doi:10.3233/SHTI240523

Enhancing Clinical Data Extraction from Pathology Reports: A Comparative Analysis of Large Language Models

Sunghyeon PARK[a,b], Wona CHOI[b,1] and InYoung CHOI[b,2]
[a] *Department of Medical Informatics, College of Medicine, The Catholic University of Korea, South Korea*
[b] *Department of Biomedicine and Health Sciences, South Korea*
ORCiD ID: Sunghyeon PARK https://orcid.org/0000-0002-2235-4358
Wona CHOI https://orcid.org/ 0000-0003-0269-6374
InYoung CHOI https://orcid.org/ 0000-0002-2860-9411

Abstract. This study evaluates the efficacy of a small large language model (sLLM) in extracting critical information from free-text pathology reports across multiple centers, addressing the challenges posed by the narrative and complex nature of these documents. Employing three variants of the Llama 2 model, with 7 billion, 13 billion, and 70 billion parameters, the research assesses model performance in both zero-shot and five-shot settings, offering insights into the impact of example-based learning. A specialized information extraction tool utilizing regular expressions for pattern identification serves as the benchmark for evaluating the models' accuracy. Conducted within a hospital's internal environment, the study emphasizes the clinical applicability of these findings. The results reveal significant variations in model performance, with the 70 billion parameter model achieving remarkable accuracy in the five-shot scenario, demonstrating the potential of sLLMs in enhancing the efficiency and accuracy of data extraction from pathology reports. The study highlights the importance of example-driven learning and the trade-offs between model size, accuracy, hallucination rates, and processing time. These findings contribute to the ongoing efforts to integrate advanced language models into clinical settings, potentially transforming patient care and biomedical research by mitigating the limitations of manual data extraction processes..

Keywords. Pathology Reports, Natural Language Processing, LLM

1. Introduction

In the domain of pathology, the significance of extracting pertinent information from pathology reports cannot be overstated, as these documents harbor critical clinical data pivotal for patient management and research. The narrative and complex nature of pathology reports often present challenges in gleaning meaningful, qualitative data efficiently, underscoring the necessity for effective information extraction techniques.

Initially, rule-based algorithms have shown promise in this domain. Martínez & Li (2011) [1], Buckley et al. (2012) [2], and Weegar & Dalianis (2015) [3] developed rule-based systems that apply predefined linguistic rules to identify and extract pertinent information from pathology reports. While these systems may lack the adaptability of machine learning models, they offer a high degree of precision in structured

[1] Corresponding Author: Wona Choi; E-mail: choiwona@gmail.com.
[2] Corresponding Author: InYoung Choi; E-mail: iychoi@catholic.ac.kr

environments, proving particularly useful in scenarios with well-defined extraction criteria.

Further advancements in this field include the use of deep learning approaches, particularly convolutional neural networks (CNNs), for information extraction from pathology reports. The studies by Yoon et al. (2016) [4], Qiu et al. (2018)[5] and Alawad et al. (2019)[6], illustrate the efficacy of CNN-based models in identifying and extracting relevant data from complex textual information. These models leverage the hierarchical structure of CNNs to capture intricate patterns in the text, thereby improving the accuracy and speed of data extraction processes.

More recently, the application of transformer-based models such as BERT has garnered significant attention. Kim et al. (2020) [7] and Zhou et al. (2022) [8] have demonstrated the superior performance of BERT-based models in extracting clinical information from free-text pathology reports. These models benefit from their ability to understand context and semantics at a deeper level, resulting in more accurate and comprehensive data extraction compared to traditional methods.

These pioneering efforts exemplify the clinical importance and the urgent need for sophisticated information extraction systems in the field of pathology. Our study aims to bridge the existing gap in knowledge by rigorously evaluating the capability of a specific language learning model (sLLM) in accurately extracting relevant information from free-text pathology reports across multiple centers. By doing so, we aim to offer a robust solution to the inefficiencies and inaccuracies inherent in manual data extraction processes, ultimately enhancing patient care and facilitating biomedical research.

2. Methods

In our study, we analyzed 793 pathology reports from a broad spectrum of 42 diseases obtained from Clinical Data Warehouse (CDW) of Catholic Medical Center (CMC). Each report was approved by the Institutional Review Board (IRB) after an ethical review, ensuring adherence to ethical standards and patient confidentiality (IRB number : KIRB-신 20231201-019).

Our primary objective was to evaluate the efficacy of large language models (LLMs) in extracting critical information, such as tumor size and location, from these complex medical documents. For this purpose, we utilized three variants of the Llama 2 model, with capacities of 7 billion (7b), 13 billion (13b), and 70 billion (70b) parameters, to assess their performance in information extraction tasks.

The models were tested in two distinct scenarios: a zero-shot setting, where the models were not provided with any prior examples, and a five-shot setting, where they were given five annotated examples to aid in inference the task. This approach allowed us to gauge the impact of example-based learning on the models' accuracy.

To benchmark the performance of the LLMs, we employed a specialized information extraction tool developed by our research team. This tool utilizes regular expressions to identify specific patterns within the text, a method previously validated in similar studies. By comparing the outputs of the LLMs against the data extracted by this tool, we were able to measure the accuracy of the models.

The study was conducted within the hospital's internal environment (CPU : AMD Ryzen Threadripper PRO 5955WX 16-cores, GPU : NVIDA RTX A6000 VRAM 48GB), leveraging the secure and controlled access to the pathology reports. This setting

provided a real-world context for our investigation, enhancing the relevance and applicability of our findings.

Our analysis focused on three key metrics to comprehensively evaluate the models' performance: accuracy, hallucination, and computational efficiency. These metrics offered insights into the models' capability to correctly identify and extract pertinent information from the pathology reports, as well as their operational feasibility in a clinical setting.

3. Results

In our study, the application of Llama 2 models to extract clinically significant information from pathology reports revealed intriguing outcomes, particularly when contrasting the performance of different model sizes under 5-shot and 0-shot conditions. Notably, the 70b model in the 5-shot scenario achieved an impressive accuracy of 97.7% for both tumor size and site, with no instances of hallucination, implying a remarkable precision in identifying relevant details without adding incorrect information. This high level of accuracy, however, came at the expense of time, requiring approximately 406 minutes to process.

Table 1. Extraction Results

		Time(min)	Size Accuracy	Size Hallucination	Site Accuracy	Site Hallucination
70b	5-shot	406	97.73	0	97.73	0
	0-shot	456	38.71	0	38.46	0
13b	5-shot	109	66.71	119	63.30	231
	0-shot	139	0	0	0	0
7b	5-shot	76	76.67	168	53.59	317
	0-shot	93	0	0	0	0

Conversely, when the 70b model operated under 0-shot conditions, the accuracy plummeted to 38.7% for size and 38.5% for the site, highlighting the significant impact of providing contextual examples on model performance. The 13b model in the 5-shot setup displayed moderate accuracy levels of 66.7% for size and 63.3% for site but encountered challenges with hallucinations, particularly for site data, where 231 instances were recorded, suggesting the introduction of incorrect information. The processing time for this model was considerably lower, at 109 minutes.

The 7b model, the smallest among the tested models, presented a balanced outcome in the 5-shot mode with a 76.7% accuracy for size and 53.6% for the site, albeit with a higher rate of hallucinations, especially for site data, where 317 instances were noted. This model required the least amount of time, taking only 76 minutes. In stark contrast, both the 13b and 7b models failed to provide accurate results under the 0-shot condition, recording a 0% accuracy rate for both size and site, which underscores the critical importance of few-shot learning in enhancing model performance. These findings underscore the nuanced trade-offs between model size, accuracy, hallucination rates, and processing time in the context of extracting vital clinical information from narrative pathology reports.

4. Discussion

The superior performance of the Llama 2 70b model in our study can be attributed to its advanced architecture, designed to capture complex patterns and contextual nuances within textual data. The model's ability to process large volumes of data and understand intricate medical terminologies likely contributes to its high accuracy in extracting relevant clinical information from pathology reports. The significant improvement observed in the 5-shot setting indicates the model's capacity to leverage minimal annotated examples to enhance its inference capabilities, emphasizing the value of few-shot learning in clinical applications.

A deeper examination of the model's internal mechanisms reveals that the 70b variant benefits from its extensive parameters and sophisticated attention mechanisms.[9] These features enable the model to maintain context over long text spans, accurately identifying and extracting specific details such as tumor size and site. Performance metrics, including precision, recall, and F1 scores, further illustrate the model's robustness in clinical information extraction tasks. Understanding these metrics is essential for medical professionals to appreciate the strengths and limitations of AI tools in clinical settings.

Our study also underscores the importance of maintaining data privacy and security in medical settings. By processing sensitive information locally, we mitigate risks associated with data breaches and ensure compliance with stringent privacy standards. This approach is crucial for maintaining patient confidentiality and trust in the use of AI technologies in healthcare.

While our study demonstrates the potential of LLMs like Llama 2 in clinical data extraction, it is not without limitations. The substantial processing time associated with larger models such as the 70b variant highlights the need for more efficient algorithms. Future research should explore techniques to optimize computational efficiency, such as model pruning or quantization, to reduce processing times without compromising accuracy. Additionally, our study primarily focused on English-language pathology reports; expanding research to include reports in other languages and formats will validate the model's versatility and robustness.

In conclusion, the deployment of advanced LLMs like Llama 2 in extracting clinical information from pathology reports represents a significant advancement in medical data management. Future research should focus on optimizing the balance between model performance and computational efficiency, exploring ways to reduce processing times without compromising accuracy. Additionally, expanding the scope of studies to include diverse pathology report formats and languages can further validate the robustness and versatility of these models, ultimately paving the way for their broader adoption in clinical practice. Ethical considerations and continuous evaluation will ensure these technologies enhance healthcare delivery while maintaining patient trust and safety.

5. Conclusions

Our study demonstrates the effectiveness of large language models (LLMs), particularly the Llama 2 70b variant, in accurately extracting critical clinical information from pathology reports. The 5-shot learning scenario significantly enhances the models' performance, as evidenced by the high accuracy rates achieved by the 70b model for both tumor size and location without any hallucinations. The findings highlight the potential

of advanced LLMs to streamline the information extraction process, thereby reducing manual effort and minimizing errors in clinical data management. However, the substantial processing time associated with larger models necessitates consideration of computational efficiency for practical deployment. This study underscores the importance of few-shot learning in improving model accuracy and suggests that incorporating LLMs into clinical workflows can substantially benefit patient care and biomedical research.

Acknowledgement

This work was supported by a fund (#2024020E7E3-00) by Research of Korea Centers for Disease Control and Prevention and supported by the National Research Foundation of Korea (NRF) grand funded by the Korea government (MSIT) (No.2019R1ASA2027588)

References

[1] Martinez D, Li Y. Information extraction from pathology reports in a hospital setting. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management. 2011 Oct; p. 1877-82.
[2] The feasibility of using natural language processing to extract clinical information from breast pathology reports. J Pathol Inform. 2012;3(1):23.
[3] Weegar R, Dalianis H. Creating a rule based system for text mining of Norwegian breast cancer pathology reports. In: Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis. 2015 Sep; p. 73-8.
[4] Yoon HJ, Ramanathan A, Tourassi G. Multi-task deep neural networks for automated extraction of primary site and laterality information from cancer pathology reports. In: Advances in Big Data: Proceedings of the 2nd INNS Conference on Big Data, October 23-25, 2016, Thessaloniki, Greece. 2017;2:195-204.
[5] Qiu J, Yoon H, Fearn P, Tourassi G. Deep learning for automated extraction of primary sites from cancer pathology reports. IEEE J Biomed Health Inform. 2018;22:244-51.
[6] Alawad M, Gao S, Qiu J, Yoon H, Christian J, Penberthy L, et al. Automatic extraction of cancer registry reportable information from free-text pathology reports using multitask convolutional neural networks. J Am Med Inform Assoc. 2019;27:89-98.
[7] Kim Y, Lee JH, Choi S, Lee JM, Kim JH, Seok J, et al. Validation of deep learning natural language processing algorithm for keyword extraction from pathology reports in electronic health records. Sci Rep. 2020;10.
[8] Zhou S, Wang N, Wang L, Liu H, Zhang R. CancerBERT: a cancer domain-specific language model for extracting breast cancer phenotypes from electronic health records. J Am Med Inform Assoc. 2022;29(7):1208-16.
[9] Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, et al. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288. 2023.