UTY at the NTCIR-18 RadNLP 2024 Task:

Possibilities and Limitations of a Hybrid Rule-Based and LLM Approach for Lung Cancer TNM Classification

Yosuke Yamagishi[1], Ryosuke Tomiyama[2], Yui Ueda[3]

[1]The University of Tokyo, Japan | [2]Japan Medical Device Corporation, Japan | [3]International University of Health and Welfare Narita Hospital, Japan

Background & Motivation

Automated extraction of TNM staging information from radiology reports is challenging due to:

- Complex clinical language interpretation
- Detailed staging criteria application
- Fine-grained classification requirements

TNM Classification Complexity

- T category:
  T0, Tis, T1mi, T1a, T1b, T1c, T2a, T2b, T3, T4
  (10 classes)
- N category:
  N0, N1, N2, N3 (4 classes)
- M category:
  M0, M1a, M1b, M1c (4 classes)

Dataset & Task

NTCIR-18 RadNLP 2024 Shared Task
Japanese Track

Basic Information

- **378 Japanese radiology reports** for lung cancer staging
- **42 unique lung cancer cases** with comprehensive documentation
- **9 board-certified radiologists** providing expert interpretations
- **8th edition JLCS criteria** (Japan Lung Cancer Society) for annotation

Dataset Characteristics

- **Multi-radiologist approach**: Each case interpreted by multiple experts to capture variability in clinical reporting
- **Fine-grained classification**: More detailed TNM subcategories compared to previous NTCIR tasks
- **Real-world clinical data**: Authentic radiology reports from actual clinical practice
- **Language-specific challenges**: Japanese medical terminology and reporting conventions

Dataset Split	Reports	Cases	Purpose
Training	108	12	We did not use
Validation	54	6	Algorithm development
Test	216	24	Final evaluation

Methodology

Hybrid Two-Stage Pipeline

Our approach combines **LLMs** with **rule-based processing** for lung cancer TNM staging.

Stage 1: Information Extraction
GPT-4o extract structured information from radiology reports
Key Features Extracted (13 items):
Tumor size, laterality, characteristics (GGO/solid)
Lymph node involvement and location
Metastasis patterns and distant spread
Pleural effusion and other findings

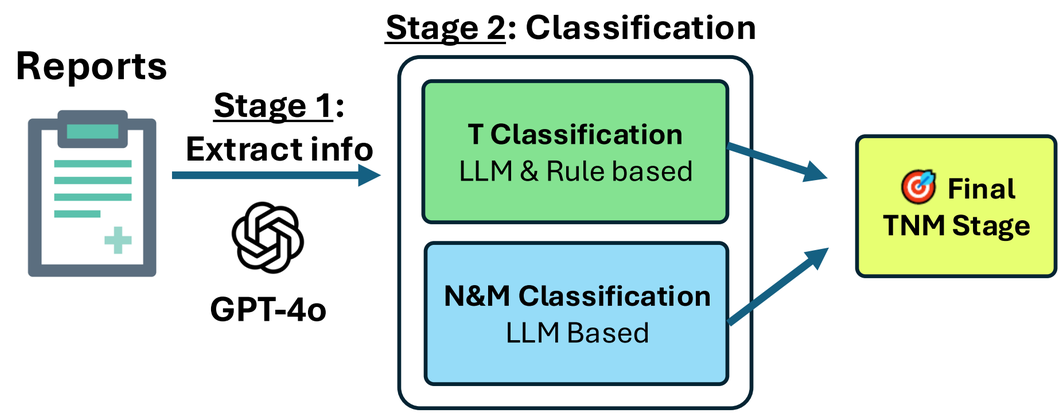
Stage 2: Classification Strategy

- **T Classification: Rule-Based**

Why: Complex size thresholds require precise rules
T0: No tumor | Tis: Pure GGO ≤30mm | T1mi: Solid ≤5mm T1a: ≤10mm | T1b: 11-20mm | T1c: 21-30mm
T2a: 31-40mm | T2b: 41-50mm
T3: >50mm or same lobe mets
T4: Adjacent structure invasion

- **N & M Classification: LLM-Based**

Why: Simpler criteria benefit from LLM flexibility
N: Direct interpretation of lymph node descriptions
M: Assessment of distant metastasis patterns



Results

Metric	Validation	Test
Joint Accuracy	0.8148	0.3889
T Accuracy	0.8704	0.4769
N Accuracy	0.9259	0.8704
M Accuracy	1.0000	0.8889

Key Observations

- Significant T-classification drop from validation to test
- Stable N/M performance across datasets
- LLM approach more robust than rule-based

Discussion

- **Key Findings**

T Classification: Strong validation → Poor test performance
N & M Classification: Consistent high performance across datasets
LLM approach generalized better than rule-based methods

- **Critical Insights**

Rule-Based Limitations
Limited validation data: Only 54 reports for development
Incomplete rules: Missing complex staging criteria

- **LLM Advantages**

Robust generalization: Maintained accuracy across datasets
Clinical flexibility: Handles diverse reporting styles
Less engineering: No extensive rule development needed

- **Implications for Clinical NLP**

Balance precision and flexibility in system design
Use component-specific strategies (different approaches for T vs N/M)
Larger datasets essential for robust rule development