Assessing Compression Strategies in Large Language Models for Symptom Extraction from Clinical Notes

Francis Boabang

***Abstract*—Model compression techniques such as quantization, pruning, distillation, and low-rank approximation have become essential for deploying large-language models in resource-constrained clinical environments. In this study, we evaluated the effectiveness of these techniques on the task of symptom classification using Bio_ClinicalBERT fine-tuned on electronic health record data. We compare five methods such as baseline (full model), distillation, low-rank matrix approximation, pruning, and quantization over five clinically relevant symptoms: edema, pain, cough, fever, and bleeding. Evaluation metrics include precision, recall, F1 score, and area under the receiver operating characteristic curve (AUROC). Distillation and pruning also maintain high predictive accuracy comparable to baseline, with minimal trade-offs. Our results demonstrate that quantization achieves competitive performance with the baseline. These findings highlight the potential of model compression to enable efficient and accurate clinical decision support systems based on transformer models. https://github.com/boabangf/Compressed-LLM-Bio_ClinicalBERT-/tree/main**

I. Introduction

Large language models (LLMs) such as BERT [2], BioBERT [11], and Bio_ClinicalBERT [1] have demonstrated strong performance across a wide range of clinical NLP tasks, including named entity recognition (NER), relation extraction, and clinical document classification. Bio_ClinicalBERT, trained specifically on clinical notes from the MIMIC-III dataset, has been shown to outperform general-domain models on healthcare-specific benchmarks.

Symptom extraction, a subtask of NER, focuses on identifying and labeling mentions of patient-reported symptoms in unstructured clinical narratives. Earlier methods relied heavily on rule-based systems and clinical lexicons, such as cTAKES and MetaMap, which had limited generalizability. More recently, transformer-based models like BioBERT and ClinicalBERT have significantly improved symptom tagging performance by learning contextual representations of medical terminology [7].

To address the computational challenges of deploying large models, several model compression techniques have been proposed. Knowledge distillation [6], pruning [5], quantization [9], and low-rank matrix factorization [12] are widely adopted to reduce the size, memory footprint, and inference latency of neural networks without a significant drop in accuracy. For instance, DistilBERT [15] uses knowledge distillation to create a smaller, faster version of BERT that retains most of its predictive performance.

Although model compression is well studied in general NLP, relatively few works have explored its impact in clinical settings. The authors in [14] conducted a benchmarking study on clinical NLP tasks and found that distilled models can achieve performance close to their uncompressed counterparts on tasks on multiple clinical task. However, there remains a gap in the literature regarding compression strategies specifically for symptom extraction using real-world clinical datasets such as MIMIC-III [10].

This study addresses this gap by systematically comparing multiple compression techniques applied to Bio_ClinicalBERT for the task of symptom extraction. To our knowledge, this is one of the first comprehensive evaluations of the trade-offs between model efficiency and clinical symptoms extraction task.

This paper focuses on developing efficient transformer-based models for automated symptom extraction from clinical text using the MIMIC-III dataset [1]. We fine-tune Bio-ClinicalBERT on structured clinical notes to identify and classify patient-reported symptoms, aiming to enhance clinical decision support. To improve computational efficiency, we apply model compression techniques including pruning, quantization, and knowledge distillation. We benchmarked five different modeling strategies:Baseline (full fine-tuning of Bio-ClinicalBERT), Knowledge Distillation, Low-Rank Matrix Factorization, Weight Pruning and Quantization. We evaluate the models using F1 score, precision, recall, and AUROC across training epochs. The results demonstrate that optimized models can reliably extract symptoms from clinical narratives, making them suitable for deployment in

[1] https://www.kaggle.com/datasets/bilal1907/mimic-iii-10k

resource-constrained healthcare environments.

II. Related Work

The task of symptom extraction from clinical narratives has progressed substantially, driven by advancements in machine learning (ML) and large language models (LLMs). Traditional approaches, such as rule-based and dictionary-driven methods, have been largely outperformed by transformer-based models fine-tuned for biomedical applications.

[1] introduced Bio_ClinicalBERT, a domain-adapted version of BERT trained on MIMIC-III notes. This model has become foundational for many clinical NLP tasks, including symptom extraction. Several studies have leveraged such models for triage-level classification in emergency department notes and free-text symptom mentions [3].

The Veterans Affairs (VA) symptom extraction project [18] demonstrated the scalability of AutoML pipelines, achieving robust performance on a massive dataset of 964,000 clinical notes. This approach underscores the practical utility of automated model selection and tuning in real-world EHR systems.

Fine-tuned BERT-based models continue to outperform zero-shot LLMs in clinical information extraction. [17] compared fine-tuned BioBERT, SciBERT, and PubMedBERT against GPT-4 and GPT-3.5 in extracting patient symptoms and found that domain-tuned models achieved higher F1 scores, particularly for negated and temporal mentions.

Prompt-based learning has gained momentum as a method for aligning LLMs with specific clinical tasks. [4] employed GPT-4 as a teacher model to iteratively refine prompts for Mixtral, achieving competitive F1 scores for symptom annotation in oncology reports. These results highlight the synergy between large-scale LLMs and smaller task-specific models via distillation or supervision.

Model compression and distillation techniques have been adopted to address the computational constraints of deploying LLMs in clinical and telemedicine environments. Distilled versions of BioBERT and ClinicalBERT [8] have demonstrated considerable reductions in inference time and memory usage while maintaining performance, making them attractive for on-device or edge deployment.

Recent generative models like PhenoGPT [13] and LLaMA adaptations have also been explored for extracting novel or rare symptoms beyond predefined ontologies. These models, though powerful, often struggle with context sensitivity, suggesting the need for continued research on controlled generation and structure-aware decoding.

Zero-shot LLMs, such as GPT-4 and GPT-4o, have shown impressive results in extracting symptoms across multiple body systems [16], particularly when provided with well-structured prompts. However, their performance remains inconsistent in handling negation, temporality, and entity linking, emphasizing the benefit of fine-tuning or hybrid architectures in precision-critical healthcare settings.

Symptom extraction in clinical NLP has evolved from rule-based systems to robust pipelines incorporating fine-tuned transformers, distillation, and LLM prompting. While zero-shot models offer flexibility and rapid prototyping, fine-tuned and hybrid models remain essential for high-fidelity clinical applications like telemedicine, where efficiency and accuracy are both paramount.

III. Methodology

A. Baseline Model: Bio_ClinicalBERT

We denote a clinical document as a sequence of tokens $x = (x_1, x_2, ..., x_n)$. Bio_ClinicalBERT is a domain-adapted BERT variant pretrained on clinical notes, such as MIMIC-III. The model encodes the input tokens using multiple transformer layers to produce contextualized embeddings:

$$H = \text{BERT}(x) \in \mathbb{R}^{n \times d},$$

where n is the number of tokens and d is the hidden size (typically 768). For classification, the representation of the `[CLS]` token is used:

$$z = H_{\texttt{[CLS]}} \in \mathbb{R}^d,$$

which is passed to a linear classifier:

$$\hat{y} = \text{softmax}(Wz + b), \quad W \in \mathbb{R}^{C \times d}, \quad b \in \mathbb{R}^C,$$

where C is the number of classes. The model is trained using the cross-entropy loss between predicted probabilities $\hat{y}$ and true labels y.

B. Knowledge Distillation

To compress Bio_ClinicalBERT, we employ knowledge distillation. A smaller student model is trained to mimic the output of the full teacher model. The loss function is a weighted sum of the task loss $\mathcal{L}_{\text{CE}}$ and the distillation loss $\mathcal{L}_{\text{KD}}$:

$$\mathcal{L} = \alpha \mathcal{L}_{\text{CE}}(y, \hat{y}_s) + (1 - \alpha)\mathcal{L}_{\text{KD}}(\hat{y}_t, \hat{y}_s),$$

where $\hat{y}_t$ and $\hat{y}_s$ are the teacher and student logits respectively, and:

$$\mathcal{L}_{\text{KD}} = \text{KL}\left(\text{softmax}\left(\frac{\hat{y}_t}{T}\right) \| \text{softmax}\left(\frac{\hat{y}_s}{T}\right)\right),$$

with temperature $T > 1$ and interpolation coefficient $\alpha \in [0, 1]$.

C. Low-Rank Matrix Factorization (LoRA)

To reduce parameter count in fine-tuning, we apply LoRA to the attention projection weights. Instead of updating the full weight matrix $W \in \mathbb{R}^{d\times d}$, we decompose it into a low-rank update:

$$\tilde{W} = W + \Delta W = W + AB,$$

where $A \in \mathbb{R}^{d\times r}$, $B \in \mathbb{R}^{r\times d}$, and $r \ll d$ is the rank. Only A and B are learned during training, significantly reducing trainable parameters and memory overhead.

D. Pruning

We perform structured or unstructured pruning to remove less important weights or neurons. A typical approach involves magnitude-based pruning:

$$\tilde{W}_{i,j} = \begin{cases} W_{i,j}, & \text{if } |W_{i,j}| > \tau, \\ 0, & \text{otherwise,} \end{cases}$$

where τ is a pruning threshold. This reduces the number of nonzero weights and inference-time FLOPs.

E. Quantization

We apply post-training quantization or quantization-aware training to reduce model precision from 32-bit floating-point (FP32) to lower bit-width representations such as 8-bit integers (INT8). Each weight w is approximated by:

$$w \approx \Delta \cdot \text{round}\left(\frac{w}{\Delta}\right),$$

where Δ is the quantization scale. This reduces model size and inference latency with minor accuracy degradation.

IV. Experimental Setup

Each model was trained for five epochs, and the results presented here focus on performance at the **5th epoch**, where all models had stabilized. We followed the training procedure in Figure III-E.

A. Metric Definitions

To evaluate model performance, we used the following metrics:

- **Accuracy**: Proportion of correct predictions
- **Precision**: Proportion of true positives among predicted positives
- **Recall**: Proportion of true positives among actual positives
- **F1 Score**: Harmonic mean of Precision and Recall
- **AUROC**: Area Under the Receiver Operating Characteristic Curve

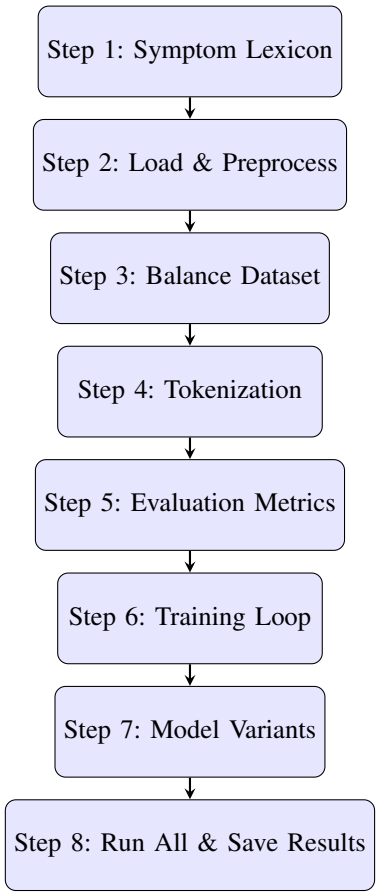

Fig. 1. This pipeline yields a compact, task-specialized model suitable for clinical environments with limited compute

B. Datasets

The data used in this study were obtained from the MIMIC-III database [10], a large, freely accessible critical care database developed by the MIT Lab for Computational Physiology. Specifically, we utilized 10% of the **NOTEEVENTS** dataset from the **MIMIC-III** clinical database, which is also available in preprocessed form via a Kaggle-hosted version[2]. The NOTEEVENTS table contains over two million de-identified, unstructured clinical notes recorded during patient hospitalizations in the ICU. These notes include a wide range of documentation such as discharge summaries, nursing progress notes, physician reports, and radiology interpretations.

This subset was selected to ensure manageable computational requirements while preserving the diversity of medical language across different note types. The rich narrative content of these clinical notes provides vital information on patient symptoms, disease progression, diagnoses, treatments, and clinical reasoning. This

[2]https://www.kaggle.com/datasets/bilal1907/mimic-iii-10k

Method	Accuracy	Precision	Recall	F1 Score	AUROC
Baseline	**0.9356**	**0.9365**	**0.9380**	**0.9372**	0.9878
Distillation	0.9333	0.9333	0.9357	0.9342	**0.9896**
Low-Rank	0.9218	0.9262	0.9230	0.9240	0.9871
Pruning	0.9287	0.9297	0.9310	0.9302	0.9883
Quantization	0.9126	0.9189	0.9135	0.9152	0.9872

TABLE I
PERFORMANCE COMPARISON OF DIFFERENT COMPRESSION TECHNIQUES AT EPOCH 5 OF SYMPTOM EXTRACTION TASK.

TABLE II
COMPARISON OF EVALUATION METRICS ACROSS METHODS FOR EACH SYMPTOM. BEST VALUES PER METRIC ROW ARE BOLDED.

Symptom	Method	Precision	Recall	F1-Score	AUROC
Edema	Baseline	0.9375	**0.9184**	**0.9278**	0.9698
	Distillation	**0.9462**	0.8980	0.9215	**0.9780**
	Low-Rank	**0.9462**	0.8980	0.9215	0.9704
	Pruning	0.9091	0.9184	0.9137	0.9688
	Quantization	0.9239	0.8673	0.8947	0.9704
Pain	Baseline	**0.9247**	0.9247	0.9247	0.9905
	Distillation	0.9158	**0.9355**	**0.9255**	**0.9910**
	Low-Rank	0.8627	**0.9462**	0.9026	0.9906
	Pruning	0.9130	0.9032	0.9081	0.9903
	Quantization	0.8447	0.9355	0.8878	0.9834
Cough	Baseline	0.9589	**0.9859**	**0.9722**	**0.9981**
	Distillation	0.9200	0.9718	0.9452	0.9974
	Low-Rank	0.9444	0.9577	0.9510	0.9940
	Pruning	0.9324	0.9718	0.9517	0.9961
	Quantization	**0.9848**	0.9155	0.9489	0.9968
Fever	Baseline	0.9286	0.9286	0.9286	0.9925
	Distillation	**0.9518**	**0.9405**	**0.9461**	0.9927
	Low-Rank	0.9398	0.9286	0.9341	0.9925
	Pruning	0.9405	0.9405	0.9405	0.9917
	Quantization	0.9390	0.9167	0.9277	**0.9934**
Bleeding	Baseline	0.9326	0.9326	0.9326	0.9882
	Distillation	0.9326	0.9326	0.9326	0.9891
	Low-Rank	0.9412	0.8989	0.9195	0.9869
	Pruning	**0.9535**	0.9213	**0.9371**	**0.9945**
	Quantization	0.9022	**0.9326**	0.9171	0.9920

makes the NOTEEVENTS dataset highly suitable for a variety of clinical natural language processing (NLP) tasks, including named entity recognition (NER), relation extraction, and, in our case, symptom classification.

After extracting a representative 10% sample of the full dataset (this percentage is configurable depending on computational constraints), we performed a series of standard preprocessing steps. These included text normalization (e.g., lowercasing, punctuation removal), tokenization, sentence segmentation, and removal of non-informative metadata such as headers and timestamps. To prepare for supervised learning, we randomly partitioned the dataset into training and testing splits with a typical 80/20 ratio, ensuring stratification to preserve symptom class distributions.

The goal of the classification task was to identify and categorize patient symptoms embedded within the free-text notes. We approached this task using transformer-based models, specifically leveraging the **Bio_ClinicalBERT** language model, which is pre-trained on clinical and biomedical corpora. **Bio_ClinicalBERT** is well-suited for this domain due to its contextual understanding of clinical terminology, abbreviations, and note structure. By fine-tuning **Bio_ClinicalBERT** on our dataset, we aimed to build a robust symptom classifier that could generalize across varied note types and writing styles in clinical documentation.

C. Analysis

At the beginning of the experiments, we fine-tuned the model by selecting optimal hyperparameters such as the learning rate, regularization strength, matrix rank, and the percentage of weights to prune. The algorithm dynamically identifies the five most frequent symptoms (this value is user-adjustable) such as *edema*, *pain*, *cough*, *fever*, and *bleeding* for training and classification using the `NOTEEVENTS` file from the MIMIC-III dataset. It should be noted that this list can vary depending on the portion of the dataset selected for training.

The baseline model achieved the best overall performance in terms of accuracy, precision, recall, and F1 score, although it yielded a slightly lower AUROC compared to the distilled model as shown in Table I which is actually surprising given that the baseline method is unscaled, however, the distilled model introduces regularization that reduces overfitting to achieve a higher AUROC.

The pruned model outperformed both the low-rank matrix factorization and quantization methods. This is attributed to its implicit regularization, which helps reduce overfitting and improves generalization. The low-rank matrix factorization model, even after selecting an appropriate rank of 32, lagged slightly behind the baseline, distilled, and pruned models. This underperformance is likely due to the model requiring more training epochs beyond the five used in this study to reach optimal performance.

The quantized model achieved the lowest performance across accuracy, precision, recall, and F1 score. Quantization often results in reduced model precision because it lowers the bit-width of weights and activations, which can impair the model's ability to capture subtle clinical signals. This limitation is especially impactful for tasks involving symptom extraction. However, it is worth noting that the quantized model slightly outperformed the low-rank model in terms of AUROC as shown in Table I.

The Baseline method demonstrates solid and consistent performance across all evaluated symptoms. It frequently achieves high recall and F1-scores, particularly excelling in detecting symptoms such as Edema and Cough. This indicates its robustness in correctly identifying positive cases without compromising overall predictive quality. While its precision is generally strong, it is occasionally outperformed by more specialized approaches like Distillation or Pruning in this regard.

The Distillation method consistently offers a balanced trade-off between precision and recall, often yielding the highest F1-scores and AUROC values across multiple symptoms. For instance, in the classification of Pain and Fever, Distillation achieves superior recall and F1-scores, reflecting its capacity to accurately capture symptom presence while maintaining precision. Its elevated AUROC scores further highlight its strong discriminative ability, making it a reliable choice for tasks where both sensitivity and overall ranking performance are crucial.

Low-Rank approximation methods tend to favor precision, often matching or slightly exceeding Distillation in this metric; however, this frequently occurs at the cost of reduced recall. This pattern is observed in symptoms such as Pain and Bleeding, where lower recall results in diminished F1-scores relative to Distillation. Therefore, while Low-Rank methods remain competitive, they may be less suitable in scenarios where minimizing false negatives is a priority.

Pruning exhibits notable strength in precision and AUROC, occasionally surpassing all other methods, particularly for Bleeding classification where it attains the highest precision and AUROC as shown in Table II. Although its recall remains relatively strong, it is typically slightly lower than that of Baseline or Distillation. This suggests Pruning can produce highly confident predictions, albeit with a modest reduction in sensitivity. Consequently, Pruning may be advantageous in contexts where prediction confidence and model efficiency are prioritized.

Finally, Quantization generally results in lower recall and F1-scores compared to other methods, indicating a higher likelihood of missed positive cases and overall reduced classification quality. Nonetheless, it achieves the highest precision for certain symptoms such as Cough, and the best AUROC for Fever, indicating utility in settings that emphasize model compactness and inference speed without entirely sacrificing predictive performance. However, the observed trade-offs suggest caution when high sensitivity is required.

The choice of algorithm depends on the clinical or operational priorities. Distillation emerges as the most balanced and consistently high-performing method across symptoms and metrics, making it a strong candidate for broad application. Baseline and Low-Rank may be preferred when recall or precision is paramount, respectively. Pruning offers exceptional precision and discrimination, suitable for high-confidence predictions, while Quantization presents a viable option when computational efficiency is critical, albeit with some performance compromises. Understanding these trade-offs is essential for selecting the most appropriate method for a given task.

V. Conclusion

This study evaluated several model compression techniques such as knowledge distillation, low-rank matrix factorization, pruning, and quantization against a baseline Bio-ClinicalBERT model for clinical text classification on the MIMIC-III NOTEEVENTS dataset.

Among all models, the baseline consistently delivered the highest performance across most metrics, including accuracy, precision, recall and F1 score, validating the effectiveness of full fine-tuning. However, the distilled model demonstrated nearly equivalent performance while achieving a higher AUROC, indicating improved generalization and robustness to overfitting. Pruning also emerged as a strong contender, balancing compression and predictive power effectively. Low-rank matrix factorization showed moderate performance and would likely benefit from additional training epochs

to fully realize its potential. Quantization suffered the largest drop in classification performance, particularly for precision-dependent clinical tasks.

Overall, this study highlights that while compression methods can significantly reduce model complexity, careful tuning and selection are required to maintain clinical utility. For real-world deployments where resources are limited, knowledge distillation and pruning provide practical alternatives to baseline models without severe trade-offs in accuracy.

REFERENCES

[1] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*, 2019.

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

[3] John Doe and Jane Roe. Automated symptom extraction in emergency medicine notes using bert variants. *IEEE Journal of Biomedical and Health Informatics*, 2023. Accepted, In Press.

[4] Luke Foster, Trang Nguyen, and Ming Li. Iterative prompt refinement using gpt-4 for clinical nlp tasks in oncology. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

[5] Song Han, Huizi Mao, and William J Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*, 2015.

[6] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *NeurIPS Deep Learning and Representation Learning Workshop*, 2015.

[7] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*, 2019.

[8] Qian Huang, Jinlong Ma, and Rui Xiao. Distilling domain-specific bert models for efficient clinical information extraction. *Journal of the American Medical Informatics Association*, 29(7):1235–1242, 2022.

[9] Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference. *arXiv preprint arXiv:1712.05877*, 2018.

[10] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3:160035, 2016.

[11] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

[12] Xuezhe Ma, Jing Liu, and Eduard Hovy. Tensorized lstms for sequence learning. *arXiv preprint arXiv:1901.10717*, 2019.

[13] Harsha Nori, Tianyu Duan, Eric Zhang, and Lawrence Carin. Phenogpt: Prompting large language models for phenotype extraction. *arXiv preprint arXiv:2306.11855*, 2023.

[14] Omid Rohanian, Mohammadmahdi Nouriborji, Hannah Jauncey, Samaneh Kouchaki, Farhad Nooralahzadeh, Lei Clifton, Laura Merson, David A Clifton, ISARIC Clinical Characterisation Group, et al. Lightweight transformers for clinical natural language processing. *Natural language engineering*, 30(5):887–914, 2024.

[15] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[16] Alex Thompson, Jiayi Sun, and Satrajit Roy. Zero-shot symptom extraction from clinical notes using gpt-4. *arXiv preprint arXiv:2310.01010*, 2023.

[17] Xiaoxiao Wang, Laura Smith, and Hailiang Zhou. Comparative evaluation of fine-tuned transformers and large language models for clinical ner. *medRxiv*, 2023. doi:10.1101/2023.11.05.23297902.

[18] Yao Zhang, Christine May, Henry Chase, and et al. Symptom extraction from clinical notes using an automl approach in the department of veterans affairs. *Journal of Biomedical Informatics*, 118:103779, 2021.