# Natural Language Processing for Extracting Diagnostic Insights from Cancer Pathology Reports

*Shreyas Kalidindi, Shrey Shah, Daniel Gabriel, Louis Vaickus, Joshua Levy*
*Emerging Diagnostic and Investigative Technologies, Department of Pathology, Dartmouth Hitchcock Medical Center*

Dartmouth Health

## ABSTRACT

- We compare TF-IDF, fine-tuned transformers, and prompted LLMs on TCGA-Reports (9,523 reports, 32 cancer types).
- Long-context transformers perform best for staging, while TF-IDF is competitive for cancer type. LLMs trail both.
- The best method depends on the task, providing a text baseline for future multimodal work.

## INTRODUCTION

**Significance of NLP in Cancer**

- Cancer is a leading cause of death, with ~2 million new cases in the U.S. each year. Pathology reports are important for diagnosis, staging, and treatment planning.
- Key information, including tumor type, TNM stage, and biomarkers, is often stored in unstructured text.
- Manually extracting this information takes time and can slow down treatment decisions.

**Limitations of Traditional Methods**

- TF-IDF treats words independently, so even when negation words are included, it cannot connect “no” with “tumor” the way a human reader would.
- Reporting styles vary across institutions, which hurts model generalization.

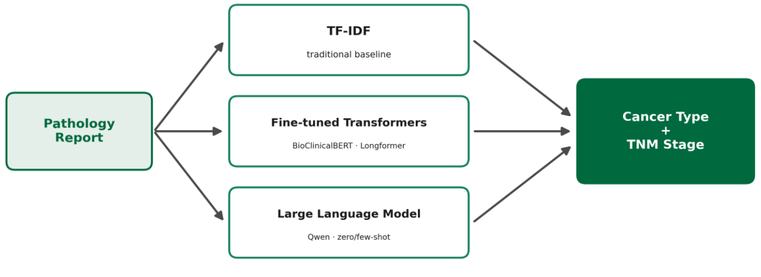
**Transformer-Based NLP**

- Models such as BioClinicalBERT use context to better understand the meaning of medical text.
- Standard BERT is limited to 512 tokens, while 64% of pathology reports are longer, motivating the use of long-context models such as Longformer and ModernBERT.

**Large Language Models (LLMs)**

- Prompted LLMs can perform tasks without being specifically trained for each task.
- Zero-shot (report + instructions only) and few-shot (with worked examples) prompting are both tested.

Figure 1: Overview of the three NLP approaches compared for extracting cancer type and TNM stage from pathology reports.




## METHODS

**Dataset**

- TCGA-Reports: 9,523 pathology reports, 32 cancer types, linked to TCGA metadata.
- Tasks: cancer-type classification and TNM staging (T, N, M predicted separately). Stratified split.

**Phase 1 - Traditional NLP**

- Preprocessing: lowercasing, tokenization, stopword removal (clinical negations kept), lemmatization.
- TF-IDF + LinearSVC (vs. logistic regression, Naive Bayes); 5-fold CV, macro-F1 as main metric.

**Phase 2 — Fine-Tuned Transformers**

- Fine-tuned BioClinicalBERT, Clinical-Longformer, and Clinical-ModernBERT on pathology reports.
- BioClinicalBERT: 512 tokens; long-context models: up to 4,096.Used class-weighted loss for cancer type and stage imbalance.
- Since 64% of reports exceed 512 tokens, long-context models were used; 3-fold cross-validation checked stability.

**Phase 3 — Large Language Models**

- Used Qwen2.5-7B-Instruct, a locally hosted 4-bit quantized LLM, with prompts containing the report and task instructions.
- Compared zero-shot prompting with few-shot prompting using examples from the training set.
- Evaluated both approaches on the same held-out test set for direct comparison with Phases 1 and 2.

## RESULTS (macro-f1)

| Method | Cancer Type | T Stage | N Stage | M Stage |
|---|---|---|---|---|
| TF-IDF + LinearSVC | 0.940 | 0.654 | 0.580 | 0.572 |
| BioClinicalBERT (512) | 0.957 | 0.570 | 0.510 | 0.500 |
| PubMedBERT | 0.970 | 0.710 | 0.750 | 0.480 |
| Llama3-Med42-8B (Zero shot) | 0.523 | 0.394 | 0.555 | 0.375 |
| Llama3-Med42-8B (Few shot) | 0.543 | 0.393 | 0.482 | 0.255 |

- Long-context transformers did not improve TNM staging substantially, though improvement on M stage was notable (0.60 with Clinical BigBird).

## RESULTS

- For cancer-type classification, the simple TF-IDF baseline (0.940) was pretty close to the transformer (0.955), showing that the added complexity of the transformer did not make a huge difference for this easier task
- The prompted LLM (Qwen, 0.617) performed worse than both other approaches. Zero-shot also did better than few-shot, likely because the longer example reports increased the number of invalid responses

## CONCLUSION

- **Potential for Clinical Impact:** Automatically pulling cancer type and staging information from unstructured reports could help cancer registries and make it easier to use information that is currently stuck in free text or large-scale research. At times, this task was more extraction rather than prediction, and so actually assessing the usefulness of text reports for clinical endpoints would serve more impactful; every cancer patient has a pathology report, but may not have other modalities like whole slide image and omics.
- **Limitations:** M-stage prediction is limited because distant metastasis is often not described in the report text.
- **Future Directions:** A multimodal approach could combine pathology report text with pathology slide images and genomic data to improve outcome prediction. Assess value of text reports compared to other modalities, and if the gap between trimodal and unimodal is recoverable it could benefit cancer prognosis around the world.
- **Data and Code Availability:** TCGA-Reports is publicly available, and the code is available on Github

edit.