Multi-Agent-Based Automated Clinical Data Extraction and Validation from Breast Cancer Pathology Reports Using Llama 3 Language Models

Sunghyeon PARK[1,3], Wona CHOI[1,2*], and InYoung, CHOI[1,3*]

[1] Department of Medical Informatics, College of Medicine, The Catholic University of Korea Seoul 06591, Republic of Korea;.

[2] Department of Nursing, Graduate School, The Catholic University of Korea, Seoul, Republic of Korea;

[3] Department of Biomedicine and Health Sciences, The Catholic University of Korea, Seoul, Republic of Korea

Introduction

- Pathology reports are essential for patient diagnosis and treatment but are typically written in free-text format, making structured data extraction difficult.
- Variations in report formats and writing styles across institutions further challenge extraction accuracy.
- Existing LLMs often perform poorly in zero-shot settings and tend to overfit under few-shot conditions.

Purpose

- This study aims to automate clinical data extraction from breast cancer pathology reports using the Llama 3.1 70B model.
- To ensure data reliability and accuracy, we developed a multi-agent system with iterative validation.

Methods

- We collected and analyzed 1,200 anonymized breast cancer pathology reports from three hospitals.
- Key fields for extraction included: Tumor Size, Tumor Site, Metastasis, Lymphatic Invasion, and pTNM Stage.
- We fine-tuned the Llama 3.1 70B model with 4-bit quantization to run on an in-house GPU environment (NVIDIA RTX A6000 with 48GB VRAM).
- The implementation was built using Python, Transformers, Langchain, and Langgraph frameworks.
- We designed a multi-agent system with two key components:
  - Extraction Agent: Identifies candidate values and their context from the text
  - Validation Agent: Assesses whether the extracted values are valid or invalid
- To increase reliability, the process is run independently three times, and the final result is determined by majority vote.

Results

- Using the Llama 3.1 70B model, we achieved 98% accuracy in extracting Tumor Size and Tumor Site under zero-shot conditions.
- Validation for additional fields—Metastasis, Lymphatic Invasion, and pTNM Stage—is ongoing using the Gemma 3 model.

Discussion

- Our findings demonstrate that recent LLMs, when combined with a multi-agent validation system, can effectively extract structured clinical data from unstructured pathology reports.
- The multi-agent approach significantly reduces both false negatives and false positives.

Conclusion

- An iterative, multi-agent validation system proves highly effective in extracting clinical data from free-text medical records.
- This approach holds strong potential for broader applications, such as standardizing medical data and supporting clinical decision-making across various healthcare settings.

References

1. Kim, Y., et al. (2020). Validation of deep learning natural language processing algorithm for keyword extraction from pathology reports in electronic health records. Scientific Reports, 10.
2. Zheng, S., et al. (2015). Support patient search on pathology reports with interactive online learning-based data extraction. Journal of Pathology Informatics, 6.
3. Atkinson-Abutridy, J. A., et al. (2004). Combining information extraction with genetic algorithms for text mining. IEEE Intelligent Systems.

Acknowledgement

This research was supported by the “Korea National Institute of Health”(KNIH) research project (2025ER050400).

<Figure1> Multi-Agent Framework for Pathology Report Information Extraction