MedDecXtract: A Clinician-Support System for Extracting, Visualizing, and Annotating Medical Decisions in Clinical Narratives

Mohamed Elgaar[†] **Hadi Amiri**[†] **Mitra Mohtarami**[‡] **Leo Anthony Celi**[‡]
[†] University of Massachusetts Lowell [‡] Massachusetts Institute of Technology
{melgaar,hadi}@cs.uml.edu mitra@csail.mit.edu lceli@mit.edu

Abstract

Clinical notes contain crucial information about medical decisions such as treatments, diagnoses and follow-ups. However, these decisions are embedded within unstructured text, making it challenging to computationally analyze clinical decision-making patterns or support clinical workflows. We present MedDecXtract: an open-source and interactive system that automatically extracts and visualizes medical decisions from clinical text. The system implements a RoBERTa-based model for identifying ten categories of medical decisions (e.g., diagnosis, treatment, follow-up) according to the Decision Identification and Classification Taxonomy for Use in Medicine (DICTUM), and provides an intuitive interface for exploration, visualization, and annotation. MedDecXtract and its source code can be accessed at https://mohdelgaar-clinical-decisions.hf.space. A video demo is available at https://youtu.be/19j6-XtIE_s.

1 Introduction

Understanding and analyzing medical decisions is crucial for improving healthcare delivery, from supporting clinical encounters to identifying system-wide patterns in care delivery. While structured data in electronic health records (EHRs) captures some clinical decisions through billing codes and order entries, the rich context and reasoning behind these decisions is primarily documented in unstructured clinical notes. These narratives contain crucial details about diagnostic hypotheses, treatment rationales, and care planning that could inform both direct patient care and healthcare policies.

Previous work has focused primarily on extracting discrete medical entities such as diagnoses, medications, and procedures (Nye et al., 2018; Lehman et al., 2019; Patel et al., 2018). However, less attention has been given to capturing the higher-level decision-making processes that meaningfully link these entities. Understanding these decisions is essential for analyzing clinical reasoning, identifying variations in care, and advancing research on medical decision-making.

To address this gap, we present **MedDecXtract**, whose primary novelty lies in its integrated system that offers a workflow for medical decision detection and extraction from clinical narratives. MedDecXtract combines: 1) automated extraction and classification of medical decisions based on the Decision Identification and Classification Taxonomy for Use in Medicine (DICTUM) framework (Ofstad et al., 2016); 2) temporal visualization of decision patterns across patient narratives; and 3) an interactive annotation interface. The extraction component is based on a fine-tuned RoBERTa model (Liu et al., 2019). As demonstrated in Section 5, specialized fine-tuned models for token classification show significantly better performance on precise span extraction for this task compared to instruction-following large language models. Thus, while LLMs represent a promising future direction, fine-tuned token classification models are currently more suitable for the task. The system uses the MedDec dataset (Elgaar et al., 2024) for training the extraction model. A key contribution is the annotation interface, which is designed to enable data annotation.

2 Related Work

Recent advances in clinical natural language processing have made significant progress in analysis of medical text (Tran et al., 2024; Nori et al., 2023; Thirunavukarasu et al., 2023). However, existing works often focus on entity (Patel et al., 2018) or relation (Nye et al., 2018) extraction, rather than higher-level decision analysis. While these tasks are important, they don't capture the complex reasoning processes documented in clinical narratives.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), pages 481–489

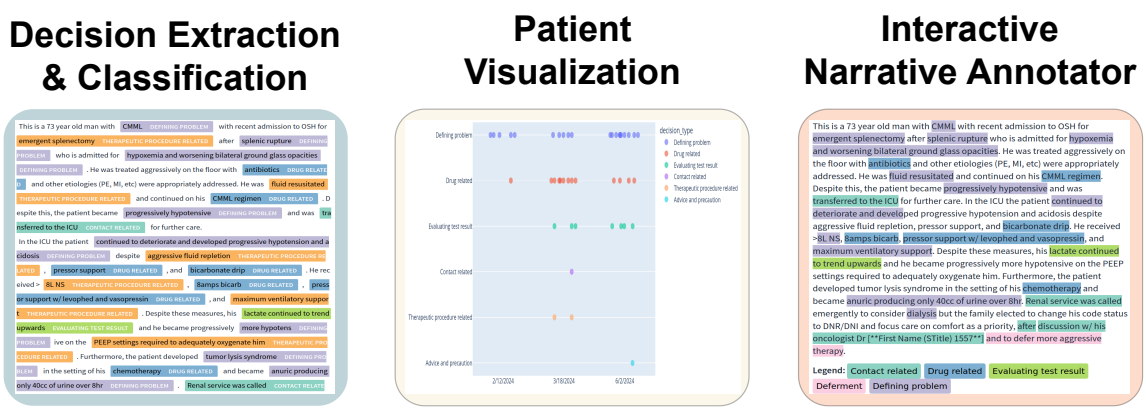

Figure 1: Overview of MedDecXtract functionalities: 1) **Decision Extraction and Classification**: Highlights key medical decisions using color-coded labels for different decision categories. 2) **Patient Visualization**: Aggregates multiple clinical notes into a timeline to visualize decision sequences over time. 3) **Interactive Narrative Annotator**: Allows manual labeling of medical decisions with support for pseudo-annotations to expedite the process.

Clinical text summarization has emerged as an important area of research to address information overload in healthcare settings (Pivovarov and Elhadad, 2015; Wang et al., 2021; Keszthelyi et al., 2023). Several approaches have been developed for summarizing clinical information, including extractive methods (Alsentzer and Kim, 2018; Liang et al., 2019) and problem-oriented summarization (Gao et al., 2022; Liang et al., 2021). Systems like HARVEST (Hirsch et al., 2014) have demonstrated the value of longitudinal patient record summarization with temporal visualization, while others have focused on query-focused summarization for specific clinical tasks (McInerney et al., 2020).

Interactive tools for clinical data exploration and visualization have also been developed, such as PatientExploreR (Glicksberg et al., 2019) for dynamic visualization of patient clinical history, ClinicalPath (Lima et al., 2022) for improving evaluation of EHRs in clinical decision-making, and CERC (Lee and Uppal, 2020) for interactive content extraction and construction. These systems highlight the importance of user-friendly interfaces for clinical data analysis, though they primarily focus on structured data or general text processing rather than specific medical decision extraction.

The conceptual foundation for clinical information summarization has been established through frameworks that emphasize the importance of problem-oriented views and temporal organization (Feblowitz et al., 2011; Adams et al., 2021). Recent work has also explored unified documentation and information retrieval systems (Murray et al., 2021), demonstrating the value of integrated approaches to clinical information management.

The Decision Identification and Classification Taxonomy for Use in Medicine (DICTUM) (Ofstad et al., 2016) provides a structured framework for categorizing clinical decisions. These categories, detailed in Table 1, cover a range of decision types from concrete actions like ordering tests (Gathering info) and prescribing medications (Drug related) to cognitive processes like formulating diagnoses (Defining problem) and setting care goals (Treatment goal).

3 System Architecture

MedDecXtract fine-tunes the transformer model RoBERTa (Liu et al., 2019), using token classification, to extract and classify decision spans into ten DICTUM categories. The model assigns IOB (Inside, Outside, Beginning) tags to each token to identify decision spans. The system processes clinical narratives using the MedDec dataset (Elgaar et al., 2024), sourced from the MIMIC-III clinical database (Pollard and Johnson III, 2016), which provides 451 annotated discharge summaries containing 1.4M tokens and 56,759 annotated medical decisions.

Medical Decision Extraction and Classification enables users to input a clinical note to receive highlighted medical decisions, categorized into predefined types according to DICTUM. To handle long clinical documents that exceed the model's input length limit, we segment the text into non-overlapping chunks. A post-processing step then merges fragmented decision spans predicted across chunk boundaries. Specifically, if two adjacent or

Table 1: Medical Decision categories in MedDec (Elgaar et al., 2024)

Category	Description
Contact related	Admit, discharge
Gathering info	Ordering test, consulting
Defining problem	Diagnosis, prognosis
Treatment goal	Quant./Qual. Goal
Drug related	Start, stop, alter
Therapeutic procedure	Start, stop, alter
Evaluating test	Positive, negative
Deferment	Transfer, wait
Advice/precaution	Advice or precaution
Legal/insurance	Sick leave, refund

overlapping text segments are predicted with the same decision category, they are merged into a single span. **Interactive Visualization**: The extracted decisions are presented through an interactive interface that enables temporal analysis and exploration. Users can track decision patterns across multiple clinical notes, filter by decision types, and generate structured summaries. **Annotation Interface**: To support ongoing improvement of decision extraction models, the system includes an annotation interface that combines automatic pre-annotation with efficient tools for expert refinement.

3.1 Model Design

MedDecXtract employs the span extraction and classification architecture introduced in Elgaar et al. (2024) for clinical decision extraction. Key innovations include: a sliding window approach for handling long documents while maintaining context and segment-level data augmentation. MedDecXtract additionally implements post-processing to merge overlapping and fragmented decision spans.

MedDecXtract is deployed using Gradio (Abid et al., 2019) to provide an interactive web interface, and is hosted on Hugging Face Spaces (Face, 2024), enabling real-time interaction and visualization. The system is designed to be lightweight; the fine-tuned RoBERTa model requires low computational resources compared to larger LLMs. Average processing time is 3.6 seconds on the hosted platform, though this varies with note length. The system is open-source, and the code is available alongside the demo on Hugging Face Spaces.

The interface is organized into three main tabs, as shown in Figure 2, corresponding to the core functionalities: Decision Extraction & Classification, Patient Visualization, and Interactive Narrative Annotator.

4 Features and Functionality

MedDecXtract implements three primary modules: (1) automated medical decision extraction and classification, (2) temporal visualization and analysis of patient histories, and (3) an interactive annotation interface for dataset creation and validation. Each module is designed to address specific challenges in clinical decision analysis.

4.1 Decision Extraction and Classification

The core extraction module employs a token classification approach using a fine-tuned RoBERTa model. The system processes clinical narratives through the following pipeline:

First, documents are tokenized and chunked into overlapping segments to handle length constraints while preserving context. Second, the model identifies decision spans and classes using token-level classification. Third, a rule-based system merges overlapping spans and resolves boundary conflicts.

The output is presented with color-coded highlighting corresponding to different decision categories, enabling rapid visual analysis of decision patterns within the text.

4.2 Temporal Analysis and Visualization

The temporal analysis module enables longitudinal study of clinical decision-making, and summarizes the decisions that have been made for a patient.

The system accepts multiple clinical notes in order to extract the decision sequences for a patient. Decisions are visualized on a temporal axis using Plotly (Inc., 2015), with customizable filters for decision categories (single or multiple selection), date ranges with flexible formatting, and demographic and clinical factors. The system also generates a structured summary, grouped by dates and decision categories. An example summary of decisions for a patient is shown in Appendix A.

4.3 Interactive Annotation Interface

The annotation module facilitates the creation of expert-labeled data through an easy-to-use web interface. The interface provides comprehensive keyboard shortcuts for efficient annotation:

The interface provides category assignment keys for different decision types: ‘c’ for contact related decisions, ‘g’ for gathering information decisions, ‘p’ for defining problem decisions, ‘t’ for treatment goal decisions, ‘d’ for drug related decisions, ‘p’

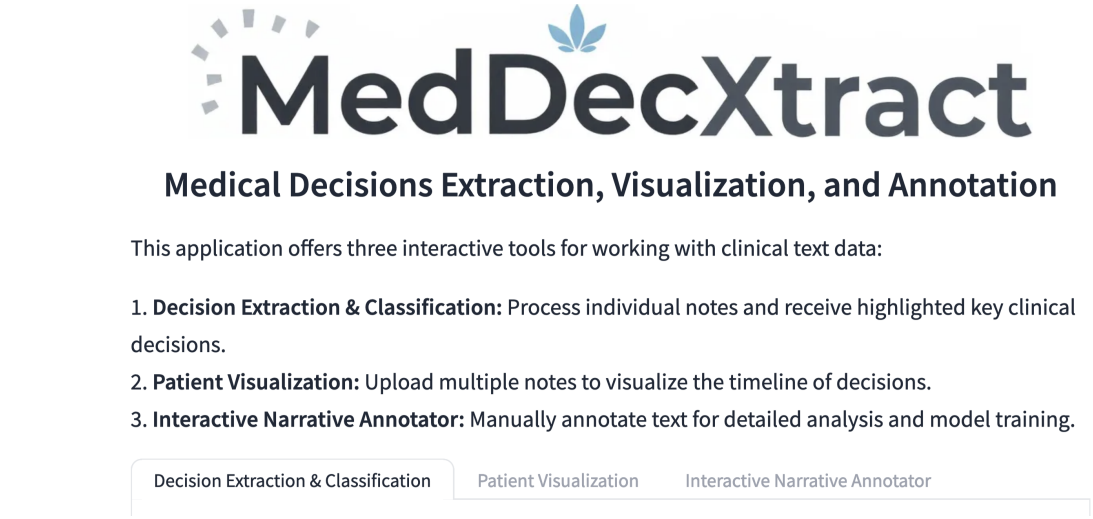

Figure 2: The header interface of MedDecXtract showing the three main tabs corresponding to the core functionalities: 1) 'Decision Extraction & Classification' for processing individual notes, 2) 'Patient Visualization' for analyzing decision sequences across multiple notes over time, and 3) 'Interactive Narrative Annotator' for manual annotation and refinement of model predictions.

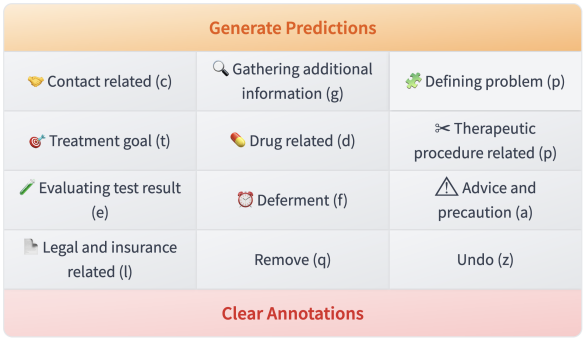

Figure 3: The annotation toolbar available in the 'Interactive Narrative Annotator' tab. It provides buttons for each of the ten DICTUM medical decision categories, along with 'Remove' (q) and 'Undo' (z) functions. Each category button displays a distinct icon and corresponds to a keyboard shortcut (shown in parentheses) for efficient annotation.

for therapeutic procedure decisions, 'e' for evaluating test result decisions, 'f' for deferment decisions, 'a' for advice and precaution decisions, and 'l' for legal and insurance related decisions. Control keys include 'q' to remove annotation from selected text and 'z' to undo the last annotation action. The user simply highlights the text and press the corresponding key to annotate the text (or remove annotation).

The system provides text selection with automatic span boundary detection, real-time visual feedback with category-specific highlighting, and undo/redo functionality for error correction. Each category is visually distinguished using a unique color scheme: Contact related (green), Gathering information (yellow), Defining problem (light purple), Treatment goal (red), Drug related (blue), Therapeutic procedure (orange), Evaluating test (light green), Deferment (pink), Advice/precaution (gray), and Legal/insurance (purple).

The annotation interface provides an intuitive toolbar (Figure 3) with distinct icons and keyboard shortcuts for each decision category. The toolbar is designed for efficient annotation through both mouse clicks and keyboard shortcuts, with additional tools for removing annotations (q) and undoing actions (z).

Figure 4 illustrates the complete annotation workflow supported by the interface, from raw text input through model-assisted pre-annotation to final human refinement and structured output export. This process enables efficient creation of high-quality training data while maintaining expert oversight.

4.4 System Documentation and Web Interface

The system is documented through the web interface, which provides comprehensive guidance across the three main tabs:

Each component includes contextual help text explaining its functionality and usage. The interface employs a modern, responsive design that adapts to different screen sizes and provides immediate visual feedback for user actions. All features are accessible through both mouse interaction and keyboard shortcuts, with tooltips providing additional guidance for complex operations.

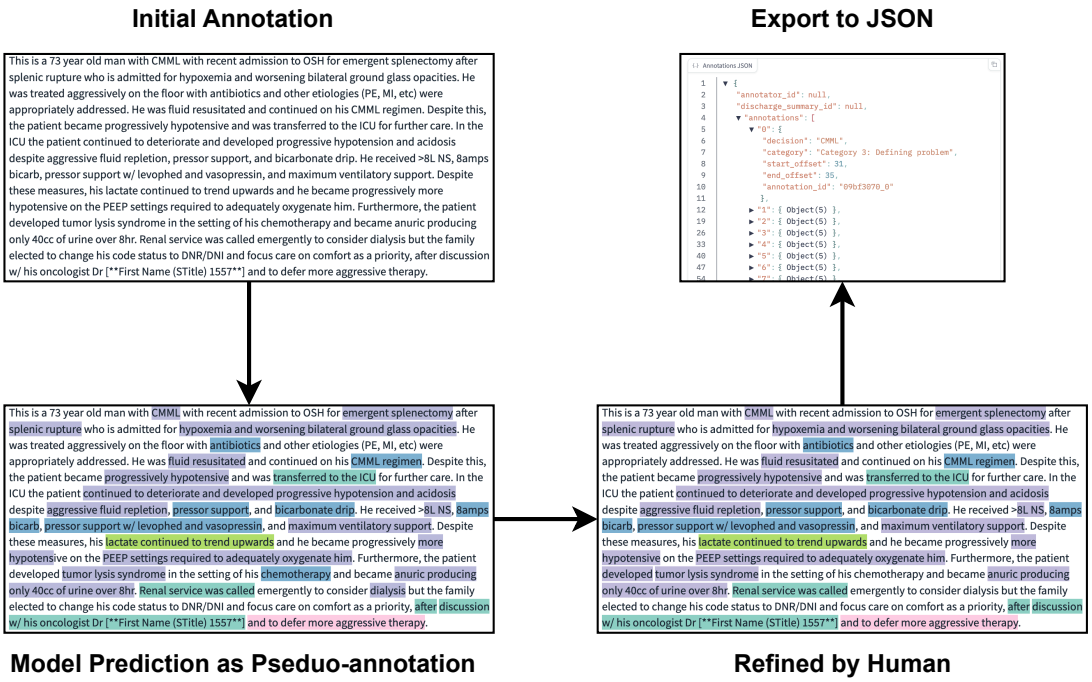

Figure 4: The annotation workflow demonstrating the progression from initial text to final annotations: (1) Initial text input, (2) Model-generated pseudo-annotations to assist the annotator, (3) Human refinement of annotations, and (4) Export of structured annotations in JSON format for further analysis or model training.

Model	Token Level (Accuracy)	Span Level (F1)
ELECTRA	78.2	34.7
BioClinicalBERT	77.8	34.5
RoBERTa	**79.9**	**34.8**
DeBERTa v3	77.4	31.9
ALBERT v2	74.6	27.8
BINDER	71.2	30.3
PIQN	69.5	28.9
DyLex	67.7	27.8
Instance-based	66.2	27.0
Llama-3.1-8B (zero-shot)	-	3.8
Llama-3.1-8B (one-shot)	-	4.8

Table 2: Token classification accuracy and span detection F1 score (exact match) of different models on MedDec. LLM results are for span extraction only and do not provide token-level accuracy.

The system also includes example clinical notes to demonstrate different decision types and annotation patterns.

5 Experimental Results

5.1 Dataset and Model Evaluation

We evaluate MedDecXtract's core extraction model (fine-tuned RoBERTa, Section 3.1) on the MedDec dataset (Elgaar et al., 2024), using its standard test split (10% of patients). MedDec is a large-scale dataset of 451 discharge summaries annotated with 56,759 medical decisions according to DICTUM, created using detailed annotation guidelines. The dataset curators reported substantial inter-annotator agreement (Cohen's Kappa $= 0.74$), ensuring data quality (Elgaar et al., 2024). Our primary evaluation metrics are token-level classification accuracy (based on IOB tags) and span-level F1 score (exact match) for the identified decision spans.

We fine-tuned the RoBERTa model using the following hyperparameters: a learning rate of 4e-5, a batch size of 8, and the AdamW optimizer. The number of training epochs was determined by monitoring performance on the validation set and selecting the best-performing checkpoint. We used a maximum sequence length of 512 tokens.

We compare our model against several strong baselines, including other fine-tuned transformer models: ELECTRA, BioClinicalBERT, DeBERTa v3, ALBERT v2; specialized span detection approaches: BINDER, PIQN, DyLex, Instance-based; and Llama-3.1-8B-Instruct (AI@Meta, 2024) as an instruction-following LLM.

As shown in Table 2, our RoBERTa-based model achieves the best performance among the tested models across both token-level classification (79.9% accuracy) and span-level detection (34.8% F1 score, exact match). The results indicate that transformer-based token classification approaches generally outperform specialized span detection models on this task. This suggests that the contextual understanding provided by transformers combined with token-level granularity is particularly beneficial for medical decision extraction, where precise boundary detection is crucial.

Among the transformer models, RoBERTa shows the strongest performance, followed closely by ELECTRA and BioClinicalBERT. The specialized span detection approaches exhibit lower performance, possibly due to the complexity and variability of medical decision spans compared to traditional NER tasks.

5.2 LLM Comparison

To evaluate the potential of large language models for medical decision extraction, we compared our fine-tuned RoBERTa model against `Llama-3.1-8B-Instruct` (AI@Meta, 2024) using zero-shot and one-shot prompting approaches.

Experimental Setup: We evaluated the LLM on 10 discharge summaries randomly selected from the MedDec test set. The LLM was prompted to extract decision spans for each of the ten DICTUM categories separately for each note using the following prompt structure:

```
[[[System]]]
Extract all substrings from the following clinical
note that contains medical decisions within the
specified category. Print each substring on a new
line. If no such substring exists, output "None".

[Clinical Note]: {Discharge summary}

# IF: one-shot setting
[[[User]]]
[Category]: {Demonstration Decision category}

[[[Assistant]]]
{Demonstrations}
# End IF

[[[User]]]
[Category]: {Target Decision category}

[[[Assistant]]]
{Response}
```

In the one-shot setting, demonstrations consist of all annotated decision spans for a single category within the same note. The demonstration category was chosen as the one with the most annotations in that specific note, excluding the target category being prompted.

Evaluation Metrics: Since LLMs generate free-form text, token-level accuracy comparable to classification models cannot be directly computed. We report span-level F1 scores based on exact match between predicted and gold standard spans. We also computed fuzzy match F1, where a match was considered positive if either span was a substring of the other and their lengths (in words) differed by no more than 10. This more lenient metric accommodates generative outputs that might be semantically similar but not identical to the gold span.

Results: As shown in Table 2, the LLM achieved span-level F1 scores of 3.8 (zero-shot) and 4.8 (one-shot) using exact match. Even with fuzzy matching, which yielded improved scores of 10.4 (zero-shot) and 17.9 (one-shot), the LLM performance remains substantially lower than the fine-tuned RoBERTa model (34.8 exact match F1).

This performance gap can be attributed to several factors: (1) challenges LLMs face with long clinical contexts (An et al., 2023), (2) the inherent difficulty in constraining free-form generative output to precisely match specific, pre-defined spans according to a structured taxonomy like DICTUM, and (3) the specialized nature of medical decision extraction which benefits from domain-specific fine-tuning.

While LLMs offer broad capabilities and excel at generative tasks, for the specific task of precise medical decision span extraction within our defined framework, fine-tuned token classification models currently provide superior accuracy and reliability. This justifies RoBERTa's use as the core extraction engine in MedDecXtract, prioritizing precision for this structured information extraction task while acknowledging LLMs as a promising direction for future exploration, potentially in hybrid systems or for related tasks like summarization or reasoning about the extracted decisions.

6 Conclusion

We presented MedDecXtract, an integrated, interactive system designed to support the extraction, visualization, and annotation of clinical decision-making documented in narrative text according to the Decision Identification and Classification Taxonomy for Use in Medicine (DICTUM). It combines automated extraction using a fine-tuned RoBERTa model, interactive temporal visualiza-

tion, and an intuitive annotation interface into a seamless workflow (Figure 1). While the RoBERTa component demonstrates superior performance for precise span extraction compared to tested instruction-following LLM approaches on this specific task (Section 5), the primary novelty lies in the synergy and integration of these components into a user-friendly web interface. The impact of this work extends to several areas including healthcare policy development, clinical education and training, and understanding of clinical decision-making processes.

Future work can investigate approaches that leverage LLM reasoning capabilities while retaining the precision of specialized models like RoBERTa, potentially informed by the datasets created using MedDecXtract's annotation tool. This could involve using multi-agent systems, or developing structured prompting strategies to better guide LLM outputs for this specific extraction task. In addition, our underlying extraction model was trained and evaluated exclusively on discharge summaries from the MIMIC-III database (Pollard and Johnson III, 2016). This may limit the generalizability of the extraction model on clinical notes of different types or from different institutions, diverse patient populations, or varying documentation styles. Future work may develop a diverse dataset of clinical notes with annotated medical decisions to improve the generalizability of the extraction model.

Ethics Statement

System Deployment: The public demo of MedDecXtract, hosted on Hugging Face Spaces, allows users to input clinical text for analysis. User-provided text is processed server-side solely for the purpose of performing inference (extraction, visualization) during the active user session. This input text is not logged, stored, or used for any other purpose beyond providing the immediate results to the user within the application interface. However, we advise users against inputting identifiable patient information into the public demo.

Dataset: The MedDec (Elgaar et al., 2024) dataset used for training and evaluation is derived from MIMIC-III (Pollard and Johnson III, 2016). Access to MIMIC-III (and subsequently MedDec) requires completion of ethics training and signing a data use agreement, ensuring responsible data handling and patient privacy protection.

References

Abubakar Abid, Ali Abdalla, Ali Abid, Dawood Khan, Abdulrahman Alfozan, and James Zou. 2019. Gradio: Hassle-free sharing and testing of ml models in the wild. In *ICML Workshop on Human in the Loop Learning (HILL)*.

Griffin Adams, Emily Alsentzer, Mert Ketenci, J. Zucker, and Noémie Elhadad. 2021. What's in a summary? laying the groundwork for advances in hospital-course summarization. *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, 2021:4794–4811.

AI@Meta. 2024. Llama 3 model card. Github repository, accessed on June 6, 2024.

Emily Alsentzer and Anne Kim. 2018. Extractive summarization of ehr discharge notes. *ArXiv*, abs/1810.12085.

Chenxin An, Shansan Gong, Ming Zhong, Mukai Li, Jun Zhang, Lingpeng Kong, and Xipeng Qiu. 2023. L-eval: Instituting standardized evaluation for long context language models. *ArXiv preprint*, abs/2307.11088.

Mohamed Elgaar, Jiali Cheng, Nidhi Vakil, Hadi Amiri, and Leo Anthony Celi. 2024. MedDec: A dataset for extracting medical decisions from discharge summaries. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 16442–16455, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Hugging Face. 2024. Hugging face.

J. Feblowitz, A. Wright, Hardeep Singh, L. Samal, and Dean F. Sittig. 2011. Summarization of clinical information: A conceptual model. *Journal of biomedical informatics*, 44(4):688–99.

Yanjun Gao, D. Dligach, T. Miller, Dongfang Xu, M. Churpek, and M. Afshar. 2022. Summarizing patients' problems from hospital progress notes using pre-trained sequence-to-sequence models. *Proceedings of COLING. International Conference on Computational Linguistics*, 2022:2979–2991.

B. Glicksberg, B. Oskotsky, P. Thangaraj, Nicholas P. Giangreco, Marcus A. Badgeley, Kipp W. Johnson, Debajyoti Datta, V. Rudrapatna, Nadav Rappoport, Mark M. Shervey, Riccardo Miotto, Theodore Goldstein, Eugenia Rutenberg, Remi Frazier, Nelson Lee, Sharat Israni, Rick Larsen, B. Percha, Li Li, J. Dudley, N. Tatonetti, and A. Butte. 2019. Patientexplorer: an extensible application for dynamic visualization of patient clinical history from electronic health records in the omop common data model. *Bioinformatics*, 35:4515 – 4518.

J. Hirsch, Jessica S. Tanenbaum, S. Gorman, Connie Liu, E. Schmitz, Dritan Hashorva, Artem Ervits, D. Vawdrey, M. Sturm, and Noémie Elhadad. 2014. Harvest,

a longitudinal patient record summarizer. *Journal of the American Medical Informatics Association : JAMIA*, 22:263 – 274.

Plotly Technologies Inc. 2015. Collaborative data science.

Daniel Keszthelyi, C. Gaudet-Blavignac, Mina Bjelogrlic, and Christian Lovis. 2023. Patient information summarization in clinical settings: Scoping review. *JMIR Medical Informatics*, 11.

Eva K. Lee and K. Uppal. 2020. Cerc: an interactive content extraction, recognition, and construction tool for clinical and biomedical text. *BMC Medical Informatics and Decision Making*, 20.

Eric Lehman, Jay DeYoung, Regina Barzilay, and Byron C. Wallace. 2019. Inferring which medical treatments work from reports of clinical trials. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3705–3717, Minneapolis, Minnesota. Association for Computational Linguistics.

Jennifer J. Liang, Ching-Huei Tsou, Bharath Dandala, Ananya Poddar, Venkata Joopudi, Diwakar Mahajan, J. Prager, Preethi Raghavan, and Michele Payne. 2021. Reducing physicians' cognitive load during chart review: A problem-oriented summary of the patient electronic record. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2021:763–772.

Jennifer J. Liang, Ching-Huei Tsou, and Ananya Poddar. 2019. A novel system for extractive clinical note summarization using ehr data. *Proceedings of the 2nd Clinical Natural Language Processing Workshop*.

Daniel Mário De Lima, Jean R. Ponciano, Agma Juci Machado Traina, Mauro M Olivatto, Claudio D. G. Linhares, Caetano Traina, Marco Antonio Gutierrez, and Jorge Poco. 2022. Clinicalpath: A visualization tool to improve the evaluation of electronic health records in clinical decision-making. *IEEE Transactions on Visualization and Computer Graphics*, 29:4031–4046.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv preprint*, abs/1907.11692.

Denis Jered McInerney, B. Dabiri, Anne-Sophie Touret, Geoffrey Young, Jan-Willem van de Meent, and Byron C. Wallace. 2020. Query-focused ehr summarization to aid imaging diagnosis. *ArXiv*, abs/2004.04645.

Luke S. Murray, D. Gopinath, Monica Agrawal, S. Horng, D. Sontag, and David R Karger. 2021. Medknowts: Unified documentation and information retrieval for electronic health records. *The 34th Annual ACM Symposium on User Interface Software and Technology*.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. *ArXiv preprint*, abs/2303.13375.

Benjamin Nye, Junyi Jessy Li, Roma Patel, Yinfei Yang, Iain Marshall, Ani Nenkova, and Byron Wallace. 2018. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 197–207, Melbourne, Australia. Association for Computational Linguistics.

Eirik H Ofstad, Jan C Frich, Edvin Schei, Richard M Frankel, and Pål Gulbrandsen. 2016. What is a medical decision? a taxonomy based on physician statements in hospital encounters: a qualitative study. *BMJ open*, 6(2):e010098.

Pinal Patel, Disha Davey, Vishal Panchal, and Parth Pathak. 2018. Annotation of a large clinical entity corpus. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2033–2042, Brussels, Belgium. Association for Computational Linguistics.

Rimma Pivovarov and Noémie Elhadad. 2015. Automated methods for the summarization of electronic health records. *Journal of the American Medical Informatics Association : JAMIA*, 22:938 – 947.

Tom J Pollard and AEW Johnson III. 2016. The mimic iii clinical database, version 1.4. *The MIMIC-III Clinical Database. PhysioNet*.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Hieu Tran, Zhichao Yang, Zonghai Yao, and Hong Yu. 2024. Bioinstruct: Instruction tuning of large language models for biomedical natural language processing. *Journal of the American Medical Informatics Association*, page ocae122.

Mengqian Wang, Manhua Wang, Fei Yu, Yue Yang, Jennifer S. Walker, and Javed Mostafa. 2021. A systematic review of automatic text summarization for biomedical literature and ehrs. *Journal of the American Medical Informatics Association : JAMIA*.

A Example Summary of Decisions

The following is an example summary of decisions for a patient:

[2/12/2024]
Defining problem
- Heart: RRR, no murmurs, rubs or gallops. Radial pulses +2 bilateral
- Gen: No acute distress, conversational,
- Lungs: Clear to ascultation bilaterally, no whee
- Psych: Well-groomed. Non-pressured speech, linear though process
- Neck: No thyromegaly, no lymphadeopa
Drug related
- Tylenol
[3/18/2024]
Drug related
- a trial of low-dose sertraline
- Improvement
- Started
- dose and
- sertraline
- 3 months
- Tylenol
Defining problem
- Gen: Appears more relaxed than the previous visit
- Psych: Appears slightly more at ease, maintains good eye contact, speech and thought process remain coherent
- Neck: No changes.
- Lungs: Clear to auscultation
- Heart: Unchanged Evaluating test result
ROS: Negative except as noted
H: No changes
Contact related
- Consider referral to therapy for additional support
Therapeutic procedure related
- breathing
- breathing
[12/29/2024]
Evaluating test result
- H: None
- PMH: No changes
- HX
- ROS: Entirely negative
Defining problem
- Gen:
- Psych:
- Looks healthy and content
- Lungs: Clear
- She feels much better and
- improvement
- SH: Stable and positive home and work environment
- Neck: No changes
- She remains active at work and home
- Maintained improvement in mental health
- Heart: Unchanged
Therapeutic procedure related
- Continue therapy and supportive measures
Drug related
- Sertraline, with a plan to taper
- gradual medication reduction
- start tapering off sertraline
- medical supervision
- in tapering off medication
- Will initiate a slow tapering process of sertraline
Contact related
- Next follow-up scheduled in 3 months to assess progress