BURExtract-Llama: An LLM for Clinical Concept Extraction in Breast Ultrasound Reports

Yuxuan Chen[†]
New York University
New York, USA
yc7087@nyu.edu

Haoyan Yang[†]
New York University
New York, USA
hy2847@nyu.edu

Hengkai Pan
Carnegie Mellon University
Pittsburgh, USA
hengkaip@andrew.cmu.edu

Fardeen Siddiqui
NYU Langone Health
New York, USA
Fardeen.Siddiqui@nyulangone.org

Antonio Verdone
NYU Langone Health
New York, USA
Antonio.Verdone@nyulangone.org

Qingyang Zhang
NYU Shanghai
Shanghai, China
qz2208@nyu.edu

Sumit Chopra
NYU Langone Health
New York, USA
Sumit.Chopra@nyulangone.org

Chen Zhao
NYU Shanghai
Shanghai, China
cz1285@nyu.edu

Yiqiu Shen*
NYU Langone Health
New York, USA
Yiqiu.Shen@nyulangone.org

ABSTRACT

Breast ultrasound plays a pivotal role in detecting and diagnosing breast abnormalities. Radiology reports summarize key findings from these examinations, highlighting lesion characteristics and malignancy assessments. However, extracting this critical information is challenging due to the unstructured nature of radiology reports, which often exhibit varied linguistic styles and inconsistent formatting. While proprietary LLMs like GPT-4 effectively retrieve information, they are costly and raise privacy concerns when handling protected health information. This study presents a pipeline for developing an in-house LLM to extract clinical information from these reports. We first utilize GPT-4 to create a small subset of labeled data, then fine-tune a Llama3-8B using this dataset. Evaluated on a subset of reports annotated by clinicians, the proposed model achieves an average F1 score of 84.6%, which is on par with GPT-4. Our findings demonstrate that it is feasible to develop an in-house LLM that not only matches the performance of GPT-4 but also offers cost reductions and enhanced data privacy.

CCS CONCEPTS

• **Computing methodologies** → **Information extraction**.

KEYWORDS

Breast Ultrasound, Radiology Reports, Clinical Information Extraction, LLM, Fine-Tuning

MCHM '24, October 28, 2024, Melbourne, VIC, Australia.

ACM ISBN 979-8-4007-1195-4/24/10
https://doi.org/10.1145/3688868.3689200

ACM Reference Format:
Yuxuan Chen[†], Haoyan Yang[†], Hengkai Pan, Fardeen Siddiqui, Antonio Verdone, Qingyang Zhang, Sumit Chopra, Chen Zhao, and Yiqiu Shen*. 2024. BURExtract-Llama: An LLM for Clinical Concept Extraction in Breast Ultrasound Reports. In *Proceedings of the 1st International Workshop on Multimedia Computing for Health and Medicine (MCHM '24), October 28, 2024, Melbourne, VIC, AustraliaProceedings of the 32nd ACM International Conference on Multimedia (MM'24), October 28-November 1, 2024, Melbourne, Australia.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3688868.3689200

1 INTRODUCTION

Ultrasonography is a medical imaging technique extensively used in diagnosing breast pathology. Traditionally, breast ultrasonography is utilized to follow up on atypical mammography findings or as an initial diagnostic tool for patients with suspected breast malignancy such as those with a palpable mass [8]. While analyzing the ultrasound for potential findings, breast radiologists construct a comprehensive report that represents their conclusions, which are then communicated to the patient, other healthcare providers, and trainees such as radiology residents. Breast radiologists use a widely accepted classification system developed by the American College of Radiology known as the Breast Imaging-Reporting and Data System (BI-RADS) to characterize the extensive amount of information associated with every lesion found [15]. This system includes guidelines for the appropriate lexicon for each discovery and the suspicion of malignancy established by the findings. Each report contains details such as the mass's location within the breast, defining characteristics, malignancy suspicion, and other critical attributes. Despite the system provided by BI-RADS, the substantial data volume within each report can be challenging to manage. Variations in institutional teaching styles, personal preferences, and clinical discrepancies further contribute to the complexity and often render the information unstructured and difficult to extract systematically.

[†] These authors contributed equally to this work.
* Corresponding author: Yiqiu Shen.

Adopting a methodology to extract this information would be useful to both physicians and patients. For instance, Clinical Decision Support Systems (CDSS) could utilize this data to offer real-time alerts and recommendations, enhancing the precision of data collection and analysis [10]. CDSS can identify patterns, trends, and correlations that might otherwise be overlooked. Transparency and traceability in extracted data also support auditing and quality assurance processes. This would enable healthcare institutions to monitor adherence to clinical guidelines and identify areas of improvement to establish a higher standard of patient care. The extraction of report data would highly benefit trainees as well. It would allow for the characterization of the cases encountered by radiology residents and the tracking of diagnostic decisions to improve their clinical reasoning skills.

Commercial LLMs such as GPT-4 [17] offer satisfactory accuracy in information retrieval tasks but can be costly for large datasets and pose privacy risks when processing medical reports containing protected health information (PHI). To address these concerns, institutions often develop in-house LLMs by fine-tuning open-source models such as LLaMA [1]. Although these models may perform worse than their proprietary counterparts, they enhance patient privacy and reduce costs. However, creating in-house LLMs is challenging due to the need for large-scale, high-quality annotated clinical datasets [16]. Additionally, the unstructured nature of radiology reports, which often vary in style and format, complicates the extraction of relevant information.

We aim to standardize a workflow for institutions to train an in-house LLM without the need for costly, large-scale manually annotated datasets. In this work, we present a pipeline for developing an in-house LLM to extract relevant clinical information from radiology reports. As depicted in Figure 1, our method begins by using a high-performance proprietary LLM like GPT-4 to create a small subset of labeled data. We then fine-tune an open-source LLM using this dataset. Evaluated on a subset manually annotated by clinicians, our model, named BURExtract-Llama, achieves high accuracy in extracting clinical information from breast ultrasound reports, comparable to GPT-4.

2 RELATED WORKS

In clinical concept extraction, methodologies have undergone substantial evolution, progressing from early rule-based systems to deep learning approaches. Initially, the domain relied heavily on rule-based systems, such as the method proposed by Friedman et al., which employs a three-phase processing approach—parsing to identify text structure, regularization to standardize terms, and encoding to map terms to a controlled vocabulary [9].

With technological growth, the introduction of machine learning models such as Conditional Random Fields [7] and Support Vector Machines [20] shifted the field towards more dynamic analysis. Furthermore, the emergence of deep learning architectures, including Recurrent Neural Networks and Long Short-Term Memory models [19], has freed humans from manual feature engineering by employing distributed word representations [18], while effectively capturing the subtle nuances and complexities inherent in clinical text [2].

The advent of the transformer [21] marked another significant milestone. Transformer-based models such as BERT [5], ALBERT [13], RoBERTa [14], and ELECTRA [3] have been explored for clinical concept extraction tasks [22]. The incorporation of self-attention mechanisms in these models enhances the long-term dependencies management, thereby providing a more sophisticated tool for this task.

Recent advancements in LLMs, such as GPT-4 [17] and Llama 3 [1], have revolutionized the field of natural language processing. Their proficiency in tasks ranging from text summarization to question answering highlights their versatility. Consequently, our research aims to investigate the fine-tuning of LLMs for clinical concept extraction, demonstrating their capability to manage complex clinical text.

Table 1: Keys and values of lesions

Keys	Values
k1=depth	Posterior, Middle, Anterior, N/A
k2=anatomical region	Retroareolar, Axillary Tail, Periareolar, Subareolar, Retropectoral, N/A
k3=lesion type	Nodule, Cyst, Mass, Lymph Node, Scar, Duct, Seroma, Post-Surgical Change, Post-Biopsy, N/A
k4=lesion shape	Oval, Round, Irregular, N/A
k5=orientation	Parallel, Non-Parallel, Other, N/A
k6=lesion margins	Circumscribed, Obscured, Angular, Microlobulated, Spiculated, Lobulated, Irregular, Septated, N/A
k7=echogenicity	Anechoic, Hyperechoic, Hypoechoic, Isoechoic, Heterogeneous, Solid, N/A
k8=calcifications	Yes, No, N/A
k9=vascularity	Absent, Present, N/A
k10=posterior features	Enhancement, Shadowing, N/A
k11=lesion subtype	Abnormal Lymph Node, Simple Cyst, Complicated Cyst, Cyst with Debris, Reactive Lymph Node, Fat Necrosis, Sebaceous Cyst, Lipoma, Cyst Cluster, Focally Ectatic Duct with Debris, N/A
k12=next step	1 Year Screening Mammogram, MRI Follow Up, 6 Months Follow-Up, 12 Months Follow-up, Fine Needle Aspiration, Ultrasound Guided Core Biopsy, Surgical Excision, N/A
k13=suspicion of malignancy	Low, Moderate, High, Benign, Probably Benign, Negative
k14=side of breast	Left, Right, N/A
k15=clock position	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, N/A
k16=distance from nipple	Numeric Value (cm)

3 METHODS

3.1 Problem Formulation

Formally, let $\mathbf{x}$ denote a breast ultrasound report, containing textual descriptions of one or multiple lesions. Our goal is to transform each report $\mathbf{x}$ into a list of JSON dictionaries $\mathbf{y} = [y_1, y_2, \ldots, y_n]$, where

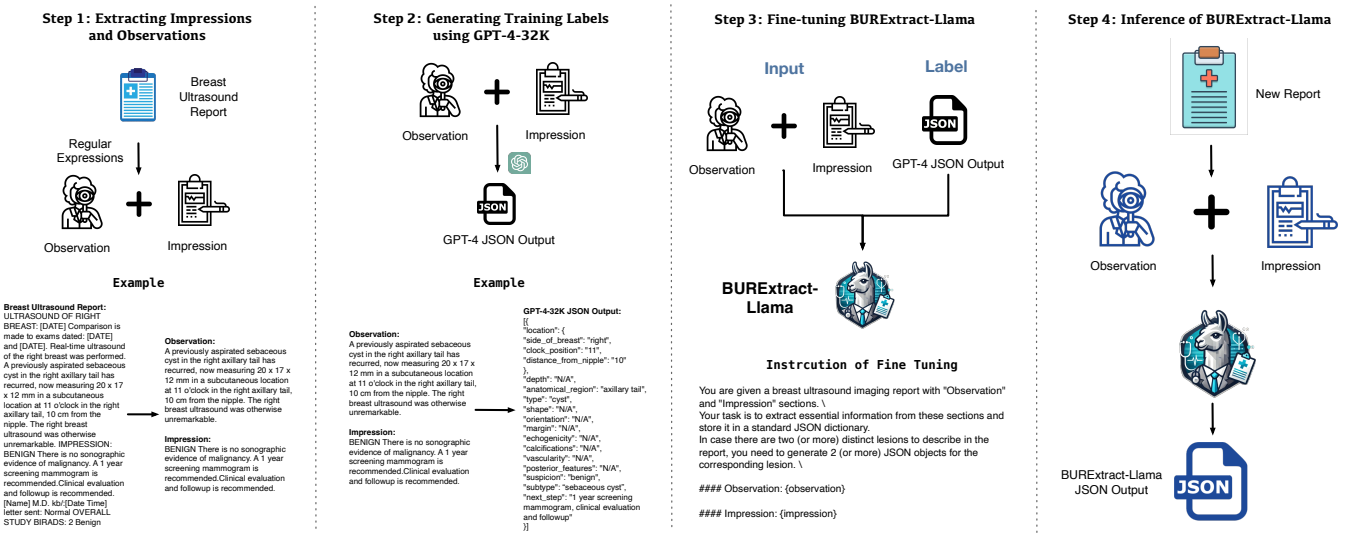


Figure 1: The pipeline for building and utilizing the BURExtract-Llama. Steps 1-3 cover the building process: extracting observations and impressions, generating training labels with GPT-4, and fine-tuning BURExtract-Llama. Step 4 shows how BURExtract-Llama infers from new reports and outputs structured JSON.

each dictionary y_i corresponds to a lesion described in $\mathbf{x}$. Each dictionary y_i is defined as $y_i = \{(k_1, v_1), (k_2, v_2), \ldots, (k_m, v_m)\}$, where $k_1, k_2, \ldots, k_m$ are the keys representing lesion attributes, and $v_1, v_2, \ldots, v_m$ are the corresponding values extracted from the report. In our study, there are 16 keys of interest, as shown in Table 1. Our objective is to train an LLM that processes $\mathbf{x}$ to generate $\mathbf{y}$.

3.2 Pipeline for building in-house LLMs

As illustrated in Figure 1, our pipeline consists of three steps: 1) extracting observations (also referred to as *findings*) and impressions from reports; 2) generating training labels using GPT-4 [17]; 3) fine-tuning Llama3-8B using Q-LoRA [4].

3.2.1 Extract Observations and Impressions. Figure 1 illustrates the components of a typical breast ultrasound report, which includes: 1) patient and examination details; 2) methodology of the ultrasound; 3) findings, detailing the notable features observed in the breast tissue; 4) impression, offering a concise summary, interpretation of findings, and recommendations for subsequent actions; and 5) disclosure. Unlike other sections whose information is available in a structured format, the observation and impression sections contain key descriptions of lesions in an unstructured manner. Consequently, we focus on extracting clinical information primarily from these two sections. We employ regular expressions to isolate the observation and impression sections, ensuring the input for the LLMs is in a clean format.

3.2.2 Generate Training Labels using GPT-4-32K. Our institute utilizes a HIPAA-compliant GPT-4 instance, allowing us to process medical records securely. Leveraging the in-context learning (ICL) capabilities [6] of GPT-4-32K [17], we generate JSON labels to fine-tune the Llama-3 model. We design a prompt template that includes 7 carefully curated examples to illustrate the expected JSON format, ensuring clarity and consistency in the generated labels.

3.2.3 Fine-Tuning. We fine-tune Llama-3-8B by utilizing QLoRA [4] on pairs of report and JSON outputs generated by GPT-4. Fine-tuning allows Llama-3-8B to perform well on our specific task by adapting our dataset. We choose QLoRA, which is a quantized version of Low-Rank Adaptation (LoRA) [11], as it can significantly reduce memory and computational requirements.

First, we quantize the model weights as $\theta_q = Q(\theta)$. Instead of performing a full fine-tuning, we only update the two low-rank matrices B and A using backpropogration, as demonstrated in (1). This approach reduces the size of the parameters that need to be updated from approximately 16GB to about 100MB compared to full fine-tuning, while maintaining model performance. The matrices B and A are updated through backpropagation based on the loss function defined in (2) and (3). This ensures that the model minimizes the discrepancy between the predicted outputs and the actual outputs at the token level.

In summary, the QLoRA method significantly reduces the number of parameters that need to be updated compared to a full fine-tuning, enhancing both memory and computational efficiency.

$$\Delta\theta_i \approx \frac{\alpha}{r} BA, \theta_i \leftarrow \theta_i + \Delta\theta_i \tag{1}$$

where $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times k}$, with rank $r \ll \min(d, k)$. r represents the LoRA attention dimension and α is for LoRA scaling.

$$p_{\theta_q}(\mathbf{s}|\mathbf{x}) = \prod_{j=1}^{t} p_{\theta_q}(s_j|\mathbf{x}, \mathbf{s}_{<j}) \tag{2}$$

$$L(\theta_q) = -\mathbb{E}_{\mathbf{x},\mathbf{s}}(\cdot|\mathbf{x}) \left[\log p_{\theta_q}(\mathbf{s}|\mathbf{x})\right] \tag{3}$$

where $\mathbf{s}$ is the output string, t is the length of string $\mathbf{s}$, $s_{<j} = \emptyset$ for $j = 1$ and $s_{<j} = [s_1, \ldots, s_{j-1}]$ for $j > 1$.

Table 2: Hyper-parameter setting.

QLoRA Parameters	Bitsandbytes Parameters	Training Parameters
LoRA attention dimension: 64	Use 4-bit precision: True	Number of training epochs: 4
Alpha parameter for LoRA scaling: 16	Compute dtype for 4-bit: float16	Initial learning rate: 2e-4
Dropout probability for LoRA: 0.1	Quantization type: nf4	Learning rate schedule: constant
	Use nested quantization: False	Optimizer: Adam
		Weight decay: 0.001
		Warmup ratio: 0.03
		Enable bf16 training: True
		Per device train batch size: 4
		Per device eval batch size: 4
		Gradient accumulation steps: 1
		Maximum gradient norm: 0.3
		Packing: False

Table 3: Performance comparison for per key matching.

	ICL of Llama3-Instruct-8B			BURExtract-Llama			ICL of GPT-4		
	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1
depth	0.798	0.906	0.849	0.853	0.935	0.892	0.839	0.934	0.884
anatomical region	0.757	0.859	0.805	0.807	0.884	0.844	0.780	0.867	0.821
lesion type	0.702	0.797	0.746	0.775	0.849	0.811	0.780	0.867	0.821
lesion shape	0.771	0.875	0.820	0.826	0.905	0.863	0.835	0.929	0.879
orientation	0.803	0.911	0.854	0.858	0.940	0.897	0.858	0.954	0.903
lesion margins	0.761	0.865	0.810	0.830	0.910	0.868	0.826	0.918	0.870
echogenicity	0.743	0.844	0.790	0.794	0.869	0.830	0.798	0.888	0.841
calcifications	0.812	0.922	0.863	0.867	0.950	0.906	0.849	0.944	0.894
vascularity	0.789	0.896	0.839	0.849	0.930	0.887	0.835	0.929	0.879
posterior features	0.789	0.896	0.839	0.862	0.945	0.902	0.853	0.949	0.899
lesion subtype	0.693	0.786	0.737	0.784	0.859	0.820	0.766	0.852	0.807
next step	0.674	0.766	0.717	0.743	0.814	0.777	0.748	0.832	0.787
suspicion of malignancy	0.564	0.641	0.600	0.615	0.673	0.643	0.638	0.709	0.671
side of breast	0.817	0.927	0.868	0.872	0.955	0.911	0.858	0.954	0.903
clock position	0.734	0.833	0.780	0.803	0.879	0.839	0.798	0.888	0.841
distance from nipple	0.743	0.844	0.790	0.807	0.884	0.844	0.803	0.893	0.845
Average	0.747	0.848	0.794	0.809	0.886	0.846	0.804	0.894	0.847

3.3 Inference

During the inference phase of BURExtract-Llama, a report undergoes processing to extract the observation and impression components. These elements are integrated with the fine-tuning instructions detailed in Step 3 of Figure 1 to execute inference, producing a structured JSON output.

4 EXPERIMENTS

4.1 Dataset

The dataset employed in this study comprises 4,000 breast ultrasound reports sourced from NYU Langone Health. To facilitate model development, we partition this dataset into three distinct subsets: 3,600 reports for training, 280 for validation, and 120 for testing, ensuring there is no overlap between them. The training and validation sets are annotated using GPT-4, while the test set is labeled by clinicians. This approach is taken to minimize potential errors associated with automated labeling, ensuring an accurate evaluation of the model's performance.

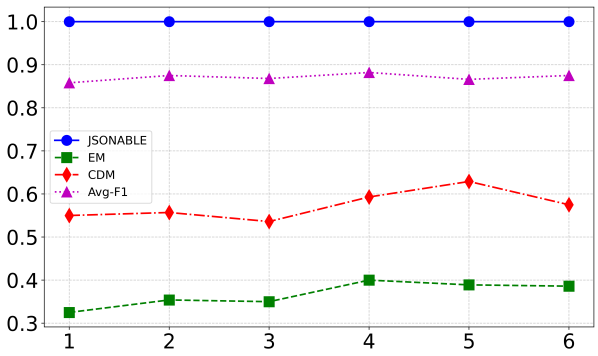
4.2 Training Detail

We include all hyper-parameters in Table 2. During the validation, we focus on the *number of training epochs* and the *LoRA attention dimension* to select the best model. As detailed in Figure 2, epoch 4 and a LoRA attention dimension of 64 yielded the best performance. Our best model was trained on an Nvidia A100 GPU for 1.5 hours.

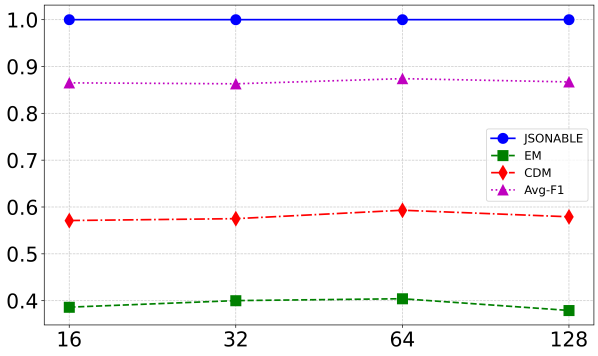
4.3 Evaluation Metrics

We employ two primary categories of evaluation metrics: per report matching and per key matching. **Per Key Matching** calculates recall, precision, and F1 score for each of 16 keys. **Per Report Matching** includes three metrics:

- **JSONable Accuracy**: The percentage of the LLM's outputs that can be converted into a valid list of dictionaries.
- **Exact Matching (EM) Accuracy**: The percentage of the LLM's outputs, after converted to JSON, that exactly match the ground truth for all 16 keys $\{k1, k2 \dots k16\}$.



(a) Epochs



(b) Attention Dimension

Figure 2: Hyperparameter optimization results. Details of the four metrics in the legend are in Section 4.3.

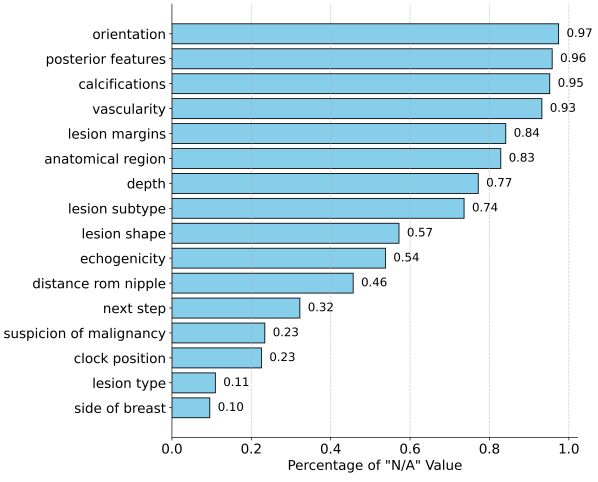


Figure 3: N/A distribution for each key in the training set.

- **Close Domain Matching (CDM) Accuracy**: The percentage of the LLM outputs, after converted to JSON, that match the ground truth for a subset of keys $\{k1, k2 \ldots k10\}$ containing categorical values critical for diagnosis.

4.4 Results

Table 3 provides a detailed comparison of precision, recall, and F1 scores for each key. The average F1 score difference between BURExtract-Llama and GPT-4's ICL is within 0.1%. Notably, our model outperforms GPT-4 in keys including "depth," "anatomical region," and "posterior features," with a maximum difference of 2.3% in "anatomical region".

As shown in Table 4, BURExtract-Llama achieves 100% structured output for the JSONable accuracy, demonstrating its ability to follow the prompt instructions and produce the desired output format. Our BURExtract-Llama outperforms Llama3-8B with ICL by 12.5% in EM and 10.0% in CDM, highlighting the benefits of fine-tuning. It matches GPT-4 in EM and is only 0.9% behind in CDM, proving BURExtract-Llama to be a viable alternative to GPT-4. Additionally, our BURExtract-Llama can infer a new report in approximately 2 seconds with vllm [12], demonstrating its low latency.

Table 4: Performance comparison for per report matching.

	ICL of Llama3-Instruct-8B	BURExtract-Llama	ICL of GPT-4
JSONable	**1.000**	**1.000**	**1.000**
EM	0.333	**0.458**	**0.458**
CDM	0.583	0.683	**0.692**

4.5 Error Analysis

We conduct an error analysis to identify instances where our model failed. Below, we provide a summary of these cases for reference.

- **Missing Lesion**: The model sometimes fails to identify all lesions mentioned within a radiology report. This limitation is reflected in the lower recall rate compared to precision, as shown in Table 3.
- **Lesion Attribute Confusion**: The model occasionally misattributes the values of one lesion to another, leading to incorrect associations.
- **Handling "N/A"**: The model sometimes predicts "N/A" for attributes with actual values, likely due to the high frequency of "N/A" in the training set, as demonstrated in Figure 3.

5 CONCLUSION

This study presents a workflow of building in-house LLM to extract relevant clinical information from radiology reports. We first utilize GPT-4 to label a small dataset and then fine-tune Llama3-8B on this labeled dataset. The fine-tuned model, BURExtract-Llama, demonstrates performance comparable to GPT-4. Still, we recognize several limitations. First, we acknowledge that the training labels generated by GPT-4 might be flawed. A future research direction could involve investigating methods to handle noisy labels to address this issue. The second limitation is the lack of external validation. Since the writing style of reports from a single hospital tends to be consistent, it is uncertain how the model would perform on data from institutions with different writing styles. Future research should include an external test set to evaluate the generalizability of the BURExtract-Llama.

ACKNOWLEDGMENTS

This work was supported in part by the Murray J. Berenson, MD Grant in Medical Education Research from the NYU Program for Medical Education Innovations and Research.

REFERENCES

[1] AI@Meta. 2024. Llama 3 Model Card. (2024). https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md

[2] Raghavendra Chalapathy, Ehsan Zare Borzeshi, and Massimo Piccardi. 2016. Bidirectional LSTM-CRF for clinical concept extraction. *arXiv preprint arXiv:1611.08373* (2016).

[3] Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555* (2020).

[4] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems* 36 (2024).

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[6] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey on in-context learning. *arXiv preprint arXiv:2301.00234* (2022).

[7] Andrea Esuli, Diego Marcheggiani, and Fabrizio Sebastiani. 2013. An enhanced CRFs-based system for information extraction from radiology reports. *Journal of biomedical informatics* 46, 3 (2013), 425–435.

[8] Karin Flobbe, Anne Marie Bosch, Alfons GH Kessels, Geerard L Beets, Patricia J Nelemans, Maarten F von Meyenfeldt, and Joseph MA van Engelshoven. 2003. The additional diagnostic value of ultrasonography in the diagnosis of breast cancer. *Archives of internal medicine* 163, 10 (2003), 1194–1199.

[9] Carol Friedman, Philip O Alderson, John HM Austin, James J Cimino, and Stephen B Johnson. 1994. A general natural-language text processor for clinical radiology. *Journal of the American Medical Informatics Association* 1, 2 (1994), 161–174.

[10] Francini Hak, Tiago Guimarães, and Manuel Santos. 2022. Towards effective clinical decision support systems: A systematic review. *PLoS One* 17, 8 (2022), e0272846.

[11] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. arXiv:2106.09685 [cs.CL] https://arxiv.org/abs/2106.09685

[12] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient Memory Management for Large Language Model Serving with PagedAttention. arXiv:2309.06180 [cs.LG] https://arxiv.org/abs/2309.06180

[13] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942* (2019).

[14] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[15] Ellen B Mendelson, Marcela Böhm-Vélez, Wendie A Berg, GJ Whitman, MI Feldman, H Madjar, et al. 2013. Acr bi-rads® ultrasound. *ACR BI-RADS® atlas, breast imaging reporting and data system* 149 (2013).

[16] David Fraile Navarro, Kiran Ijaz, Dana Rezazadegan, Hania Rahimi-Ardabili, Mark Dras, Enrico Coiera, and Shlomo Berkovsky. 2023. Clinical named entity recognition and relation extraction using natural language processing of medical free text: A systematic review. *International Journal of Medical Informatics* 177 (2023), 105122.

[17] OpenAI. 2024. GPT-4-32K. https://openai.com. Accessed: 2024-07-04.

[18] Cheng Peng, Xi Yang, Zehao Yu, Jiang Bian, William R Hogan, and Yonghui Wu. 2023. Clinical concept and relation extraction using prompt-based machine reading comprehension. *Journal of the American Medical Informatics Association* 30, 9 (2023), 1486–1493.

[19] Alex Sherstinsky. 2018. Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network. *CoRR* abs/1808.03314 (2018). arXiv:1808.03314 http://arxiv.org/abs/1808.03314

[20] Buzhou Tang, Hongxin Cao, Yonghui Wu, Min Jiang, and Hua Xu. 2013. Recognizing clinical entities in hospital discharge summaries using Structural Support Vector Machines with word representation features. In *BMC medical informatics and decision making*, Vol. 13. Springer, 1–10.

[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[22] Xi Yang, Jiang Bian, William R Hogan, and Yonghui Wu. 2020. Clinical concept extraction using transformers. *Journal of the American Medical Informatics Association* 27, 12 (2020), 1935–1942.