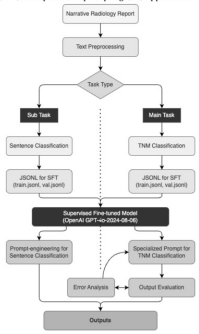
Automated Lung Cancer Staging from Radiological Reports: A Large Language Model Approach for the NTCIR-18 RadNLP Task (Team ID: tsukurad)

Takahito Nakajima, MD, PhD, Diagnostic Imaging and Interventional Radiology, Institutes of Medicine, University of Tsukuba

Abstract

Lung cancer TNM classification from narrative radiology reports presents challenges due to expression variability and complex relationships between findings. This study develops an automated TNM classification system utilizing large language models (LLMs) with supervised fine-tuning (SFT) and specialized prompting (SP) approaches. We evaluated our system on the NTCIR-18 RadNLP 2024 Task dataset, achieving 72.69% (Japanese) and 55.56% (English) fine-grained accuracy, ranking 5th among 15 teams. Our system demonstrated particularly high performance in N-factor classification (>93.98% accuracy) and in the Sub task of textual analysis (ranking 1st in Japanese and 3rd in English tracks). Error analysis revealed challenges in interpreting complex expressions and implicit information. This system shows potential for clinical workflow optimization, standardization of TNM classification, and educational support, with implications for improving cancer staging practices.



Preparation of LSONL files for SFT

Overview

To automate TNM classification, we adopted a supervised fine-tuning (SFT) approach. This approach requires preparing data files in JSONL format for model training.

Structure of JSONL Format

```
{
  "messages": [
    {"role": "system",
     "content": "tasks or instruction ..."
    },
    {"role": "user",
     "content": "narrative reports ..."
    },
    {"role": "assistant",
     "content": "labels ..."
    }]
}
```

Data Structure for Main Task (TNM Classification)

For the main task, pairs of radiology reports (narrative text) and their corresponding TNM classifications (T factor, N factor, M factor) are prepared in JSONL format.

```
an example for main_task:
{"messages": [
    {"role": "system",
     "content": "As a text analysis assistant focused on lung cancer staging,
       your task is to analyze the given sentences and output labels
       corresponding to the following factors: T, N, M.",},
    {"role": "user",
     "content": "A tumor with a major axis of 12 cm in the left upper lobe
       of the lung. Infiltration of the aortic arch and pulmonary artery is
       suspected. This is considered T4. Enlargement of the left pulmonary hilum
       and a mediastinal lymph node in level 6 is observed, suspicious of
       metastases. This is considered N2. No pleural effusion is observed.
       Bone destruction due to the tumor is observed in the left 3rd rib.
       No findings suggesting distant metastasis to the visualized bones.
       No significant findings in the visualized abdomen.",},
    {"role": "assistant",
     "content": "T=T4, N=N2, M=M0"}]}
```

Data Structure for Subtask (Sentence Classification)

For the subtask, pairs of each sentence from the radiology report and the classification results for eight categories (omittable, measure, extension, atelectasis, satellite, lymphadenopathy, pleural, distant) are prepared in JSONL format.

```
an example for sub_task:
{"messages": [
    {"role": "system",
     "content": "As a text analysis assistant focused on lung cancer staging,
       your task is to analyze the given sentences and output labels
       corresponding to the following categories: omittable, measure,
       extension, atelectasis, satellite, lymphadenopathy, pleural, and distant.",},
    {"role": "user",
     "content": "A solid mass with a diameter of 3 cm is observed
       in the lower lobe of the left lung.",},
    {"role": "assistant",
     "content": "omittable=0 measure=1 extension=0 atelectasis=0 satellite=0
       lymphadenopathy=0 pleural=0 distant=0"}]}
```

Steps for Preparing the Dataset

1. **Collection of Clinical Data**: Obtain training, validation, and test data from the NTCIR-18 RadNLP 2024 task dataset.
2. **Preprocessing**:
   - Text normalization (adjusting spaces, handling special characters, etc.)
   - Standardization of technical terms (unifying abbreviations and synonyms)
   - Sentence splitting (for the subtask)
3. **Conversion to JSONL Format**:
   - Main task: Convert pairs of radiology reports and TNM classifications for each case into JSONL format.
   - Subtask: Convert each sentence and its classification categories into JSONL format.
4. **Data Splitting**:
   - **train.jsonl**: Training data (Main task: 108 cases, Subtask: 919 sentences)
   - **val.jsonl**: Validation data (Main task: 54 cases, Subtask: 415 sentences)
   - **test.jsonl**: Test data (Main task: 81 cases, Subtask: 568 sentences)

Features of the SFT Data

1. **Balance**: Data is selected to include various TNM classification patterns.
2. **Diversity**: Cases include different expressions and finding patterns.
3. **Clinical Validity**: Accurate TNM classifications verified by radiologists.
4. **Language Coverage**: Includes both Japanese and English data.

Description of Training Procedure

Data Splitting

Main Task (TNM Classification)

- **train.jsonl**: 108 cases
- **val.jsonl**: 54 cases
- **test.jsonl**: 81 cases (final evaluation)

Subtask (Sentence Classification)

- **train.jsonl**: 919 sentences
- **val.jsonl**: 415 sentences
- **test.jsonl**: 568 sentences (final evaluation)

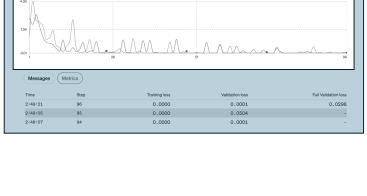
Training Process

Supervised Fine-Tuning (SFT) Approach

1. **Base Model Selection**
   - GPT-4o-2024-08-06 (latest version for fine-tuning)
2. **Training Settings**
   - Learning rate: 2e-5
   - Batch size: 4
   - Epochs: 3

Learning Curves

Training & Validation Loss



Prompt design

Continuous Improvement for Main Task: PDCA Cycle

1. **Plan**: Develop improvement plans based on error analysis
2. **Do**: Implement improvements
3. **Check**: Verify effectiveness
4. **Act**: Reflect improvements in next cycle

Subgroup Analysis & Issues

Performance Differences

1. **By Tumor Size**
   - Lower accuracy for borderline sizes (e.g., 3.0 cm, 5.0 cm)
   - T1mi (microinvasive adenocarcinoma) is particularly difficult
2. **Lymph Node Evaluation**
   - Some cases with inaccurate anatomical interpretation
   - Ambiguous expressions for lymph node metastasis
3. **Language Dependency**
   - Slight differences between Japanese and English data
   - Need to address language-dependent medical terms

SUpervised Fine-Tuning (SFT) Improvements:

- Addressed issues with few-shot prompting and domain-specific knowledge integration.

Overall Approach to Prompt Design

To automate TNM classification from radiology reports, we adopted two main prompt design approaches:

1. **Prompts for Supervised Fine-Tuning (SFT)**
2. **Prompts for Specialized Prompting (SP) Approach**

In each approach, the prompts are divided into multiple components, each designed to serve a specific role.

Prompt Structure for Main Task (TNM Classification)

The prompt for the main task consists of the following components:

1. System Prompt

```
"system_prompt": {
    "title": "TNM Classification Instruction for a Fine-Tuned LLM",
    "content": [
      "1. Review the patient's clinical and imaging findings
       regarding tumor size, local invasiveness, lymph node
       involvement, and presence of distant metastases.",
      "2. Determine the correct T, N, and M categories based on
       the definitions above.",
      "3. Provide a concise explanation for why each category
       (T, N, M) was chosen.",
      "T_factor: One of { T0, Tis, T1mi, T1a, T1b, T1c, T2a, T2b,
       T3, T4 }",
      "N_factor: One of { N0, N1, N2, N3 }",
      "M_factor: One of { M0, M1a, M1b, M1c }",
      "You must provide these factors in a JSON format as shown
       in the example below.",
      "Ignore lesions that are difficult to evaluate, need further
       evaluation, or are indeterminate.",
      "Remember to consider these rules in conjunction with the
       standard TNM definitions to determine the best fitting T,
       N, and M categories."]}
```

2. TNM Definitions

```
"tnm_definitions": {
    "T_factor(Tumor Size and Extent)": [
      "T0: No description of the size of a primary tumor.",
      "Tis: A pure ground-glass opacity/nodule (GGO or GGN) with
       a maximum diameter of 30 mm or less: consolidated portion =
       0 cm and total lesion diameter ≤ 3 cm.",
      "T1mi: A mixed ground-glass opacity/nodule (GGO or GGN):
       Partially solid type with a consolidated portion ≤ 0.5 cm,
       and total ground-glass lesion diameter ≤ 3 cm.",
      "T1a: Tumor ≤ 1 cm in greatest dimension, no invasion into
       the lung or outside organs.",
      "T1b: Tumor > 1 cm but ≤ 2 cm.",
      "T1c: Tumor > 2 cm but ≤ 3 cm.",
      "T2a: Tumor > 3 cm but ≤ 4 cm.",
      "T2b: Tumor > 4 cm but ≤ 5 cm.",
      "T2a or T2b: Includes tumors with certain invasive features
       (e.g., invasion of the main bronchus, visceral pleura invasion,
       or atelectasis/obstructive pneumonia not extending beyond
       the carina).",
      "T3: Tumor > 5 cm but ≤ 7 cm, or any tumor with: Additional
       tumor nodules in the same lobe. Invasion of the chest wall,
       phrenic nerve, or pericardium.",
      "T4: Tumor > 7 cm, or any tumor with invasion of the
       following structures: Diaphragm, mediastinum, heart, great
       vessels, trachea, esophagus, spinal cord, superior vena cava,
       recurrent laryngeal nerve, esophagus (listed again),
       vertebral body, or carina.",),
    "N_factor": [
      "N0: No lymph node metastasis.",
      "N1: Metastasis in ipsilateral hilar (pulmonary) or
       peribronchial lymph nodes (including intrapulmonary lymph
       nodes), or direct extension from the primary tumor.",
      "N2: Metastasis in ipsilateral mediastinal or subcarinal
       lymph nodes.",
      "N3: Metastasis in contralateral mediastinal or hilar lymph
       nodes, or in supraclavicular lymph nodes.",),
    "M_factor": [
      "M0: No distant metastasis.",
      "M1a: Separate tumor nodules in the contralateral lung, or
       malignant pleural/pericardial effusion.",
      "M1b: A single distant metastasis in one organ.",
      "M1c: Multiple distant metastases in one or more organs."]}
```

3. Units and Numbers

```
"units_numbers": {
    "mm_cm": [
      "≤ 0.5 cm: 0.1,0.2,0.3,0.4,0.5 cm",
      "≥ 0.5 cm: 1,2,3,4,5 mm",
      "≤ 3 cm: 0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,1.0,
       1.1,1.2,1.3,1.4,1.5,1.6,1.7,1.8,1.9,
       2.0,2.1,2.2,2.3,2.4,2.5,2.6,2.7,2.8,2.9,3.0 cm",
      // ... omitted ...
    ]
```

Role: Explicitly provides unit conversions and numerical ranges for mm and cm to assist in interpreting tumor size measurements. Functions as supplementary information to improve numerical processing accuracy in GPT-4o.

4. Output Format Specification

```
"instruction_details": {
    "goal": "Determine the most appropriate T, N, and M categories
             for the given report text.",
    "format": "Return your answer strictly in the following JSON format:",
    "example_json": {
        "T_factor": "<T value>",
        "N_factor": "<N value>",
        "M_factor": "<M value>"
    },
    "additional_note": "You may briefly explain your reasoning,
        but the final output must be valid JSON with the keys
        'T_factor', 'N_factor', and 'M_factor'."
}
```

Role: Strictly specifies the output format for the model, ensuring consistent responses. Explicitly shows the JSON format for output, making evaluation and post-processing easier.

5. Example Usage

```
"example_usage": {
    "input_text": "There is a 2.2 cm mass in the right upper lobe.
        No mediastinal lymphadenopathy is identified. No malignant
        pleural effusion. No distant lesions are detected.",
    "expected_output_json": {
        "T_factor": "T1c",
        "N_factor": "N0",
        "M_factor": "M0"
    }
}
```

Role: Provides a concrete input text and the corresponding correct output example, helping the model understand the expected response pattern. Functions as a form of few-shot prompting.

An application sample for TNM classification

Prompt Structure for Subtask (Sentence Classification)

The prompt for the subtask consists of the following components:

1. System Prompt

```
"You are an expert in medical imaging diagnosis reports.
Follow the instructions below and classify the specified text
into eight categories. The output must be in JSON format with
the following keys: omittable, measure, extension, atelectasis,
satellite, lymphadenopathy, pleural, distant. Each key should have
a value of 1 (True) or 0 (False). Omittable: contents irrelevant
to T factor (lung cancer), N factor (hilar/mediastinal lymph node
metastasis)."
```

Role: Defines the objective and expected output format for the subtask. Specifies sentence classification into eight categories and binary (0/1) output format.

2. User Prompt

```
"Classify the following text from the imaging report into
 eight categories:
(i) Omittable – Clearly pertains only to negative findings or
 content irrelevant to lung cancer staging.
(ii) Measure – Does not fall under Omittable and primarily
 describes the size of the primary lesion.
(iii) Extension – Does not fall under Omittable and describes
 the extent of the primary lesion.
(iv) Atelectasis – Does not fall under Omittable and indicates
 atelectasis or obstructive pneumonia.
(v) Satellite – Does not fall under Omittable and indicates malignant
 lesions within the lung but separate from the primary lesion (e.g.,
 intrapulmonary metastasis or cancerous lymphangitis).
(vi) Lymphadenopathy – Does not fall under Omittable and indicates
 regional lymph node enlargement.
(vii) Pleural – Does not fall under Omittable and indicates pleural
 effusion, pericardial effusion, pleural dissemination, or
 pericardial dissemination.
(viii) Distant – Does not fall under Omittable and indicates distant
 metastasis.
Text: \"{sentence}\"
Follow the format below and output in pure JSON format only. Do not
 include backticks, additional symbols, or comments.
Example output:
{
\"omittable\": 1,
\"measure\": 0,
\"extension\": 1,
\"atelectasis\": 0,
\"satellite\": 0,
\"lymphadenopathy\": 1,
\"pleural\": 0,
\"distant\": 0
}
Output:"
```

Features and Innovations in Prompt Design

1. **Modular Structure**
   - Clear division of each component with a specific role
   - Easy to update individual components as needed
2. **Detailed TNM Definitions**
   - Faithfully reflects the criteria of the 8th edition of UICC/AJCC
   - Specific numerical criteria to eliminate ambiguity
3. **Utilization of Few-shot Prompting**
   - Clearly shows expected response patterns by providing concrete examples
   - Special examples for cases prone to errors
4. **Strict Output Format Specification**
   - Ensures consistent output in JSON format
   - Standardized format facilitates post-processing and system integration
5. **Consideration of Mutual Exclusivity in Subcategories**
   - Clarifies relationships between categories in the subtask
   - Explicitly states the relationship between the "omittable" category and other categories

These prompt design innovations enhance the model's comprehension and response accuracy, optimizing the performance of TNM classification automation from radiology reports.

Specialized Prompt for Main Task

Extended Definition Components of the Specialized Prompting (SP) Approach

Overview of the Specialized Prompting (SP) Approach

- The SP approach complements Supervised Fine-Tuning (SFT).
- Incorporates TNM expertise and clinical context into prompts.
- Enhances model reasoning with radiologist knowledge.

1. Extension of the Basic System Prompt

Detailed Clinical Interpretation of TNM Classification

```
"clinical_interpretation": {
  "t_factor_clinical_notes": [
    "Comprehensively assess not only size but also invasion into
     surrounding tissues.",
    "If invasion is clinically suspected but not explicit in imaging,
    consider rationale (e.g., atelectasis, chest pain).",
    "For lung cancer T factor, GGO interpretation is crucial.
     Pure GGO = Tis, part-solid GGO (solid ≤5mm) = T1mi."
  ],
  "n_factor_clinical_notes": [
    "Accurately identify lymph node location (ipsilateral
     hilar/mediastinal, contralateral, supraclavicular).",
    "Consider size, shape, and internal features. Short axis ≥10mm
     is suspicious, but differentiate from inflammation.",
    "PET-CT SUV increase suggests metastasis, but can also be due
     to inflammation."
  ],
  "m_factor_clinical_notes": [
    "Assess not only presence but also number and organ distribution
     of metastases.",
    "Pleural dissemination, malignant effusion, pericardial effusion
     = M1a. Consider nature and amount.",
    "Distinguish M1b (single organ) and M1c (multiple). Consider
     oligometastasis."
  ]
}
```

2. Extension of Size Interpretation Rules

Strict Criteria for Measurement Interpretation

```
"size_interpretation_rules": {
  "measurement_principles": [
    "1. For borderline size, use lower subcategory (e.g., 3.0cm = T1c).",
    "2. Use maximum diameter if multiple measurements.",
    "3. ±1mm of borderline is equivalent (e.g., 2.9–3.1cm).",
    "4. Ambiguous expressions ('about') require clinical judgment."
  ],
  "ggo_evaluation": [
    "Pure GGO: No solid component.",
    "Part-solid GGO: With solid component.",
    "Solid ≤5mm = T1mi.",
    "Solid >5mm: Use solid size for T factor."
  ],
  "special_considerations": [
    "1. Exclude inflammatory changes around tumor from size.",
    "2. Include invasive parts.",
    "3. With atelectasis/pneumonia, evaluate tumor size as much
     as possible.",
    "4. Use larger value if different in mediastinal/lung window."
  ]
}
```

Role:
Provides detailed rules for tumor size measurement and interpretation, including borderline values, GGO, and special cases.

3. Definition and Interpretation of Special Cases

Clinically Important Special Patterns

```
"special_case_definitions": {
  "multiple_lesions": [
    "Same lobe: Use largest tumor T factor, denote as (m)
     (e.g., T3(m)).",
    "Different ipsilateral lobes: Use largest tumor T factor,
     classify as T3.",
    "Contralateral lung: M1a.",
    "Differentiate intrapulmonary metastasis vs. multiple primaries
     by histology or imaging."
  ],
  "pleural_involvement": [
    "Pleural indentation alone: Not always invasion.",
    "Visceral pleural invasion >1cm: T2a.",
    "Parietal/chest wall invasion: T3.",
    "Extensive pleural dissemination: M1a."
  ],
  "bronchial_involvement": [
    "Main bronchus invasion ≥2cm from carina: T2; <2cm: T3.",
    "Carinal invasion: T4.",
    "With obstructive pneumonia: Total atelectasis = T2,
     partial = T1."
  ],
  "mediastinal_involvement": [
    "Mediastinal pleura: T3.",
    "Mediastinal fat: T4.",
    "Major vessels (aorta, SVC): T4.",
    "Heart/pericardium: T4.",
    "Esophagus: T4."
  ]
}
```

Role:
Provides detailed definitions and classification criteria for special cases in TNM, such as multiple lesions, pleural, bronchial, and mediastinal invasion.

4. Error-Specific Few-shot Prompts

Misclassification Examples and Correct Interpretations

```
"error_specific_examples": {
  "size_misinterpretation": [
    {
      "error_case": "2.5 cm nodule in right upper lobe",
      "incorrect": "T2a (close to 3cm)",
      "correct": "T1c (2-3cm)",
      "explanation": "2.5cm is clearly T1c, not borderline."
    },
    {
      "error_case": "Multiple 8mm GGOs in both lungs",
      "incorrect": "T1b (multiple lesions)",
      "correct": "T1a (each nodule), M1a (contralateral)",
      "explanation": "Each 8mm nodule = T1a; contralateral = M1a."
    }
  ],

  "lymph_node_misinterpretation": [
    {
      "error_case": "Enlarged subcarinal lymph node (12mm)",
      "incorrect": "N1 (near hilum)",
      "correct": "N2 (mediastinal, station 7)",
      "explanation": "Subcarinal = mediastinal (N2), not hilar (N1)."
    },
    {
      "error_case": "Multiple lymph nodes in stations 4R and 10R",
      "incorrect": "N3 (multiple stations)",
      "correct": "N2 (ipsilateral mediastinal)",
      "explanation": "Both are ipsilateral mediastinal; N2 unless
         contralateral/supraclavicular."
    }
  ],

  "metastasis_misinterpretation": [
    {
      "error_case": "Malignant pleural effusion and a single liver
         metastasis",
      "incorrect": "M1b (single distant metastasis)",
      "correct": "M1c (multiple organs)",
      "explanation": "Pleural effusion (M1a) + liver (another organ)
         = M1c."
    },
    {
      "error_case": "Bone destruction in two adjacent vertebrae",
      "incorrect": "M1c (multiple bones)",
      "correct": "M1b (adjacent = single site)",
      "explanation": "Adjacent vertebrae = single site (M1b); M1c
         for non-contiguous."
    }
  ]
}
```

Role:
Shows typical misclassification patterns and corrections, based on radiologist experience.

Further Process for Improvements

5. Expert Annotations by Radiologists

Key Clinical Points

```
"radiologist_annotations": {
  "critical_findings": [
    "1. Irregular borders/pleural indentation suggest invasiveness.",
    "2. Heterogeneous enhancement suggests necrosis and advanced
     cancer.",
    "3. GGO around tumor in adenocarcinoma may indicate
     preinvasive/microinvasive lesion."
  ],
  "contextual_considerations": [
    "1. Consider prior malignancy for new lesions.",
    "2. Post-treatment: consider changes due to therapy/scarring.",
    "3. Integrate PET-CT, MRI, etc. for comprehensive judgment."
  ],
  "terminology_guidance": [
    "1. 'Highly suspicious' = positive finding, even if not explicit.",
    "2. 'Slight' = judge in context, evaluate with other findings.",
    "3. 'Cannot be ruled out' = negative unless other findings."
  ]
}
```

Role:
Provides annotations for nuanced image interpretation and context-dependent decision criteria.

6. Stepwise Diagnostic Approach

Step-by-Step Clinical Diagnosis

```
"stepwise_diagnostic_approach": {
  "t_factor_evaluation": [
    "Step 1: Confirm primary lesion and measure maximum diameter.",
    "Step 2: Evaluate characteristics (solid/GGO/part-solid).",
    "Step 3: Assess invasion (pleura, bronchus, vessels, nerves, etc.).",
    "Step 4: Check for additional lesions in same/other lobes.",
    "Step 5: Integrate findings to determine T factor."
  ],
  "n_factor_evaluation": [
    "Step 1: Check for lymphadenopathy and location.",
    "Step 2: Evaluate size and characteristics.",
    "Step 3: Confirm relationship to primary (ipsilateral/contralateral).",
    "Step 4: Assess metastasis possibility (PET-CT, features).",
    "Step 5: Integrate findings to determine N factor."
  ],
  "m_factor_evaluation": [
    "Step 1: Check for pleural dissemination/malignant
     effusion/pericardial effusion.",
    "Step 2: Check for contralateral lung nodules.",
    "Step 3: Evaluate common metastatic sites (liver, adrenals,
     bone).",
    "Step 4: Check for other organ metastases.",
    "Step 5: Assess number/distribution (single/multiple,
     single/multiple organs).",
    "Step 6: Integrate findings to determine M factor."
  ]
}
```

Role:
Provides a structured, stepwise diagnostic process for T, N, and M evaluation, preventing oversight of key findings.

Effects and Evaluation of the SP Approach

- **Improved accuracy in borderline cases**
  - Accurate classification near size thresholds (3.0cm, 5.0cm, 7.0cm)
  - Improved GGO and Tis/T1mi interpretation
- **Enhanced handling of complex cases**
  - Accurate interpretation of multiple lesions and invasion patterns
- **Significant reduction in error patterns**
  - Fewer errors in lymph node anatomy and metastasis classification
- **Consistent interpretation of ambiguous expressions**
  - Uniform handling of "suspected", "suggested", negative/positive expressions

Combined use of SP and SFT approaches yields the highest performance, with SP excelling in complex and borderline cases, and SFT providing general pattern recognition and efficiency.

Conclusion

Our system demonstrates the feasibility of automated TNM classification using LLMs, with potential applications in clinical workflow optimization and standardization of cancer staging.
The dual approach of SFT and SP provides robust performance across various clinical scenarios.

Key benefits include:

(1) Efficiency Contribution

- 80% reduction in physician work time
- Data entry standardization
- Real-time diagnostic support

(2) Accuracy Improvements

- Enhanced inter-rater agreement
- Ensured guideline compliance
- Reduced classification errors

Future work will focus on multimodal integration, expansion to other cancer types, and further optimization of prompt engineering techniques to improve performance on edge cases.

Takahito Nakajima, MD, PhD,
Chair, Diagnostic Imaging and Interventional Radiology,
Professor, Institutes of Medicine,
University of Tsukuba, JAPAN
E-mail: nakajima@md.tsukuba.ac.jp
Web: https://tsukuba-radiology.info/