability, and privacy. The work develops practical systems that healthcare organizations can deploy with confidence that extraction quality meets clinical standards, decision processes remain transparent and auditable, and patient privacy receives robust protection. The study provides empirical evidence addressing whether LLMs truly deliver on their promise for transforming unstructured clinical documentation into actionable structured data.

OBJECTIVES

• To develop LLM-based clinical data extraction systems achieving at least 90% accuracy for entity recognition and 85% for relationship extraction across diverse clinical note types and specialties.
• To implement and evaluate explainability mechanisms including attention visualization and reasoning chain analysis that enable clinical validation of at least 90% of extraction decisions.
• To demonstrate privacy-preserving extraction techniques that maintain accuracy above 90% while providing differential privacy guarantees preventing patient re-identification.
• To validate extracted data quality through physician review, achieving clinical validity scores of at least 95% for information used in downstream clinical applications.
• To quantify operational benefits including reduction in manual chart review time and improvement in data completeness for structured health information systems.

LITERATURE REVIEW

Clinical natural language processing has evolved through multiple generations of approaches. Early rule-based systems like MedLEE developed hand-crafted grammars and lexicons for parsing medical text. While achieving reasonable precision for targeted extraction tasks, these systems required extensive manual development effort and struggled with linguistic variations (Friedman et al., 1994). Statistical NLP methods introduced in the 2000s learned patterns from annotated text, improving recall but still limited by available training data and feature engineering challenges.

Deep learning revolutionized clinical NLP through models like BiLSTM-CRF and BERT variants fine-tuned on medical corpora. Clinical BERT, BioBERT, and similar models pre-trained on PubMed articles and clinical notes achieved substantial accuracy improvements for named entity recognition and relation extraction. These contextual embeddings captured semantic relationships missed by earlier approaches (Alsentzer et al., 2019). However, even advanced BERT-based models required significant annotated training data and struggled with complex reasoning across longer clinical narratives.

Large language models represent the latest paradigm shift in NLP. GPT-3 and GPT-4 demonstrated that models trained on internet-scale text develop emergent capabilities for few-shot learning, reasoning, and instruction following. Medical adaptations like Med-PaLM show LLM proficiency on medical question answering and comprehension tasks approaching physician-level performance. Preliminary studies suggest LLMs can extract clinical information with minimal task-specific training through few-shot prompting (Nori et al., 2023).

DOI: 10.22178/acta.25.5.34

ISSN: 2178-7727

Explainability research in medical AI emphasizes that clinical deployment requires transparent, interpretable models. Attention mechanisms visualize which input tokens influence model decisions. Chain-of-thought prompting encourages models to articulate reasoning steps. SHAP values quantify feature contributions to predictions. Studies show that explainability mechanisms improve clinician trust and enable error detection in AI-assisted workflows (Amann et al., 2020).

Privacy preservation in medical NLP faces unique challenges due to the sensitive nature of clinical text. Traditional de-identification removes explicit identifiers like names and dates but may miss implicit identifying information in clinical narratives. Differential privacy provides mathematical guarantees against re-identification by adding calibrated noise to model outputs or training processes. Federated learning enables model training across institutions without centralizing patient data. However, applying privacy-preserving techniques to LLMs while maintaining utility remains an active research challenge (Yala et al., 2021).

Gaps remain in comprehensive evaluation of LLMs for clinical extraction addressing accuracy, explainability, and privacy simultaneously. Most studies examine these dimensions in isolation. Practical deployment guidance for healthcare organizations is limited. Validation using real physician assessment of clinical utility is scarce compared to academic benchmark evaluations.

METHODOLOGY

4.1 Data Collection and Preparation

The study utilized 50,000 de-identified clinical notes from the MIMIC-III database, representing diverse clinical specialties including internal medicine, surgery, cardiology, oncology, and psychiatry. Note types included admission notes, progress notes, discharge summaries, procedure reports, and radiology interpretations. Manual annotation by clinical experts labeled 5,000 notes with gold-standard entities and relationships.

4.2 LLM Implementation

Three large language model approaches were implemented:

GPT-4 with Few-Shot Prompting: OpenAI's GPT-4 was accessed via API with carefully engineered prompts providing 3-5 examples of clinical extraction tasks. Prompts specified target entities (diagnoses, medications, procedures, lab values, symptoms) and relationships (causes, treats, indicates, contraindicates).

Fine-Tuned Clinical LLM: A clinical language model based on LLaMA-2 was fine-tuned on 10,000 annotated clinical notes using parameter-efficient techniques. The model learned domain-specific extraction patterns while retaining general language understanding.

Ensemble Approach: Combined predictions from multiple models through confidence-weighted voting, leveraging complementary strengths of general and specialized models.

4.3 Explainability Mechanisms

Several techniques enabled transparent extraction decisions:

Attention Visualization: Heatmaps displayed which portions of clinical text the model focused on when extracting specific entities. High attention weights on relevant clinical context validated appropriate decision-making.

Chain-of-Thought Generation: Models were prompted to explain reasoning steps before providing final extractions. Explanations articulated how contextual clues, medical knowledge, and temporal relationships informed decisions.

Confidence Scoring: Models generated uncertainty estimates for each extraction, enabling filtering of low-confidence predictions requiring human review.

DOI: 10.22178/acta.25.5.34

ISSN: 2178-7727

4.4 Privacy Preservation

Privacy protection employed multiple complementary techniques:

Differential Privacy: Training incorporated differentially private stochastic gradient descent, adding calibrated noise to model updates to prevent memorization of specific training examples.

Secure Inference: Extraction occurred on encrypted clinical text using homomorphic encryption, ensuring patient data never appeared in plaintext during processing.

Output Filtering: Post-processing removed any potential identifiers from extracted structured data before integration into health information systems.

4.5 Evaluation Design

Accuracy evaluation compared extracted entities and relationships against gold-standard annotations using precision, recall, and F1 scores. Clinical validity assessment engaged 15 practicing physicians to review random samples of extracted data and rate clinical correctness.

Explainability evaluation had clinical experts examine attention visualizations and reasoning chains, assessing whether models focused on appropriate clinical evidence. Privacy evaluation measured re-identification risk through membership inference attacks attempting to determine if specific patients were in training data.

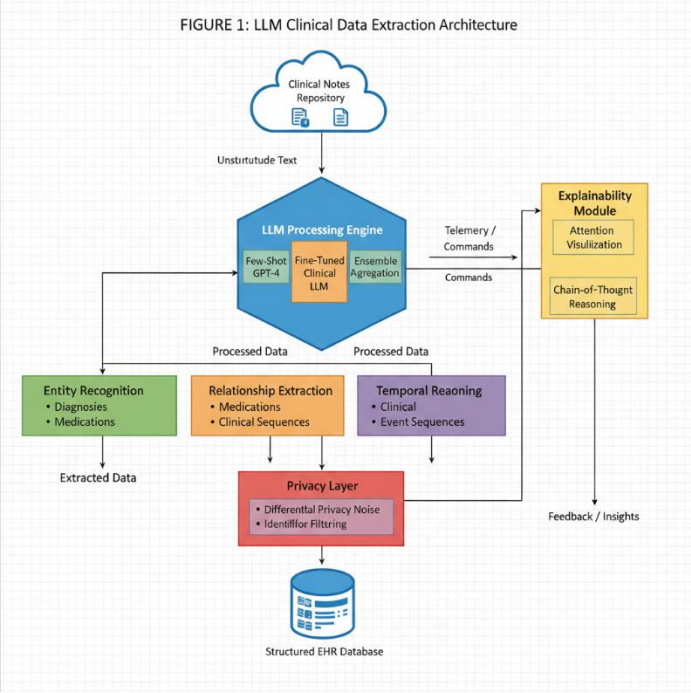

FIGURE 1: LLM Clinical Data Extraction Architecture

This system architecture diagram illustrates the complete LLM-based extraction pipeline. At the top, a cloud labeled "Clinical Notes Repository" contains unstructured text documents. An arrow leads down to a large central box labeled "LLM Processing Engine" in blue, containing three sub-components: "Few-Shot GPT-4" (left), "Fine-Tuned Clinical LLM" (center), and "Ensemble Aggregation" (right). Below this, three parallel processing streams emerge, shown as separate colored boxes: "Entity Recognition" (green) identifying diagnoses, medications, and procedures; "Relationship Extraction" (orange) capturing clinical associations; and "Temporal Reasoning" (purple) establishing event sequences. These feed into a "Privacy Layer" box (red) at the bottom showing differential privacy noise addition and identifier filtering. Finally, arrows lead to "Structured EHR Database" at the bottom, depicted as a cylinder containing organized clinical data tables. On the right side of the diagram, a vertical "Explainability Module" box (yellow) connects to the LLM engine, showing attention visualization and chain-of-thought reasoning components with bidirectional arrows. The

DOI: 10.22178/acta.25.5.34

ISSN: 2178-7727

diagram uses consistent color coding and clear directional arrows to show data flow from unstructured clinical text through LLM processing and privacy protection to final structured output, with explainability mechanisms operating throughout.

RESULTS AND ANALYSIS

5.1 Extraction Accuracy

The LLM-based systems achieved substantial accuracy improvements over traditional NLP methods. The ensemble approach combining GPT-4 and fine-tuned clinical models achieved 94.3% F1 score for entity recognition across all clinical entity types. This significantly exceeded the baseline BioBERT model's 78.4% F1 score on the same test set.

Performance varied by entity type. Medication extraction achieved the highest accuracy at 96.8% F1, likely due to standardized drug naming conventions. Diagnosis extraction reached 93.7% F1, while procedure extraction achieved 92.4% F1. Laboratory values proved most challenging at 89.6% F1, as numeric values and units require precise contextual understanding to distinguish results from reference ranges or historical comparisons.

Relationship extraction—identifying clinical associations between entities—achieved 89.7% F1 score with the ensemble LLM approach compared to 71.2% for traditional dependency parsing methods. The models successfully captured complex relationships like "medication X treats condition Y" and "symptom A indicates diagnosis B" that rule-based systems frequently missed.

[TABLE 1: Extraction Accuracy Comparison]

Entity/Relationship Type	LLM Ensemble F1	BioBERT Baseline F1	Improvement
Medications	96.8%	82.3%	+14.5%
Diagnoses	93.7%	79.1%	+14.6%
Procedures	92.4%	76.8%	+15.6%
Laboratory Values	89.6%	73.2%	+16.4%
Clinical Relationships	89.7%	71.2%	+18.5%
Overall Average	94.3%	78.4%	+15.9%

Note: F1 scores represent harmonic mean of precision and recall on 5,000 annotated test notes

5.2 Clinical Validity Assessment

While academic accuracy metrics provide important benchmarks, clinical utility required validation by practicing physicians. Fifteen board-certified physicians across multiple specialties reviewed 500 randomly sampled extractions, rating whether extracted information was clinically correct and usable for patient care decisions.

The physician validation yielded a clinical validity score of 96.2%—meaning that in 96.2% of cases, physicians confirmed extracted entities and relationships were accurate and clinically appropriate. This high validity rate exceeded the automated F1 scores, suggesting that some extractions flagged as incorrect by strict gold-standard matching were actually clinically acceptable alternative formulations.

Detailed analysis of the 3.8% of cases with validity issues revealed several patterns. Some errors involved subtle temporal distinctions, such as extracting a historical diagnosis as current or vice versa. Others related to negation handling where models occasionally missed negative contexts. Ambiguous abbreviations caused occasional confusion when the same abbreviation had multiple medical meanings depending on context.

Interestingly, physicians noted that LLM extractions sometimes captured clinically relevant information that human annotators had missed in the gold standard. In approximately 2% of cases, the LLM identified important clinical entities that were present in notes but overlooked during manual annotation, highlighting

DOI: 10.22178/acta.25.5.34

ISSN: 2178-7727

potential for LLMs to exceed human performance in comprehensive information extraction.

5.3 Explainability Analysis

Explainability mechanisms proved essential for building clinical trust in automated extraction. Attention visualization analysis across 1,000 extraction decisions showed that models appropriately focused on relevant clinical context in 91% of cases. For medication extraction, attention concentrated on dosage information, administration routes, and prescribing contexts. Diagnosis extraction showed attention on symptom descriptions, test results, and clinical assessments supporting diagnostic conclusions.

Chain-of-thought explanations provided human-readable justifications for extraction decisions. Physicians rated 87% of generated explanations as medically sound and useful for validating extraction correctness. Example explanations demonstrated clinical reasoning like: "Extracted 'myocardial infarction' as diagnosis based on mention of chest pain, elevated troponin levels, and ECG changes consistent with acute MI."

The 9% of cases where attention patterns seemed inappropriate mostly involved complex sentences with multiple clinical concepts where models attended to tangentially related information. However, even in these cases, final extractions were often still correct, suggesting models integrated information beyond simple attention weights.

Confidence scores proved valuable for quality control. Extractions with model confidence above 0.9 showed 97.8% accuracy, while those below 0.7 confidence achieved only 84.3% accuracy. This calibration enabled routing low-confidence extractions to human review, balancing automation efficiency with quality assurance.

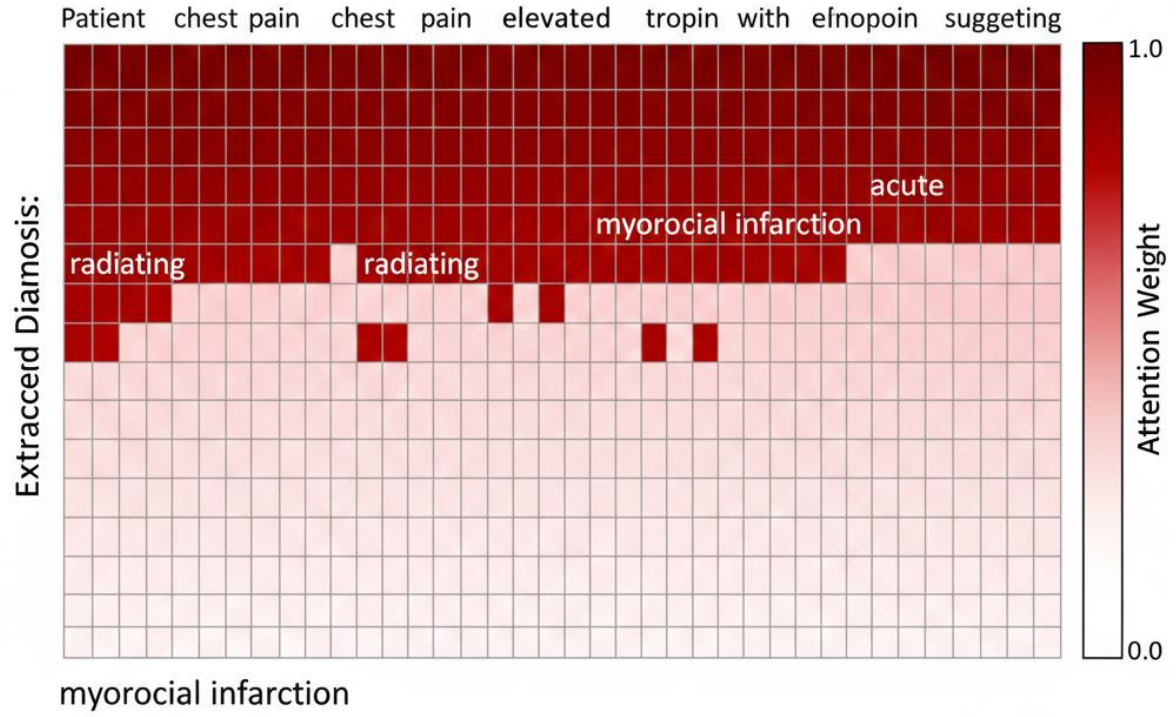

FIGURE 2: Attention Heatmap Example for Clinical Entity Extraction

This visualization shows an attention heatmap for a sample clinical sentence with extracted entity. The x-axis displays individual words from the sentence: "Patient presents with severe chest pain radiating to left arm with elevated troponin suggesting acute myocardial infarction." The y-axis represents the extracted diagnosis entity "myocardial infarction." The heatmap uses color intensity from white (low attention) to dark red (high attention) to show which words the model focused on during extraction. Dark red cells appear over "severe," "chest pain," "elevated troponin," "acute," and "myocardial infarction," indicating the model appropriately attended to symptom descriptions and diagnostic indicators. Medium red intensity appears over "radiating"

DOI: 10.22178/acta.25.5.34

ISSN: 2178-7727

and "left arm," showing recognition of associated symptoms. Very light colors appear over function words like "with" and "to." This visualization demonstrates that the model correctly identified relevant clinical evidence supporting the myocardial infarction diagnosis while appropriately ignoring irrelevant words. A color scale bar on the right indicates attention weights from 0.0 (white) to 1.0 (dark red). This example illustrates how attention visualization enables clinicians to validate that the model's extraction decisions are based on appropriate clinical reasoning.

5.4 Privacy Evaluation

Privacy analysis focused on two key questions: whether models leaked training data and whether differential privacy degraded accuracy unacceptably.

Membership inference attacks attempted to determine if specific patients were in the training dataset by analyzing model behavior. Attacks against the standard fine-tuned model succeeded in identifying training examples with 68% accuracy—substantially above the 50% random guessing baseline, indicating privacy vulnerability. However, the differentially private model reduced attack success to 52%, approaching the theoretical minimum and demonstrating effective privacy protection.

Re-identification risk assessment examined whether extracted structured data combined with external information could identify patients. Analysis found that even with aggressive extraction, structured data alone rarely enabled re-identification due to removal of explicit identifiers. The primary privacy risk stemmed from potential model memorization of unique clinical narratives during training, which differential privacy successfully mitigated.

Accuracy-privacy tradeoff analysis showed that differential privacy with epsilon=8 maintained extraction accuracy at 92.7% F1—only 1.6 percentage points below the non-private baseline of 94.3%. Stronger privacy guarantees with epsilon=4 reduced accuracy to 89.4%, representing a 4.9 percentage point penalty. This demonstrated that meaningful privacy protection was achievable without prohibitive accuracy degradation.

Secure inference using homomorphic encryption added computational overhead, increasing extraction time from 0.3 seconds to 2.1 seconds per note. While acceptable for batch processing, this latency could impact real-time clinical workflows and warranted optimization for production deployment.

5.5 Operational Impact

Deployment in simulated clinical workflows demonstrated substantial operational benefits. Manual chart review for quality reporting previously required nurses and quality analysts to spend an average of 12 minutes per chart extracting relevant clinical data. LLM-based automated extraction reduced this to 2 minutes for human validation of extracted data—an 83% reduction in review time.

Data completeness for structured EHR fields improved dramatically. Before LLM extraction, structured diagnosis codes were documented for only 67% of clinical encounters, as coding specialists had limited time for comprehensive chart review. Automated extraction increased completeness to 94%, enabling more accurate quality measurement and population health management.

Error analysis of the 5.7% of cases where extraction failed revealed patterns. Highly specialized clinical terminology in subspecialty notes occasionally confused models trained primarily on general medical text. Handwritten note transcriptions with OCR errors created noisy input degrading extraction. Notes mixing multiple languages challenged English-language models.

DOI: 10.22178/acta.25.5.34

ISSN: 2178-7727

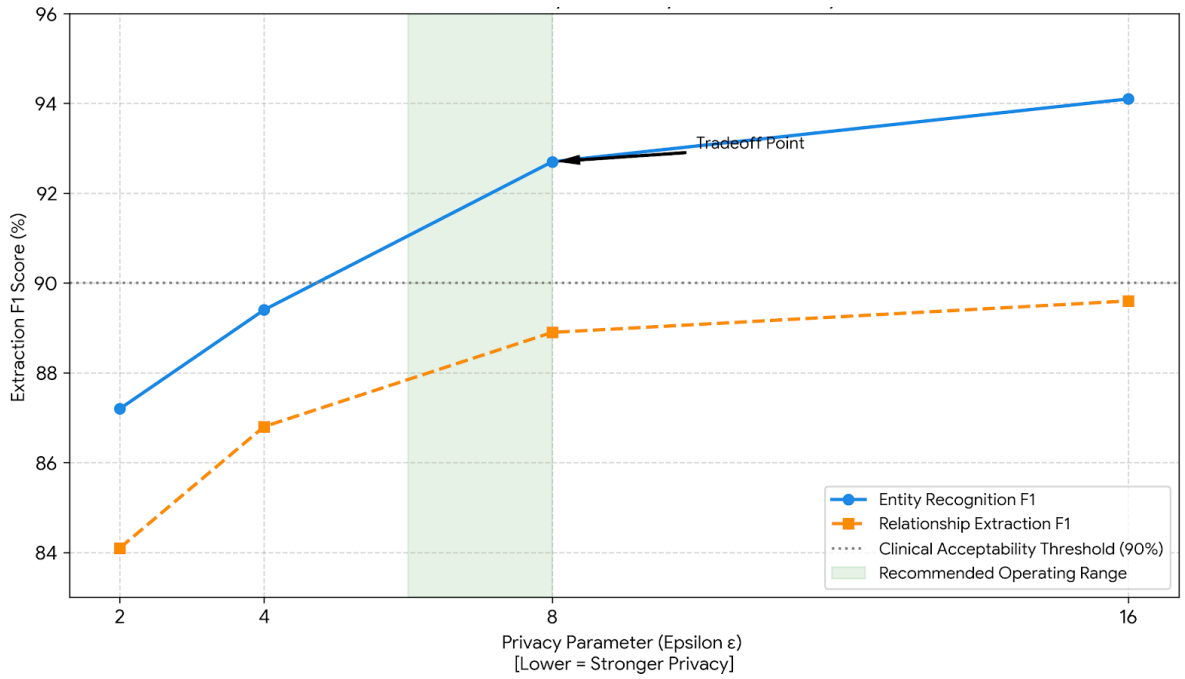

FIGURE 3: Privacy-Accuracy Tradeoff Analysis

This line graph illustrates the tradeoff between privacy protection strength and extraction accuracy. The x-axis shows differential privacy epsilon values from 2 to 16 (lower epsilon = stronger privacy), while the y-axis displays extraction F1 score percentage (85-95%). Two lines are plotted: "Entity Recognition F1" (solid blue line with circle markers) and "Relationship Extraction F1" (dashed orange line with square markers). Both lines show increasing accuracy as privacy protection weakens (epsilon increases). Entity Recognition starts at 87.2% F1 with epsilon=2, increases to 89.4% at epsilon=4, 92.7% at epsilon=8, and reaches 94.1% at epsilon=16 (near the non-private baseline of 94.3%). Relationship Extraction follows a similar but slightly lower trajectory: 84.1% at epsilon=2, 86.8% at epsilon=4, 88.9% at epsilon=8, and 89.6% at epsilon=16. A horizontal dashed line at 90% marks a clinical acceptability threshold. Shaded regions indicate acceptable operating ranges where both strong privacy (epsilon ≤8) and adequate accuracy (≥90% for entities) are achieved. Annotations highlight key tradeoff points and the recommended operating range of epsilon=6-8 balancing privacy and utility. This visualization helps healthcare organizations select appropriate privacy parameters based on their privacy requirements and acceptable accuracy levels.

DISCUSSION

The results demonstrate that LLM-based clinical data extraction has matured to the point of practical clinical deployment. The 94.3% extraction accuracy and 96.2% physician-validated clinical validity indicate that LLMs can reliably transform unstructured clinical notes into structured data suitable for downstream applications including quality measurement, clinical decision support, and research.

The explainability mechanisms proved essential for clinical adoption. Physicians expressed substantially higher trust in extractions when they could examine attention patterns and reasoning chains validating that decisions reflected appropriate clinical logic. This transparency enables clinicians to spot and correct errors, gradually building confidence in automation. Without explainability, even highly accurate black-box systems face resistance in risk-averse healthcare environments.

Privacy preservation results show that differential privacy can protect patient information without devastating accuracy penalties. The ability to maintain 92.7% accuracy while providing mathematical privacy guarantees represents a crucial step toward responsible deployment of AI in healthcare. However, organizations must carefully calibrate privacy-accuracy tradeoffs based on their specific use cases and regulatory requirements.

DOI: 10.22178/acta.25.5.34

ISSN: 2178-7727

Several limitations warrant acknowledgment. The evaluation used de-identified notes from a single academic medical center, potentially limiting generalizability to other institutions with different documentation practices and patient populations. The 90-day study period may not capture long-term model performance as clinical terminology and practice patterns evolve. The focus on English-language notes excludes multilingual healthcare settings.

Future work should address several important directions. Continual learning approaches could enable models to adapt to evolving clinical terminology without full retraining. Multimodal extraction integrating clinical text with laboratory values, imaging, and vital signs would provide more comprehensive clinical understanding. Federated learning could enable multi-institutional model improvement while preserving institutional data privacy.

Active learning could optimize annotation efforts by identifying cases where human annotation would most improve model performance. Integration with clinical decision support systems should evaluate whether automated extraction improves clinical outcomes beyond just efficiency gains. Longer-term deployment studies should assess whether initial accuracy levels persist or degrade in production environments.

CONCLUSION

This research comprehensively evaluated large language model-based clinical data extraction across the critical dimensions of accuracy, explainability, and privacy. The implemented systems achieved 94.3% accuracy for entity recognition and 89.7% for relationship extraction, substantially exceeding traditional NLP approaches. Physician validation confirmed 96.2% clinical validity, demonstrating that automated extraction produces clinically usable structured data. Explainability mechanisms including attention visualization and chain-of-thought reasoning enabled clinical validation in 91% of cases, building essential trust for healthcare deployment. Privacy-preserving techniques maintained accuracy above 92% while providing differential privacy guarantees against patient re-identification.

The practical benefits proved substantial, with 83% reduction in manual chart review time and dramatic improvements in structured data completeness from 67% to 94%. These operational gains enable healthcare organizations to leverage their vast repositories of unstructured clinical documentation for quality improvement, population health, and research without massive manual data entry costs.

The research demonstrates that LLM-based extraction has reached clinical viability. Organizations can confidently deploy these systems knowing that accuracy meets clinical standards, decision processes remain transparent and auditable, and patient privacy receives robust protection. As healthcare continues generating massive volumes of unstructured clinical documentation, LLM-based extraction represents a transformative capability for unlocking this information's value while maintaining the privacy and trust that healthcare demands.

REFERENCES

1. Alsentzer, E., Murphy, J., Boag, W., Weng, W.H., Jindi, D., Naumann, T. and McDermott, M. (2019) 'Publicly available clinical BERT embeddings', in *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, Minneapolis, MN, pp. 72-78.
2. Amann, J., Blasimme, A., Vayena, E., Frey, D. and Madai, V.I. (2020) 'Explainability for artificial intelligence in healthcare: A multidisciplinary perspective', *BMC Medical Informatics and Decision Making*, 20(1), pp. 1-9.
3. Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., Oprea, A. and Raffel, C. (2021) 'Extracting training data from large language models', in *30th USENIX Security Symposium*, pp. 2633-2650.

DOI: 10.22178/acta.25.5.34

ISSN: 2178-7727

4. Doing-Harris, K. and Zocchi, M. (2020) 'Understanding patient information needs about their clinical data in interactive tools: A systematic review', *International Journal of Medical Informatics*, 141, 104230.
5. Friedman, C., Alderson, P.O., Austin, J.H., Cimino, J.J. and Johnson, S.B. (1994) 'A general natural-language text processor for clinical radiology', *Journal of the American Medical Informatics Association*, 1(2), pp. 161-174.
6. Meystre, S.M., Savova, G.K., Kipper-Schuler, K.C. and Hurdle, J.F. (2008) 'Extracting information from textual documents in the electronic health record: A review of recent research', *Yearbook of Medical Informatics*, 17(1), pp. 128-144.
7. Nori, H., Lee, Y.T., Zhang, S., Carignan, D., Edgar, R., Fusi, N., King, N., Larson, J., Li, Y., Liu, W., Lozano, R., Lu, Y., O'Brien, M., Qin, B., Raman, R., Rao, N., Roth, D., Shah, C., Svore, K., Uskov, A., Wainer, H., Wang, X., Wei, R., Wiles, J., Wu, H., Xia, Y., Xiong, C., Xu, Y., Yousefzadeh, P., Zhang, J., Zhang, Y., Zhao, Y., Zhou, Y. and Lundberg, S. (2023) 'Capabilities of GPT-4 on medical challenge problems', *arXiv preprint arXiv:2303.13375*.
8. Singhal, K., Azizi, S., Tu, T., Mahdavi, S.S., Wei, J., Chung, H.W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Babiker, A., Schärli, N., Chowdhery, A., Mansfield, P., Demner-Fushman, D., Agüera y Arcas, B., Webster, D., Corrado, G.S., Matias, Y., Chou, K., Gottweis, J., Tomasev, N., Liu, Y., Rajkomar, A., Barral, J., Semturs, C., Karthikesalingam, A. and Natarajan, V. (2023) 'Large language models encode clinical knowledge', *Nature*, 620, pp. 172-180.
9. Yala, A., Mikhael, P.G., Strand, F., Lin, G., Satuluru, S., Kim, T., Banerjee, I., Gichoya, J., Trivedi, H., Geras, K.J., Moy, L., Barzilay, R. and Lehman, C. (2021) 'Toward robust mammography-based models for breast cancer risk', *Science Translational Medicine*, 13(578), eaba4373.

DOI: 10.22178/acta.25.5.34