LLM-Based Two-stage Reasoning for TNM Staging: NECMedDX at the NTCIR-18 RadNLP Task

Soma Onishi, Daisaku Shibata, and Masanori Tsujikawa
NEC Corporation
Kawasaki, Japan
{soma-onishi, daisaku-shibata, tujikawa}@nec.com

Ryo Ishii, Junya Tominaga, and Hideki Ota
Tohoku University Hospital
Sendai, Japan
{ryo.ishii.d3, junya.tominaga.a6, hideki.ota.d6}@tohoku.ac.jp

Method

Our method employs a two-stage reasoning process for accurate TNM staging of lung cancer using radiology reports. We utilize a structured chain-of-thought (CoT) and iterative reviews to ensure consistency and transparency.

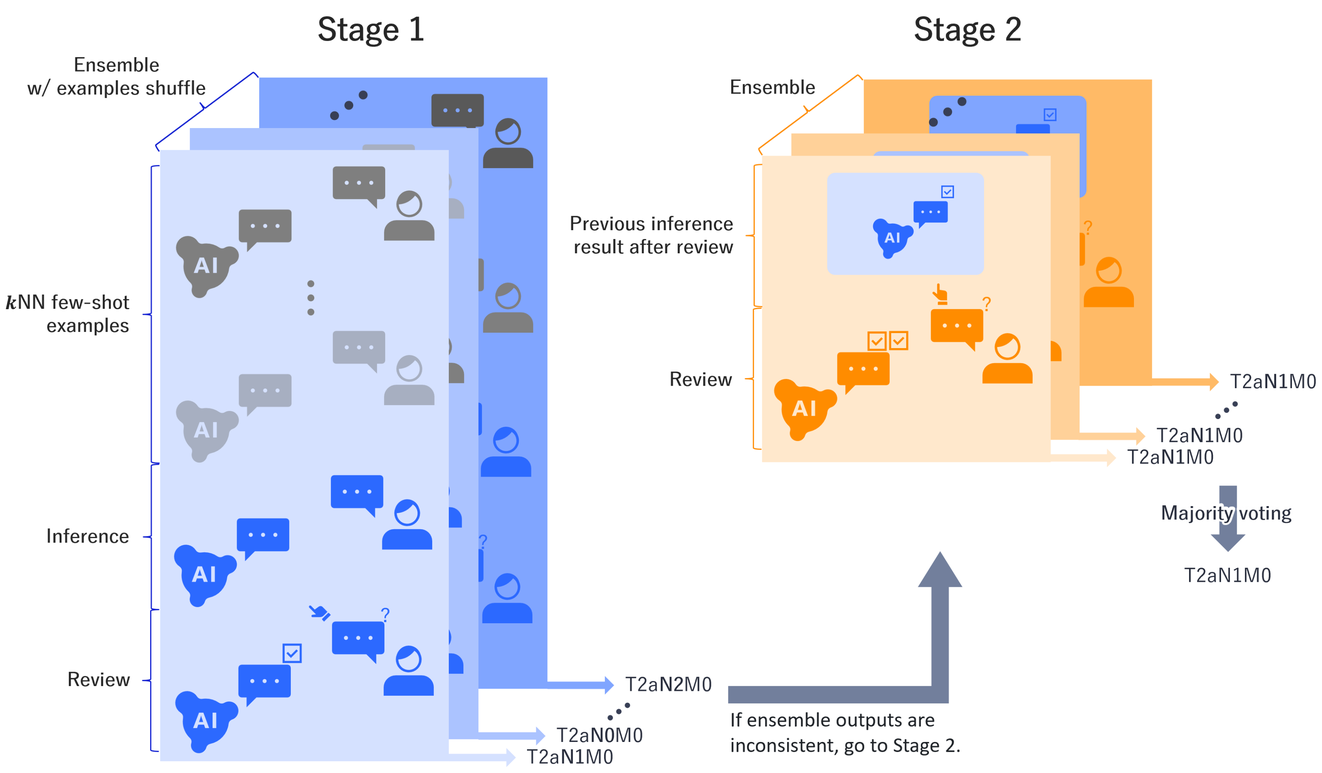
Purpose: Reduce inference errors through iterative review and majority voting, ensuring robust TNM classification.

Stage 1 (Initial Inference + Self-Review)

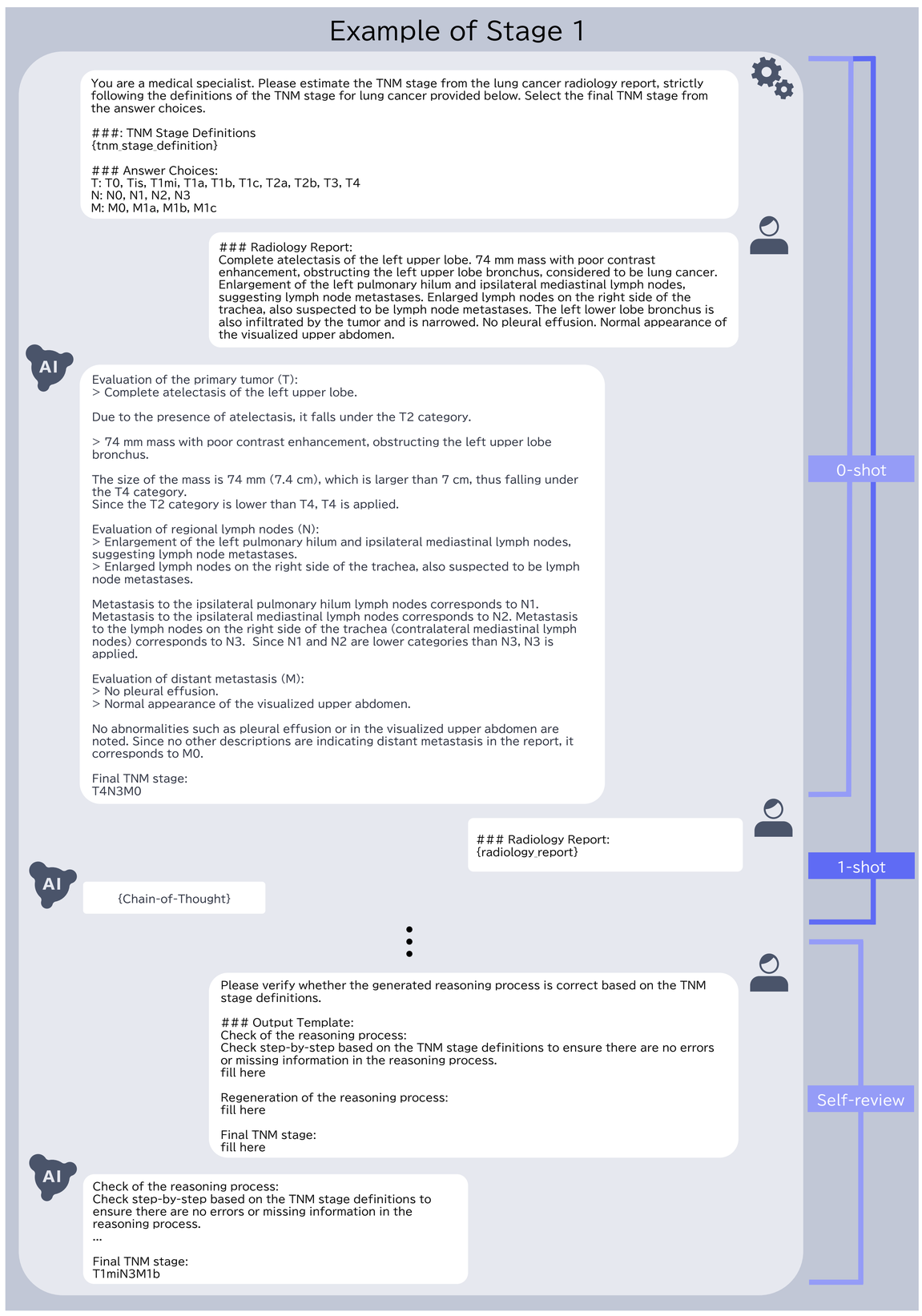
- **kNN few-shot learning with CoT:** Embed reports using text-embedding-3-large, retrieve top-10 similar cases, and generate reasoning (CoT) via the LLM.
- **Self-review:** The LLM revisits its own reasoning based on TNM definitions, detecting and correcting mistakes.
- **Ensemble check:** If five inferences produce ≥2 different predicted stages, they are deemed inconsistent, prompting Stage 2.

Stage 2 (Second Review + Majority Vote)

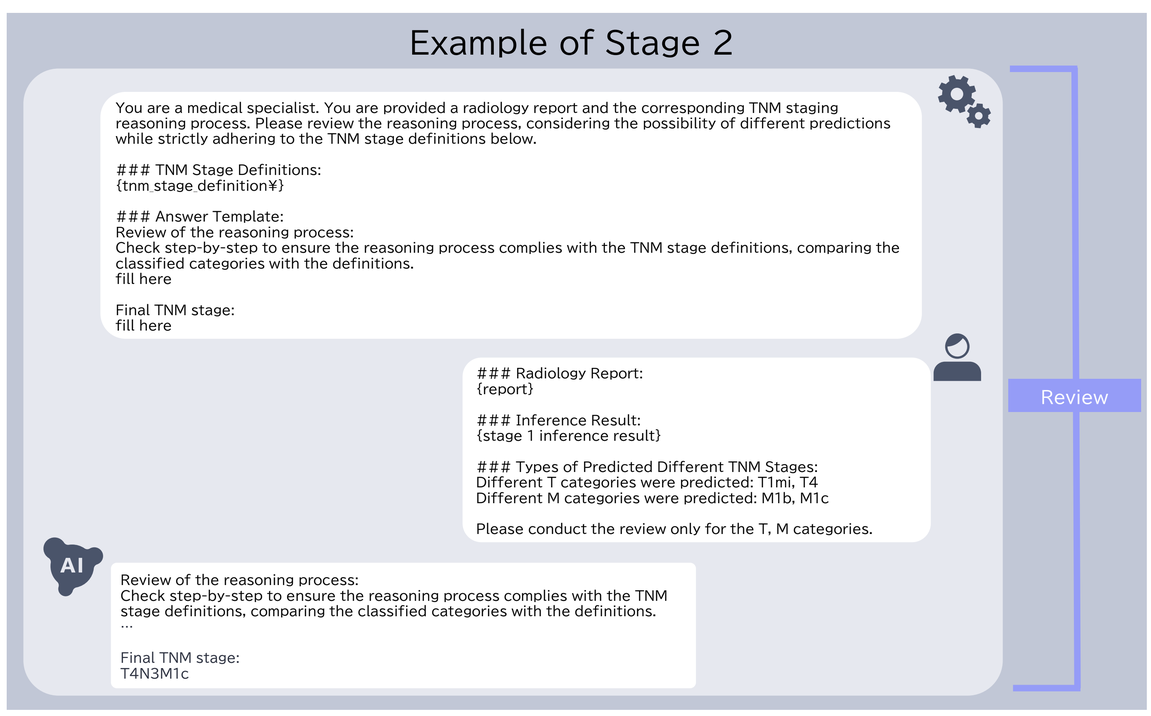
- **Second review:** The LLM re-evaluates Stage 1's reasoning as if it were provided by a different model/human, ensuring a different perspective.
- **Majority voting:** Final TNM stage is determined by majority vote of five re-reviewed outputs, enhancing agreement and accuracy.



Example of Stage 1



Example of Stage 2



Experiment

Setting: We conducted group k-fold cross-validation (k=18) to prevent data leakage across nine radiology reports for each case. A total of 18 lung-cancer cases were manually distinguished from the dataset.

Comparison Approaches: Zero-shot, +10-shot, +Ensemble, +Review (Stage 1), +Review (Stage 2)

LLM Selection: Claude 3.5 Sonnet

Results & Discussion

Results: The proposed method raised fine joint accuracy from 66.7% (Zero-shot) to 82.7%. Notably, T accuracy improved by 13 points, highlighting the effectiveness of iterative reviews.

Discussion: N accuracy dropped with +10-shot, suggesting possible CoT design issues. Adopting newer reasoning models (like OpenAI o-series, DeepSeek R1) can further enhance TNM-stage predictions.