LLM-Driven Clinical Trial Matching for Lung Cancer Patients: An Explainable Approach

Vittoria Peppoloni[1,*], Giuseppe Leone[1], Laura Mazzeo[1,2], Alberto Ferrarin[1,2], Vanja Miskovic[1,2], Giuseppe Lo Russo[1], Paolo Baili[3], Arsela Prelaj[1,2] and Federica Corso[1,*]

[1]*Medical Oncology Department 1, Fondazione IRCCS Istituto Nazionale dei Tumori di Milano, Milan, Italy*
[2]*Department of Electronic, Information and Bioengineering, Politecnico di Milano, Milan, Italy*
[3]*Department of Epidemiology and Data Science, Fondazione IRCCS Istituto Nazionale dei Tumori di Milano, Milan, Italy*

Abstract

Matching lung cancer patients to clinical trials remains a labor-intensive and error-prone process. We present MedMatch, an explainable, on-premises system leveraging Large Language Models (LLMs) for automated patient-trial matching. We evaluated 35 lung cancer patients from IRCCS Istituto Nazionale dei Tumori, achieving 80% trial matching accuracy. As part of the pipeline, the system first extracts 12 clinical parameters from Electronic Health Records with 84.9% overall accuracy using LLaMA 3.1 8B, then performs trial matching using Gemma 3 27B. Performance varied from perfect for demographics (100%) to more challenging for complex features such as mutations (71%) and line of therapy (77% accuracy, 36% F1-score). Benchmarking on a 10-patient subset showed that LLaMA 3.1 8B outperformed three alternative models, including the domain-specific MedLLaMA 2 (87.4% vs. 50% accuracy). Incorporating few-shot prompting with oncologist-curated examples further improved LLaMA's performance. However, hallucination analysis revealed unreliable behavior in cases with missing data, with hallucination rates ranging from 15.4% for previous treatments to 100% for ECOG status. MedMatch addresses these challenges by combining structured extraction with layered explainability through JSON outputs, eligibility justifications, and PDF evidence highlighting, while preserving data privacy via local deployment.

Keywords

Large Language Models, Electronic Health Records, Explainability, Clinical Trial Matching

1. Introduction

Lung cancer is biologically complex, with high molecular and phenotypic heterogeneity that demands increasingly personalized therapeutic approaches. Clinical trials are critical to advancing treatment, yet identifying eligible patients remains inefficient: clinicians must manually review patient records against complex eligibility criteria, a process that is both time-consuming and error-prone.

Several digital solutions have attempted to address this challenge. Early rule-based systems required structured inputs and failed to capture the narrative richness of real-world records. Commercial platforms such as IBM Watson for Oncology, ClinicalTrials.ai, and Mendel.ai have explored natural language processing approaches, but they have faced criticism for limited validation, lack of transparency, and privacy concerns due to cloud-based deployment. These shortcomings underline the need for accurate, explainable, and privacy-preserving systems.

Large Language Models (LLMs) represent a transformative opportunity. They can extract nuanced information from unstructured text and support clinical decision-making [1], but their use in trial matching raises key challenges: ensuring accuracy, avoiding hallucinations, maintaining explainability, and protecting sensitive data.

In this work we present MedMatch, an on-premises, explainable system for processing Electronic Health Records (EHRs) of lung cancer patients. MedMatch extracts clinically relevant features, matches patients to appropriate trials, and provides transparent explanations with visual evidence highlighting.

EXPLIMED 2025 - Second Workshop on Explainable Artificial Intelligence for the medical domain - 25-30 October 2025, Bologna, Italy
*Corresponding author.
vittoria.peppoloni01@gmail.com (V. Peppoloni); federica.corso@istitutotumori.mi.it (F. Corso)

By combining accuracy, interpretability, and local deployment, our system addresses the key limitations of existing approaches.

2. Related Work

Automated trial matching has been studied through commercial and academic efforts. Rule-based systems such as the NCI Clinical Trials Search required manual data entry, while IBM Watson for Clinical Trial Matching attempted to use NLP but was discontinued after limited clinical success. More recent platforms (Mendel.ai, Deep 6 AI) improve automation but remain cloud-based and opaque, offering little transparency or benchmarking [2, 3].

Academic research has explored machine learning and LLM approaches. BERT-based models achieved moderate accuracy but required extensive manual annotation, while GPT-3 and GPT-4 have been tested for criteria extraction, with promising but incomplete results, particularly regarding hallucination control and explainability [4, 5].

Beyond trial matching, LLMs have demonstrated strong potential in healthcare tasks such as note summarization, diagnostic support, and information extraction. However, studies also document limitations including bias and high hallucination rates, underscoring the need for rigorous validation and interpretable outputs. Recent work on explainability has explored both traditional methods (LIME, SHAP) and prompt-based strategies such as chain-of-thought and structured explanation templates [6]. Building on these foundations, MedMatch introduces schema-constrained extraction and evidence highlighting, ensuring traceability from raw EHR text to eligibility decision.

Table 1 summarizes how MedMatch differs from representative systems. Unlike commercial cloud-based tools, it is deployed on-premises, enforces schema-driven feature extraction, and provides multi-level explainability through JSON outputs, criteria analysis, and PDF highlighting.

System	Deployment	Explainability	Privacy	Benchmarking
IBM Watson Oncology	Cloud	Limited	No	No
Mendel.ai	Cloud	Black-box	No	Limited
Deep 6 AI	Cloud	Minimal	No	Limited
MedMatch (ours)	**On-premises**	**JSON + criteria + highlight PDF**	**Yes**	**LLM benchmarking**

Table 1
Comparison of different clinical AI systems.

3. System Architecture

MedMatch employs a modular pipeline architecture with four components:

- a **database management** system managing the collection of clinical trials, providing efficient storage and retrieval mechanisms through SQLAlchemy ORM [7] with a PostgreSQL [8] backend, as configured through Flask's application context.
- an **input data processing** module handling document parsing, text normalization, and preparation for LLM processing. It accepts both PDFs and text entries, leveraging pdfplumber [9] for text extraction from PDF documents.
- a **feature extraction** engine leveraging LLaMA 3.1 8B to identify clinically relevant information from unstructured text, converting narrative clinical descriptions into a structured JSON format that captures key clinical attributes, while highlighting in the original PDF the source text segments used for feature extraction.

- a **trial matching** and explanation module with Gemma 3 27B matching the extracted patient features against the eligibility criteria of the available trials, generating multi-level explanations for each match recommendation.

These components are seamlessly integrated within a web-based interface, enabling clinicians to interact with the system, visualize matching results, and explore detailed explanations. MedMatch is implemented as a Flask web application, utilizing an Ollama server for LLM deployment. The model operates locally, ensuring data privacy by avoiding any external data transmission, while GPU acceleration provides fast and efficient processing.

3.1. Feature Extraction Pipeline

The feature extraction module processes unstructured EHR narratives to extract twelve clinically relevant parameters:

- **Demographics**: Age, Gender
- **Disease characteristics**: Diagnosis, Stage
- **Performance status**: ECOG Performance Status, PD-L1 expression
- **Genomic alterations**: Mutations
- **Metastatic involvement**: Brain Metastasis
- **Therapies**: Line of Therapy, Previous Treatments, Concomitant Treatments
- **Comorbidities**

Feature extraction relies on structured prompting with explicit schema definitions, guiding the language model to produce a standardized JSON object. The schema enforces strict constraints (e.g., `"stage"` limited to I–IV) while allowing the use of `"not mentioned"` when information is absent. For transparency, the LLM also records the source text for each feature, enabling direct verification within the original PDF (Figure 1). All attributes are semantically aligned with trial eligibility requirements: mutation data are restricted to actionable genes (e.g., *EGFR*, *KRAS*, *MET*), PD-L1 expression is categorized by clinical cutoffs (<1%, 1–49%, >50%), and previous treatments are constrained to those relevant for lung cancer trials. By returning only the JSON object, the system minimizes hallucinations and extraneous content, ensuring robust, machine-readable outputs for downstream matching.

3.2. Trial Matching and Explainability

In the trial matching step, the LLM utilizes the JSON representation of a patient's features to compare them against the inclusion and exclusion criteria of each trial in the database. For efficiency, the list of available trials is processed in batches, a parameter that can be adjusted through system settings. For each trial, the system generates a JSON output containing:

- Trial ID and title retrieved from the database
- A match score (0-100), calculated using a scoring system that starts with 100 points and subtracts 10 points with each unmet inclusion criterion and subtracts 50 points with each violated exclusion criterion.
- An overall recommendation ("Eligible" or "Not Eligible"), determined by a threshold score of 70.
- A detailed criteria analysis, specifying which inclusion and exclusion criteria were met or violated.
- A brief summary explanation clarifying the reasoning behind the eligibility decision.

Trials with a match score above 70 are displayed on the platform in ascending order.

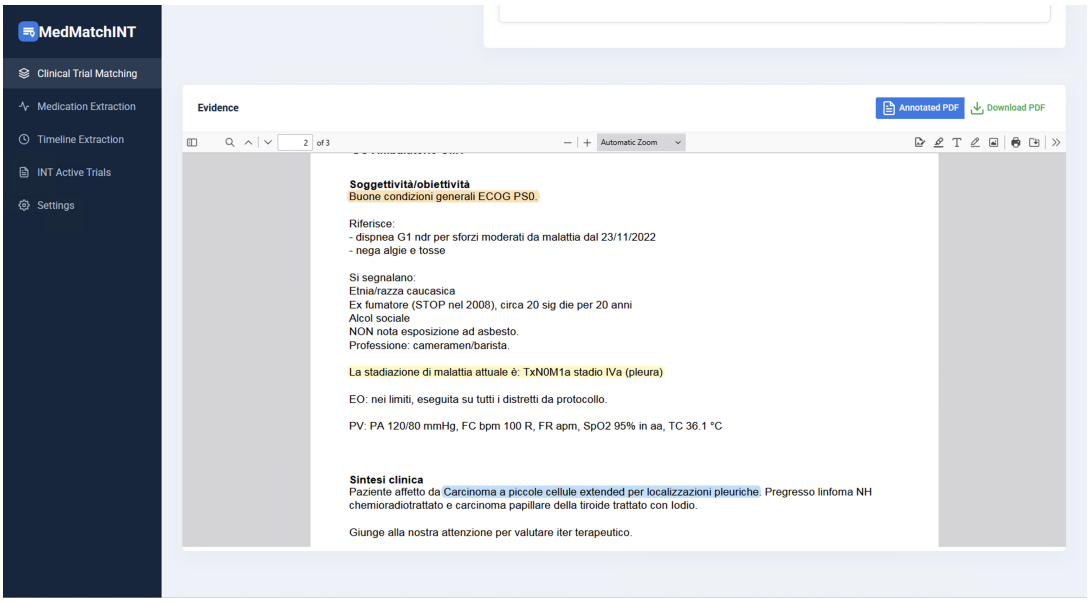

Figure 1: Example of the highlight functionality. Colored overlays on the patient's clinical PDF indicate the exact evidence supporting extracted features such as diagnosis and staging classifications.

4. Materials and Methods

4.1. Study Cohorts

We analyzed EHRs from 35 lung cancer patients recruited at IRCCS Istituto Nazionale dei Tumori (INT), divided into:

- **Benchmarking set**: 10 patients for model comparison and prompt optimization
- **Validation set**: 25 patients for full pipeline evaluation

The trial database comprised 17 lung cancer studies currently active at our institution, programmatically retrieved from ClinicalTrials.gov [10] using their NCT identifiers. For each trial, key information (phase, title, description, eligibility criteria, demographic restrictions, and recruitment status) was extracted, standardized, and stored in a local relational database. This resource formed the basis for trial matching and was also made accessible through a dedicated page within the web application.

4.2. Experimental Design

The study was organized as a series of targeted experiments, each designed to isolate and assess a specific component of the pipeline. We began by benchmarking four LLMs for feature extraction (LLaMA 3.1 8B, DevStral 24B, Mistral 7B, and MedLLaMA 2) on the 10-patient set, in order to compare general-purpose and domain-specific models. Based on these results, the best model was further evaluated with two prompting strategies: zero-shot and few-shot prompting, the latter enriched with examples crafted by an oncologist.

Once the optimal feature extraction setup was identified (LLaMA 3.1 8B with few-shot prompting), we assessed trial matching by comparing Gemma 3 27B and Qwen 2.5 32B on the benchmarking set. The best-performing configuration (LLaMA + Gemma) was then validated on the 25-patient validation set to evaluate end-to-end performance in a realistic setting.

To better understand system behavior, we conducted two additional analyses. First, an ablation study removed the feature extraction step, directly matching trials from raw PDFs to assess the added value of structured extraction. Second, a hallucination analysis examined whether the model correctly reported features as "not mentioned" rather than generating unsupported outputs.

This ablation design allowed us to progressively benchmark components, optimize prompting, validate the integrated pipeline, and probe system robustness under challenging conditions.

4.3. Evaluation Methodology

The evaluation framework assessed both the accuracy of feature extraction and the trial matching performance. Feature extraction was evaluated by comparing LLM-extracted features with expert-annotated ground truth data. Trial matching performance was measured by comparing the system's eligibility decisions with expert assessments, focusing on both the accuracy of predictions and the clarity of generated explanations.

5. Results

5.1. LLM Benchmarking for Feature Extraction

We compared the performance of four privacy-preserving LLMs. The heatmap below (Figure 2) shows comparative performance across them on the 10-patient benchmarking set. LLaMa 3.1 8B achieved the highest average accuracy (87.4%), followed by Mistral 7B (84.1%), DevStral 24B (74.1%), and MedLLaMa 2 (50%).

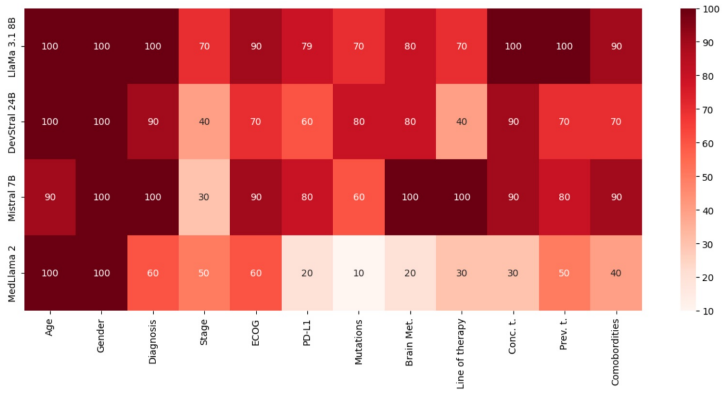

Figure 2: Comparison of the four models in extracting all the features.

5.2. Prompt Engineering Impact

The radar plot (Figure 3) compares zero-shot versus few-shot prompting strategies on the benchmarking cohort using LLaMA 3.1 8B. Few-shot prompting, incorporating oncologist-crafted examples, improved overall accuracy from 80.8% to 87.4% (+6.6% relative improvement). The most substantial gains were observed in complex features: metastases detection improved from 70 to 80 (+10%), concomitant treatment from 60 to 100 (+40%), and previous treatments from 70 to 100 (+30%). Simple demographic features showed minimal improvement.

5.3. Trial Matching Model Comparison

With LLaMA 3.1 8B established as the feature extractor, we compared Gemma 3 27B and Qwen 2.5 32B as trial matching engines on the 10-patient benchmarking set. By fixing the extracted features, the evaluation isolated the performance of the matching logic itself. Table 2 summarizes the results.

Table 2
Trial matching performance (10 patients × 17 trials).

Model	Matching Accuracy (%)
Gemma 3 27B	90
Qwen 2.5 32B	60

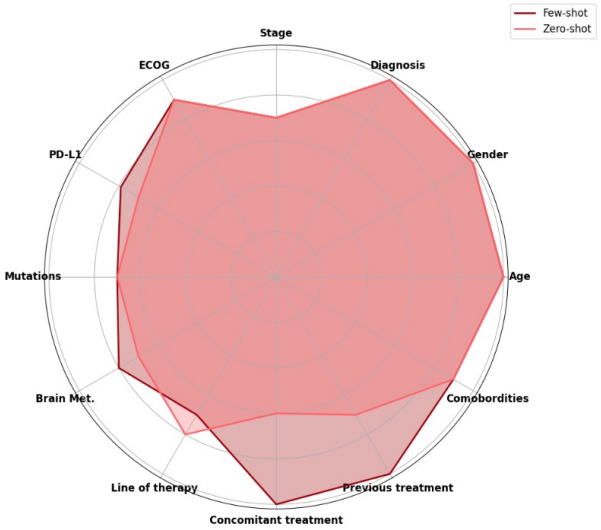

Figure 3: Zero-shot vs Few-shot prompting performance (LLaMA 3.1 8B, n=10).

Table 3
Performance Metrics by Clinical Feature.

Clinical Feature	Accuracy (%)	Macro F1-Score (%)
Gender	100	100
Age	100	100
Diagnosis	88	59
Stage	85	68
PD-L1	77	78
ECOG	91	47
Brain Metastasis	77	53
Mutations	71	57
Line of Therapy	77	36
Concomitant Treatment	85	74
Previous Treatments	80	55
Comorbidities	88	74

Gemma 3 27B outperformed Qwen 2.5 32B, correctly identifying eligible trials in 9 of 10 cases and producing coherent, criterion-based explanations. Its outputs consistently traced inclusion and exclusion criteria, yielding interpretable justifications. Qwen 2.5 32B, by contrast, achieved 6 correct matches, one of which contained partially incorrect eligibility assumptions, while in the four mismatched cases it hallucinated inclusion/exclusion criteria.

5.4. Full Pipeline Validation

The optimized configuration (LLaMA 3.1 8B with few-shot prompting + Gemma 3 27B) was validated on the full set of 35 patients. Table 3 shows feature extraction performance.

The model reached an overall accuracy of 84.9% in feature extraction and 80% accuracy in trial matching. Demographic features (gender, age) were extracted with perfect accuracy (100%). Disease characterization showed strong accuracy, with diagnosis at 88% and stage at 85%; however, they report the corresponding F1-scores (59% and 68%). ECOG performance status reached 91% accuracy but only 47% F1. Treatment-related variables varied: concomitant treatments achieved balanced performance (85% accuracy, 74% F1), while previous treatments reached 80% accuracy but a lower F1-score (55%). Line of therapy was particularly challenging, with 77% accuracy but the lowest F1-score (36%). Among complex features, PD-L1 expression achieved the highest F1-score (78%) with 77% accuracy. Brain

metastasis detection yielded moderate results (77% accuracy, 53% F1), and genetic mutations showed the lowest accuracy overall (71%).

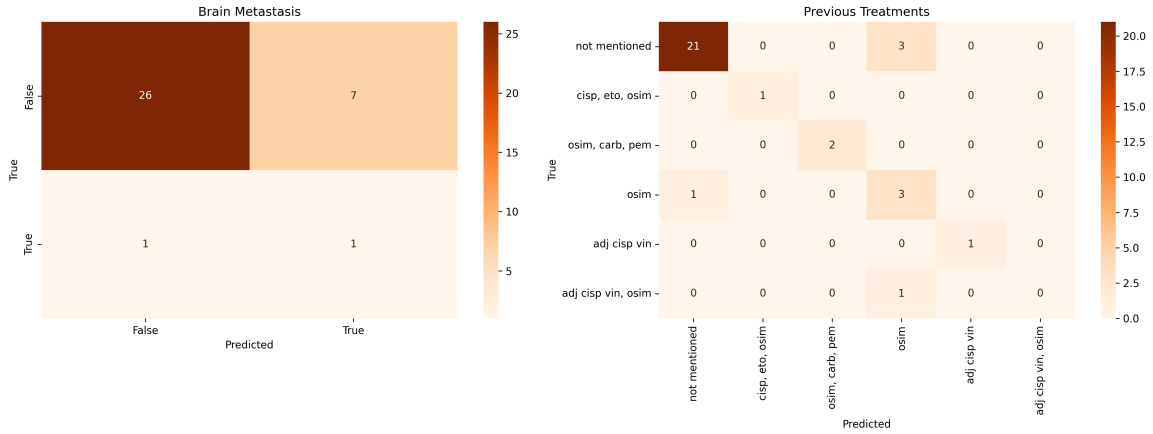

Figure 4: Confusion matrices of perfomance of the extraction of Brain Metastasis and Previous Treatments features via Llama 3.1 8B. osim: osimertinib; adj: adjuvant; cisp: cisplatin; vin: vinorelbine; carb: carboplatin; pem: pembrolizumab, eto: etoposide.

Of the 35 patients, 28 (80.4%) were correctly matched to eligible trials with the highest match score. An additional 2 patients (5.7%) were correctly matched, but their eligible trials did not receive top scores. One patient (2.8%) was not matched due to feature extraction errors, and 4 (11.1%) were incorrectly classified as ineligible despite meeting trial criteria.

Importantly, the system provides clear and logical explanations for each eligibility determination, generating comprehensive schema that detail why a patient qualifies or does not qualify for a trial. An example of this explanatory output is illustrated in Figure 5, which demonstrates how the system highlights specific criteria contributing to the matching decision.

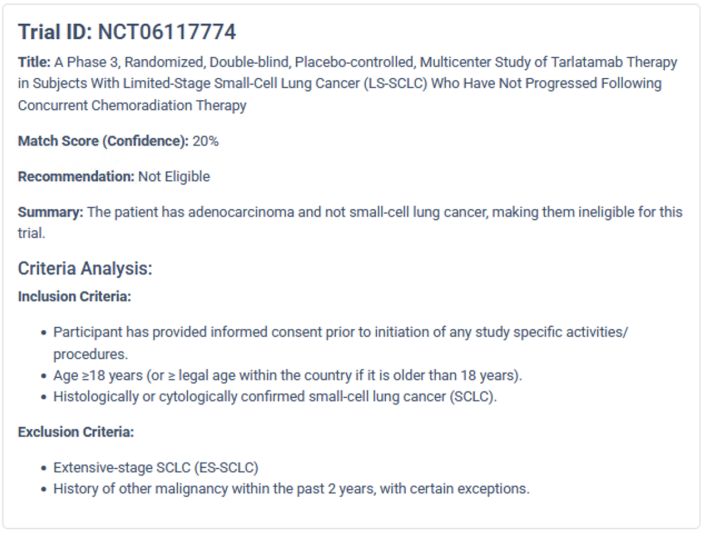

Figure 5: Extract of MedMatch interface showing trial mismatch explanation for a patient with adenocarcinoma being correctly identified as ineligible for a small-cell lung cancer trial, with detailed criteria analysis provided.

5.5. Ablation Study: Impact of Structured Feature Extraction

To quantify the contribution of structured feature extraction, we compared the full pipeline against direct PDF-to-matching (Table 4).

Table 4
Ablation Study Results (n=35).

Metric	With Feature Extraction	Direct PDF Matching
Matching Accuracy (%)	80	65
Processing Time (min)	5	1

While direct PDF matching reduced processing time from five minutes to one minute per patient, it led to a substantial loss in performance, with accuracy dropping by 15%.

5.6. Missing Data Hallucination

We investigated whether the model could correctly handle missing information in the EHR by returning the label `not mentioned`. In such cases, the model frequently generated plausible but incorrect values: a behavior commonly referred to as hallucination in the LLM literature. This issue was particularly evident for ECOG, PD-L1, mutations and previous treatments. Table 5 reports hallucination rates for these features when the ground truth indicated missing information.

Table 5
Hallucination analysis for missing information (n=35).

Clinical Feature	Missing in GT (n,%)	Correctly Predicted as MI	Hallucinations (n,%)
ECOG	2 (5.7%)	0	2 (100%)
PD-L1	10 (28.5%)	6	4 (40%)
Mutations	13 (37.1%)	7	6 (46.1%)
Previous Treatments	26 (74.2%)	22	4 (15.4%)

6. Discussion

This study introduced an explainable LLM-based system designed to address the challenge of matching lung cancer patients with appropriate clinical trials. The system was developed with the objective of streamlining the trial matching process by automatically extracting relevant clinical features from EHRs and identifying eligible clinical trials while providing transparent explanations for the recommendations. Our evaluation on the full dataset of 35 patients demonstrates 80% matching accuracy, highlighting the feasibility of achieving high performance without compromising explainability or data privacy.

Our initial benchmarking revealed notable disparities in feature extraction performance across LLMs. LLaMA 3.1 8B emerged as the most effective model, reaching 87.4% accuracy on the 10-patient benchmark set. Surprisingly, it substantially outperformed MedLLaMA 2, a domain-specific model fine-tuned on medical literature, which achieved only 50%. This result challenges the assumption that domain specialization guarantees better clinical performance. Instead, it suggests that general-purpose models with broader training corpora and more advanced architectures may generalize better to the fragmented, variable language of real-world EHRs than models trained on curated biomedical texts.

Prompt engineering also emerged as a critical lever for performance. We observed a 6.6% gain in overall accuracy using few-shot prompting, particularly when prompts were crafted by oncologists. These carefully designed examples enhanced the model's ability to reason over complex clinical features, such as comorbidities or treatment history, more effectively than over simple demographic fields, highlighting the importance of domain expertise not just in annotation but in prompt design as well.

At the model level, significant variation in reasoning ability was observed. For instance, Qwen 2.5 32B underperformed Gemma 3 27B, not only in accuracy but also in reliability. Qwen frequently hallucinated eligibility criteria, underscoring that raw parameter size is not a sufficient indicator of clinical utility. Hallucinations in this context represent more than just noise, they pose tangible safety risks in clinical decision-making.

Our validation revealed a clear performance hierarchy: demographics and ECOG status achieved near-perfect extraction due to standardized reporting, while complex features like mutations and line of therapy showed concerning accuracy gaps. The exceptionally low F1-score for line of therapy (36%) indicates fundamental difficulties in temporal reasoning: determining treatment sequences requires understanding information spanning multiple documents, a task that challenges current LLM architectures.

The ablation study further reinforced this limitation. Removing the intermediate feature extraction step led to a 15% drop in trial matching accuracy, clearly demonstrating that current LLMs cannot reliably parse complex eligibility criteria directly from raw text. This challenges the trend toward fully end-to-end LLM pipelines and supports the inclusion of structured intermediate representations as a core architectural element.

Our hallucination analysis revealed a critical safety concern with dramatic variability by feature type. The model's tendency to invent ECOG scores (100% hallucination rate) or assume molecular testing was performed when absent poses direct clinical risks. A hallucinated EGFR mutation could falsely suggest targeted therapy eligibility, while an invented performance status could inappropriately exclude patients from trials. These patterns represent fundamental limitations requiring mitigation strategies before clinical deployment.

Despite these challenges, the system demonstrated robust trial matching performance, suggesting resilience to individual extraction errors through redundant eligibility criteria and weighted scoring. The multi-layered explainability module illustrated in Figure 5 offers multi-level transparency, combining match scores, criteria-level justifications, and evidence highlighting. This layered approach marks a significant step forward from black-box systems and aligns directly with clinician demands for interpretability.

To conclude, the local deployment architecture successfully balanced performance with privacy concerns, demonstrating that modern LLMs can be effectively deployed within institutional boundaries while maintaining competitive performance, though requiring substantial computational resources.

7. Limitations

The present study has several significant limitations that must be considered when interpreting the results. The limited sample size (35 patients) restricts the generalizability of our findings. The single-institution source of the data introduces potential biases in documentation structure and content. The predominance of specific histological subtypes and disease stages may have influenced overall system performance. Additionally, the lack of significant demographic diversity limits our ability to assess system behavior across different populations.

The LLMs demonstrate some systematic difficulties processing complex temporal information and managing clinical ambiguities. The tendency to generate information not present in the original data represents a significant risk in clinical settings. The system also shows limitations in understanding particularly complex or subjective eligibility criteria, such as those related to comorbidities or functional status. Despite efforts to implement explainability mechanisms, the black-box nature of the model limits the ability to directly address the causes of specific errors. Local deployment, while privacy-preserving, entails significant computational requirements that might limit adoption in resource-constrained clinical settings.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT for grammar and spelling checks. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

Acknowledgments

We thank all the patients who accepted to actively participate in the following studies: INT 23 22, INT 46 23, INT 68 24, INT 76 24, INT 192 23, INT 196 22, INT 225 24, INT 239 24, INT 247 23, INT 251 23, INT 270 23. This work was supported by Fondazione IRCCS 'Istituto Nazionale dei Tumori'.

Data Availability

The datasets presented in this article are not readily available because of patients' privacy protection. Requests to access the datasets should be directed to the corresponding author.

Online Resources

The sources for this paper are available via GitHub.

References

[1] S. M.S., An overview of revolutionizing lung cancer management with ai: Current advances and future prospects, International Journal of Pharmaceutical Sciences 3 (2025) 884–909. URL: https://www.ijpsjournal.com/article/An+Overview+of+Revolutionizing+Lung+Cancer+Management+with+AI+Current+Advances+and+Future+Prospects, accessed: 2025-05-21.

[2] D. Calaprice-Whitty, K. Galil, W. Salloum, A. Zariv, B. Jimenez, Improving clinical trial participant prescreening with artificial intelligence (ai): A comparison of the results of ai-assisted vs standard methods in 3 oncology trials, Therapeutic Innovation & Regulatory Science 54 (2020) 69–74. URL: https://doi.org/10.1007/s43441-019-00030-4. doi:10.1007/s43441-019-00030-4, epub 2020 Jan 6.

[3] Z. Jie, Z. Zhiying, L. Li, A meta-analysis of watson for oncology in clinical application, Scientific Reports 11 (2021) 5792. URL: https://doi.org/10.1038/s41598-021-84973-5. doi:10.1038/s41598-021-84973-5.

[4] L. Tang, Z. Sun, B. Idnay, et al., Evaluating large language models on medical evidence summarization, npj Digital Medicine 6 (2023). URL: https://doi.org/10.1038/s41746-023-00896-7. doi:10.1038/s41746-023-00896-7.

[5] M. Agrawal, S. Hegselmann, H. Lang, Y. Kim, D. Sontag, Large language models are few-shot clinical information extractors, in: Y. Goldberg, Z. Kozareva, Y. Zhang (Eds.), Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 1998–2022. URL: https://aclanthology.org/2022.emnlp-main.130/. doi:10.18653/v1/2022.emnlp-main.130.

[6] J. Wei, X. Wang, D. Schuurmans, M. Bosma, E. H. Chi, Q. Le, D. Zhou, Chain of thought prompting elicits reasoning in large language models, CoRR abs/2201.11903 (2022). URL: https://arxiv.org/abs/2201.11903. arXiv:2201.11903.

[7] M. Bayer, Sqlalchemy: The database toolkit for python, https://www.sqlalchemy.org/, 2025. Version 2.0.

[8] The PostgreSQL Global Development Group, Postgresql: The world's most advanced open source relational database, https://www.postgresql.org/, 2025. Version 15.

[9] B. Welsh, pdfplumber: A python library for extracting information from pdf files, https://github.com/jsvine/pdfplumber, 2023. Version 0.7.7.

[10] U. N. L. of Medicine, Clinicaltrials.gov api, https://clinicaltrials.gov/data-api/api, 2025. Accessed: 2025-05-21.