BIOPSY - Biomarkers In Oncology: Pipeline for Structured Yielding

Sanya A. Chetwani
Kognitic, Inc.
sanya@kognitic.com

Jaseem Mahmmdla
Kognitic, Inc.
jaseem@kognitic.com

Abstract

In clinical science, biomarkers are crucial indicators for early cancer detection, prognosis, and guiding personalized treatment decisions. Although critical, extracting biomarkers and their levels from clinical texts remains a complex and underexplored problem in natural language processing research. In this paper, we present BIOPSY, an end-to-end pipeline that integrates a domain-adapted biomarker entity recognition model, a relation extraction model to link biomarkers to their respective mutations, a biomarker-type classifier, and finally, a tailored algorithm to capture biomarker expression levels. Evaluated on 5,000 real-world clinical texts, our system achieved an overall F1 score of 0.86 for oncology and 0.87 for neuroscience domains. This reveals the ability of the pipeline to adapt across various clinical sources, including trial records, research papers, and medical notes, offering the first comprehensive solution for end-to-end, context-aware biomarker extraction and interpretation in clinical research.

1 Introduction

Traditional approaches in biomedical text mining (Zhu et al., 2013) have primarily focused on extracting entities such as drugs, diseases, genes, and treatments. However, the complexity of clinical research extends far beyond these basic elements. Progress in this domain increasingly depends on capturing more nuanced attributes that directly influence trial designs, outcomes, and treatment efficacy. One such critical attribute is the biomarker (Califf, 2018).

A biomarker is a measurable indicator of a biological condition or process. In the context of cancer (Wu and Qu, 2015), biomarkers are proteins, genes, or molecules in humans that indicate the presence, stage, and subtype of the disease. A mutation is a specific change in this gene sequence or structure. Both of these combined, are often used to determine a patient's eligibility for specific treatments, assess prognosis, or monitor response to therapy. In this work, we use the term *biomarker entity* to refer broadly to biomarkers mentioned in clinical texts (e.g., EGFR, HER2, PD-L1) as well as their corresponding mutations (e.g., Exon 19, T790, L858R).

Biomarkers have become essential for advancing precision medicine (Santoshi and Sengupta, 2021), enabling more targeted and effective therapies. Professionals involved in drug development, clinical trial design, and pharmaceutical intelligence significantly rely on biomarker entities to make informed decisions about patient eligibility, therapeutic targets, and study endpoints. As a result, identifying and analyzing biomarkers and their mutations has become a focal point in the clinical and pharmaceutical industries.

The extraction of biomarker entities is notably challenging, particularly when identical terms may denote distinct medical concepts within clinical texts. For instance, a term might denote a drug target (Torchilin, 2000), which is typically a protein in the human body that serves as the binding site for a therapeutic agent, or a biomarker, which reflects a biological state or condition of the protein. Consider the Human Epidermal Growth Factor Receptor 2 (HER2) protein (Gutierrez and Schiff, 2011). In the sentence *"elevated HER2 expression correlates with poor prognosis"*, HER2 functions as a biomarker, indicating disease severity. In contrast, in *"using trastuzumab to inhibit HER2 in first-line patients"*, HER2 acts as a drug target, serving as the molecular site for therapeutic intervention. Disambiguating such cases is crucial for accurate information extraction.

Hence, we propose BIOPSY (Biomarkers In Oncology: Pipeline for Structured Yielding), a one-stop solution designed to help clinical experts with fast, accurate, and structured biomarker insights on clinical texts. BIOPSY combines a domain-

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 2313–2321
November 4-9, 2025 ©2025 Association for Computational Linguistics

adapted Named Entity Recognition (NER) (Keraghel et al., 2024) system, a relation extraction model for linking biomarkers to their respective mentioned mutations, a large language model using few-shot prompting (Brown et al., 2020) for entity type classification, and a tailored algorithm for extracting biomarker expression levels. Additionally, a custom-built post-processing algorithm aggregates the extracted information and presents it in an industry-standard format, approved by clinical experts.

2 Related Works

Biomedical information extraction (Perera et al., 2020) has witnessed substantial progress in recent years, with dedicated tools and systems (Houssein et al., 2021) addressing cancer-specific entity recognition (Zhou et al., 2022), stratification, and biomarker-level detection. Prior research has primarily focused on developing systems that extract well-established biomedical entities such as diseases, drugs, genes/proteins, and clinical procedures. However, these systems are often limited in scope, targeting a narrow set of biomarkers or specific scoring expressions, and thus fail to generalize across the broader biomarker landscape required for comprehensive clinical trial analysis.

Few of the well established tools in this domain are ScispaCy (Neumann et al., 2019) and CLAMP (Soysal et al., 2017). The ScispaCy library offers a deployment-ready framework for extracting clinical entities (genes, diseases, chemicals, etc.) making the use of spaCy (Montani et al., 2023) models retrained on biomedical literature sourced from PubMed [1] and GENIA (Kim et al., 2003). It is a fast, robust system often preferred for its low latency and easy-to-implement framework. CLAMP being a versatile tool for NER, is also preferred for its fast performance, but like ScispaCy, lacks coverage for biomarker-specific text mining.

Some of the recent works employ transformer-based architecture to push the NER performance forward. BERN2 (Sung et al., 2022), building upon BERN (Kim et al., 2019) utilizes a hybrid approach combining rule-based and neural components, improving inference and accuracy over BERN. Both, BERN2 and BioReX (Gao et al., 2024) rely on standard medical ontologies, addressing NER along with entity normalization, with BioReX placing special focus on oncology pathological reports. DeepPhe-CR (Hochheiser et al., 2023) on the other hand reports end-to-end clinical concept pipelines that include NER and eligibility screening, but perform well only on structured registries, as opposed to free-text eligibility criteria.

However, there have been approaches that take a deep dive into biomarker-targeted extractions, placing importance on this entity to serve the evolving needs of the clinical/pharmaceutical industries. Works like Holmes et al. (2021), Pironet et al. (2021), and Lin et al. (2024) focus on biomarker extraction for breast cancer biomarkers like PD-L1 and HER2. These works demonstrate good performance on biomarker-specific tasks but remain constrained by factors like biomarker variety, scoring methods, and language. They incorporate rule-based and classifier-driven approaches that achieve high F1 scores but remain limited to specific cancers, with benchmarking on a limited set of 5-6 biomarkers and require substantial domain engineering for expansion.

Recent efforts have explored large language models and domain-specific transformers for cancer biomarker extraction. Alkhoury et al. (2025) proposed a system to extract PD-L1 testing details from unstructured Electronic Health Records (EHR), achieving high accuracy across institutions. Cohen et al. (2025) introduced CancerBERT, a breast cancer-specific language model that outperformed general biomedical models in phenotype extraction. While both studies demonstrate the strengths of domain adaptation, they face limitations in generalization: the former is narrowly focused on PD-L1 and institution-specific formats, while the latter is trained solely on breast cancer corpora. These constraints underscore the need for broader, multi-biomarker systems that generalize across diseases and documentation styles.

3 Methodology

In this section, we walk through the design and development of BIOPSY, an end-to-end system for extracting and interpreting biomarker data from clinical text. BIOPSY is a modular pipeline built with each component tailored to our four key tasks. The first component recognizes and extracts biomarkers and mutations, followed by the second component, which performs semantic relation extraction between the biomarker and mutation entities. The third component is employed to classify the extracted biomarker and mutation (if captured) into

[1] https://pubmed.ncbi.nlm.nih.gov

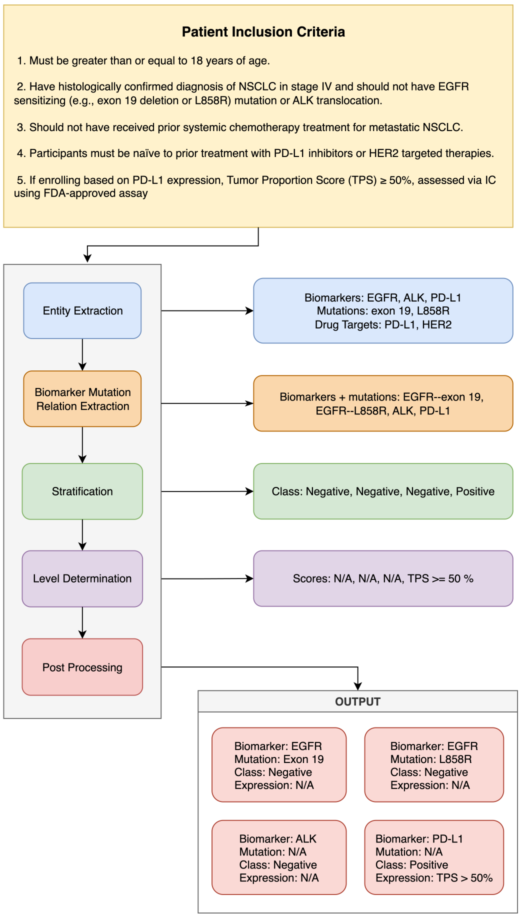
Patient Inclusion Criteria
1. Must be greater than or equal to 18 years of age.
2. Have histologically confirmed diagnosis of NSCLC in stage IV and should not have EGFR sensitizing (e.g., exon 19 deletion or L858R) mutation or ALK translocation.
3. Should not have received prior systemic chemotherapy treatment for metastatic NSCLC.
4. Participants must be naïve to prior treatment with PD-L1 inhibitors or HER2 targeted therapies.
5. If enrolling based on PD-L1 expression, Tumor Proportion Score (TPS) ≥ 50%, assessed via IC using FDA-approved assay
Entity Extraction
Biomarkers: EGFR, ALK, PD-L1
Mutations: exon 19, L858R
Drug Targets: PD-L1, HER2
Biomarker Mutation Relation Extraction
Biomarkers + mutations: EGFR--exon 19, EGFR--L858R, ALK, PD-L1
Stratification
Class: Negative, Negative, Negative, Positive
Level Determination
Scores: N/A, N/A, N/A, TPS >= 50 %
Post Processing
OUTPUT
Biomarker: EGFR
Mutation: Exon 19
Class: Negative
Expression: N/A
Biomarker: EGFR
Mutation: L858R
Class: Negative
Expression: N/A
Biomarker: ALK
Mutation: N/A
Class: Negative
Expression: N/A
Biomarker: PD-L1
Mutation: N/A
Class: Positive
Expression: TPS > 50%

Figure 1: BIOPSY Pipeline

one of three classes ("positive", "negative", and "assessment"). The last component of the pipeline is responsible for extracting the biomarker expression or biomarker level in the human body.

We begin by describing the data generation process used to create our proprietary dataset, which plays a pivotal role in training and fine-tuning the components of the pipeline. Figure 1 represents the construct of this pipeline along with the input, intermediate outputs, and the final output of the process.

3.1 Dataset Sourcing and Generation

Training Dataset: The specialized nature of this field, along with its complexity and limited prior research, has resulted in a significant lack of labeled and open-source datasets. To address this gap, we curate a handcrafted dataset of 5,000 real-world oncology-focused abstracts, sampled from ClinicalTrials.gov [2] and PubMed. Each abstract was carefully annotated to highlight biomarker, mutation, and drug target entities, ensuring balanced

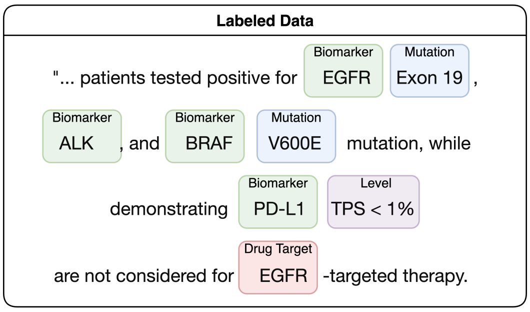

Figure 2: Snippet of our proprietary dataset.

representation across all known oncology biomarkers and mutations. Building on this foundation, we employed few-shot prompting and fine-tuning with GPT-4o (OpenAI et al., 2024) to generate 50,000 synthetic samples and labels in batches. To ensure consistency in quality, we test sentence complexity, class representation, and entity richness across the dataset. The Flesch-Kincaid readability metric (Gopal et al., 2021) is first applied for baseline filtering. A domain-aware vocabulary density test, using the Unified Medical Language System (UMLS) (Humphreys et al., 1993) meta-thesaurus, is then conducted to retain only those samples that match the medical density of the real-world data. This is followed by class balancing to ensure uniform distribution across categories. This process results in a final proprietary dataset of 37,000 samples, which have been quality checked by the clinical experts. This data is used to train our NER models for the biomarker entity extraction task.

Testing Dataset: To address the linguistic complexity of real-world clinical texts, we utilize our handcrafted dataset of 5,000 manually labeled samples [3] focused on oncology. These datasets were sourced and labeled to reflect the authentic clinical trial language and entity usage, serving as rigorous benchmarks for in-domain evaluation. Additionally, to assess the generalization of the pipeline, an independent hand-labeled dataset of 2,000 samples focused on neuroscience, which is a distinct therapeutic area in clinical terms, was created. This external dataset facilitates the evaluation of the pipeline's adaptability and robustness when applied to a different therapeutic domain, thereby demonstrating its potential utility in an industry characterized by shifting priorities driven by regulatory requirements, evolving global health needs, and commercial considerations.

[2] https://clinicaltrials.gov

[3] https://github.com/SanyaCodes/BIOPSY-Dataset-5K

3.2 Biomarker Entity Extraction

Multiple NER models were evaluated by training on synthetically generated biomedical data, followed by systematic hyperparameter tuning to optimize performance across entity types as shown in Appendix A. Among the models tested, the GLiNER biomed-large-v1.0 model (Zaratiana et al., 2024) achieved the highest overall F1 score of 0.88, indicating strong generalization to complex clinical language.

Further qualitative analysis of its outputs revealed that the model, after training, was capable of reliably distinguishing between key biomedical entity categories, including biomarkers, mutations, and drug targets. In particular, GLiNER exhibited robust recognition of context-specific mentions, such as oncology-related genes (e.g., EGFR, KRAS, ALK) and mutational variants (e.g., T790M, Exon 20), even when phrased ambiguously or embedded in longer clinical expressions.

The model's ability to detect rare or low frequency biomarker terms, which are often underrepresented in manually curated corpora, suggests that synthetic training data can help expand entity coverage. This is especially valuable for downstream tasks such as biomarker stratification, trial matching, and precision oncology applications, where entity diversity and disambiguation are critical.

3.3 Biomarker Mutation Relation Extraction

While GLiNER effectively detects biomarkers and mutations, a dedicated component is required to extract semantic relations between them. For example, in texts listing multiple biomarkers and mutations, such as *"the patient must be tested positive for EGFR Exon 19, ALK, BRAF V600, and HER2 mutations"*, accurate relation extraction (Fraile Navarro et al., 2023) is essential.

To address this, a relation extraction model based on ensemble learning and an attention mechanism is fine-tuned, following the methodology proposed by Jia et al. (2024). Their approach combines multiple biomedical language models such as BioBERT (Lee et al., 2020), BlueBERT (Peng et al., 2019) and PubMedBERT (Gu et al., 2021) as base classifiers. These models are trained independently on our relation-labeled dataset, and their output probabilities are aggregated using an attention-based stacking mechanism. A meta-classifier then integrates these weighted predictions to produce the final relation label between biomarkers and their respective mutation. This architecture enables robust and context-aware relation extraction across complex biomedical texts, reporting an F1 score of 0.87 as shown in Appendix B.

3.4 Biomarker Stratification

Inferring patient stratification from clinical texts based on biomarker status poses a variety of semantic and contextual challenges. More specifically, the clinical trial texts often express biomarker stratification using nuanced phrasing under the Trial Criteria section. For example, a criterion such as *"patients should not have any EGFR sensitizing mutation to qualify for enrollment"* implies that the patient should test EGFR-Negative to be included in this trial since there is just one layer of negation in the sentence with respect to the biomarker. There can be multiple layers of negation; for instance, the text could read *"Patients will be excluded if no EGFR sensitizing mutation is found"*. This sentence contains two layers of negation nested in the sentence with respect to the EGFR biomarker, hence implying that the patient should be EGFR-Positive to be included in this study. Further complexities arise in cases where specific mutation profiles are referenced under inclusion and exclusion in a single sentence. A great example of this is *"patients should carry the EGFR Exon 19 mutation but not the other sensitizing mutations, including but not limited to T790 and Exon 20"*. This text implies that to qualify for the trial, a patient should be marked by EGFR Exon 19-Positive, EGFR Exon 20-Negative, and EGFR T790-Negative. The third category is assessment, wherein a patient should be marked EGFR-Assessment and undergo testing to qualify for the trial. The sentence would generally read *"Patient must provide a tumor tissue sample for biopsy to test for EGFR status"*. Such specificity necessitates deep reasoning beyond general biomarker presence.

This step not only presents a convoluted context around the mentioned biomarker that needs to be understood by a model, but also a wide variety of linguistic complexities, such as nested negation, the presence of multiple biomarker groups in a single text piece, and so on. This presents the need for a context-aware model that has been trained on a wide variety of data and incorporates the understanding of different styles of human writing. Large language models and some advanced context-aware classification models are able to solve this challenge. Based on the tuning and experimentation with our datasets, Llama 3.1 70B (Grattafiori et al., 2024) model demonstrates this capability with 0.85 F1 score as seen in Appendix C. In our observation, the recent capabilities of LLMs remain unmatched due to their deep understanding of language, large training datasets, and increasing context retention ability.

3.5 Biomarker Level Determination

It is often seen in clinical texts that for biomarkers like PD-L1 and HER2, among a few others, the authors of trials do not explicitly state whether the patient allowed to be enrolled in the study should test positive or negative. They provide qualification scores instead, for example, *"PD-L1 Tumor Proportion Score (TPS) should be greater than 50%"* or *"the patient must test HER2 IHC 3+"* that implicitly convey that the patient must be PD-L1 Positive and HER2 Positive, respectively. On the other hand, a phrase like *"PD-L1 < 1%"* will imply that the patient is PD-L1 Negative. In an ever-expanding field like oncology, there exist 150-200 biomarkers recognized by the U.S. Food and Drug Administration (FDA) and 10-15 major scoring techniques used by clinical specialists worldwide, making adaptability crucial.

The process begins with tokenizing the text using spaCy (v3.7), followed by sentence segmentation to facilitate more efficient context processing. Each sentence is subsequently parsed using a constituency parser, such as BENEPAR (Kitaev et al., 2018), to extract immediate noun phrases. During experimentation, noun phrases comprising biomarker names and scoring methods were observed to dominate those containing numerical values in the constituency parse tree. Additionally, regular expressions were utilised to identify candidate phrases containing numerical scores, as commonly found in clinical texts. QuantityIE is then employed (Wang et al., 2023) to traverse the constituency parse tree and identify the smallest enclosing noun phrase that contains a numerical value along with its candidate context. A second filtering pass is applied to retain only those candidate spans where the numerical value is a nummod (numerical modifier) or amod (adjective modifier) of a biomarker token. The quantified biomarker entity is also verified to have a direct syntactic dependency with the numerical modifier (nummod) in the parse tree. For instance, "50%" is a nummod of clinical scoring methods like "TPS" and is compounded by "PD-L1", resulting in the "PD-L1 TPS" span.

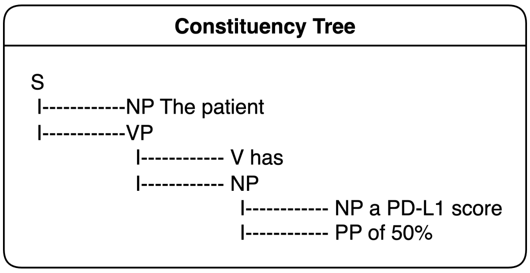

Figure 3: An example of a constituency parse tree.

These spans are normalized using a predefined mapping to ensure domain-specific formatting and consistent interpretation of the final output. Identified relations follow a structured format, such as: "span": "PD-L1 TPS", "relation": ">", "value": "50%". In a subsequent step, an additional curated mapping is applied to determine whether the combination of "relation" and "value" indicates a Positive or Negative expression state for the biomarker. This method achieves 0.89 F1 score on our hand-labeled dataset.

3.6 Post Processing

To refine the extractions, stratification, and biomarker levels, standardized mappings are utilized to resolve biomarkers and their mutations into their respective official terminology. This is followed by the creation of structured tuples in the format *(biomarker, mutation, stratification, score)*. In instances where stratification is not explicitly stated and thus undetermined by the LLM, domain-specific logic is applied to infer the class type based on the extracted score levels.

4 Results

Through experimentation on various tasks, we have chosen the best-trained and fine-tuned components to build the pipeline. Since there are not many research works that solve this end-to-end industry-relevant problem spread across all biomarkers and therapeutic areas, to the best of our knowledge, we test this pipeline on the handcrafted test set in Oncology and Neuroscience settings.

We observe that the model demonstrates exceptional capability in learning the context around words since it identifies biomarkers of the neuroscience domain as well as the oncology domain (refer Table 1).

To further evaluate the contribution of domain-specific fine-tuning, we compared BIOPSY against

Dataset	Precision	Recall	F1
Oncology	0.85	0.88	0.86
Neuroscience	0.91	0.84	0.87

Table 1: Comparative Evaluation of NER Models for Biomarker Entity Recognition

a direct GPT-4o baseline with and without any specialized training data. Tables 2 and 3 present the comparative performances across oncology and neuroscience domains, respectively.

While GPT-4o exhibits excellent zero-shot capabilities, our trained pipeline outperforms both the untrained baseline pipeline and standalone GPT-4o across all metrics. This performance gap highlights that, although large language models provide reasonable starting points, the complexity of clinical applications necessitates specialized training. These findings validate our decision to invest in domain-specific adaptation, where accurate and context-aware interpretation is critical for clinical deployment. The results also demonstrate the tangible value added by our proprietary dataset, which enables precise biomarker identification and relation extraction in real-world clinical text.

Model	P	R	F1
GPT-4o	0.69	0.78	0.73
BIOPSY (untrained)	0.77	0.73	0.75
BIOPSY (fine-tuned)	**0.85**	**0.88**	**0.86**

Table 2: Oncology domain: comparison of GPT-4o, baseline BIOPSY, and fine-tuned BIOPSY models.

Model	P	R	F1
GPT-4o	0.78	0.71	0.74
BIOPSY (untrained)	0.74	0.68	0.71
BIOPSY (fine-tuned)	**0.91**	**0.84**	**0.87**

Table 3: Neuroscience domain: comparison of GPT-4o, baseline BIOPSY, and fine-tuned BIOPSY models.

5 Conclusion

This paper presents BIOPSY, an end-to-end pipeline for clinical biomarker extraction that integrates entity recognition, mutation linking, stratification, and level inference. By combining curated datasets, transformer-based models, large language models, and syntax-guided extraction, BIOPSY achieves high accuracy across diverse biomarker types and therapeutic domains. Our evaluation across oncology and neuroscience demonstrates the value of domain-specific fine-tuning, highlighting BIOPSY's robustness and adaptability. Addressing a pressing need in clinical research, it provides a scalable and interpretable solution for biomarker-centric text mining, with strong potential for deployment in trial design, drug development, and precision oncology. Future work will extend this pipeline to additional therapeutic areas and explore automatic dataset generation for continual fine-tuning.

6 Limitations

While BIOPSY achieves strong performance across diverse clinical tasks and therapeutic domains, there remain opportunities to further enhance its scope and adaptability. A portion of the training data is synthetically generated using few-shot prompting with LLMs, which—though carefully curated and domain-filtered—can be further enriched with larger real-world datasets to capture greater linguistic diversity. Additionally, while BIOPSY generalizes effectively to oncology and neuroscience, extending its evaluation to additional therapeutic areas will help validate its cross-domain robustness.

7 Acknowledgment

This work was developed as part of the biomedical research and development at Kognitic, Inc., a platform used by pharmaceutical teams focused on competitive intelligence, business development and commercial strategy. The BIOPSY pipeline was developed to support biomarker translational insights for drug development, trial recruitment, and trial success estimation. We acknowledge the contributions of the data science and clinical science team at Kognitic in building and supporting this work. More information about the platform is available at www.kognitic.com.

References

Nour Alkhoury, Maqsood Shaik, Ricardo Wurmus, and Altuna Akalin. 2025. Enhancing biomarker based oncology trial matching using large language models. *npj Digital Medicine*, 8(1).

Parminder Bhatia, Busra Celikkaya, and Mohammed Khalilia. 2018. Joint entity extraction and assertion detection for clinical text. *arXiv preprint arXiv:1812.05270*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33.

Robert M Califf. 2018. Biomarker definitions and their applications. *Experimental biology and medicine*, 243(3):213–221.

Wendy W Chapman, Will Bridewell, Paul Hanbury, Gregory F Cooper, and Bruce G Buchanan. 2001. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of biomedical informatics*, 34(5):301–310.

Aaron B. Cohen, Blythe Adamson, Jonathan Kelly Larch, and Guy Amster. 2025. Large language model extraction of pd-l1 biomarker testing details from electronic health records. *AI in Precision Oncology*, 2(2):57–64.

David Fraile Navarro, Kiran Ijaz, Dana Rezazadegan, Hania Rahimi-Ardabili, Mark Dras, Enrico Coiera, and Shlomo Berkovsky. 2023. Clinical named entity recognition and relation extraction using natural language processing of medical free text: A systematic review. *International Journal of Medical Informatics*, 177:105122.

Weiting Gao, Xiangyu Gao, Wenjin Chen, David J. Foran, and Yi Chen. 2024. Biorex: Biomarker information extraction inspired by aspect-based sentiment analysis. In *Advances in Knowledge Discovery and Data Mining*, pages 129–141, Singapore. Springer Nature Singapore.

Revathi Gopal, Mahendran Maniam, Noor Alhusna Madzlan, Siti Shuhaida binti Shukor, and Kanmani Neelamegam. 2021. Readability formulas: An analysis into reading index of prose forms. *Studies in English Language and Education*, 8(3):972–985.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23.

Carolina Gutierrez and Rachel Schiff. 2011. Her2: biology, detection, and clinical implications. *Archives of pathology & laboratory medicine*, 135(1):55–62.

Harry Hochheiser, Sean Finan, Zhou Yuan, Eric B. Durbin, Jong Cheol Jeong, Isaac Hands, David Rust, Ramakanth Kavuluru, Xiao-Cheng Wu, Jeremy L. Warner, and Guergana Savova. 2023. Deepphe-cr: Natural language processing software services for cancer registrar case abstraction. *JCO Clinical Cancer Informatics*, (7):e2300156. PMID: 38113411.

Benjamin Holmes, Dhananjay Chitale, Joshua Loving, Mary Tran, Vinod Subramanian, Anna Berry, Matthew Rioth, Raghu Warrier, and Thomas Brown. 2021. Customizable natural language processing biomarker extraction tool. *JCO Clinical Cancer Informatics*, (5):833–841.

Essam H. Houssein, Rehab E. Mohamed, and Abdelmgeid A. Ali. 2021. Machine learning techniques for biomedical natural language processing: A comprehensive review. *IEEE Access*, 9:140628–140653.

B. L. Humphreys, A. T. McCray, and D. A. B. Lindberg. 1993. The unified medical language system. *Methods of Information in Medicine*, 32(04):281–291.

Yaxun Jia, Haoyang Wang, Zhu Yuan, Lian Zhu, and Zuo-lin Xiang. 2024. Biomedical relation extraction method based on ensemble learning and attention mechanism. *BMC Bioinformatics*, 25.

Imed Keraghel, Stanislas Morbieu, and Mohamed Nadif. 2024. A survey on recent advances in named entity recognition. *arXiv preprint arXiv:2401.10825*.

Donghyeon Kim, Jinhyuk Lee, Chan Ho So, Hwisang Jeon, Minbyul Jeong, Yonghwa Choi, Wonjin Yoon, Mujeen Sung, and Jaewoo Kang. 2019. A neural named entity recognition and multi-type normalization tool for biomedical text mining. *IEEE Access*, 7:73729–73740.

Jin-Dong Kim, Tomoko Ohta, Yutaka Tateisi, and Jun'ichi Tsujii. 2003. Genia corpus-a semantically annotated corpus for bio-textmining. *Bioinformatics*.

Nikita Kitaev, Steven Cao, and Dan Klein. 2018. Multilingual constituency parsing with self-attention and pre-training. *arXiv preprint arXiv:1812.11760*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Xin Lin, Kuan Kang, Pan Chen, Zhaoyang Zeng, Guiyuan Li, Wei Xiong, Mei Yi, and Bo Xiang. 2024. Regulatory mechanisms of pd-1/pd-l1 in cancers. *Molecular Cancer*, 23(1).

Ines Montani, Matthew Honnibal, Matthew Honnibal, Adriane Boyd, Sofie Van Landeghem, and Henning Peters. 2023. explosion/spacy: v3.7.2: Fixes for apis and requirements.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and robust models for biomedical natural language processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy. Association for Computational Linguistics.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets. In *Proceedings of the 2019 Workshop on Biomedical Natural Language Processing (BioNLP 2019)*, pages 58–65.

Nadeesha Perera, Matthias Dehmer, and Frank Emmert-Streib. 2020. Named entity recognition and relation detection for biomedical information extraction. *Frontiers in cell and developmental biology*, 8:673.

Antoine Pironet, Hélène A. Poirel, Tim Tambuyzer, Harlinde De Schutter, Lien van Walle, Joris Mattheijssens, Kris Henau, Liesbet Van Eycken, and Nancy Van Damme. 2021. Machine learning-based extraction of breast cancer receptor status from bilingual free-text pathology reports. *Frontiers in Digital Health*, 3.

Seneha Santoshi and Dipankar Sengupta. 2021. Artificial intelligence in precision medicine: A perspective in biomarker and drug discovery. *Artificial Intelligence and Machine Learning in Healthcare*, pages 71–88.

Ergin Soysal, Jingqi Wang, Min Jiang, Yonghui Wu, Serguei Pakhomov, Hongfang Liu, and Hua Xu. 2017. Clamp – a toolkit for efficiently building customized clinical natural language processing pipelines. *Journal of the American Medical Informatics Association*, 25(3):331–336.

Mujeen Sung, Minbyul Jeong, Yonghwa Choi, Donghyeon Kim, Jinhyuk Lee, and Jaewoo Kang. 2022. Bern2: an advanced neural biomedical named entity recognition and normalization tool. *Bioinformatics*, 38(20):4837–4839.

Vladimir P Torchilin. 2000. Drug targeting. *European journal of pharmaceutical sciences*, 11:S81–S91.

Zixiang Wang, Tongliang Li, and Zhoujun Li. 2023. Unsupervised numerical information extraction via exploiting syntactic structures. *Electronics*, 12(9).

Li Wu and Xiaogang Qu. 2015. Cancer biomarker detection: recent achievements and challenges. *Chem. Soc. Rev.*, 44:2963–2997.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2024. GLiNER: Generalist model for named entity recognition using bidirectional transformer. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5364–5376, Mexico City, Mexico. Association for Computational Linguistics.

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 207–212, Berlin, Germany. Association for Computational Linguistics.

Sicheng Zhou, Nan Wang, Liwei Wang, Hongfang Liu, and Rui Zhang. 2022. Cancerbert: a cancer domain-specific language model for extracting breast cancer phenotypes from electronic health records. *Journal of the American Medical Informatics Association*, 29(7):1208–1216.

Fei Zhu, Preecha Patumcharoenpol, Cheng Zhang, Yang Yang, Jonathan Chan, Asawin Meechai, Wanwipa Vongsangnak, and Bairong Shen. 2013. Biomedical text mining and its applications in cancer research. *Journal of Biomedical Informatics*, 46(2):200–211.

A Model Selection Study for Biomarker Entity Extraction

Several biomedical NER models, originally trained on general clinical entities, were fine-tuned on a proprietary biomarker-specific dataset. Their ability to differentiate between biomarkers, mutations and drug targets, a critical distinction in clinical research, was evaluated. The study aimed to assess each model's capacity to capture the contextual cues essential for accurate entity recognition. The evaluation was conducted on a proprietary test set, and the results are summarised in Table 4. GLiNER biomed-large-v1.0 demonstrated superior performance, achieving the highest F1 score. Addition-

Model	Precision	Recall	F1
TinyBERN	0.71	0.68	0.69
BERN2	0.76	0.74	0.75
LLama 3.1 70B	0.79	0.76	0.77
Qwen2.5 72B	0.78	0.75	0.76
GLiNER biomed-large-v1.0	**0.86**	**0.91**	**0.88**
GLiNER biomed-NER	0.84	0.83	0.83

Table 4: Comparative Evaluation of NER Models for Biomarker Entity Recognition

Model	Precision	Recall	F1
Rule-based Classifier	0.61	0.59	0.60
BiLSTM + Attention	0.72	0.70	0.71
LLama 3.1 70B	0.84	0.85	0.84
Ensemble + Attention	**0.86**	**0.88**	**0.87**

Table 5: Comparative Evaluation of Relation Extraction Models for Biomarker–Mutation Associations

Model	Precision	Recall	F1
Rule-based	0.61	0.53	0.56
Llama 3 8B	0.72	0.69	0.70
Llama 3.1 70B	**0.87**	**0.83**	**0.85**
Qwen2.5 72B	0.84	0.85	0.84

Table 6: Comparative Evaluation of rule-based, neural and LLM models for Biomarker Stratification

ally, off-the-shelf large language models were evaluated using few-shot learning techniques. These models showed promising results in recognizing biomarker entities without requiring further fine-tuning, but struggled with disambiguating closely related biomarkers and drug targets.

B Model Selection Study for Biomarker Mutation Relation Extraction

To identify the most effective relation extraction model, several approaches were evaluated. These included rule-based classifiers, bidirectional LSTM architectures (Zhou et al., 2016), attention-based ensemble models, and LLM-based methods, particularly using LLama 3.1 70B. These models were fine-tuned using a custom-labeled dataset of biomarker–mutation pairs derived from clinical text. The primary evaluation objective was to assess the model's ability to accurately infer true biomarker–mutation associations, especially in the presence of negations, nested clauses, and domain-specific phrasing. Table 5 presents the comparative evaluation results across all tested models.

C Model Selection Study for Biomarker Stratification

Accurate classification of biomarkers into positive, negative (Bhatia et al., 2018), and assessment categories requires a strong understanding of contextual cues, often dispersed across multiple sentences and phrases. Traditional approaches, such as rule-based systems (Chapman et al., 2001) and ontology-driven methods, typically struggle with complex cases involving nested negations or distributed context. Table 6 presents the precision, recall, and F1 scores of the LLMs selected and evaluated during this study.

D Post-Processing: Human Expert Collaboration

For better interpretability, we include an example illustrating how post-processing integrates domain expertise to normalize biomarker scoring outputs.

Programmed death-ligand 1 (PD-L1) is frequently reported in clinical text using two metrics: Tumor Proportion Score (TPS), or the percentage of tumor cells showing PD-L1 expression, and Combined Positive Score (CPS), or the ratio of PD-L1–positive tumor and immune cells to all viable tumor cells. In cases where the stratification is not explicitly mentioned in the text but a score is measured on the TPS or CPS scales, we make use of domain knowledge to interpret the scores extracted. For instance, in the sentence "Patients with PD-L1 $TPS \geq 50\%$ were considered for the study, while $TPS < 1\%$ were excluded." the model output before post-processing is:

- PD-L1, stratification = null, $score \geq 50\%$
- PD-L1, stratification = null, $score < 1\%$

But after the Post-Processing step, the output contains the inferred stratification:

- PD-L1, stratification = positive, $score \geq 50\%$
- PD-L1, stratification = negative, $score < 1\%$

The post-processing step applies clinical thresholds derived from domain literature and expert consultation to interpret categorical outcomes. Specifically, a TPS or CPS score greater than or equal to 1 is considered positive, whereas a TPS or CPS score less than 1 is considered negative.

These rules are specific to our downstream use case. We implemented them as configurable mappings within the post-processing module, hence they can be refined within the post-processing module, depending upon the end user's specific needs/interpretations. The resulting framework preserves expert transparency, interpretability, and modularity while enabling automated consistency across datasets.