Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

OncoTrack: LLM-based Structured Extraction and Progression Summarization from Longitudinal Oral Cancer Records

Nishith P

bl.en.u4cse22138@bl.students.amrita.edu

Amrita Vishwa Vidyapeetham

Krishnakumar Thankappan

Amrita Institute of Medical Sciences and Research Centre

Lakshmi Ravunniarth Menon

Amrita Institute of Medical Sciences and Research Centre

Sagun Kumari

Amrita Institute of Medical Sciences and Research Centre

Nagesh K. Subbanna

Amrita Vishwa Vidyapeetham

Article

Keywords: Clinical Decision Support, Data Extraction, Fine-Tuning, Large Language Models, Longitudinal Clinical Records, Natural Language Processing, Oral Cancer, Progression Summarization

Posted Date: November 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10540617/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

OncoTrack: LLM-based Structured Extraction and Progression Summarization from Longitudinal Oral Cancer Records

Nishith P[1*], Krishnakumar Thankappan[3], Lakshmi Ravunniarth Menon[3], Sagun Kumari[3], Nagesh K. Subbanna[2*]

[1]Department of Computer Science and Engineering, Amrita School of Computing, Amrita Vishwa Vidyapeetham, Bengaluru, India.
[2]Department of Artificial Intelligence, Amrita School of Artificial Intelligence, Amrita Vishwa Vidyapeetham, Bengaluru, India.
[3]Department of Head and Neck Surgery and Oncology, Amrita Institute of Medical Sciences, Amrita Vishwa Vidyapeetham, Kochi, India.

*Corresponding author(s). E-mail(s): bl.en.u4cse22138@bl.students.amrita.edu; nageshks@am.amrita.edu;

Abstract

Oral cancer poses a significant global health challenge, with patient management heavily reliant on prior case notes and reports. Manual extraction of critical clinical features from unstructured patient records is often time-consuming, labour intensive, expensive, inconsistent, and prone to human error, hindering timely insights into disease progression and effective clinical decision-making. The proposed work, "OncoTrack", introduces an AI-powered pipeline to automatically extract relevant clinical data from oral cancer documents and generate progression summaries. The system leverages a fine-tuned Large Language Model (LLM), specifically Meta-Llama-3.1-8B, for extracting 24 clinically relevant tracking features from longitudinal records. In the initial 90:10 experiment on a curated 100-document AIMS cohort, the model achieved approximately **87.83% ± 1.81** accuracy and **86.80%±2.00** accepted-match precision after **10** epochs of Low-Rank Adaptation (LoRA) fine-tuning. The evaluation was then extended to a broader, heterogeneous document pool (total processed: 674), where the retained 339-instance held-out analysis produced an overall average normalized field accuracy of **79.97%** and precision of **73.45%** after excluding sparse or inference-heavy fields from the aggregate score; across all 24 fields, average accuracy was **70.37%**. The weakest

temporal field was 'lastFU', which achieved **0.00** accuracy in the initial test set and approximately **9.44%** accuracy in the expanded evaluation. The 'lastFU' was not present in most of the documents scanned by the LLM which explains the low accuracy. Subsequent summarization of extracted features is performed by Google's Gemma 3 4b model. The web application and RAG-based assistant are presented as research-prototype components that require independent evaluation, security validation, and clinical workflow assessment before deployment.

Keywords: Clinical Decision Support, Data Extraction, Fine-Tuning, Large Language Models, Longitudinal Clinical Records, Natural Language Processing, Oral Cancer, Progression Summarization

1 Introduction

Oral cancer is associated with substantial morbidity and mortality, particularly when diagnosis occurs at advanced stages [1]. Clinical management depends on longitudinal assessment of disease status, including primary tumor stage, nodal status, pathological risk factors, treatment timelines, recurrence, and follow-up outcomes documented across multiple visits. A large portion of this information is embedded in unstructured progress notes, pathology reports, and discharge summaries [2]. Variation in terminology, abbreviations, and documentation style across clinicians and centres makes reliable manual abstraction difficult.

Manual abstraction of these records is time-intensive and susceptible to inter-reviewer variability and transcription error [3, 4]. Clinicians must review long chronological narratives to identify key variables such as tumor staging, pathological findings, and treatment dates [5]. This documentation burden can delay tabulation of longitudinal evidence needed for treatment planning and follow-up decisions [2].

To mitigate these issues, the field of medical informatics has turned to Natural Language Processing (NLP). Early solutions involved rule-based systems and traditional machine learning models [6]. While these approaches offered some degree of automation, they had significant limitations. Rule-based systems are brittle and struggle to generalize across the varied language of clinical notes, while traditional ML models require large, painstakingly annotated datasets, which are expensive and difficult to create in specialized domains like oncology [3].

The advent of Large Language Models (LLMs) has introduced a transformative approach. Models like BERT and GPT, pre-trained on massive text corpora, have shown remarkable capabilities in understanding complex language [7, 8]. However, applying these general-purpose models directly to highly specialized domains like oncology often yields suboptimal performance due to a lack of domain-specific knowledge and a tendency to "hallucinate" information [9].

To address this gap, this study investigates whether a domain-adapted open-source LLM can reliably extract structured oral-cancer variables from longitudinal clinical text and support downstream progression summaries. In this manuscript, "longitudinal" refers to patient records containing temporally distributed clinical events and follow-up notes; we do not perform formal statistical time-series modeling, temporal forecasting,

state-transition modeling, or trajectory prediction. We define two hypotheses: *H1*, parameter-efficient fine-tuning of Meta-Llama-3.1-8B can produce stable, schema-conformant extraction across the predefined 24 clinical features (Table 1); and *H2*, extraction performance will be higher for explicitly stated categorical fields than for inference-heavy temporal fields (e.g., last follow-up).

Accordingly, we develop and evaluate OncoTrack, a pipeline that combines LoRA-based fine-tuning for feature extraction with a separate summarization module. This paper details the dataset, annotation schema, model adaptation strategy, system architecture, and quantitative/qualitative evaluation.

Table 1 Targeted Clinical Features and Mappings for Data Extraction

No.	Feature Name	Mapping / Description	No.	Feature Name	Mapping / Description
1.	age	Numerical value	13.	skininv	{no: 0, yes: 1} (Skin Invasion)
2.	sex	{male: 1, female: 2}	14.	ENS	{no: 0, yes: 1} (Extracapsular Nodal Spread)
3.	subsite	{tongue: 1, BM: 2, FOM: 3, alveolus: 4, hard palate: 5, RMT: 6, lips: 7}	15.	Surgery	{SND 1-3: 1, SND 1-4: 2, MRND: 3, RND: 4}
4.	pathology	{WDSCC: 1, Verrucous: 2, MDSCC: 3, PDSCC: 4, others: 5, ca in situ: 6}	16.	lateralitysurgery	{ipsilateral: 1, bilateral: 2}
5.	clinicalT8thed	{T1: 1, T2: 2, T3: 3, T4a: 4, T4b: 5}	17.	DOS	Date of Surgery (e.g., MM/DD/YYYY)
6.	nodal8thed	{No nodes: 0, N1: 1, N2a: 2, N2b: 3, N2c: 4, N3a: 5, N3b: 6}	18.	adjRx	{no: 0, yes: 1} (Adjuvant Therapy)
7.	pt8thed	{T1: 1, T2: 2, T3: 3, T4a: 4, T4b: 5}	19.	date of LR	Date of Local Recurrence (e.g., MM/DD/YYYY)
8.	pN8thed	{N0: 0, N1: 1, N2a: 2, N2b: 3, N2c: 4, N3a: 5, N3b: 6}	20.	Nodalrecurrencelevel	{ipsilateral: 1, contralateral: 2}
9.	PNI	{no: 0, yes: 1} (Perineural Invasion)	21.	Nodallevel	Numerical or text value
10.	LVI	{no: 0, yes: 1} (Lymphovascular Invasion)	22.	date of Nodalrecurrence	Date (e.g., MM/DD/YYYY)
11.	Margins	{negative: 0, close: 1, positive: 2}	23.	lastFU	Date of Last Follow-Up (e.g., MM/DD/YYYY)
12.	Boneinv	{no: 0, yes: 1} (Bone Invasion)	24.	stateFU	{NED: 1, AWD: 2, DWD: 3, Dead due to other causes: 4}

Abbreviations: BM – Buccal Mucosa, FOM – Floor of Mouth, RMT – Retromolar Trigone, WDSCC – Well-Differentiated Squamous Cell Carcinoma, MDSCC – Moderately Differentiated Squamous Cell Carcinoma, PDSCC – Poorly Differentiated Squamous Cell Carcinoma, SND – Selective Neck Dissection, MRND – Modified Radical Neck Dissection, RND – Radical Neck Dissection, NED – No Evidence of Disease, AWD – Alive With Disease, DWD – Died With Disease.

1.1 Literature Survey

The automated extraction of structured information from unstructured clinical text is a well-established field in medical informatics [2, 10, 11], evolving from early rule-based methods to modern deep learning approaches. This survey examines the key methodologies, their shortcomings, and the context in which our proposed work, OncoTrack, offers a superior solution.

1.1.1 Traditional NLP and Early Machine Learning

Early approaches to clinical information extraction predominantly used rule-based systems and traditional machine learning (ML) models for extraction of different types of relevant information from clinical data such as bone metastasis [6] or general clinical information [10]. Rule-based systems, which rely on handcrafted patterns and domain-specific lexicons, are highly interpretable. However, their performance is context-dependent: they can work well for constrained templates but often degrade with heterogeneous note styles, local abbreviations, and evolving documentation practices, and they require substantial maintenance effort [2, 12].

Traditional ML models, such as support vector machines (SVMs) and conditional random fields (CRFs) [13], offered improved adaptability. However, they rely heavily on handcrafted features and large annotated datasets, as in the works reported in [3] and [14]. Creating such datasets remains a costly, time-consuming, and expert-driven bottleneck, particularly in specialized oncology settings.

1.1.2 Transformers and Pre-trained Language Models

The introduction of Transformer-based models like BERT (Bidirectional Encoder Representations from Transformers) [15] revolutionized NLP. By pre-training on vast amounts of general text, these models learn rich, contextualized representations of language. This led to the development of domain-specific variants like BioBERT, ClinicalBERT [16], and CancerBERT [17], which were pre-trained or further fine-tuned on biomedical and clinical corpora. These models significantly outperformed their predecessors on tasks such as named entity recognition and relation extraction in clinical texts [11, 18].

The shortcoming of these early Transformer models, when fully fine-tuned, is their substantial computational cost in terms of GPUs and time, which can be a barrier for many research groups and clinical institutions [19, 20].

1.1.3 Zero-Shot and Few-Shot Learning with LLMs

More recent, even larger models like GPT-3 [21], GPT-4 [22], and ChatGPT [23] demonstrated impressive zero- and few-shot learning capabilities. These models can perform tasks with only a natural language description or a few examples in the prompt, without any gradient updates. Studies have shown their feasibility in extracting structured information from radiology and pathology reports with minimal setup [24, 25].

The primary shortcoming of these very large, generalist models is their propensity for "hallucination"—generating factually incorrect or nonsensical information—and

a lack of deep, specialized domain knowledge [9, 26]. While prompt engineering can mitigate this to an extent [27], their reliability for high-stakes clinical applications remains a concern. Certain features may be continuously ignored by the techniques in the medical context [24, 26] when the number of training epochs is comparatively small. Furthermore, their performance can be brittle and highly sensitive to the position of information within a long context, often struggling to use information "lost in the middle" [28].

1.1.4 Parameter-Efficient Fine-Tuning (PEFT)

To bridge the gap between resource-intensive full fine-tuning and the unreliability of zero-shot prompting, Parameter-Efficient Fine-Tuning (PEFT) methods have emerged. Techniques such as Low-Rank Adaptation (LoRA) [19] allow adaptation of large-parameter LLMs (typically billions of parameters) by training only a small fraction of the weights. This dramatically reduces computational and memory requirements, making fine-tuning more accessible [7, 20]. Importantly, freezing most parameters does not by itself eliminate hallucination; output reliability still depends on domain-specific training data, prompt constraints, and evaluation design. Studies have demonstrated that fine-tuning with LoRA can achieve performance comparable to, or even exceeding, full fine-tuning while being significantly more efficient [8, 29].

1.2 Rationale for the OncoTrack Design

The OncoTrack design combines a moderately sized open-source model (Meta-Llama-3.1-8B) with LoRA to balance extraction performance, computational feasibility, and deployment practicality:

- **Compared with traditional methods:** The approach reduces dependence on manually encoded rules and extensive handcrafted features by leveraging representations learned during LLM pre-training.
- **Compared with zero-shot prompting:** Domain-specific fine-tuning improves schema consistency and feature-level reliability for oral-cancer extraction. In our experiments with the base Meta-Llama-3.1-8B model under zero-shot prompting, outputs were frequently non-deterministic and structurally inconsistent (Figure 1), including markdown blocks, hallucinated fields, inconsistent date values, and variable key formatting. Such behavior requires complex post-processing and reduces reliability for clinical data pipelines.
- **Compared with full fine-tuning:** LoRA reduces memory and compute requirements, enabling adaptation on constrained hardware. In our setup, full fine-tuning was not feasible on 16GB GPUs even with memory optimizations.

Overall, this design is intended to provide a reproducible and resource-aware framework for structured information extraction in oral oncology.

The remainder of this paper is organized as follows. Section 2 details our methodology, covering the dataset, the feature extraction task, the LLM fine-tuning pipeline, and the proposed system architecture. Section 3 provides specific implementation details regarding the tools, libraries, and configurations used. In Section 4, we present and



Figure 1 Examples of non-deterministic and erroneous output from the zero-shot Llama-3.1-8B model on the same patient profile. Note the structural inconsistencies, markdown code blocks, hallucinated features (Verrucous, WDSCC, MDSCC, PDSCC, others, ca in situ, BM, FOM, RMT, lips), and varied formatting for feature names and missing values.

analyze the quantitative and qualitative results of our experiments. Finally, Section 5 concludes the paper with a summary of our findings and discusses potential avenues for future research.

2 Methodology

The study protocol was reviewed and approved by the institutional review board, the Ethics Committee of Amrita School of Medicine (ECASM), Amrita Vishwa Vidyapeetham (Approval No.: ECASM-AIMS-2025-405; Date: November 6, 2025). Since this was a retrospective analysis of de-identified patient records, the requirement for informed consent was waived by the institutional review board. All methods were performed in accordance with the relevant guidelines and regulations and under the applicable institutional ethical oversight.

The OncoTrack project develops a system for automated data extraction and summary generation for oral cancer patients. This section details the dataset used, the features targeted for extraction, the LLM fine-tuning pipeline, and the overall system architecture.

2.1 Dataset

The initial development dataset consisted of 100 historical oral-cancer patient documents from Amrita Institute of Medical Sciences (AIMS). To protect patient privacy and adhere to ethical guidelines, all documents were de-identified before use by removing personally identifiable information (PII), such as names and medical record numbers. These records, written by different clinicians, already showed meaningful stylistic and vocabulary variability. This curated 100-document cohort was used for fine-tuning and early validation under a 90:10 train-test split.

After this controlled experiment, evaluation was extended to a much wider and more variegated pool of documents from multiple sources (674 total processed records). Following quality checks, the expanded corpus was handled as two non-overlapping subsets. These quality checks were introduced to identify cases in which the source transcript did not contain sufficient evidence for deterministic comparison with the manually curated labels; they were used for evaluation reporting rather than for model training.

The first subset comprised 339 records retained for the primary expanded held-out evaluation. In this subset, 'NodalrecurrenceLevel' and 'nodallevel' were frequently absent, and 'lastFU' remained strongly inference-heavy; therefore, these sparse or inference-heavy fields were excluded from aggregate score reporting in expanded-set summaries. Their individual field scores are still reported to avoid masking these failure modes.

The second subset comprised 335 records excluded from headline benchmarking because source transcripts showed severe text-label misalignment or missing evidence. The exclusion signals (not mutually exclusive) included 191 cases where ground-truth age did not match regex-extracted age from text and 106 cases where age was absent in the document. Because these exclusions account for nearly half of the expanded pool, the retained-set results should be interpreted as performance on records with stronger text-label support rather than as an all-comers deployment estimate. A fully prospective benchmark would require independent re-annotation, explicit "not documented" or "conflicting evidence" labels, and adjudication of cases with multiple tumor events or ambiguous time points.

2.2 Feature Extraction Task

The core task is to extract 24 predefined clinical features relevant to oral cancer progression. These features are listed in Table 1, which also details the numerical mappings used for standardization.

2.3 LLM Fine-Tuning Pipeline

The fine-tuning process, illustrated in Figure 2, was designed to adapt an open-source LLM for the specific task of extracting the 24 oral cancer features.

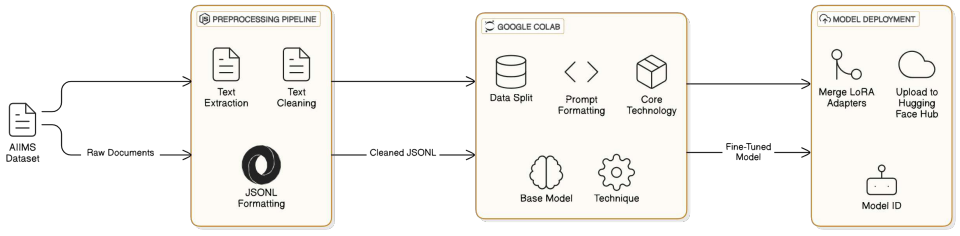


Figure 2 Overview of the LLM Fine-Tuning Pipeline.

2.3.1 Data Preprocessing

The raw text from the 100 patient documents was first extracted. A javascript code was developed to clean the text by removing unnecessary characters such as excessive newlines, tab characters, and carriage return/newline combinations. The cleaned text served as the "input" for fine-tuning. The corresponding "output" for each document was a manually curated semicolon-separated list of the 24 features and their values.

2.3.2 Instruction and Prompt Formatting

A detailed system instruction was crafted to guide the LLM during fine-tuning and inference: *"Extract the necessary information from the given text, which contains a patient's oral cancer data. Use the provided mappings (as in Table 1) to ensure accurate and efficient extraction. Format the output as a semi-colon separated list of key-value pairs, strictly adhering to the following structure: feature1:value1;feature2:value2;... Example: age:40;sex:1;subsite:2;Pathology:3;... Ensure precise and efficient extraction by applying the given mappings accurately."* The full list of mappings and full forms for abbreviations were included as part of this instruction.

The preprocessed data (instruction, input document text, and target output string) was then formatted into a JSONL file, where each line is a JSON object. For the fine-tuning process, this data was further transformed into the Alpaca prompt format where the placeholders are filled with the system instruction, the cleaned document text, and the target semicolon-separated feature string, respectively.

2.3.3 Model Selection and Fine-Tuning

Initial experiments were conducted with Google's Gemma2 9B [30], Meta's Llama-3.1-8B [31], and Mistral AI's Mistral-Nemo-Base-2407 [32]. Tokenizer analysis revealed that Gemma's tokenizer resulted in significantly longer token sequences (up to 16k tokens) for the input documents compared to Llama and Mistral (up to 14k tokens). Due to Google Colab's GPU memory constraints (15GB on the free tier), fine-tuning Gemma2 was not feasible, leading to "CUDA: out of memory" errors.

Consequently, Meta-Llama-3.1-8B (4-bit quantized version: 'Meta-Llama-3.1-8B-bnb-4bit') and Mistral AI's Mistral-Nemo-Base-2407 [32] were chosen as the primary models for fine-tuning, performed on Google Colab Pro with an A100 GPU (40GB VRAM). The Unsloth library was used for efficient fine-tuning using LoRA and 4-bit quantization. The models were fine-tuned for various epochs (2, 5, and 10). The 10-epoch fine-tune yielded the best performance, achieving approximately $87.83\% \pm 1.81$ accuracy and $86.80\% \pm 2.00$ precision on the 10% test set. The maximum sequence length was set according to the model's tokenizer requirements for the dataset, around 14k tokens for Llama.

2.3.4 Low-Rank Adaptation (LoRA) Explained

Full fine-tuning of an LLM, which involves updating all of its billions of parameters, is computationally prohibitive. Low-Rank Adaptation (LoRA) [19] is a parameter-efficient fine-tuning (PEFT) technique that overcomes this barrier. The core idea

behind LoRA is the hypothesis that the change in model weights during adaptation has a low "intrinsic rank".

Instead of updating the original weight matrix W_0, LoRA freezes W_0 and injects a pair of trainable, low-rank matrices, A and B, into each target layer. The weight update ΔW is represented by the product of these smaller matrices, i.e., $\Delta W = BA$. The modified forward pass becomes $h = W_0x + BAx$. During training, only the parameters of A and B are updated, drastically reducing the number of trainable parameters (by up to 10,000 times) and the GPU memory requirement (by up to 3 times) [19]. For inference, the learned weights can be merged back into the original model ($W = W_0 + BA$), introducing no additional latency. This makes LoRA highly efficient for both training and deployment.

2.4 Summary Generation

After feature extraction, the OncoTrack system generates a summary of the patient's cancer progression. This involves:

1. **Reverse Mapping:** The numerically mapped feature values (e.g., 'sex:1') are converted back to their textual labels (e.g., 'sex: male').
2. **Narrative Generation:** These labeled features are then fed to a lightweight LLM, Google's Gemma3 4b model [33], selected for resource-efficient open-weight text generation [34], which is instructed to weave them into a grammatically coherent and readable narrative summary describing the patient's condition and progression over time. This summarization component was implemented as a prototype and was not independently benchmarked for factual consistency or hallucination rate in the present study. This model is also hosted on the Kaggle server.

2.5 System Architecture

The OncoTrack system is designed with a modular architecture to support experimentation and maintainability, as depicted in Figures 3 and 4. The implementation described here should be interpreted as a retrospective research prototype rather than a validated clinical or patient-facing deployment.

The architecture comprises:

1. **Frontend:** A web application built using Streamlit serves as the user interface (Figure 6). In the prototype, it allows authorized research or clinical users to:
   - Upload patient documents (e.g., Word files).
   - The system then uses regular expressions to identify and extract relevant "Progress Notes" sections from the uploaded document.
   - Send the extracted progress notes to the backend server for feature extraction via a POST request.
   - View the 24 extracted features displayed in a table.
   - Download the extracted features as a CSV file.
   - View the AI-generated summary of the patient's history.
   - Interact with an AI assistant via text or voice in their native language in a demonstration setting.

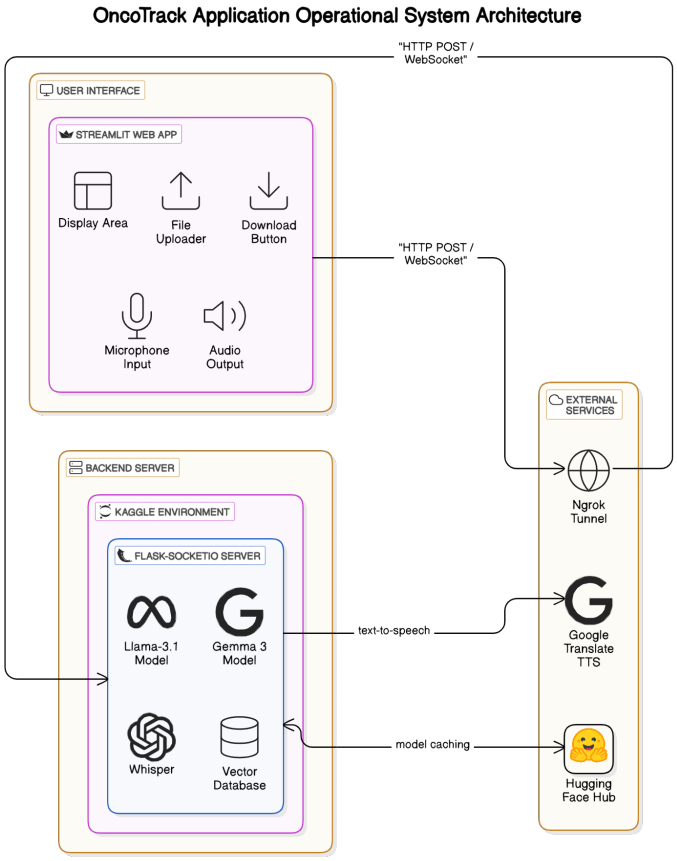


Figure 3 OncoTrack Operational System Architecture. The Streamlit Web interface accepts the files, and the extracted and cleaned textual material is sent to the Flask server. This server runs the fine tuned LLM model, then extracts the features, and generates the textual summary, which is then sent back to the front end.

2. **Backend Server:** The backend is a Flask server hosted on a Kaggle Notebook environment, leveraging Kaggle's GPU resources. Its responsibilities include:
   - Receiving progress notes from the frontend.
   - Loading the fine-tuned Meta-Llama-3.1-8B model from Hugging Face [31] for feature extraction.
   - Performing inference to extract the 24 clinical features.

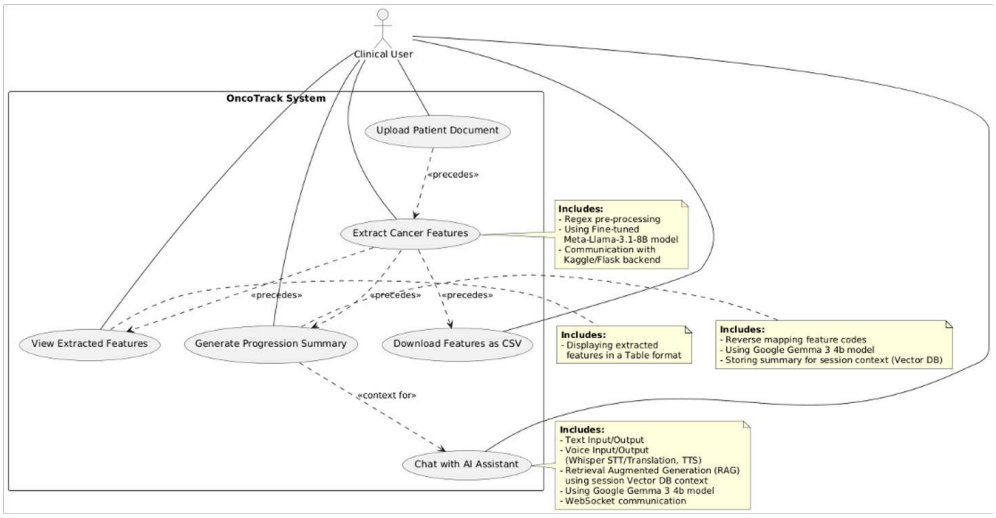


Figure 4 OncoTrack Use Case Diagram showing interactions between a Clinical User and the system functionalities.

- Returning the extracted features (JSON object) to the frontend.
- Utilizing Google's Gemma3 4b model [33] for generating patient summaries based on the extracted features.
- Handling real-time chat interactions. This involves:
  - Receiving text or audio input from the user.
  - Using OpenAI Whisper [35] for Speech-to-Text (STT), translation (if non-English), and language detection.
  - Performing Retrieval Augmented Generation (RAG) using the generated summary and historical data (stored session-wise in a ChromaDB vector database) as context for the Gemma3 model [33].
  - Generating a response with Gemma3.
  - Using Google Translate Text-to-Speech (TTS) for voice output in the user's native language.

The connection to the Kaggle server is tunneled using Ngrok, providing a publicly accessible FQDN URL for API and WebSocket communication during prototype testing. Data is handled session-wise and is ephemeral in this implementation. However, systematic validation of authentication, encryption, audit logging, retention/deletion policies, vector-store isolation, audio retention, and third-party service exposure has not yet been completed. Consequently, this stack is not presented as ready for routine clinical deployment or direct patient-facing use.

The user workflow is illustrated in Figure 5.

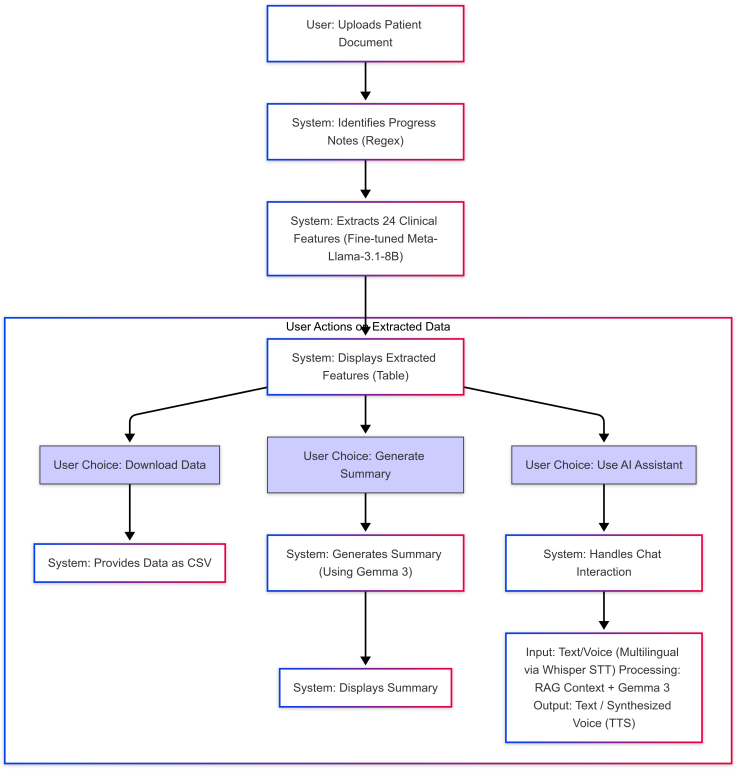


Figure 5 User Workflow and System Actions in OncoTrack.

3 Implementation Details

This section provides specific details about the implementation of the preprocessing of the data, fine-tuning of the model, and the deployment of the backend for the OncoTrack system.

3.1 Data Preprocessing and Prompt Formatting

A suite of Javascript code was developed to handle the initial data pipeline. A script first extracts raw text from the source '.docx' files using the 'mammoth' library. It then cleans the text by standardizing whitespace and removing extraneous characters. Another script formats the cleaned text, the system instruction, and the manually curated feature strings into a JSONL file. This file is then converted into the Alpaca prompt format required for training with the Unsloth library.

3.2 Fine-Tuning Environment and Tools

- **Platform:** Fine-tuning was conducted on Google Colab Pro, utilizing a single NVIDIA A100 GPU with 40GB of VRAM.

- **Frameworks and Libraries:** The core of the fine-tuning process was built on PyTorch. We used the **Unsloth** library for its highly efficient, memory-optimized implementation of LoRA, which enables training large models on consumer-grade or free-tier cloud GPUs. Other key libraries include Hugging Face's 'transformers' for model and tokenizer access, 'peft' (Parameter-Efficient Fine-Tuning) for LoRA implementation, 'accelerate' for hardware optimization, and 'bitsandbytes' for 4-bit quantization.

3.2.1 LoRA Configuration

- **Rank (r):** 16
  This determines the number of trainable parameters. A rank of 16 is a standard, recommended value that offers a good balance between model capacity and the risk of overfitting.
- **Alpha (α):** 16 (scaling factor)
  This acts as a scaling factor for the LoRA updates. Setting alpha equal to the rank is a common and reliable baseline practice to ensure the updates are appropriately weighted.
- **Target Modules:** LoRA adapters were applied to all linear layers of the Transformer architecture. *Applying LoRA to all major linear layers in both the attention and Multi-Layer Perceptron (MLP) blocks is a crucial practice recommended for achieving performance comparable to full fine-tuning.*
- **Dropout:** 0
  Set to 0 as per Unsloth's optimization guidelines, which state that dropout provides limited regularization benefits for short fine-tuning runs and that setting it to 0 enables a faster training implementation.
- **Bias:** "none"
  Configured to "none" to prevent the training of bias terms, a recommended practice that reduces memory usage and speeds up training with negligible impact on performance.

3.2.2 Training Arguments

- **Epochs:** 10
  While common guidelines suggest 1-3 epochs to prevent overfitting, 10 epochs were chosen as this yielded the best performance on the test set, likely due to the complexity of the feature extraction task from noisy clinical text.
- **Learning Rate:** 2e-3
  A relatively high learning rate was used, which is higher than the typical starting point of 2e-4. This aggressive rate can lead to faster convergence, which proved effective in this training run.
- **Optimizer:** `adamw_8bit`
  A memory-efficient version of the AdamW optimizer, suitable for training with 4-bit quantization.
- **Batch Size:** 2 (per device) with **Gradient Accumulation Steps:** 4
  This results in an effective batch size of 8 (2 x 4), which falls within the recommended

range for stable training. This standard technique manages VRAM usage by processing data in smaller micro-batches before performing a weight update.

- **Maximum Sequence Length:** 14336 tokens
  A large sequence length was essential to process the lengthy, unabridged clinical documents without truncating critical patient information.

3.3 Backend and System Deployment

- **Backend Server:** A Flask server runs on a Kaggle Notebook, exposed to the internet via an Ngrok tunnel for prototype testing. This setup provides access to GPU resources for inference but would require institutional hosting, authentication, transport encryption, and security review before use with identifiable clinical data.
- **LLM Serving:** The Gemma3 4b model [33] for summarization and chat is served locally on the Kaggle instance using Ollama. This ensures low-latency responses for interactive components.
- **AI Assistant (RAG):** The chat functionality is implemented using a Retrieval-Augmented Generation (RAG) pipeline. For each session, patient data and generated summaries are embedded using 'nomic-embed-text' [36] and stored in a ChromaDB vector database. When a user asks a question, relevant context is retrieved from the database to augment the prompt sent to the Gemma3 model. This design is intended to improve contextual grounding, but the present study did not independently quantify answer accuracy, hallucination rate, speech-recognition quality, translation quality, or cross-session contamination risk.
- **User Interface:** The frontend is a Streamlit application (Figure 6) that communicates with the backend via HTTP requests for extraction/summarization and WebSocket connections for real-time chat.

4 Results and Analysis

This section presents the quantitative and qualitative results of the LLM fine-tuning experiments, analyzing the performance of the selected model on the feature extraction part of the oral cancer database.

4.1 Fine-Tuning Performance

The primary challenge for fine-tuning was the extensive context length of the clinical documents. A tokenizer analysis of the 100-document AIMS dataset revealed that the maximum combined token length (instruction + input + output) was approximately 14,000 tokens for Llama-based models. The frequency distribution in Figure 7 illustrates that a substantial number of documents require a large context window, justifying the need for hardware with significant VRAM, such as the A100 GPU. In the later expanded evaluation set, some records exceeded this limit and required an increased inference-time context window, as discussed in Section 4.4.

The performance of the fine-tuned Meta-Llama-3.1-8B and Mistral-Nemo-Base-2407 models was evaluated on the test set, comprising 10 documents, after 2, 5, and 10 epochs. The results, in terms of normalized field accuracy and accepted-match

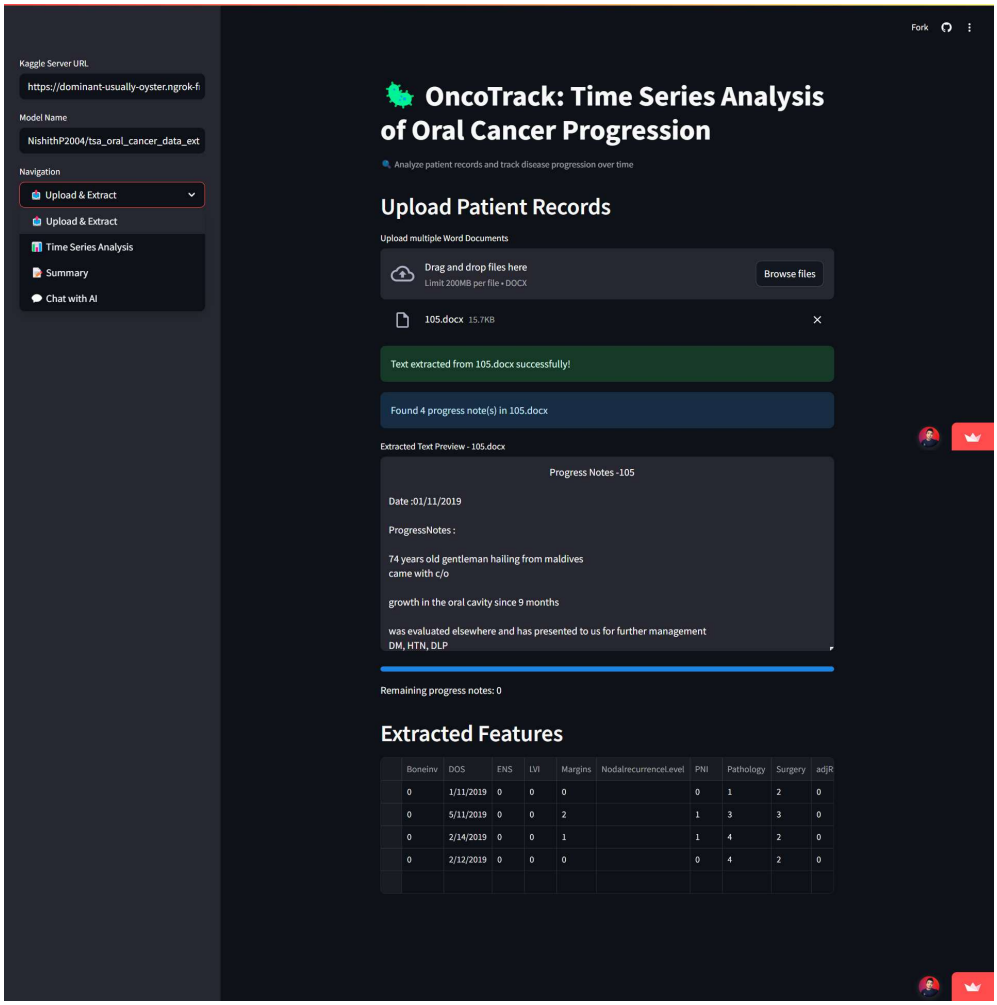


Figure 6 OncoTrack Streamlit User Interface Mockup.

precision, are summarized in Table 2. Accuracy denotes the proportion of evaluated field predictions that exactly matched the normalized reference value. Accepted-match precision denotes correctness among outputs that passed format checks and were accepted for field-level comparison; it should not be interpreted as conventional clinical precision across all possible positive and negative classes.

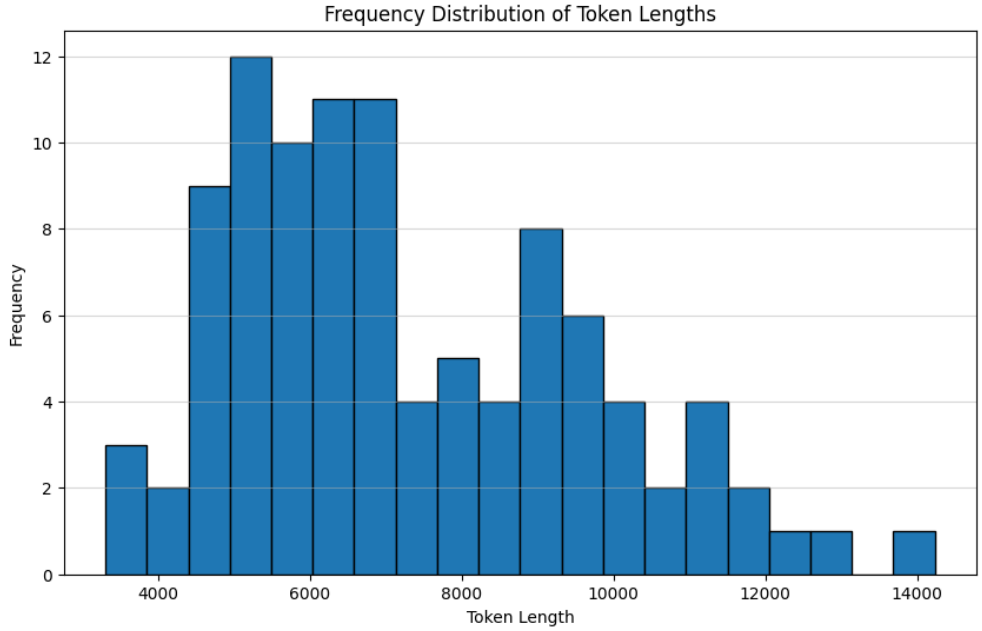


Figure 7 Frequency distribution of token lengths for the 100-document AIMS dataset using a Llama-family tokenizer.

Table 2 Performance Comparison of Fine-Tuned Llama-3.1-8B and Mistral-Nemo-2407 Models on the 10-Document Test Set

Feature	2 Epochs				5 Epochs				10 Epochs			
	Llama-3.1-8B		Mistral-Nemo		Llama-3.1-8B		Mistral-Nemo		Llama-3.1-8B		Mistral-Nemo	
	Acc.	Prec.	Acc.	Prec.	Acc.	Prec.	Acc.	Prec.	Acc.	Prec.	Acc.	Prec.
age	0.60	0.55	0.50	0.40	0.50	0.50	0.70	0.65	1.00	1.00	0.60	0.55
sex	0.90	0.81	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
subsite	0.40	0.55	0.60	0.90	0.50	0.33	0.90	0.95	0.70	0.75	0.90	0.90
Pathology	0.80	0.80	0.70	0.79	0.80	0.80	0.80	0.80	0.90	0.91	0.80	0.80
clinicalT8thed	0.30	0.63	0.50	0.63	0.50	0.63	0.50	0.36	1.00	1.00	0.70	0.85
nodal8thed	0.40	0.41	0.60	0.85	0.50	0.53	0.60	0.83	0.50	0.43	0.70	0.93
pt8thed	0.50	0.51	0.50	0.60	0.70	0.67	0.60	0.51	0.60	0.73	0.50	0.67
pN8thed	0.40	0.80	0.40	0.73	0.60	0.60	0.90	0.82	0.80	0.72	0.70	0.63
PNI	0.60	0.60	0.90	1.00	0.90	0.92	0.90	0.92	0.90	0.92	0.90	0.92
LVI	0.80	0.80	0.70	0.78	1.00	1.00	0.90	0.93	0.90	0.93	0.90	0.93
Margins	0.80	0.80	0.80	0.80	0.90	0.81	0.90	0.81	0.90	0.90	0.90	0.90
Boneinv	0.90	0.95	0.80	0.95	0.90	0.95	0.90	0.95	0.90	0.95	0.90	0.95
skininv	1.00	1.00	0.90	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
ENS	0.70	0.81	0.50	0.28	0.40	0.38	0.80	0.80	0.90	0.92	0.80	0.80
Surgery	0.50	0.92	0.50	0.90	0.90	0.81	0.90	0.81	0.90	0.81	0.80	0.80
lateralitysurgery	0.10	0.01	0.20	0.91	0.90	0.95	1.00	1.00	1.00	1.00	1.00	1.00
DOS	0.70	0.70	0.30	0.30	0.70	0.70	0.60	0.60	1.00	1.00	0.60	0.60
adjRx	0.80	0.80	0.80	0.90	0.80	0.85	1.00	1.00	1.00	1.00	0.80	0.85
dateofLR	0.90	0.81	0.90	0.90	0.90	0.90	0.90	0.81	0.90	0.90	0.90	0.81
Nodalrecurrencelevel	N/A	N/A	N/A	N/A	0.90	0.81	0.90	0.81	0.90	0.81	0.90	0.81
nodallevel	0.90	0.81	0.80	0.80	0.90	0.81	0.90	0.81	0.90	0.81	0.90	0.81
dateofNodalrecurrence	0.90	0.81	0.80	0.80	0.90	0.81	0.90	0.81	0.90	0.81	0.90	0.81
lastFU	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
stateFU	0.60	0.36	0.70	0.65	0.60	0.45	0.80	0.65	0.70	0.67	0.80	0.65
Overall (Avg.)	**0.63**	**0.66**	**0.63**	**0.73**	**0.74**	**0.72**	**0.80**	**0.78**	**0.88**	**0.89**	**0.79**	**0.79**

N/A indicates the feature was not consistently present or predicted in the 2-epoch test runs.

A comparative analysis of the Llama and the Mistral models reveals distinct learning characteristics. While Mistral-Nemo demonstrated competitive, and in some metrics superior, performance at earlier training stages (e.g., higher overall accuracy at 5 epochs), Meta-Llama-3.1-8B benefited significantly more from extended training. The Llama model's performance surged after 10 epochs, achieving perfect scores on several key features like 'age', 'clinicalT8thed', and 'DOS', where Mistral's performance was notably lower. Conversely, Mistral maintained a more consistent, albeit lower, performance across epochs for complex categorical features like 'subsite' and 'nodal8thed'. This suggests that while Mistral may be a more stable learner with fewer training cycles, Llama-3.1 possesses a higher capacity for learning the specific nuances of this dataset when trained for longer. The shared, complete failure of both models on the 'lastFU' feature in this initial 10-document test set highlights the inherent difficulty of inferential tasks over simple extraction.

The Meta-Llama-3.1-8B model fine-tuned for 10 epochs demonstrated the best overall performance, with an average accuracy of $87.83\% \pm 1.81$ and an average precision of $86.80\% \pm 2.00$. This model was selected as the primary data extraction engine for OncoTrack. The results clearly show that performance improves with more training epochs, with the 10-epoch Llama model outperforming all other configurations.

4.2 Extended Evaluation on a Larger Held-Out Set

To address concerns regarding statistical power and cross-source robustness, the selected 10-epoch fine-tuned Meta-Llama-3.1-8B model was re-evaluated on the retained larger held-out set of 339 instances drawn from the expanded multi-source pool. For this analysis, accuracy was computed as the proportion of predictions matching the ground truth after normalization. Two normalization rules were applied during scoring: (i) 'age' predictions within ± 1 year of the ground truth were considered correct, and (ii) date-related fields were counted as correct if the predicted date matched the ground truth under any valid equivalent date format. Precision was defined for the information-extraction setting as $\mathrm{TP}/(\mathrm{TP} + \mathrm{FP})$, where true positives were cases in which the system extracted a value and that value matched the ground truth after the defined normalization or tolerance rules, and false positives were cases in which the system extracted a value but the value was incorrect, outside the acceptable tolerance window, or present when the ground truth was empty. Thus, precision measures the trustworthiness of the system when it decides to output a value, rather than the proportion of all records with complete clinical correctness. Because 'NodalrecurrenceLevel' and 'nodallevel' were absent in a large fraction of records, and 'lastFU' remained strongly inference-heavy, these fields were excluded from the aggregate expanded-set score; the resulting overall averages were 0.80 for accuracy and 0.73 for precision. Including all 24 fields reduced the average accuracy to 0.70.

Table 3 Feature-wise Performance of Fine-Tuned Meta-Llama-3.1-8B on 339 Held-Out Instances

Feature	**Accuracy**	**Precision**	**Feature**	**Accuracy**	**Precision**
age	0.96	0.96	skininv	0.97	0.97
sex	0.83	0.83	ENS	0.83	0.83
subsite	0.74	0.74	Surgery	0.70	0.70
Pathology	0.81	0.81	lateralitysurgery	0.83	0.83
clinicalT8thed	0.62	0.62	DOS	0.76	0.76
nodal8thed	0.58	0.58	adjRx	0.83	0.83
pt8thed	0.70	0.70	dateofLR	0.75	0.10
pN8thed	0.78	0.79	NodalrecurrenceLevel	0.00	0.00
PNI	0.83	0.83	nodallevel	0.00	0.00
LVI	0.92	0.92	dateofNodalrecurrence	0.88	0.17
Margins	0.85	0.85	lastFU	0.09	0.09
Boneinv	0.88	0.88	stateFU	0.74	0.74
Overall Avg. (excluding NodalrecurrenceLevel, nodallevel, lastFU)				**0.80**	**0.73**

As shown in Table 3, extraction remains strongest for explicitly stated fields (for example, 'age', 'skininv', and 'LVI'). Performance remains weaker for inference-heavy or sparse fields ('lastFU', 'NodalrecurrenceLevel', and 'nodallevel'), reinforcing that temporal reasoning and less frequent recurrence descriptors are still the main bottlenecks. The low precision for recurrence-date fields such as 'dateofLR' and 'dateofNodalrecurrence' indicates that, when the extraction logic produced a value for these fields, many outputs did not match the reference under the defined tolerance rules. In particular, the 'lastFU' field achieved only 0.00 accuracy in the initial 10-document test set and approximately 9.44% accuracy in the retained 339-instance evaluation, showing that the current pipeline does not yet solve complex last-follow-up inference. In the included 339 subset, there were 6 cases where prediction matched extracted text while ground truth did not (58 prediction-text matches and 52 triple matches).

4.3 Data Variability and Exclusion Analysis

The separate 335-record exclusion subset shows substantial heterogeneity in documentation quality and alignment between document text and manually curated labels. This is the principal reason these records were excluded from primary benchmarking.

Age-based audit on the excluded 335 records highlighted strong transcript inconsistency: ground-truth age mismatched extracted textual evidence in 191 cases, and age was absent in 106 records. At the same time, only 38 cases had ground-truth age aligned with extracted text, while 128 cases showed the opposite pattern (prediction aligned with extracted text but ground truth did not), indicating non-trivial annotation noise rather than purely model failure.

Clinical-stage consistency checks for 'clinicalT8thed' further reinforced dataset variability. Across all 674 processed documents, direct text-vs-ground-truth validation yielded 140 matches, 153 mismatches, and 381 records with no detectable clinicalT pattern in text. In the excluded 335 subset, there were 7 cases where prediction matched extracted text while ground truth did not, derived from 50 prediction-text matches and 43 triple matches.

For comparison with the retained evaluation subset, the included 339 records had 6 cases where prediction matched extracted text while ground truth did not, derived from 58 prediction-text matches and 52 triple matches. These mismatch-recovery pockets suggest that the model can sometimes flag potentially plausible values in places where the reference labels are inconsistent with source text, but they require expert adjudication before being treated as corrected labels.

4.4 Robustness Improvements for the Large-Set Evaluation

During large-set inference, five recurrent failure modes were observed: context-window overflow, schema breakdown (free-text paragraphs appearing as dictionary keys), occasional repetitive generation loops, key-name inconsistencies (for example, 'DOS' vs 'dos'), and rogue data types/blank rows. These issues were mitigated through the following updates:

1. **Increased context window:** The inference configuration was updated from `max_seq_length = 14336` to `max_seq_length = 33792` to avoid truncation of instruction and formatting prompts.
2. **Schema validation with retries:** A schema check was introduced to enforce the expected 24 output keys, with a retry mechanism (`max_retries = 2`) for malformed generations. Samples that failed all retries were skipped instead of injecting structurally invalid rows into the final analysis. Because skipped samples were not counted as field-level errors in the reported retained-set table, the expanded-set scores may overestimate performance relative to a strict all-generations analysis.

These safeguards improved output consistency and enabled stable evaluation on the larger cohort while preserving strict downstream tabulation requirements. Future reporting should include valid-output rate, retry counts, skipped-sample counts, and sensitivity analyses that count unrecoverable malformed generations as errors.

4.5 Qualitative Analysis of Data Extraction

A qualitative review of the model's predictions on the test set provides deeper insights into its capabilities and limitations:

- **Categorical Feature Performance:** The model excelled at extracting categorical features that had clear textual indicators in the notes. Features like 'sex', 'PNI', 'LVI', and 'Margins' consistently achieved high accuracy and precision, demonstrating the model's ability to map textual evidence to predefined categories correctly.
- **Numerical and Date Extraction:** The model exhibited some difficulty with precise numerical values, particularly for 'age' and date-related features like 'DOS' and 'lastFU'. Often, the predicted age or date was off by ± 1 unit (e.g., predicting age 54 when it was 55, or DOS 3/12/2018 instead of 3/13/2018). To account for clinically acceptable near-matches, the expanded evaluation used a tolerance of ± 1 year for 'age' and normalized matching for equivalent date formats.
- **Implicit Information and Complex Inference:** The most significant challenge was the extraction of implicit information. The 'lastFU' (Date of Last Follow-Up) feature remained the most difficult (9.44% accuracy in the 339-instance evaluation), because the model must infer the latest follow-up date from multiple chronologically distributed notes rather than extract a directly labeled value. In contrast, explicitly mentioned date fields such as 'dateofNodalrecurrence' improved substantially after normalization-based matching.
- **Staging-Specific Error Modes in the 10-Document Test Set:** The relatively low 10-epoch Llama accuracy for 'pt8thed' (0.60) and 'nodal8thed' (0.50) reflects the fact that these fields often require more than direct categorical extraction. For 'pt8thed', the target mapping requires explicit AJCC 8th-edition pathological T substages ('T1', 'T2', 'T3', 'T4a', or 'T4b'), but several source documents record advanced tumors only as 'T4', such as 'Stage - pT4N3b' or 'Stage: p T4N2b', without specifying 'T4a' versus 'T4b'. In such cases, the model is forced to choose between two mapped values without sufficient textual evidence. Additional errors arose from intra-document contradictions, such as discharge summaries listing 'pT3?T4aN2bM0' while final histopathology reports state 'pTNM stage: pT3N2b', and from multiple

tumor timelines in the same record, where an original primary tumor, a recurrence, and a second primary tumor each carry different stage descriptions. For 'nodal8thed', the main challenge was distinguishing the required clinical N stage from more prominent pathological N-stage statements. Clinical staging is frequently embedded mid-paragraph as a provisional impression (for example, '?cT1N0Mx'), whereas pathological staging is often presented in clearer final-report headers, causing the model to select the wrong stage source. A further source of error is the transition from AJCC 7th- to 8th-edition criteria: older records from 2011-2015 may explicitly state stages such as 'pT3N2cMx', while extranodal extension or perinodal spread would require reinterpretation as 'N3b' under 8th-edition rules. These cases demonstrate that the lower scores for 'pt8thed' and 'nodal8thed' are driven by missing sub-classifications, conflicting chronology, and edition-specific clinical inference rather than simple failure to recognize stated labels.

- **Data Quality Review:** An unexpected observation was the model's potential to flag annotation–text inconsistencies for human review. In several test cases (approximately 9/10), the model's prediction for features like 'clinicalT8thed' and 'nodal8thed' differed from the ground truth. Upon manual re-verification of the source document, the model's prediction appeared better supported by the available text, while the original manual annotation appeared inconsistent. For example, in one of the test documents, the 'clinicalT8thed' had a ground truth value of $T4a$, but the manual extraction had a value of $T3$. The algorithm predicted the text-supported value. This finding should be interpreted cautiously: the review was not an independent blinded dual-expert adjudication, and agreement with a local text span does not guarantee that the model selected the correct time point, tumor event, or staging system.
  In the retained 339-record subset, 'clinicalT8thed' mismatch-recovery was observed in 6 cases where prediction matched text but ground truth did not.
  In the excluded 335-record subset, the expanded audit showed 128 cases where age predictions aligned with extracted text while the ground truth value did not align. For 'clinicalT8thed', mismatch-recovery was observed in 7 cases where prediction matched text but ground truth did not.

These findings indicate that while the fine-tuned LLM is proficient at structured data extraction based on explicit textual cues, its ability to perform complex numerical, staging-context, and temporal inference needs further development. They also show that noisy reference labels and heterogeneous documentation can materially affect evaluation, motivating future use of independent re-annotation and adjudication.

4.6 Limitations

Several limitations should be considered when interpreting these results. First, the initial fine-tuning cohort was small (100 documents with a 10-document test split), so the early performance estimates may be sensitive to overfitting and sampling variability. The later 339-instance retained evaluation improves statistical coverage but does not eliminate this concern because 335 of 674 expanded records were excluded from headline benchmarking due to text-label misalignment or missing evidence.

These excluded records represent real-world documentation heterogeneity; therefore, retained-set performance should not be interpreted as all-comers clinical performance.

Further, in many of the excluded documents, many of the fields are not present at all. Consequently, we have no way of evaluating if the ground truth provided to us is correct when it comes to these fields. In documents where 5 or more fields did not match the ground truth, we excluded them.

Second, the reported precision values in the expanded held-out evaluation use an information-extraction definition, $\mathrm{TP}/(\mathrm{TP} + \mathrm{FP})$, where the denominator includes instances in which the system extracted a value. They should be read alongside field-level accuracy, sparse-field failures, and the all-field average accuracy. In future studies, we will report additional measures such as valid-output rate, field-level coverage, exact-match accuracy without tolerance, recall, F1-score where class structure permits, and sensitivity analyses that count skipped or malformed generations as errors.

Third, the output format in the current implementation is a semicolon-delimited key–value string with downstream schema checks and retries. Although this was sufficient for the present prototype, a stricter structured-output interface such as JSON schema validation or Pydantic-style validation would be preferable for production-grade tabulation and clearer accounting of malformed outputs.

Fourth, the summarization, RAG-based chat, multilingual speech, translation, and TTS components were not independently evaluated for factual consistency, hallucination rate, retrieval grounding, cross-session isolation, or speech/translation accuracy. They are therefore described as prototype components, not as validated clinical decision-support functions.

Finally, the deployment stack involving Kaggle Notebook, Ngrok, ChromaDB, Whisper, and Google TTS was used for research prototyping. Clinical deployment would require institutional infrastructure, access control, encryption, audit logging, retention and deletion controls, vector-store isolation, third-party data-processing review, and prospective security/privacy validation.

5 Conclusion and Future Work

5.1 Conclusion

The OncoTrack project demonstrates the feasibility of using fine-tuned, open-source Large Language Models for automated extraction of oral-cancer progression variables from unstructured records. In the initial curated AIMS cohort (100 documents; 90:10 split), the LoRA-fine-tuned Meta-Llama-3.1-8B model achieved approximately $87.83\% \pm 1.81$ accuracy and $86.80\% \pm 2.00$ accepted-match precision, confirming strong performance for structured extraction under controlled conditions.

In the expanded multi-source evaluation (674 processed records), 339 instances were retained for primary benchmarking and yielded an overall average normalized field accuracy of 79.97% and precision of 73.45% after excluding sparse/inference-heavy fields ('NodalrecurrenceLevel', 'nodallevel', and 'lastFU'); across all 24 fields, average accuracy was 70.37%. The precision values in the expanded table use $\mathrm{TP}/(\mathrm{TP} + \mathrm{FP})$ over instances where the system extracted a value and should not be interpreted as evidence of complete clinical reliability.

In the initial dataset, 'lastFU' was present in the ground truth provided separately to us but was not present in the documents scanned by the LLM. In other words, the 'lastFU' was manually added to the ground truth but were not manually taken from the documents themselves. The documents themselves do not contain the 'lastFU' date. For most of the documents in the expanded dataset, the 'lastFU' date was not present in the documents.

The remaining 335 records were excluded from headline scoring because transcript evidence was frequently inconsistent with annotation labels (for example, substantial age mismatches and missing age evidence), highlighting real-world data heterogeneity and potential selection bias in retained-set estimates. A key practical outcome is that the model may help flag annotation–text inconsistencies for subsequent human review: beyond the manually reviewed test cases (approximately 9/10) where model outputs appeared better supported by source text than the annotations, the expanded audit showed systematic mismatch-recovery patterns (including 128 age cases where prediction aligned with extracted text while ground truth did not). Overall, OncoTrack offers a practical framework for scalable clinical-data structuring while also exposing annotation quality issues that are critical for reliable deployment.

5.2 Future Work

Future development of the OncoTrack project will focus on addressing current limitations and expanding its capabilities.

1. **Enhancing Model Performance:**
   - **Improved Inference:** We will explore multi-step reasoning pipelines and more sophisticated prompt engineering techniques. For example, to extract the 'lastFU', the model could first be prompted to extract all dates and their contexts, and a second prompt could ask it to identify the latest one.
   - **Model Comparison and Ablation:** We will evaluate zero-shot and few-shot prompting, encoder-based clinical NLP baselines, and newer open-weight LLMs as they become available and more accessible, comparing their performance on the same benchmark. We will also perform LoRA hyperparameter ablations, including rank, learning rate, target modules, and epoch count.
   - **Structured Output Validation:** We will replace semicolon-delimited generation with stricter JSON schema validation or Pydantic-style field validation to improve reproducibility and provide clearer reporting of malformed outputs.
2. **Dataset Expansion and Generalizability:** To enhance robustness, the training dataset will be expanded with more patient documents from diverse sources. This will help the model generalize better to different writing styles and clinical settings. Future datasets should include explicit labels for "not documented", "insufficient evidence", and "conflicting evidence" rather than forcing deterministic mappings for unsupported fields.
3. **Reference-Standard Adjudication:** Future evaluation will use independent dual clinical review, adjudication of disagreements, and inter-rater agreement

reporting, especially for staging fields, multiple tumor timelines, recurrence events, and follow-up dates.

4. **Clinical Integration and Evaluation:** A crucial next step is to conduct a formal clinical validation study. This will involve oncologists using the OncoTrack system in a controlled real-world setting to evaluate its usability, impact on workflow efficiency, and the clinical relevance of the generated summaries and extracted data. The summarization and RAG/chat modules will require separate human evaluation for factual consistency, grounding, hallucination, and safety before any clinical use.
5. **Security and Privacy Validation:** Before deployment, the prototype stack must be moved to institutionally governed infrastructure with authentication, encryption, access logs, retention/deletion policies, vector-store isolation, third-party processing review, and audit procedures for text, embeddings, summaries, audio, and chat logs.
6. **Multimodal Data Integration:** Expand beyond text to incorporate radiology reports, pathology images, genomic profiles, and laboratory data. Develop pipelines that can link structured and unstructured data for richer patient profiles and more comprehensive clinical insights.
7. **Interoperability with Clinical Systems:** Seamless integration with electronic health records, tumor board platforms, and cancer registries. Enable bi-directional exchange of structured data to support both clinical decision-making and research use cases.
8. **Real-Time Clinical Decision Support:** Extend functionality from retrospective data summarization to prospective decision support (eg. treatment recommendations, trial eligibility screening). Pilot OncoTrack as a digital assistant in multidisciplinary tumor boards.

By pursuing these avenues, OncoTrack can evolve from a proof-of-concept project into a powerful and indispensable tool for the oncology community, ultimately contributing to improved patient outcomes.

Conflict of Interest Statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author Contributions

NP, NKS, KT, LRM, and SK contributed to the conception and design of the study. NP implemented the model pipeline, conducted experiments, analyzed results, and drafted the manuscript. NKS supervised the technical methodology and evaluation. KT, LRM, and SK contributed clinical expertise, data interpretation, and manuscript review. All authors contributed to manuscript revision and approved the submitted version.

Funding

The authors declare that no specific funding was received for this work.

Acknowledgments

The authors thank Amrita Institute of Medical Sciences and Amrita Vishwa Vidyapeetham for institutional support during the development and evaluation of OncoTrack.

Data Availability Statement

The clinical datasets analyzed for this study are not publicly available because they contain sensitive patient information and are subject to institutional ethics and privacy restrictions. De-identified aggregate results supporting the conclusions are included in the manuscript.

Ethics Statement

The study protocol was reviewed and approved by the Ethics Committee of Amrita School of Medicine, Amrita Vishwa Vidyapeetham (Approval No.: ECASM-AIMS-2025-405; Date: November 6, 2025). The requirement for written informed consent was waived because this was a retrospective analysis of de-identified patient records.

References

[1] Thankappan, K. Artificial intelligence and machine learning in head and neck oncology. *J. Head Neck Physicians Surg.* **10**, 117–120 (2022).

[2] Chen, D., Alnassar, S. A., Avison, K. E., Huang, R. S. & Raman, S. Large language model applications for health information extraction in oncology: Scoping review. *JMIR Cancer* **11**, e65984 (2025). URL https://doi.org/10.2196/65984.

[3] Zając, H. D., Avlona, N. R., Kensing, F., Andersen, T. O. & Shklovski, I. Ground truth or dare: Factors affecting the creation of medical datasets for training ai. *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society* 351–362 (2023).

[4] Thankappan, K. *et al.* Oral cancer screening using mobile phone-based(mhealth) approach versus conventional oral examination approach, protocol of a cluster randomized study with cost-effectiveness analysis. *Int. J. Surg. Protoc.* **23**, 1–5 (2020).

[5] Mirian, C. *et al.* Metrics of pn-staging in oral squamous cell carcinoma: An analysis of 1,905 patients. *Eur. J. Cancer* **150**, 33–41 (2021).

[6] Kanemaru, N., Yasaka, K., Fujita, N., Kanzawa, J. & Abe, O. The fine-tuned large language model for extracting the progressive bone metastasis from unstructured radiology reports. *Journal of Imaging Informatics in Medicine* **38** (2024).

[7] Gema, A. P., Minervini, P., Daines, L., Hope, T. & Alex, B. Parameter-efficient fine-tuning of llama for the clinical domain (2024). URL https://arxiv.org/abs/2307.03042. arXiv:2307.03042.

[8] Guluzade, A. *et al.* Elmtex: Fine-tuning large language models for structured clinical information extraction. a case study on clinical reports (2025). URL https://arxiv.org/abs/2502.05638. arXiv:2502.05638.

[9] Pal, A., Umapathi, L. K. & Sankarasubbu, M. Jiang, J., Reitter, D. & Deng, S. (eds) *Med-HALT: Medical domain hallucination test for large language models.* (eds Jiang, J., Reitter, D. & Deng, S.) *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, 314–334 (Association for Computational Linguistics, Singapore, 2023). URL https://aclanthology.org/2023.conll-1.21/.

[10] Wang, Y. *et al.* Clinical information extraction applications: A literature review. *J. Biomed. Inform.* **77**, 34–49 (2018).

[11] Wu, S. *et al.* Deep learning in clinical natural language processing: a methodical review. *J. Am. Med. Inform. Assoc.* **27**, 457–470 (2020).

[12] Savova, G. K. *et al.* Use of natural language processing to extract clinical cancer phenotypes from electronic medical records. *Cancer Res.* **79**, 5463–5470 (2019).

[13] Lo Vercio, L. *et al.* Supervised machine learning tools: a tutorial for clinicians. *Journal of Neural Engineering* **17**, 062001 (2020). URL https://dx.doi.org/10.1088/1741-2552/abbff2.

[14] Zhao, Z., Jin, Q., Chen, F., Peng, T. & Yu, S. A large-scale dataset of patient summaries for retrieval-based clinical decision support systems. *Sci. Data* **10**, 909 (2023).

[15] Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Burstein, J., Doran, C. & Solorio, T. (eds) *BERT: Pre-training of deep bidirectional transformers for language understanding.* (eds Burstein, J., Doran, C. & Solorio, T.) *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)*, 4171–4186 (Association for Computational Linguistics, Minneapolis, Minnesota, 2019). URL https://aclanthology.org/N19-1423/.

[16] Alsentzer, E. *et al.* Rumshisky, A., Roberts, K., Bethard, S. & Naumann, T. (eds) *Publicly available clinical BERT embeddings.* (eds Rumshisky, A., Roberts, K., Bethard, S. & Naumann, T.) *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, 72–78 (Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019). URL https://aclanthology.org/W19-1909/.

[17] Zhou, S., Wang, N., Wang, L., Liu, H. & Zhang, R. Cancerbert: a cancer domain-specific language model for extracting breast cancer phenotypes from electronic health records. *Journal of the American Medical Informatics Association* **29**, 1208–1216 (2022). URL https://doi.org/10.1093/jamia/ocac040.

[18] Kehl, K. L. *et al.* Assessment of deep natural language processing in ascertaining oncologic outcomes from radiology reports. *JAMA Oncol.* **5**, 1421–1429 (2019).

[19] Hu, E. J. *et al.* Lora: Low-rank adaptation of large language models. *ICLR* **1**, 3 (2022).

[20] Goswami, J., Prajapati, K. K., Saha, A. & Saha, A. K. Parameter-efficient fine-tuning large language model approach for hospital discharge paper summarization. *Applied Soft Computing* **157**, 111531 (2024). URL https://www.sciencedirect.com/science/article/pii/S1568494624003053.

[21] Brown, T. *et al.* Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. & Lin, H. (eds) *Language models are few-shot learners.* (eds Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. & Lin, H.) *Advances in Neural Information Processing Systems*, Vol. 33, 1877–1901 (Curran Associates, Inc., 2020). URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[22] Adams, L. C. *et al.* Leveraging GPT-4 for post hoc transformation of free-text radiology reports into structured reporting: A multilingual feasibility study. *Radiology* **307**, e230725 (2023).

[23] Choi, H. S., Song, J. Y., Shin, K. H., Chang, J. H. & Jang, B.-S. Developing prompts from large language model for extracting clinical information from pathology and ultrasound reports in breast cancer. *Radiat Oncol J* **41**, 209–216 (2023). URL http://www.e-roj.org/journal/view.php?number=1575.

[24] Hu, D., Liu, B., Zhu, X., Lu, X. & Wu, N. Zero-shot information extraction from radiological reports using chatgpt. *International Journal of Medical Informatics* **183**, 105321 (2024). URL https://www.sciencedirect.com/science/article/pii/S1386505623003398.

[25] Truhn, D. *et al.* Extracting structured information from unstructured histopathology reports using generative pre-trained transformer 4 (gpt-4). *The Journal of Pathology* **262**, 310–319 (2024). URL https://pathsocjournals.onlinelibrary.wiley.com/doi/abs/10.1002/path.6232.

[26] Huang, J. *et al.* A critical assessment of using ChatGPT for extracting structured data from clinical notes. *NPJ Digit. Med.* **7**, 106 (2024).

[27] Wang, J. *et al.* Prompt engineering for healthcare: Methodologies and applications (2024). URL https://arxiv.org/abs/2304.14670. arXiv:2304.14670.

[28] Liu, N. F. *et al.* Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics* **12**, 157–173 (2024). URL https://aclanthology.org/2024.tacl-1.9/.

[29] Tan, R. S. Y. C. *et al.* Inferring cancer disease response from radiology reports using large language models with data augmentation and prompting. *J. Am. Med. Inform. Assoc.* **30**, 1657–1664 (2023).

[30] Team, G. Gemma (2024). URL https://www.kaggle.com/m/3301.

[31] Meta. Llama-3.1-8b (2024). URL https://huggingface.co/meta-llama/Llama-3.1-8B.

[32] Mistral AI Team. Model card for mistral-nemo-base-2407 (2024). URL https://huggingface.co/mistralai/Mistral-Nemo-Base-2407.

[33] Team, G. Gemma 3 (2025). URL https://goo.gle/Gemma3Report.

[34] LMArena. https://lmarena.ai/. LMArena (formerly Chatbot Arena) is a public, web-based platform that evaluates large language models (LLMs) through anonymous, crowd-sourced pairwise comparisons. Users enter prompts for two anonymous models to respond to and vote on the model that gave the better response, in which the model's identities are revealed. Users can also choose models to test themselves.

[35] Radford, A. *et al.* Robust speech recognition via large-scale weak supervision (2022). URL https://arxiv.org/abs/2212.04356.

[36] Nussbaum, Z., Morris, J. X., Duderstadt, B. & Mulyar, A. Nomic embed: Training a reproducible long context text embedder (2024). arXiv:2402.01613.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- AdobeScan13Nov2025.pdf