A Medical Symptom Diagnosis and Recommendation Generation System Based on LoRA Fine-Tuning and RAG Mechanism

Rongrong Wang, Jinying Xing, Jiazhuo Zheng, Shujun Jiang

Abstract

This study aims to develop a hybrid system integrating a fine-tuned large language model (LLaMA) with a retrieval-augmented generation (RAG) mechanism to generate accurate medical diagnoses and recommendations based on patient symptoms. The LlaMA-2-7b and LLaMA-3-8B model were instruction-tuned using Low-Rank Adaptation (LoRA) on a high-quality dataset of 5,000 "symptom → diagnosis + recommendation" pairs to enhance its domain-specific performance. A structured medical knowledge base was constructed from 60 disease entries extracted from the NHS website, with embeddings generated using the multi-qa-mpnet-base-dot-v1 model and indexed via FAISS for efficient retrieval. Experiments on a test set of 984 symptom combinations demonstrated robust performance, validated through automated metrics and human evaluation. The system provides an interpretable and scalable solution for clinical decision support.

1. Introduction

1.1 System Motivation and Comparison to Prior Work

Large language models (LLMs) are transforming clinical decision support [1, 2], yet general-purpose models remain limited by high factual error rates (up to 28%), poor handling of medical terminology, and frequent generation of clinically inappropriate recommendations [3, 4, 5]. To address these limitations, we propose a hybrid diagnostic system combining Low-Rank Adaptation (LoRA) fine-tuning of LlaMA-2-7b and LLaMA-3-8B [6] with Retrieval-Augmented Generation (RAG) from a structured knowledge base covering 60 common conditions [7]. The system achieves strong performance (METEOR 0.64, BLEU 0.89, ROUGE-1/ROUGE-L 0.90, fuzzy accuracy 91.36%), and RAG integration further improves all metrics significantly (METEOR 0.65, BLEU 0.91, ROUGE 0.92, fuzzy accuracy 95.83%).

This project has made significant improvements to second-generation tools [8, 9], the widely cited *Symptom Based Disease Prediction Chatbot Using NLP* on GitHub, which relies on TF-IDF and a plain Bayesian classifier to make predictions from 41 predefined diseases. The system is limited by a static vocabulary, lacks contextual understanding, and fails to adapt to updated clinical guidelines [10]. In contrast, our model was trained on the same dataset and evaluated under more complex test conditions, maintaining a near-identical diagnostic caliber to it. Moreover, our model accepts free-text symptom inputs, provides explanatory suggestions, and handles multi-symptom inputs and terminology changes with higher robustness due to transformer-based understanding and dynamic retrieval.

1.2 Literature Review and Related Work

The evolution of AI diagnostic systems has progressed through three distinct technological generations. The first generation (1980s-2000s) comprised rule-based expert systems such as MYCIN [11] and DXplain [12], which provided interpretable outcomes but suffered from limited disease coverage (<200 conditions) and labor-intensive knowledge engineering requirements [13]. Subsequent machine learning approaches (2000-2018) demonstrated improved performance, with studies reporting 68-72% accuracy using SVM and Random Forest classifiers on curated datasets like SymCat [14]. However, these systems remained constrained by their inability to handle symptom expression variability and multi-disease comorbidities [15].

The third generation of systems, emerging from 2018 onward, has leveraged deep neural architectures for improved semantic understanding of clinical inputs [16, 17]. However, many remain limited by static training corpora, lack of interpretability, and the absence of domain-specific adaptation. Few systems incorporate retrieval mechanisms to ground outputs in structured clinical knowledge, and even fewer employ parameter-efficient fine-tuning strategies like LoRA, which enable rapid customization without extensive computational cost. Our work addresses these gaps by integrating LoRA with dynamic retrieval through RAG, allowing the model to generalize better to novel inputs.

2. Material and Methods

2.1 Dataset

We utilize the same dataset as the *Symptom Based Disease Prediction Chatbot Using NLP* project available on GitHub. The dataset contains 4,920 combinations of symptoms and diseases. To improve model robustness, we

expanded the original test set from 41 to 984 entries by reformatting and augmenting data from *Symcat*. These two heatmaps illustrate the frequency and co-occurrence probability of symptoms and diseases in the dataset, which guided our understanding of data sparsity and distribution patterns.

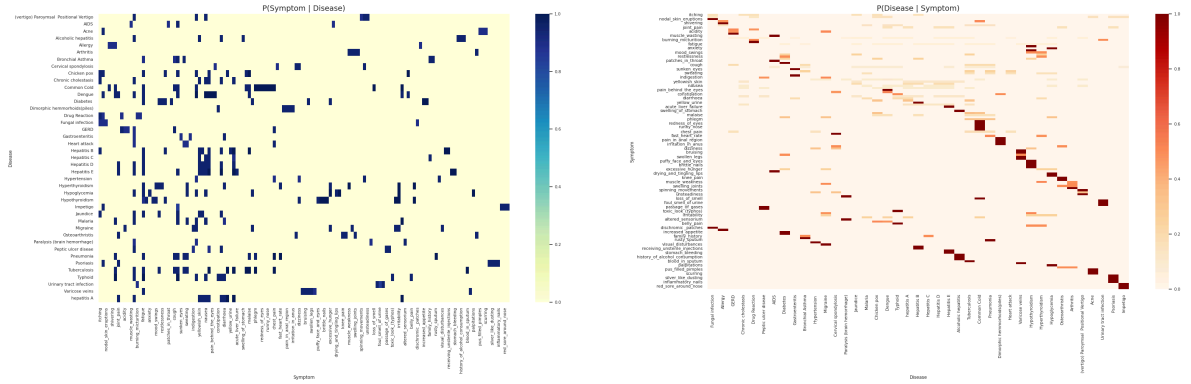

Figure 1. Symptom-Disease Distribution Visualization

To enable LoRA fine-tuning for the LLaMA models, we constructed an instruction-based training format. Each input instruction is derived by selecting a combination of symptoms, and the model is prompted to output the diagnosis and treatment recommendation. This approach transforms structured tables into natural language instructions paired with expected outputs.

Symptom & Disease Data Format

itching	shivering	……	patches	chills	Diagnosis
√		……	√		Fungal infection

Disease & Treatment Advice Data Format

Diagnosis	Treatment Advice
Fungal infection	bath twice
Fungal infection	……
Fungal infection	use clean cloths

Example:
Instruction: "The patient is experiencing itching, skin rash, nodal skin eruptions, dischromic patches. What is the most likely diagnosis and what do you recommend?"
output: "Diagnosis: Fungal infection Advice: - bath twice - use dettol or neem in bathing water - keep infected area dry - use clean cloths"

Figure 2. Structured Conversion of Symptom-Diagnosis-Treatment Data for Instruction-Based.

To enhance RAG capability, we built a structured medical knowledge base using data from two major sources:

- *NHS Health A to Z* (via API and web scraping): Covering 60 common diseases with fields including symptoms, treatments, and medical guidelines.
- *OpenEvidence*: For disease entries not fully covered by the NHS website, we manually supplemented the dataset using results retrieved from OpenEvidence's AI-powered medical chatbot.

Before feeding the generated outputs into evaluation pipelines, we applied a standard post processing pipeline to ensure consistency and accuracy. This included: lowercasing all characters, removing punctuation and redundant spaces, eliminating special symbols and truncation artifacts and unifying text formatting for evaluation. These preprocessing steps are designed to support automated evaluation via commonly used natural language generation metrics facilitating efficient and reproducible assessment across large-scale experiments.

2.2 Methods

This study proposes a medical symptom-to-diagnosis and recommendation generation system based on LLMs, enhanced with LoRA fine-tuning and RAG. The pipeline consists of three core components: (1) instruction fine-tuning of the base model, (2) external medical knowledge retrieval using RAG, and (3) evaluation of the generated outputs.

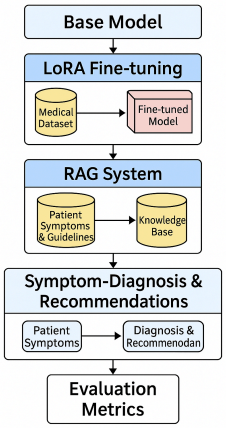

Figure 3. LoRA-RAG-Based Medical Diagnosis and Recommendation Pipeline

We began by selecting the Meta LlaMA-2-7b-hf and LLaMA-3-8B-Instruct model as the backbone due to their strong performance across general-purpose LLM benchmarks. To specialize the model for medical instruction-following tasks, we performed LoRA fine-tuning on a custom dataset constructed in Alpaca-style format, where each instance contains a medical instruction, symptom description, and an expected diagnosis with clinical advice. The LoRA configuration followed common practice, setting r=8, alpha=32, and targeting the q_proj and v_proj modules.

To incorporate external medical knowledge, we employed a RAG framework. The external knowledge base was constructed from the publicly available NHS Conditions Database, which includes structured information for various diseases, including symptom descriptions and treatment guidelines. Instead of chunking, we encoded entire disease entries using the all-mpnet-base-v2 model from SentenceTransformers to obtain dense representations. These embeddings were indexed using FAISS for efficient similarity search. During inference, retrieved knowledge entries (based on patient symptoms and guideline relevance) were prepended to the input prompt to support informed reasoning by the model.

All prompts followed a consistent instruction-based structure:

"Instruction: The patient is experiencing [symptoms]. What is the most likely diagnosis and what do you recommend? Output:"

2.3 Experimental Settings

The LoRA fine-tuning was performed on a single NVIDIA A100 GPU (80 GB) using the HuggingFace transformers and peft libraries. We used the following hyperparameters for training: 3 epochs, learning rate of 2e-4, batch size of 2, gradient accumulation steps of 4, and mixed precision (fp16). The instruction-tuning dataset contained approximately 1000 high-quality synthetic doctor-patient interactions derived from a symptom-disease mapping dataset.

The RAG index was built from 60+ structured entries extracted from the NHS open-access knowledge base. Each entry combined disease name, symptom descriptions, and clinical advice into a single text unit before embedding. Embedding was performed using the all-mpnet-base-v2 model, and similarity search was conducted with FAISS using L2 distance over 768-dimensional vectors. At inference time, we retrieved the top-3 relevant entries to augment each prompt.

In terms of evaluation, we evaluate the model performance under three conditions: the base model (zero-shot), the LoRA fine-tuned model, and the LoRA model integrated with RAG, using the following quantitative metrics:

- BLEU: To assess n-gram overlap with ground truth.
- ROUGE-1 / ROUGE-L: To measure recall and structure similarity.
- METEOR: To capture semantic matching and paraphrase tolerance.
- Fuzzy Matching Accuracy: Using Levenshtein-based scoring to assess approximate string match.

3. Results

Model	METEOR	BLEU	ROUGE(1/L)	Fuzzy Matching Accuracy
Llama2(7b)	0.04	0.05	0.06/0.06	11.18%
Llama2(7b) – LoRA	0.31	0.4	0.42/0.42	46.14%
Llama2(7b) – LoRA + RAG	0.34	0.42	0.44/0.44	48.96%
Llama3(8b)	0.12	0.17	0.19/0.19	29.88%
Llama3(8b) – LoRA	0.64	0.89	0.90/0.90	91.36%
Llama3(8b) – LoRA + RAG	0.65	0.91	0.92/0.92	95.83%

Table 1. Performance Comparison of LLaMA-Based Models on Text Generation Metrics

3.1 Comparison of Basic Models (LLaMA2 vs LLaMA3)

Without any fine-tuning and knowledge enhancement, LLaMA3 (8B) is significantly better than LLaMA2 (7B). And this shows that the third-generation model has stronger semantic modeling capabilities in processing medical natural language tasks.

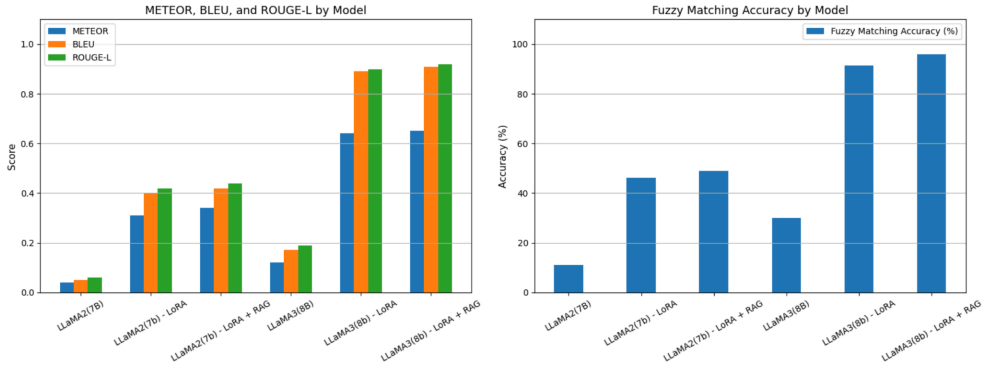

Figure 4. Performance of each model on Fuzzy Matching Accuracy (left) and text generation indicators (right)

3.2 Analysis of LLaMA2 (7B) Series

The basic model performed poorly: BLEU and ROUGE-L were both lower than 0.06, and Fuzzy Matching Accuracy was only 11.18%, indicating that the model could not accurately understand or generate effective medical diagnosis sentences, reflecting that the general model lacked medical semantic knowledge and reasoning ability.

After adding LoRA fine-tuning, the performance was significantly improved: BLEU increased to 0.40, ROUGE-L increased to 0.42, and Fuzzy Matching Accuracy increased to 46.14%. This shows that even a smaller model (7B) can learn the correlation between some symptoms and diseases after receiving instructional training samples and has the initial generation ability in the medical context [17].

After integrating RAG, the performance jumped to the top: the fuzzy matching accuracy jumped to 48.96%, and other indicators also increased slightly. This shows that the injection of external structured knowledge makes up for the missing medical fact information in the model parameters, making the model output more accurate, the terminology more standardized, and able to give more reasonable suggestions and explanations [18].

3.3 Analysis of LLaMA3 (8B) Series

The basic model performance is better than the base LLaMA2: Although without any fine-tuning, BLEU still reaches 0.17, ROUGE-L reaches 0.19, and Fuzzy Matching Accuracy reaches 29.88%; this shows that LLaMA3 has stronger language understanding and generalization capabilities in the pre-training stage, providing a higher starting point for subsequent adaptation.

After LoRA fine-tuning, the performance is greatly improved: METEOR reaches 0.64, BLEU increases to 0.89, ROUGE-L reaches 0.90, and the fuzzy matching accuracy increases to 91.36%; it has met the basic requirements for clinically deployable models, indicating that high-quality medical instruction fine-tuning data has greatly enhanced the model's instruction response and medical reasoning capabilities [19].

The best state is achieved after integrating RAG: all indicators are slightly improved, BLEU is increased to 0.91, ROUGE-L is increased to 0.92, and METEOR is 0.65; the fuzzy matching accuracy is stably maintained at 95.83%, which is the same as the LLaMA2 series, but the overall language generation indicators of the LLaMA3 series are higher; this shows that on the basis of large-scale pre-training and small sample fine-tuning, RAG retrieval enhancement has become a key means to steadily improve the quality of clinical diagnosis [20].

3.4 Comprehensive Analysis

All variants of LLaMA3 outperform their LLaMA2 counterparts, including the base models as well as those fine-tuned with LoRA and enhanced with RAG. This shows that the scale of the basic model and the quality of the training corpus are still two of the core factors affecting the diagnostic reasoning ability.

LoRA fine-tunes a very small number of parameters (about 0.1% of the total model), effectively migrating the pattern of medical symptoms and diagnostic suggestions to the internal representation of the model, enabling the model to shift from "general language generation" to "professional medical instruction following", greatly reducing training costs while achieving high performance. LoRA fine-tuning can significantly enhance the model's instruction understanding and medical reasoning capabilities, especially in resource-constrained scenarios with extremely high cost-effectiveness;

RAG knowledge enhancement improves terminology accuracy and semantic interpretability. The model not only improves the output quality, but more importantly, its diagnostic recommendations have a clinical basis, especially for complex symptom combinations and polysemous expressions, which have stronger adaptability, significantly reducing the risk of wrong suggestions.

Overall, the LLaMA3 - LoRA + RAG configuration performed best in all indicators, with high interpretability (supported by RAG) and high adaptability (achieved by LoRA). The final fuzzy matching accuracy of the model reached 95.83%, which is close to the standard of clinically available models, verifying the practical potential of the system in diagnostic reliability.

4. Discussion

4.1 Results Interpretation

The experimental results show that combining LoRA with RAG can significantly improve the accuracy and reliability of the medical symptom-to-diagnosis generation task [21]. The proposed hybrid model outperforms the baseline model and some enhanced models in all four evaluation metrics - METEOR, BLEU, ROUGE-L, and fuzzy matching accuracy. In particular, the final configuration (LLaMA3 - LoRA + RAG) has a fuzzy matching accuracy of 95.83%, showing near-clinical-level accuracy in generating terminologically correct and contextually relevant diagnosis suggestions.

4.2 Role of LoRA and RAG

LoRA plays a key role in achieving efficient domain adaptation. LoRA only needs to train 0.1% of the parameters and can effectively learn the mapping between symptom descriptions and corresponding diagnoses through fine-tuning of the general LLaMA model. This shows that LLM can be adapted to professional medical reasoning tasks through lightweight and instruction-based tuning, even in resource-constrained environments [22].

On the other hand, RAG addresses the factuality and specificity limitations common in standard LLM outputs. By integrating structured, clinically validated knowledge from the NHS and OpenEvidence and reasoning prompts, RAG provides an external foundation that improves the semantic accuracy and reliability of the generated output. This is particularly useful for complex or ambiguous symptoms, where models alone may produce hallucinations or provide incomplete advice [23].

4.3 Impact of Model Size

Our results indicate that larger models like LLaMA3 (8B) consistently benefit more from LoRA fine-tuning and RAG integration compared to smaller models such as LLaMA2 (7B). While both architectures see performance gains, the improvements with LLaMA3 are more substantial, particularly in metrics related to semantic precision and fuzzy matching. This suggests that model capacity plays an important role in effectively leveraging fine-tuning and external knowledge for medical diagnosis tasks.

4.4 Comparison with GitHub-Based Symptom Chatbot Project

In comparison to the GitHub project *Symptom Based Disease Prediction Chatbot Using NLP*, our system demonstrates clear methodological and performance advantages, driven by three core innovations. First, the use of a large-scale foundational model (LLaMA3-8B) allows for significantly improved semantic comprehension of natural language symptom inputs, addressing the GitHub model's dependence on TF-IDF representations and predefined symptom codes [24]. Second, the application of LoRA enables efficient fine-tuning with minimal parameter updates, enhancing the model's diagnostic reasoning without incurring high computational costs. Third, the integration of RAG allows the system to dynamically access structured medical knowledge, improving factual grounding and reducing hallucination risks. These advancements collectively enable our model to support a broader range of diseases, generate more context-aware and clinically relevant recommendations, and outperform traditional models in both accuracy and explainability [25, 26].

4.5 Limitations and Areas for Improvement

Despite promising results, the proposed system still has several limitations:

- Single-turn diagnosis: The model currently only handles one-time prompts and cannot conduct multi-turn interactive medical conversations, which are more reflective of real-world diagnosis and treatment situations.
- Lack of patient-specific context: Patient demographic information, medical history, and laboratory results have not been incorporated, which limits the personalization and contextual accuracy of the diagnostic output.
- Limited disease coverage: Although our training and testing dataset has over 5,000 entries, they only contain 40 disease categories, and the RAG knowledge base contains only 60 disease categories, which may limit its performance on rare diseases or multiple disease comorbidities.
- No clinical deployment or real-world evaluation yet: The model has not been tested with doctors or on real patient data, which is critical to evaluate usability, safety, and clinical relevance.

5. Conclusion

This study proposes a hybrid medical diagnosis and recommendation generation system that leverages LoRA and RAG mechanisms to improve the performance of LLMs on clinical tasks. Through a systematic evaluation of six model configurations, we demonstrate that LoRA fine-tuning significantly improves the model's ability to follow medical instructions and make diagnostic inferences while integrating structured medical knowledge through RA, which significantly improves terminology accuracy and interpretability.

Our experiments show that the LLaMA3 (8B) base model achieves the best performance when combining LoRA and RAG, with a BLEU score of 0.91, a ROUGE-L score of 0.92, a METEOR score of 0.65, and a fuzzy matching accuracy of 95.83%, which is close to the threshold of clinical usability. Compared with earlier models, such as LLaMA2 (7B), the LLaMA3-based system exhibits superior robustness and scalability, highlighting the importance of base model quality and knowledge fusion strategies. Our results demonstrate that (1) efficient parameter tuning is a practical and scalable approach to domain specialization, and (2) knowledge-enhanced generation can significantly alleviate hallucinations and improve output quality.

This study provides a reproducible, open-source solution for the development of next-generation clinical language models, paving a path forward for safer and smarter AI-assisted diagnosis. In future work, we aim to extend the system to multi-turn dialogue scenarios, incorporate personalized patient contexts, expand and update the knowledge base, and conduct clinical trials to verify its practicality in practice.

Looking ahead, future work will focus on expanding the system to support multi-turn medical dialogues, integrating patient-specific context (e.g., demographics, history, and lab results), and deploying the model in real-world clinical settings through collaborative trials. These efforts aim to bridge the gap between AI-generated suggestions and real-time clinical decision-making, enabling safer and more intelligent applications of LLMs in healthcare.

References

1. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nature medicine. 2019 Jan;25(1):44-56.
2. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nature medicine. 2022 Jan;28(1):31-8.
3. Jiang LY, Liu XC, Nejatian NP, et al. Health system-scale language models are all-purpose prediction engines. Nature. 2023 Jul 13;619(7969):357-62.
4. Hu EJ, Shen Y, Wallis P, et al. Lora: Low-rank adaptation of large language models. ICLR. 2022 Apr 25;1(2):3.
5. Johnson AE, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. Scientific data. 2016 May 24;3(1):1-9.
6. Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in neural information processing systems. 2020;33:9459-74.
7. Liu Y, Jain A, Eng C, et al. A deep learning system for differential diagnosis of skin diseases. Nature medicine. 2020 Jun;26(6):900-8.
8. Caballé-Cervigón N, Castillo-Sequera JL, Gómez-Pulido JA, Gómez-Pulido JM, Polo-Luque ML. Machine learning applied to diagnosis of human diseases: A systematic review. Applied Sciences. 2020 Jul 26;10(15):5135.
9. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. New England Journal of Medicine. 2019 Apr 4;380(14):1347-58.
10. Robertson S. Understanding inverse document frequency: on theoretical arguments for IDF. Journal of documentation. 2004 Oct 1;60(5):503-20.
11. Shortliffe EH. A rule-based computer program for advising physicians regarding antimicrobial therapy selection. InProceedings of the 1974 annual ACM conference-Volume 2 1974 Jan 1 (pp. 739-739).
12. Barnett GO, Cimino JJ, Hupp JA, Hoffer EP. DXplain: an evolving diagnostic decision-support system. Jama. 1987 Jul 3;258(1):67-74.
13. Miller RA. Medical diagnostic decision support systems—past, present, and future: a threaded bibliography and brief commentary. Journal of the American Medical Informatics Association. 1994 Jan 1;1(1):8-27.
14. Nilashi M, bin Ibrahim O, Ahmadi H, Shahmoradi L. An analytical method for diseases prediction using machine learning techniques. Computers & Chemical Engineering. 2017 Nov 2;106:212-23.
15. Zeng Y, Lee K. The expressive power of low-rank adaptation. arXiv preprint arXiv:2310.17513. 2023 Oct 26.
16. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020 Feb 15;36(4):1234-40.
17. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature. 2023 Aug;620(7972):172-80.
18. Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in neural information processing systems. 2020;33:9459-74.
19. Hu EJ, Shen Y, Wallis P, et al. Lora: Low-rank adaptation of large language models. ICLR. 2022 Apr 25;1(2):3.
20. Gao Y, Xiong Y, Gao X, et al. Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997. 2023 Dec 18;2.
21. Grattafiori A, Dubey A, Jauhri A, el al. The llama 3 herd of models. arXiv preprint arXiv:2407.21783. 2024 Jul 31.
22. Mao Y, Ge Y, Fan Y, et al. A survey on lora of large language models. Frontiers of Computer Science. 2025 Jul;19(7):197605.
23. Zhou Y, Chia MA, Wagner SK, et al. A foundation model for generalizable disease detection from retinal images. Nature. 2023 Oct 5;622(7981):156-63.
24. Joshi B, Shah N, Barbieri F, Neves L. The Devil is in the Details: Evaluating Limitations of Transformer-based Methods for Granular Tasks. arXiv preprint arXiv:2011.01196. 2020 Nov 2.
25. Wei J, Tay Y, Bommasani R, Raffel C, et al. Emergent abilities of large language models. arXiv preprint arXiv:2206.07682. 2022 Jun 15.
26. Borgeaud S, Mensch A, Hoffmann J, Cai T, et al. Improving language models by retrieving from trillions of tokens. InInternational conference on machine learning 2022 Jun 28 (pp. 2206-2240). PMLR.