

From Symptoms to Syndromes: Development and Validation of Fine-Tuning Transformer Architectures for Genetic Neurological Disease Diagnoses

Stanford CS224N Custom Project

Abstract

This project investigates the efficacy of transforming state-of-the-art LLMs (Llama-3.1, Qwen3, and Meditron) into expert classifiers through a multi-stage training pipeline. We employ Domain-Adaptive Pre-training (DAPT) on the OMIM database followed by Parameter-Efficient Fine-Tuning (PEFT) using Low-Rank Adaptation (LoRA). To overcome data scarcity, we developed a three-tiered synthetic data generation methodology incorporating paraphrasing to simulate real-world clinical documentation. Our results show that fine-tuned models significantly outperform zero-shot baselines. Finally, we implement an ensemble architecture using ChatGPT-5.4 as a "Meta-Judge" on top of fine-tuned models to resolve discrepancies between local experts, providing a scalable framework to bridge the gap toward clinical-level diagnostic accuracy.

1. Key Information to include:

Members: Yi Li, lyi@stanford.edu, Ximing Gao, ximg95@stanford.edu, Anushka Rawat, anushkar@stanford.edu

External collaborators: Yuyi Zhang, YuyiZhang@stanfordhealthcare.org

Mentor: Advit Deepak

Sharing project: Explored this project with various ML training methods in CS229, but different approach

Late Days (if you are using any):

Team member Late days remaining

Team Member	Late Days Remaining
Yi Li	5
Ximing Gao	3
Anushka Rawat	4
Total Pooled	12

Team size: 3; Late days applied for the report: 9 (3 day from each teammate)

2 Introduction

The "diagnostic odyssey"—period between the first onset of symptoms and an accurate diagnosis—for rare genetic neurological conditions currently averages up to 12 years. This delay is the consequence of the overwhelming scale of the domain; it is difficult for clinicians to maintain expert knowledge of the approximately 2,000 known genetic neurological syndromes, many of which present with overlapping or subtle phenotypic profiles. While general-purpose LLMs have shown utility in broad medical tasks, they remain unreliable for this specialized, high-stake domain due to frequent hallucinations and a lack of depth regarding rare phenotypic signatures.

We plan to design a tool to assist clinicians in diagnosing rare genetic neurological disorders. Our research addresses this gap by asking: *Can modern, pre-trained LLMs be effectively adapted to function as highly accurate classifiers for high-cardinality medical domains?* We frame this as a large-scale, multi-class text classification task where the system receives clinical phenotypes as input and outputs a ranked list of the most probable diagnoses from a curated set of 1,716 genetic neurological diseases.

This work contributes a tiered training methodology. First, we infuse models with medical knowledge through Domain-Adaptive Pre-training (DAPT) on the raw Online Mendelian Inheritance in Man (OMIM) corpus. Second, we evaluate the efficiency of Low-Rank Adaptation (LoRA) in capturing the nuances of a 1,716-class label space, and address bottleneck of data scarcity through sophisticated synthetic case generation that utilizes paraphrasing to reflect the semantic variance of physician notes. Finally, we introduce a hierarchical ensemble architecture where local fine-tuned "experts" (Llama, Qwen, and Meditron) provide candidates to a high-reasoning "Meta-Judge" (ChatGPT-5.4). Through this approach, we aim to demonstrate that specialized fine-tuning can transform general-purpose transformers into viable tools for rare disease discovery.

3 Related Work

Title: GPAD: a natural language processing-based application to extract the gene-disease association discovery information from OMIM. BMC Bioinformatics. 2024 Feb 27;25(1):84.

Year: 2024; URL: <https://pubmed.ncbi.nlm.nih.gov/38413851/>: Recent literature has highlighted the potential of NLP systems to unlock structured insights from the vast, unstructured repositories of rare disease data. Related

previous work include GPAD (Gene-Disease Association Discovery) system, which addresses the challenge of extracting knowledge from the OMIM database. While OMIM is the definitive reference for Mendelian disorders, its heterogeneous, free-text narrative style makes systematic discovery tracking difficult. GPAD demonstrated that task-specific NLP pipelines can outperform general-purpose LLMs in extracting structured gene-disease association metadata, identifying trends like the surge in discoveries following exome sequencing.

Our work builds on these insights but shifts the objective from information extraction to direct clinical diagnosis. While systems like GPAD rely on syntactic and rule-based patterns tailored to OMIM's writing conventions, we explore the generative and reasoning capabilities of transformers to generalize across varied clinical documentation styles. Furthermore, we address a major limitation noted in existing medical NLP research: the lack of interpretability and the susceptibility to hallucinations in generalist models. By employing a Meta-Judge ensemble architecture, we seek to combine the deep, narrow domain expertise of fine-tuned local models with the broad reasoning capabilities of flagship LLMs.

4. Approach

Our project frames the diagnostic challenge as a large-scale, multi-class text classification task. The system receives a string of clinical phenotypes as input and outputs a ranked list of the most probable diagnoses from a curated set of 1,716 genetic neurological diseases.

4.1 Base Models:

We leverage a heterogeneous suite of state-of-the-art LLMs to capture diverse architectural biases and pre-training distributions: (a) **Llama-3.1-8B**: Selected as a high-performance, general-purpose industry standard. (b) **Qwen-3-8B**: Incorporated for its strong reasoning capabilities and diverse training data. (c) **Meditron-7B**: An open-source LLM specifically pre-trained on medical corpora (e.g., PubMed, clinical guidelines) to provide domain-specific grounding. (d) **GPT-5.4 (ChatGPT)**: Integrated via API to serve as a high-reasoning baseline and the "Meta-Judge" for our ensemble architecture. All local experimentation is implemented using the Hugging Face transformers and peft libraries.

4.2 Data Acquisition and Synthetic Cases Generation:

Training data is curated from the **Online Mendelian Inheritance in Man (OMIM)** database, the definitive gold-standard reference for human genetic disorders. To address the significant **data scarcity** and privacy constraints associated with real-world rare disease records, we developed a two-pronged data extraction strategy:

(1) Knowledge Corpus Extraction: We retrieved the general textual descriptions from OMIM to serve as a comprehensive knowledge base. This corpus is used for **Domain-Adaptive Pre-training (DAPT)**, allowing the models to internalize the complex medical nomenclature and mechanisms of neurological disorders.

(2) Synthetic Case Generation: We leveraged the Clinical Synopsis section of each OMIM entry—which contains structured lists of signs and symptoms—to generate high-fidelity synthetic patient profiles. By transforming these raw phenotypic medical phrases into various style matching clinical notes, we ensure our models are exposed to the full clinical spectrum of the 1,716 target syndromes.

4.3 Training Methodology: We evaluate three progressive fine-tuning strategies to transform these general-purpose LLMs into specialized diagnostic classifiers:

- **Strategy 1: Classification Head Replacement & Parameter-Efficient Fine-Tuning (PEFT):** We replace the standard language modeling head with a new linear classification layer (dimension: 1,716) followed by a softmax function. To maintain computational efficiency, we employ **Low-Rank Adaptation (LoRA)** to inject trainable matrices into the transformer attention blocks, allowing for task-specific adaptation while freezing the backbone parameters.
- **Strategy 2: Domain-Adaptive Pre-training (DAPT):** To bridge the gap between general linguistic knowledge and rare-disease expertise, we implement a two-stage pipeline:
 - (1) Stage 1 (Knowledge Infusion):** We perform continued pre-training on the base models using the raw OMIM genetic neurological diseases text corpus. This allows the models to internalize specialized medical nomenclature and complex phenotype-genotype associations.
 - (2) Stage 2 (Task Adaptation):** We apply Strategy 1 (LoRA) to these domain-adapted backbones.
- **Strategy 3: Supervised Fine-Tuning (SFT) via Paraphrasing:** To ensure the model learns medical concepts rather than specific linguistic templates, we use an LLM-based augmentation pipeline. We paraphrase symptoms into 3 various styles and randomly sampled to generate synthetic cases, forcing the

model to generalize across different ways a doctor might describe the same syndrome. We randomly checked 20 synthetic cases to make sure it's reasonable quality.

- **Strategy 4: Ensemble Modeling Approach and Architecture Building:** To maximize diagnostic accuracy and minimize individual model bias, we implement an Ensemble model designed as follows:
 - (1) **Probability Extraction:** For a given list of clinical symptom input, we extract the Top-50, 100, and 200 logit probabilities from our fine-tuned Llama, Qwen, and Meditron models.
 - (2) **Consensus Reasoning (The Meta-Judge):** The patient symptoms and the aggregated candidate lists from the local experts are fed into **GPT-5.4**. The model acts as a "Meta-Judge," tasked with resolving discrepancies and applying high-level clinical reasoning to provide a final, calibrated **Top-20 ranking**.

5. Experiments:

5.1 Data:

Our experimental framework utilizes a combination of expert-curated knowledge bases and sophisticated synthetic data generation for training, with evaluation performed on a validation set of real-world clinical cases.

(1) **Training Cohort/Knowledge Base:** The ground truth for our study is derived from the **Online Mendelian Inheritance in Man (OMIM)**, (<https://www.omim.org/>), a definitive public database of human genes and genetic disorders. From this repository, we extracted a structured knowledge base linking **1,716 genetic neurological diseases** to a vast library of phenotypic features.

(2) **Synthetic Training Data Generation:** To address the scarcity of rare disease records, we developed a synthetic generation pipeline to produce a training set of approximately **85,000 cases**.

(3) **Validation Data Cohort:** To assess real-world generalizability, we curated an external validation cohort consisting of (1) 100 published genetic neurological cases extracted from **PubMed**, and (2) 96 de-identified cases based on the local **Stanford genetic epilepsy database**. This ensures our evaluation reflects clinical reality rather than just synthetic patterns.

5.2 Evaluation Method:

- **Top-k Accuracy (k=1, 5, 10, 20):** This remains our primary metric, as providing a ranked list of potential diagnoses is the most clinically relevant application for rare disease screening.
- **Oracle Accuracy:** We track the percentage of cases where the correct diagnosis appears within the combined candidate pool (Top-200) of our diverse ensemble models (Llama, Qwen, and Meditron). This represents the theoretical performance ceiling for our Meta-Judge architecture.
- **Failure Independence Analysis:** For our ensemble evaluations, we measure the degree of uncorrelated errors between the local specialized models and ChatGPT-5.4 to quantify the "diversity gain" achieved by combining different architectures.

5.3 Experimental details:

(1) **Synthetic data generation methodologies:** To overcome the scarcity of real-world rare disease data, we developed 3 progressive methodologies for generating synthetic patient profiles from the OMIM clinical synopses. Each method aims to better simulate clinical variability while addressing the "long-tail" distribution of phenotypic symptoms.

Method 1: Weighted Proportional Sampling: Our initial approach focused on basic clinical simulation by sampling known phenotypes associated with each of the 1,716 target syndromes.

- a) **Proportional Selection:** For each disease, we randomly sampled 30% to 70% of its recorded symptoms to create a single case, simulating how patients often present with only a subset of a syndrome's full phenotypic spectrum.
- b) **Cardinal Weighting:** We implemented a weighting mechanism using a reference neurologic corpus to prioritize "anchor" symptoms (e.g., Central Nervous System indicators), giving them a higher probability of being selected for a generated profile. (2:1 ratio)
- c) **Baseline Noise:** To simulate real-world diagnostic complexity, we injected 1–3 "false positive" symptoms (noise) not associated with the target disease into each case.

Method 2: Constrained Range/input symptoms length and Paraphrase Mapping + Method 1: The second iteration refined the linguistic diversity and structural constraints of the data to prevent the model from over-fitting to specific sentence patterns on top of method 1:

- a) **Strict Length Constraints:** We introduced configurable minimum (2) and maximum (12) symptom limits per case to ensure synthetic profiles matched the typical density of clinical notes.

- b) **Paraphrase Injection:** We mapped raw medical terms to a dictionary of synonymous clinical descriptions. For every selected symptom, the script randomly chose between the technical term or its paraphrased equivalent (e.g., "seizures" vs. "convulsive episodes").

Method 3: Dynamic Upsampling and Rare Signal Saturation + Method 2: Our most advanced methodology addresses the "long-tail" challenge where rare symptoms may not appear frequently enough for the model to learn them effectively based on top of method 2:

- a) **Dynamic Case Boosting:** We identified "rare diseases"—those containing symptoms appearing in fewer than five total diagnoses—and doubled their synthetic case count from 1:50 to 1:100 to provide the model with more training examples for difficult classes.
- b) **Signal Saturation (Target Floor):** To ensure every symptom reached a "learnable" threshold, we implemented a saturation floor. If a rare symptom appeared fewer than 100 times across the entire dataset, the script systematically retrofitted it into eligible synthetic cases until the target frequency was met.
- c) **Tiered Noise Injection:** Noise levels were dynamically adjusted based on the complexity of the primary diagnosis; cases with more symptoms received a higher density of "false positive" noise to maintain a realistic signal-to-noise ratio. (eg: 1-2 false positive symptoms for original symptoms >10, 0-1 for original symptoms <10).

(2) Domain-Adaptive Pre-training (DAPT): To provide the base models with latent medical knowledge, we perform **Continued Pre-training** on Llama-3.1-8B and Qwen1.5-7B.

Corpus: The models are exposed to the raw text corpus from the **OMIM database**, which contains detailed clinical synopses and phenotype-genotype relationships.

Implementation: We utilize a causal language modeling (CLM) objective with a sequence length of 1,024 tokens. This stage "medicalizes" the vocabulary and internalizes the specialized nomenclature of rare syndromes before task-specific training.

(3) Supervised Post-training & LoRA Fine-Tuning: Following DAPT, we adapt the models into specialized classifiers by replacing the language modeling head with a linear layer corresponding to the 1,716 disease classes.

PEFT Strategy: We employ **Low-Rank Adaptation (LoRA)** to efficiently update the model. Based on preliminary findings that lower ranks were "stiff," we increased the capacity to **rank (r) = 32** and **alpha = 64**, targeting all linear layers including attention and feed-forward blocks.

Optimization: Training utilizes a **cosine learning rate scheduler** with a warmup ratio of 0.1 and **label smoothing** (0.1) to improve generalization across the sparse symptom space.

(4) Data Augmentation via Paraphrasing: Synthetic symptoms are often too structured to reflect the messy reality of clinical documentation. We implement an **LLM-driven paraphrasing strategy** /Llama-3.1-8B-Instruct—which has superior generative capabilities compared to the base llama 3.1 8B model—to rewrite each of our unique clinical symptoms into various styles. These variations are designed to mimic a real-world input from physicians. By exposing the models to 3 diverse descriptions for the same clinical syndrome, we encourage the architecture to internalize the underlying phenotypic signals rather than memorizing fixed sentence patterns.

(5) Ensemble Modeling: To break the performance ceiling of individual models, we developed a stacked ensemble architecture that leverages the diverse strengths of multiple base learners.

- a) **Level-0 Base Learners:** We extract the Top-200, 100 or 50 diagnostic candidates and their corresponding logit probabilities from our three fine-tuned local models (Llama, Qwen, and Meditron). While the Top 200 list provides the highest "Oracle Accuracy", we also experimented on different tiers of input differential list, and aim to balance this against the potential for increased noise to degrade the final reasoning.
- b) **Level-1 Meta-Judge:** These candidates, alongside the original patient symptoms, are fed into ChatGPT-5.4.
- c) **Consensus Mechanism:** ChatGPT acts as a "Meta-Judge," using its superior zero-shot reasoning to resolve conflicts between the specialized local models.

5.4 Results:

(1) Characteristics of Knowledge Database and Training Data Generation: Our dataset is derived from the OMIM knowledge base, from which we identified 1,716 diseases and their associated phenotypes.

Analysis of the phenotype data reveals a long-tail distribution, as shown in Table 1, where a vast number of symptoms are associated with only a single disease. This sparsity makes classification a significant challenge. (Table 1: Frequency of phenotypic symptoms across the 1,716 diseases in our knowledge base, Fig1, supplement Fig 1 in appendix).

Training Data: To mitigate the scarcity of real-world cases, we generated a synthetic dataset for each disease at 1:50 – 1:100 ratio. For each disease, we created synthetic patient profiles by randomly sampling 30-70% of its known symptoms. This mimics clinical variability and provides robust training data.

(2) Baselines and Evaluation Metrics

Our primary baseline is the zero-shot performance of general-purpose LLMs. Table 2 shows results from baseline models (without training) with patient symptoms. Our evaluation metrics are Top-k Accuracy (k=1, 5, 10, 20), which is important for assessing performance across rare classes. Example prompt: “Match the patient symptoms to the exact ONE OMIM diagnosis name.”

Symptom Frequency	# of Unique Symptoms
1	9687
46032	4110
11-100	609
101-200	42
201-400	25
>400	3

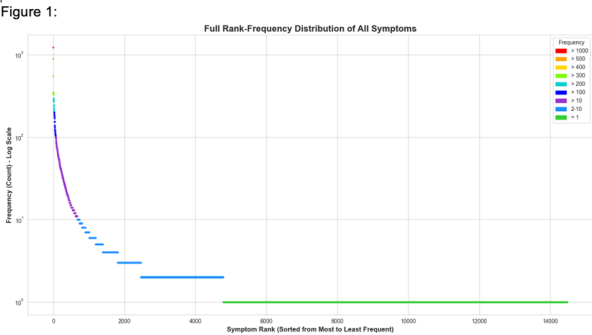


Table 2: Zero-Shot Baseline Performance

Model	Top-1 Acc.	Top-5 Acc.	Top-10 Acc.	Top-20 Acc.
Baseline (Qwen3-8B)	2.04%	5.10%	5.10%	5.10%
Baseline (Llama-3.1-8B)	0.00%	0.51%	0.51%	0.51%
Baseline (Meditron-7B, 4-bit)	1.50%	1.50%	1.50%	2.60%
Baseline (ChatGPT 5.4)	14.80%	31.62%	36.73%	39.8%

(3) Standalone Model Fine-Tuning Results

(a) Diagnostic Performance of the Llama-3.1-8B Classifier:Our initial experiments focused on establishing a working pipeline with Llama-3.1-8B. The results demonstrate a significant performance trajectory as we transitioned from standard fine-tuning to a domain-aware, multi-stage training pipeline.

- (i) Prompt Format: Inputs are structured for clarity and token efficiency: “Diagnose the genetic neurological disorder based on the following symptoms: [symptom 1], ..., [symptom n].”
- (ii) LoRA Configuration: We implemented LoRA with a rank r=16/32 and lora_alpha=32/64, applied to the attention block's query and value layers.
- (iii) Training: The model was trained for one epoch (with effort to decrease run time), using the AdamW optimizer with a learning rate of 2e-4 on a single NVIDIA T4 GPU.

The table below summarizes the Top-k accuracy improvements achieved throughout the project's lifecycle, which demonstrates improvement over the zero-shot baseline (Table 3), validating that domain-specific fine-tuning is a promising direction.

Table 3: Comparative Performance of Llama-3.1-8B Across Experimental Trials

Model	Top 1 Acc.	Top 5 Acc.	Top 10 Acc.	Top 20 Acc.
Approach 1: Baseline LoRA (Rank=16, Weighted Proportional Sampling)	5.85%	15.96%	21.28%	23.94%
Approach 2: DAPT + LoRA (Rank=32) + Multi-Style Paraphrasing (Synthetic Method 1)	6.90%	17.60%	20.20%	27.70%
Approach 3: Approach 2 + Validation Set Paraphrasing (Synthetic Method 2, Rank=32, 2 Epochs)	6.90%	19.10%	21.30%	30.3%.
Approach 4: Approach 3 + Rare Symptom Balancing & Signal Saturation (Synthetic Method 3, Rank=64, Epochs 1)	5.30%	18.10%	22.30%	25.00%

(b) Diagnostic Performance of the Qwen3-8B model: Similar to the working pipeline of Llama-3.1-8B, we test the Qwen3-8B model for comparison. Our experiments with Qwen3-8B explored the contrast between direct sequence classification fine-tuning and multi-stage domain pretraining pipelines. The results revealed

critical, model-specific dynamics regarding how injecting specialized medical knowledge can unexpectedly impact foundational reasoning capabilities.

- (i) Prompt Format: Inputs were structured for clarity and consistency with the synthetic dataset generation: "Diagnose the genetic neurological disorder based on the following symptoms: [symptom 1], ..., [symptom n]."
- (ii) LoRA Configuration: We implemented LoRA with a rank $r=16$ and $\text{lora_alpha}=32$. To maximize adaptation capacity, adapters were applied comprehensively across all primary linear projection layers (q_proj , k_proj , v_proj , o_proj , $gate_proj$, up_proj , $down_proj$).
- (iii) Training: The model was trained for 3 epochs using the AdamW optimizer with a learning rate of $2e-4$. Training was executed on NVIDIA RTX 5090 GPU, bfloat16 precision and batch size of 16 (batch size 4 with 4 gradient accumulation steps) to optimize VRAM utilization.

Table 4 below summarizes the Top-k accuracy across the Qwen3-8B experimental lifecycle. Unlike the Llama-3.1-8B trajectory, the Qwen3-8B results demonstrate that the most direct fine-tuning approach yielded the highest generalization to the PubMed and Stanford Clinic validation cohort.

Table 4: Comparative Performance of Qwen3-8B Across Experimental Trials

Model	Top 1 Acc.	Top 5 Acc.	Top 10 Acc.	Top 20 Acc.
Experiment 1: Baseline LoRA (rank=16, lora_alpha = 32)	6.63%	12.76%	19.90%	29.59%
Experiment 2: DAPT + LoRA (rank=16, lora_alpha = 32) + Multi-Style Paraphrasing (Synthetic Method 2)	1.02%	7.65%	10.71%	15.31%
Experiment 3: DAPT + LoRA (rank=16, lora_alpha = 32) + Multi-Style Paraphrasing (Synthetic Method 2) + Paraphrasing Validation Cohort	0.51%	2.04%	3.57%	7.14%

(c) Meditron-7B model: We further evaluated Meditron-7B, a biomedical domain-adapted LLM, to assess whether a medically grounded base model could improve rare neurological syndrome classification. Compared with general-purpose models, Meditron is pre-trained on biomedical corpora, providing stronger domain grounding for phenotype-based diagnosis. Hybrid retrieval improves diagnostic accuracy by combining semantic and lexical signals to generate more complete candidate syndrome lists. This combination is particularly effective for rare disease diagnosis, where symptom descriptions may vary linguistically but still refer to the same underlying phenotype.

Table 5: Comparative Performance of Meditron-7B Across Experimental Trials

Model	Top 1 Acc.	Top 5 Acc.	Top 10 Acc.	Top 20 Acc.
Experiment 1: Hybrid retrieval (full size OMIM - 29,485)	10.71%	%	32.65%	38.78%
Experiment 2: Hybrid retrieval (neurologic only - 1,667)	8.7%	21.4%	31.1%	37.8%
Experiment 3: Hybrid retrieval + RAG Inference	8.16%	%	29.59%	35.20%
Experiment 4: Hybrid retrieval + LoRA (rank=16, lora_alpha = 32)	8.16%	%	29.59%	35.20%

(4) Ensemble Modeling utilizing ChatGPT as Meta-Judge: We developed an ensemble strategy that leverages a large-scale general-purpose LLM — ChatGPT 5.4 — as a "Meta-Judge" to re-rank and synthesize the candidate lists produced by our 3 fine-tuned specialist models: Llama-3.1-8B, Qwen3-8B, and Meditron-7B. We provide ChatGPT 5.4 with the patient symptoms alongside the top-N (up to top 200) candidate lists from all 3 fine-tuned models. The model is prompted to act as a Senior Clinical Geneticist performing a multi-step diagnostic workflow: (1) symptom analysis to identify key clinical features, (2) overlap review, where diagnoses appearing in two or more model lists are explicitly flagged as high-confidence candidates and given extra weight, (3) evaluation of unique candidates from individual lists, and (4) final ranking into a top-20 output list.

A key design question is how many candidates to feed from each model. Providing more candidates increases the oracle ceiling (the maximum possible accuracy), but also introduces noise that may dilute the Meta-Judge's signal. To characterize this trade-off, we evaluated the ensemble across four input search depths: top-20, top-50, top-100, and top-200 candidates from each of the three models. (Results see Table 6 below)

Table 6: Performance of Ensemble Architecture (ChatGPT Meta-Judge) across Input search depth

Model	Top 1 Acc.	Top 5 Acc.	Top 10 Acc.	Top 20 Acc.
Ensemble Input (Top-200 Candidates from 3 models)	16.33%	32.14%	38.27%	48.47%
Ensemble Input (Top-100 Candidates from 3 models)	17.35%	33.67%	40.82%	47.96%
Ensemble Input (Top-50 Candidates from 3 models)	16.33%	33.16%	40.31%	47.96%

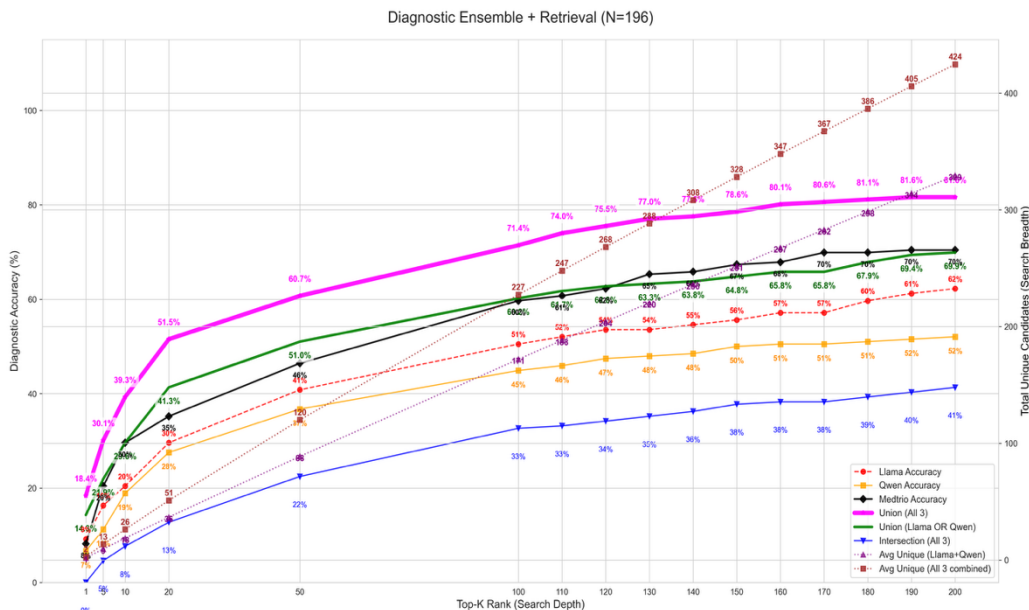
Ensemble Input (Top-20 Candidates from 3 models)	15.31%	31.12%	38.27%	44.90%
Ensemble Input (Top-200 Candidates from 3 models- RRF)	16.33%	32.14%	38.78%	44.90%
Ensemble Input (Top-200 Candidates from 3 models+ in text prompt)				
Ensemble Input (Top-200 Candidates from Midtron&llama models)	17.35%	29.08%	35.71%	45.41%
Ensemble Input (CoT + Overlap Highlighting (Llama+Meditron))	16.84%	33.67%	42.86%	50.51%

Results discussion in more details see the appendix due to page limit.(Supplement Result section)

6 Analysis:

(1) Ensemble Performance/Oracle accuracy from stand-alone models: llama, Qwen and Meditron

We compared the cumulative accuracy of the ensemble model against standalone configurations across Top-K ranks 1 to 200 (N=196). The ensemble model (bold pink line in Figure 2 on the right) consistently outperformed all individual models, achieving a Top-1 accuracy of 14.3% and rising sharply to 60.7% at Top-50. Accuracy continued to climb to 71.4% (Top-100) and 81.6% (Top-200), demonstrating a strong synergistic effect from combining Meditron, Llama 3, and Qwen. Notably, 18.4% of cases remained unresolved at Top-200, likely representing high-complexity cases outside the current models' collective knowledge base.



Model Pairwise Synergies: Analysis of diagnostic overlap at varying thresholds revealed that the shared success rate among all three models nearly doubled from 22.4% (n=44) at Top-50 to 41.3% (n=81) at Top-200 (Supp. Fig. 5-8). This convergence suggests that broader prediction windows allow models to reach a consensus on a common set of resolvable cases. Despite this core overlap, distinct model-specific strengths emerged: Meditron consistently provided the highest number of unique successful solves, identifying 23 cases (11.7%) at Top-200 that neither Llama 3 nor Qwen resolved. The strongest synergy occurred between Llama 3 and Meditron, which co-solved an additional 13.3% (n=26) of cases at Top-200 that Qwen failed to capture.

(2) Summary of Top-k Accuracy from best performance of Three Stand-alone models and Ensemble model with ChatGPT as “Mega-Judge” on top of 3 fine-tuned model:

Model	Top 1 Acc.	Top 5 Acc.	Top 10 Acc.	Top 20 Acc.
Llama3 model (Approach 3: DAPT + paraphrase synthetic database + paraphrase validation data + post training/LORA=32, epoch=2)	6.9%	19.1%	21.3%	30.3 %.
Qwen model (Baseline+LoRA, rank=16, lora_alpha = 32)	6.6%	12.8%	19.9%	29.6%
Meditron7B model (Hybrid retrieval, full size OMIM - 29,485)	10.7%	%	32.7%	38.8%
Ensemble model with ChatGPT 5.4 (CoT + Overlap Highlighting (Llama+Meditron))	16.8%	33.7%	42.9%	50.5%

(3) Failure Independence Analysis: To quantify the diversity gain of our ensemble, we conducted a Failure Independence Analysis on the 196 validation cases, examining how the error sets of our models overlap at different candidate pool depths. ChatGPT-5.4 standalone achieved 39.8% Top-20 accuracy. Despite comparable individual accuracies, their error sets showed substantial independence: the error overlap (Jaccard similarity) between ChatGPT and Llama was only 35.7%, and between ChatGPT and Qwen only 46.4%. This low degree of correlated error confirms that the models capture fundamentally different phenotypic nuances —

ChatGPT leverages broad clinical reasoning while the fine-tuned models exploit deep, narrow domain-specific patterns learned from the OMIM training corpus. (Supp. Fig 9).

Comparing the ensemble Meta-Judge (top-100 input, 3 models) directly against ChatGPT-5.4 standalone, the ensemble correctly diagnoses 99 cases (50.5%) at Top-20, compared to 78 (39.8%) for ChatGPT alone. Of these, 55 cases (28.1%) are correctly diagnosed by both approaches, while 39 cases (19.9%) are uniquely rescued by the ensemble — cases where the fine-tuned models provided candidates that ChatGPT could not generate on its own. Conversely, 23 cases are correctly diagnosed by ChatGPT standalone but missed by the ensemble, suggesting that the Meta-Judge occasionally fails to promote the correct diagnosis from within the candidate lists. These failure cases represent a key target for future prompt optimization.

7 Conclusion:

Our results demonstrate that specialized fine-tuning and ensemble architectures can effectively transform general-purpose transformers into powerful diagnostic tools for rare genetic neurological disorders. Our models fine-tuned on synthetic data derived from the OMIM database significantly outperform zero-shot baselines, with our ensemble architecture achieving a peak Top-200 accuracy of 50.51%. Through our experimental trials, we identified that while Domain-Adaptive Pre-training (DAPT) and Supervised Fine-Tuning (SFT) with LoRA are effective for knowledge infusion in models like Llama-3.1-8B, they must be balanced carefully against the risk of forgetting and overfitting to synthetic linguistic patterns, as observed in our Qwen3-8B experiments. Our findings confirm that semantic variety—achieved through paraphrasing—is important for generalizing across the high variance of real-world clinical documentation. Furthermore, the implementation of a LLM (ChatGPT-5.4 model) "Meta-Judge" on top of DAPT & fine-tuned models showed improvement in resolving discrepancies between local experts, leveraging pairwise synergies between models like Llama 3 and Meditron to break the performance ceilings of standalone architectures.

Moving forward, our primary objective is to enhance the model's diagnostic precision and robust generalization by optimizing both the architectural configurations and the quality of the training cohort. To address the limitations observed in our experiments, (1) one future direction is the generation of a high-quality, clinical-grade training cohort: (a) we will implement symptom clustering using LLM embeddings to identify and group semantically redundant Human Phenotype Ontology (HPO) terms, thereby reducing input dimensionality without sacrificing clinical depth. (b) To improve the signal-to-noise ratio, we would try to perform unique symptoms pruning following expert review, removing diagnostic classes with insufficient training examples to focus learning capacity on the most statistically significant syndromes. (c) To ensure data integrity, we will conduct a manual quality audit of 100 synthetic cases to verify alignment with gold-standard OMIM synopses. (2) We aim to scale our supervised fine-tuning (SFT) through multi-style paraphrasing strategy. By transforming each synthetic case into different clinical styles—Formal Medical Reporting, Narrative Patient History, and Sparse HPO Bullets—we can better train the model to generalize across the high semantic variance found in real-world documentation.

Team contributions (Required for multi-person team)

Yi: Llama 3 experiments, paraphrasing, synthetic case generation, held-out test validation.

Ximing: Baseline experiment for Qwen3-8B and Llama3.1-8B-base, Qwen3-8B fine tuning experiments.

Anushka: Data extraction pipeline, baseline experiment for ChatGPT 5.4, GPT 5.4 Judge experiments.

Yuyi: Baseline experiment for Meditron 7B 4-bit, Hybrid Retrieval and Meditron 7B fine running experiments.

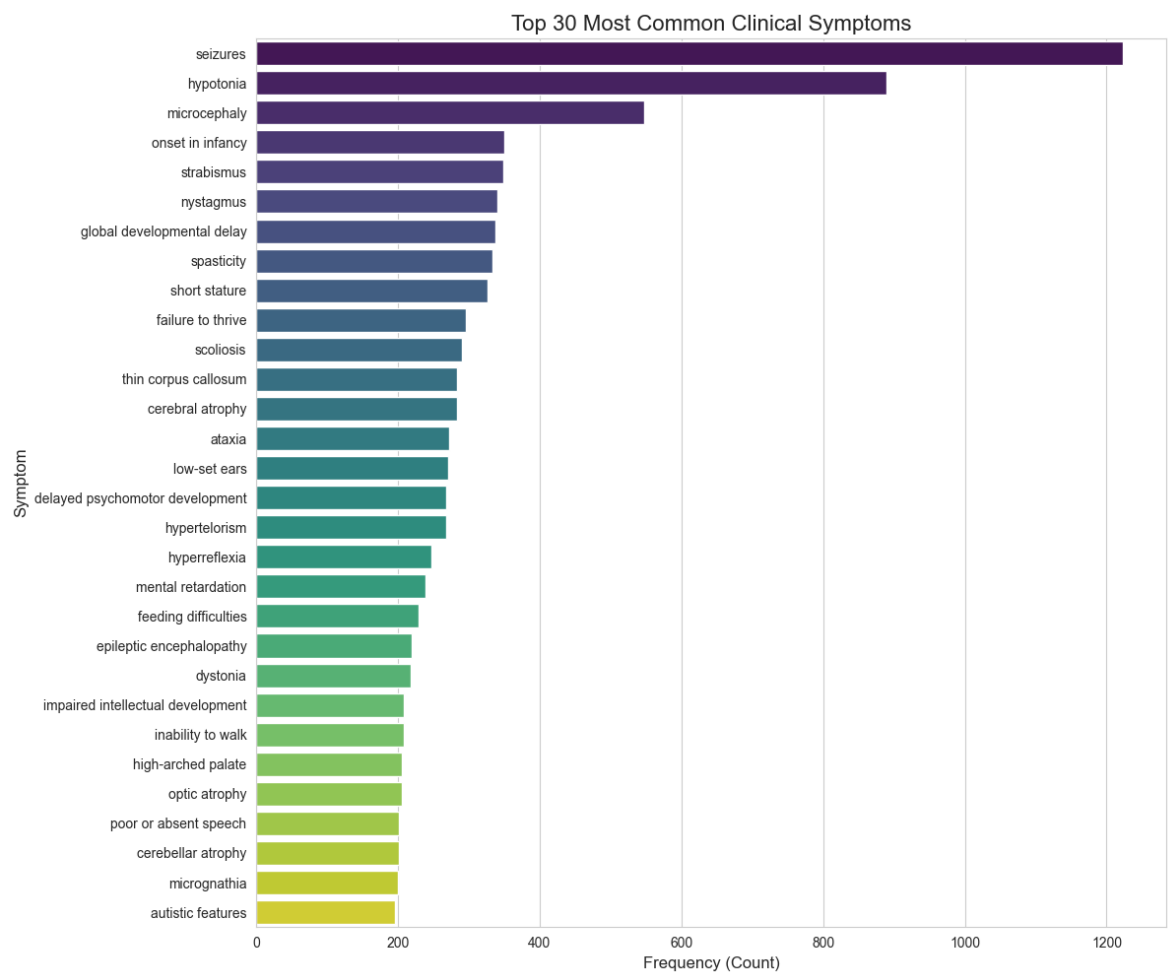
References:

1. Rahit KMTH, Avramovic V, Chong JX, Tarailo-Graovac M. GPAD: A Natural Language Processing-Based Application to Extract the Gene-Disease Association Discovery Information From OMIM. BMC Bioinformatics. 2024.
2. Collaborators GBDNSD. Global, regional, and national burden of disorders affecting the nervous system, 1990-2021: a systematic analysis for the Global Burden of Disease Study 2021. Lancet Neurol. Apr 2024;23(4):344-381.
3. Rastin C, Schenkel LC, Sadikovic B. Complexity in Genetic Epilepsies: A Comprehensive Review. Int J Mol Sci. Sep 27 2023;24(19)doi:10.3390/ijms241914606
4. Siegel IJ, Vaithilingam SL, Hartig MM, Patty EC, Mantsch LE, Garrison SR. Diagnostic delays in rare genetic disorders with neuropsychiatric manifestations: A systematic review. Eur J Med Genet. Jun 2025;75:105016.
5. Knowles JK, Helbig I, Metcalf CS, et al. Precision medicine for genetic epilepsy on the horizon: Recent advances, present challenges, and suggestions for continued progress. Epilepsia. Oct 2022;63(10):2461-2475.
6. Hager P, Jungmann F, Holland R, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. Nat Med. Sep 2024;30(9):2613-2622.
7. Kafkas S, Abdelhakim M, Althagafi A, et al. New The Application of Large Language Models to the Phenotype-Based Prioritization of Causative Genes in Rare Disease Patients. Scientific Reports. 2025.
8. Li PH, Sun YY, Juan HF, et al. A Large Language Model Framework for Literature-Based Disease-Gene Association Prediction. Briefings in Bioinformatics. 2024.
9. Young CC, Enichen E, Rivera C, et al. Diagnostic Accuracy of a Custom Large Language Model on Rare Pediatric Disease Case Reports. American Journal of Medical Genetics. Part A. 2025.

Appendix:

- Supplement Figures 1-9
- Supplement Results Section

Supplement Figure 1: Top 30 most common clinical symptoms in the genetic neurological disorders based on OMIM database



Supplement Figure 2: Comparative Training Performance of Llama-3.1-8B Across Methodological Approaches (Approach 2-4)

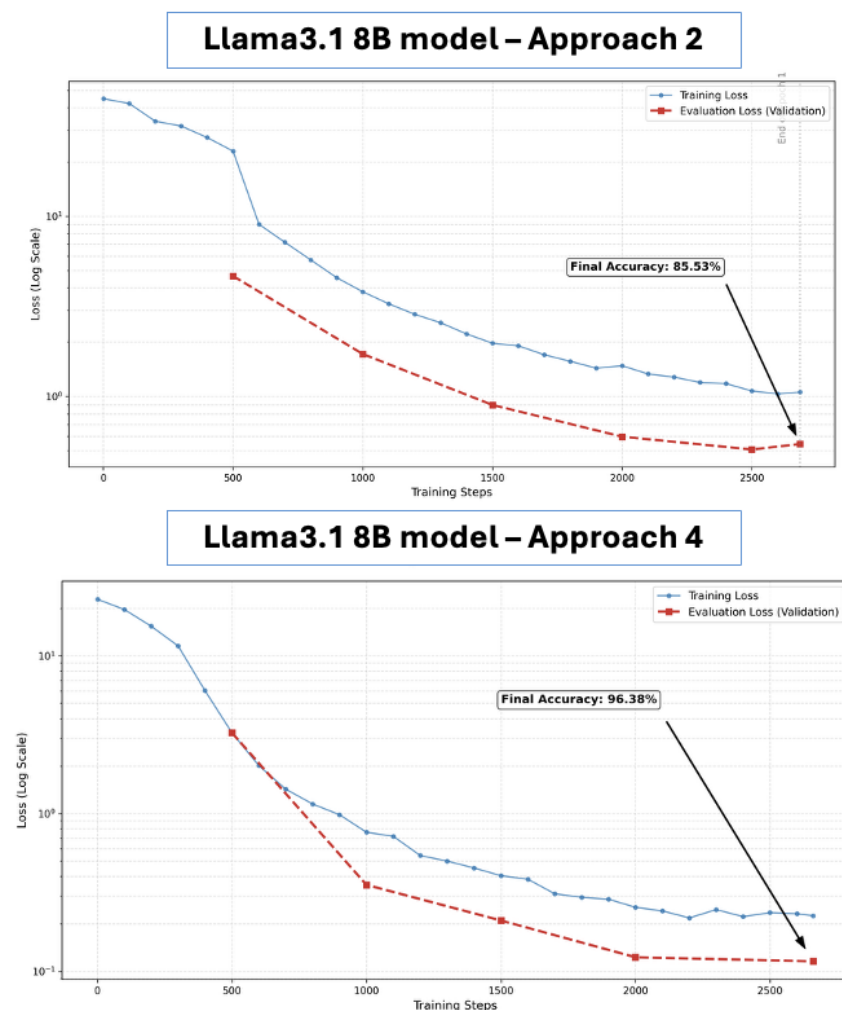
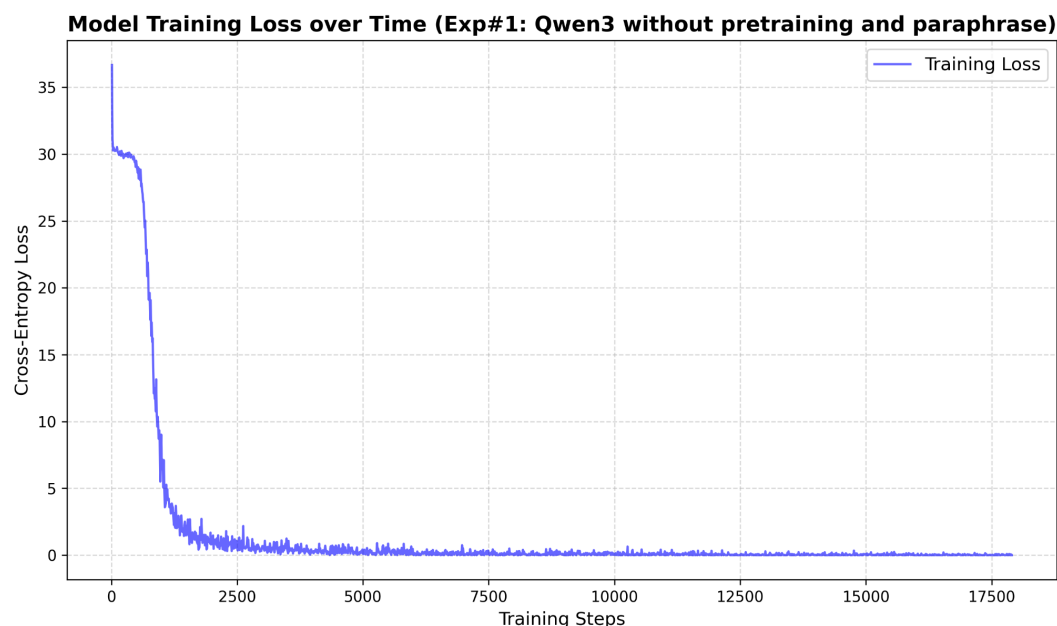


Figure Legend: This figure illustrates the optimization dynamics and validation accuracies for two distinct experimental configurations: Approach 2 (top) and Approach 4 (bottom). Approach 3 figure was already shown in supplement result section as the best performance model.

Approach 2 (Knowledge Infusion & Initial SFT): This configuration utilized Domain-Adaptive Pre-training (DAPT) followed by Supervised Fine-tuning on a synthetic database with a LoRA rank of $r=32$. The training loss (blue) and evaluation loss (red, dashed) exhibit a steady logarithmic decline, achieving a final validation accuracy of 85.53% on the synthetic test split at the conclusion of one epoch (approximately 2,700 steps).

Approach 4 (Expanded Capacity & Rare Symptom Balancing): This model represents the most advanced individual configuration, utilizing a higher LoRA capacity ($r=64$) and a training cohort expanded to include signal saturation for rare diseases. While the model achieved a significantly higher synthetic validation accuracy of 96.38%, subsequent evaluation on real-world clinical cases showed a performance plateau (Top-20 accuracy of 25.0%), suggesting that the increased capacity may have led to overfitting on synthetic patterns rather than generalized medical reasoning.

Supplement Figure 3: Training Performance of Qwen3-8B for Exp 1



Supplement Figure 4: Training Performance of Qwen3-8B for Exp 2 and 3

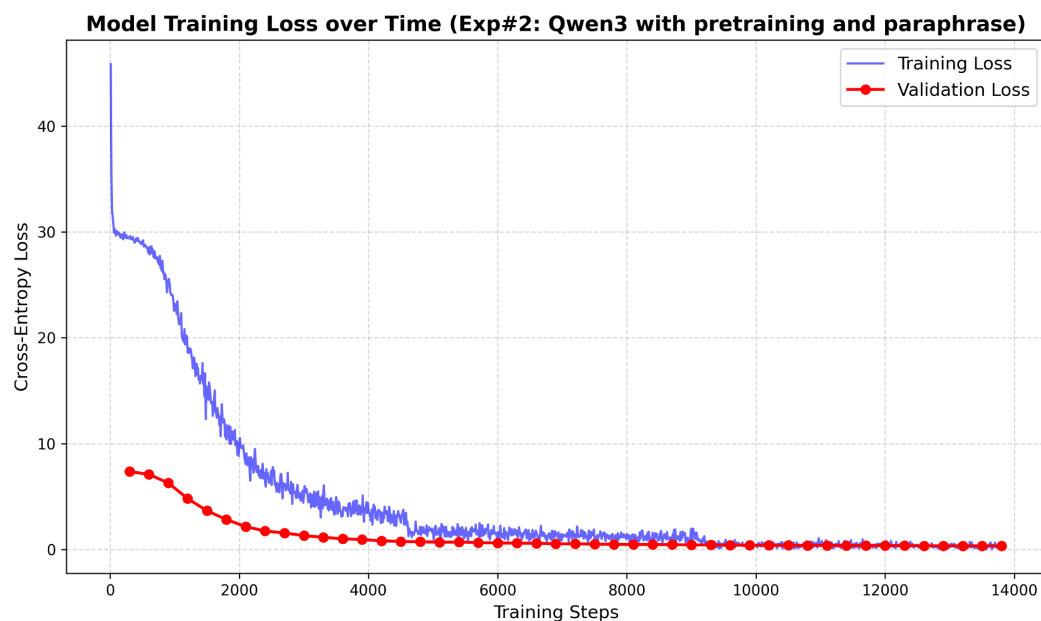


Figure Legend: This figure illustrates the optimization dynamics for two distinct experimental configurations utilizing the Qwen3-8B classifier: Experiment 1 (top) and Experiment 2 & 3 (bottom).

Experiment 1: This configuration utilized direct Supervised Fine-Tuning on the pristine base model using a synthetic database of raw OMIM symptoms with a LoRA rank of $r=16$. The training loss (blue) exhibits rapid, stable convergence over nearly 18,000 steps.

Experiment 2 & 3: This configuration introduced unsupervised domain pre-training followed by SFT on a paraphrased synthetic dataset. The training loss (blue) and evaluation loss (red) converge smoothly, culminating in an exceptionally high synthetic validation accuracy of 92.0% (eval_loss: 0.334) at approximately 14,000 steps. However, the model's performance on real-world PubMed and clinical cases (1.02% and 0.51% Top-1 accuracy) indicates that the model overfit to the artificial linguistic patterns of the paraphrased training data rather than developing generalized medical reasoning.

Supplement Figure 4: Training Performance of Meditron 7B, 4-bit for Exp 4

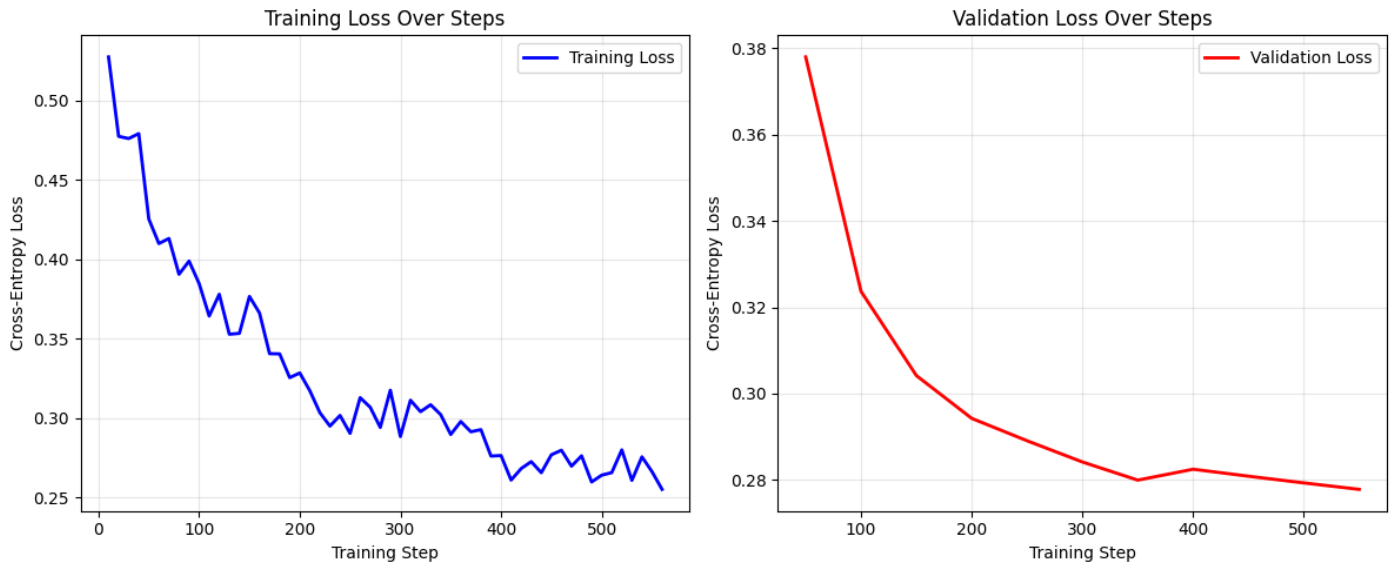


Figure Legend: This figure illustrates the optimization dynamics of the Meditron-7B model during LoRA fine-tuning under a 4-bit quantization setup. The left panel shows the training loss trajectory, while the right panel shows the validation loss measured on the held-out synthetic validation set.

Experiment 4 (Hybrid Retrieval + LoRA Adaptation): In this configuration, Meditron-7B was adapted using parameter-efficient LoRA fine-tuning on the synthetic OMIM-derived dataset. Training loss (blue) demonstrates a stable and consistent decline from approximately 0.53 to 0.26 over ~550 training steps, indicating effective optimization of the classification objective. Validation loss (red) follows a similar downward trend, decreasing from 0.38 to 0.28, suggesting that the model successfully learns the mapping between phenotypic symptom strings and the target syndrome labels without severe overfitting during training. However, despite the smooth convergence observed in the training curves, downstream evaluation on real-world clinical cases shows only modest improvements in diagnostic ranking performance. This result suggests that while the model successfully fits the synthetic training distribution, overall diagnostic performance is primarily driven by the quality of the hybrid retrieval stage rather than additional parameter-efficient fine-tuning of the Meditron backbone.

Supplement Figure 5: Venn Chart Analysis of Top- K Performance Across Meditron, Llama, and Qwen Models:

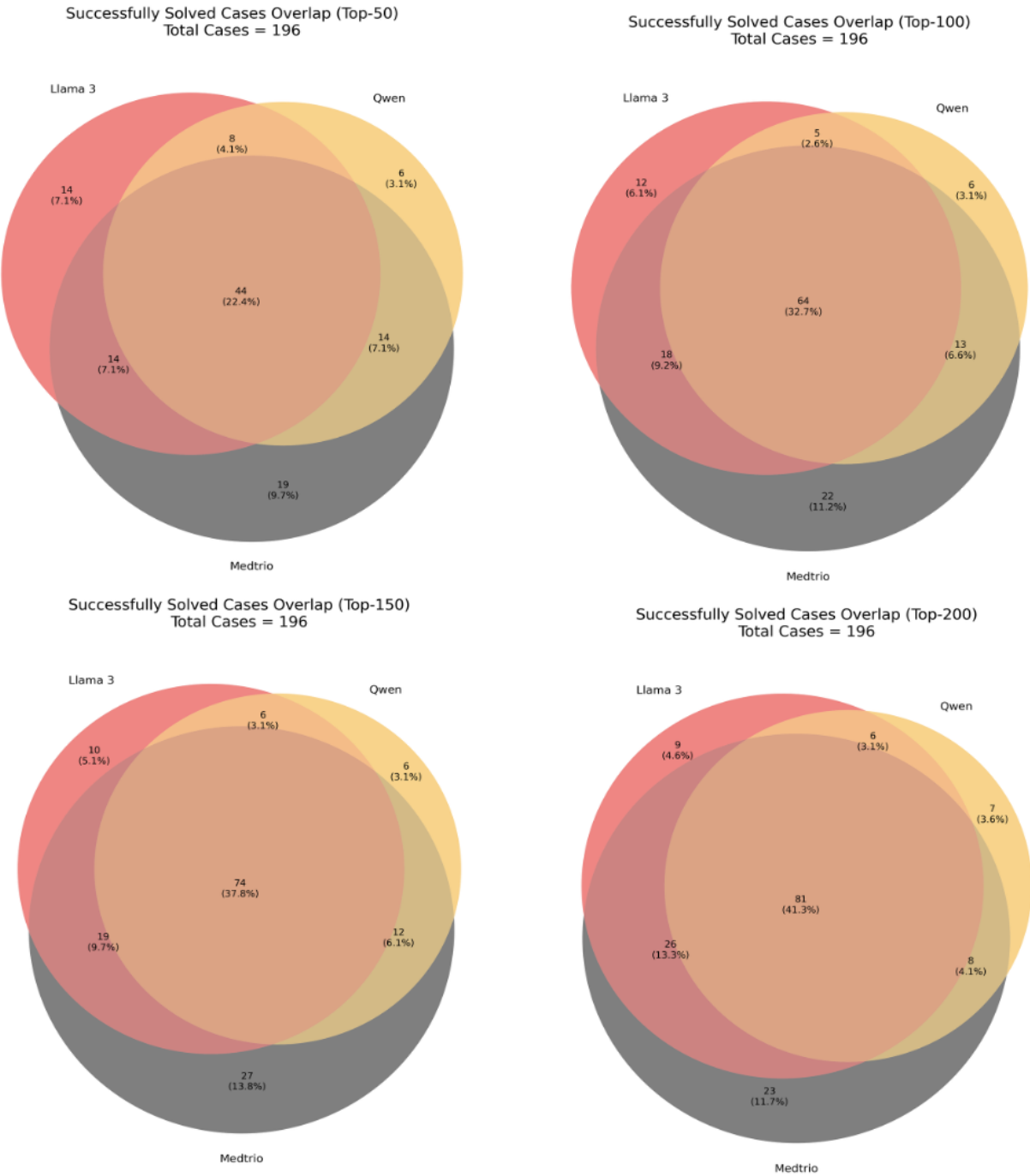


Figure legend: Venn diagrams showing the intersection of correct predictions among three standalone models— Meditron, Llama, and Qwen—at a performance threshold of top 50, 100, 150, 200 diagnoses. Overlapping regions indicate shared successful predictions, while non-overlapping areas highlight model-specific strengths. Data are derived from the validation set.

Supplement Figure 6: Performance Convergence and Unique Diagnostic Yield of Qwen and Meditron

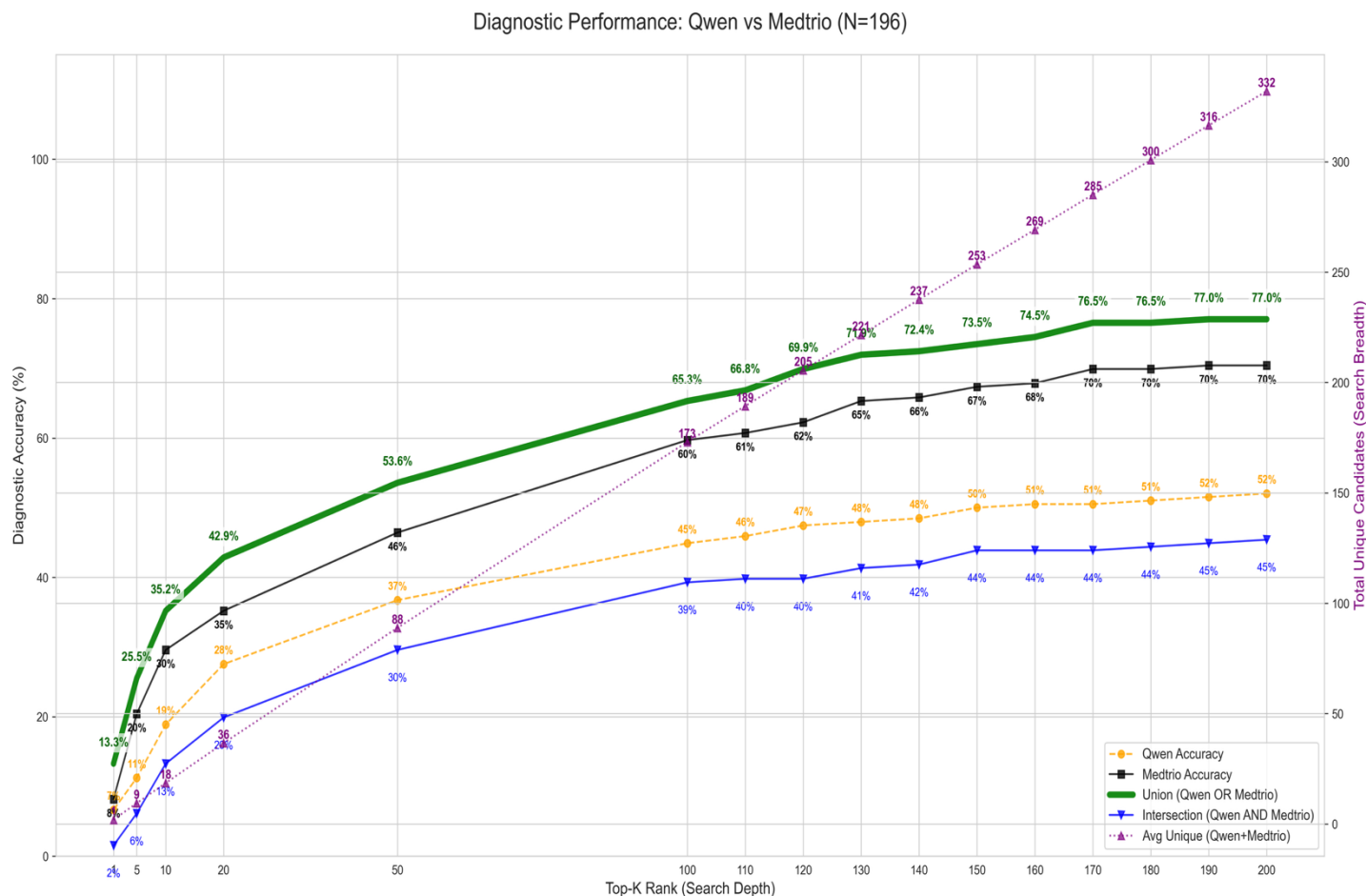


Figure Legend: Cumulative diagnostic accuracy (%) for Qwen (orange dashed line) and Meditron (black line) plotted against Top-k search depth. The union (solid green line) represents the combined accuracy where either model correctly identifies the diagnosis, reaching 77.0% at Top-200. The intersection (blue line) indicates cases resolved by both models, showing a convergence of 45% at Top-200. The secondary y-axis (purple dotted line) tracks the total unique diagnostic candidates (Search Breadth) across the cohort (N=196).

Supplement Figure 7: Synergistic Accuracy and Overlap Analysis of Llama 3 and Meditron Models.

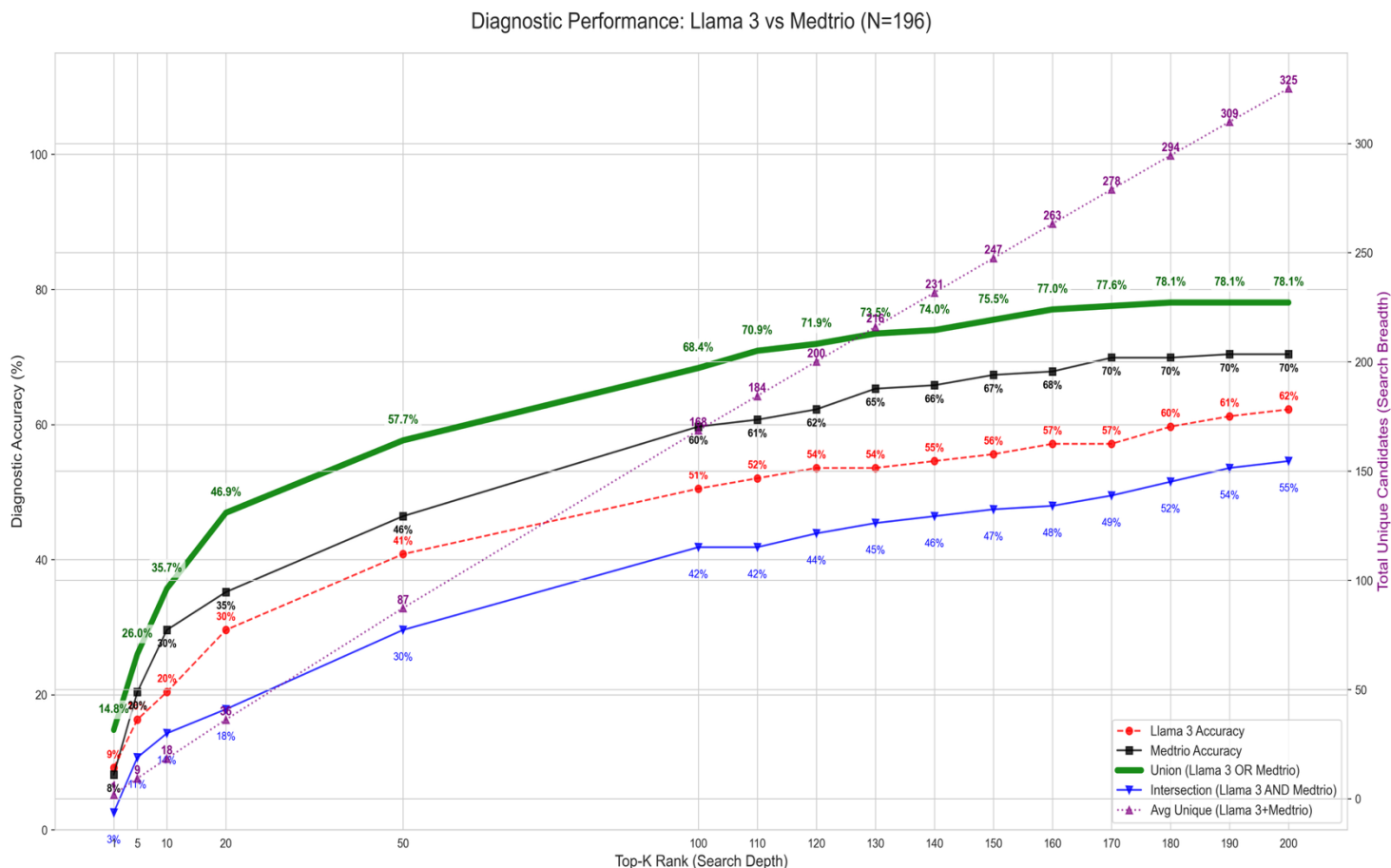


Figure Legend: Cumulative accuracy trajectories for Llama 3 (red dashed line) and Meditron (black line). **The combination of these two models (Union, green line) achieves the highest observed performance among pairwise comparisons, reaching 78.1% accuracy at Top-200.** The intersection (blue line) demonstrates **a high degree of diagnostic agreement, with 55% of cases correctly identified by both models at the maximum search depth.** Total unique candidates generated by the pair are indicated by the purple triangles, reflecting the breadth of the differential search.

Supplement Figure 8: Differential Accuracy and Shared Success of Llama 3 and Qwen Standalone:

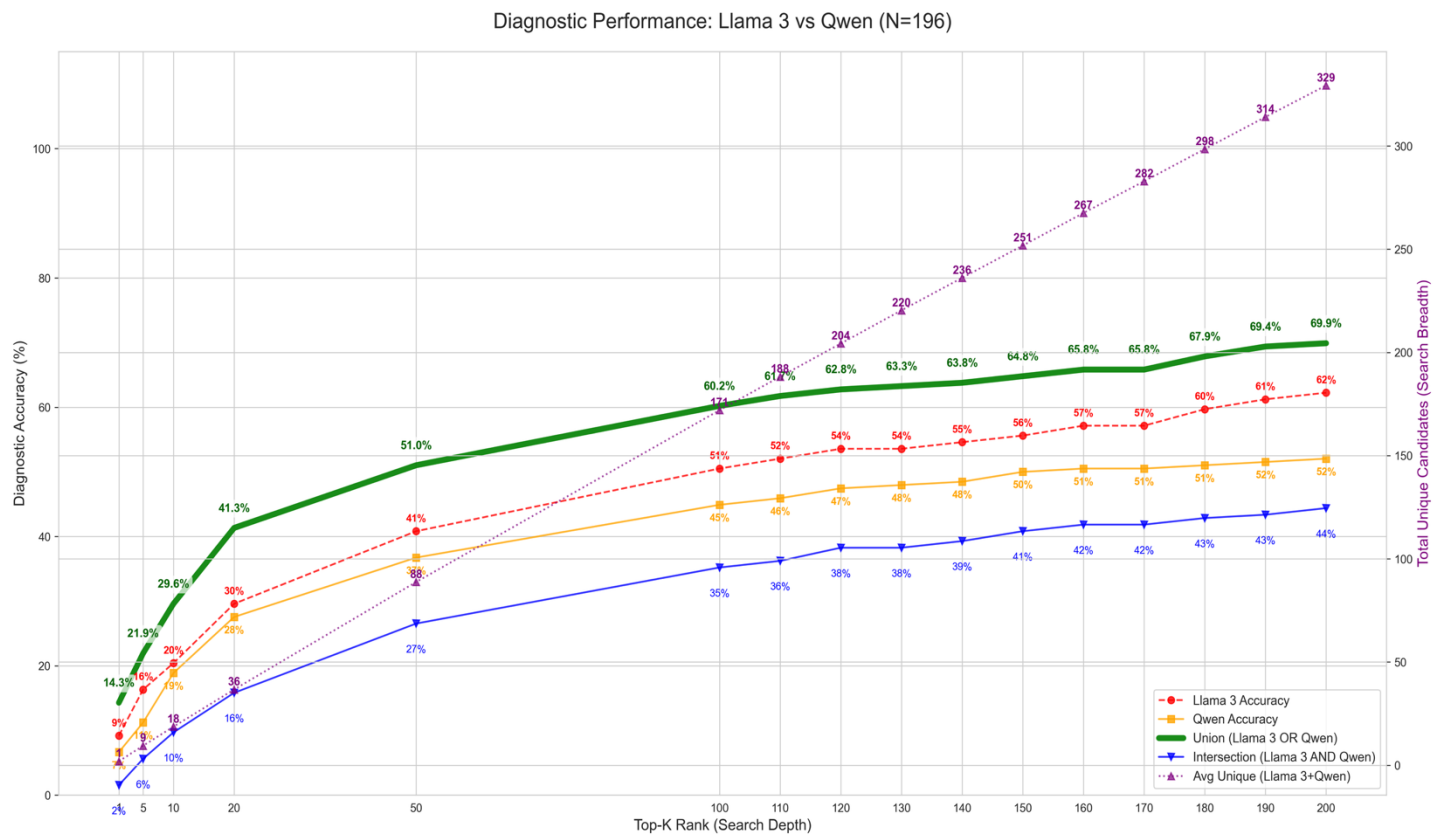
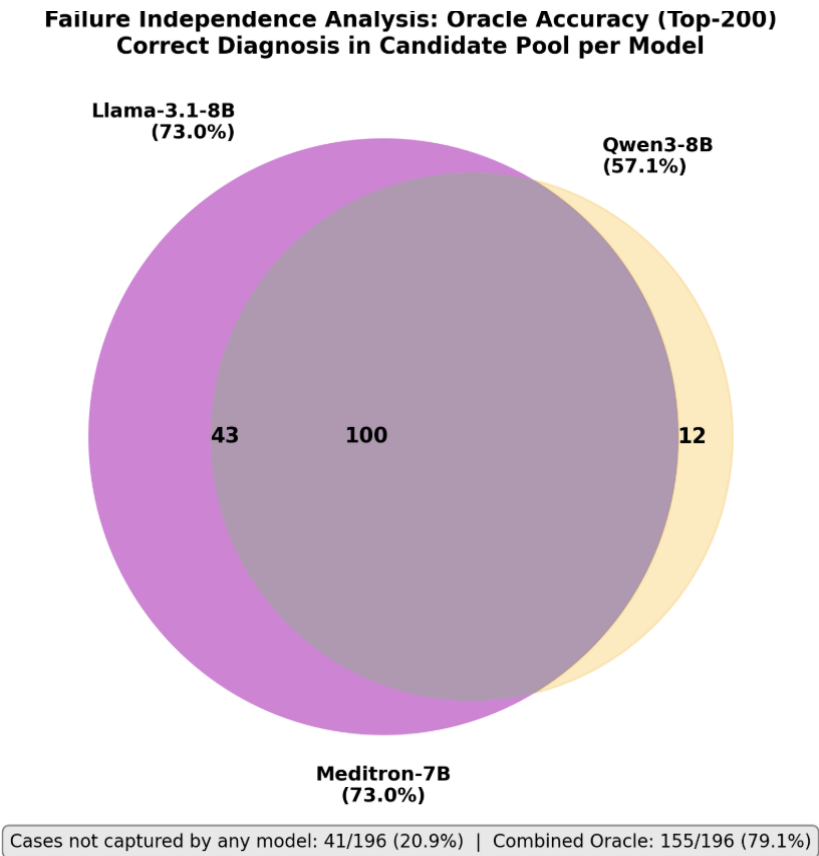


Figure Legend: Cumulative diagnostic performance of Llama 3 (red dashed line) compared to Qwen (orange line). Llama 3 consistently maintains a performance lead across all Top-K ranks, contributing to a combined union accuracy (green line) of 69.9% at Top-200. The intersection (blue line) reveals that 44% of cases were successfully resolved by both models at K=200. The lower union performance relative to Figures 1 and 2 highlights the critical role of Meditron in expanding the diagnostic reach of the ensemble.

Supplement Figure 9: Failure Independence Analysis of Fine-Tuned Model Ensemble (Top-200 Oracle Accuracy)



A Venn diagram analysis of the Top-200 oracle reveals the structure of this complementarity across the three fine-tuned models. Of the 155 cases where at least one model surfaces the correct diagnosis, 100 cases (51.0%) are captured by all three models, 43 cases (22.0%) are captured only by Llama and Meditron jointly, and 12 cases (6.1%) are uniquely captured by Qwen alone. No cases are uniquely captured by only Llama or only Meditron individually, indicating that these two models cover a highly overlapping but not identical diagnostic space, while Qwen provides a complementary signal on a distinct subset of cases. 41 cases (20.9%) remain unresolved by any of the three fine-tuned models, representing the current ceiling of our candidate generation pipeline.

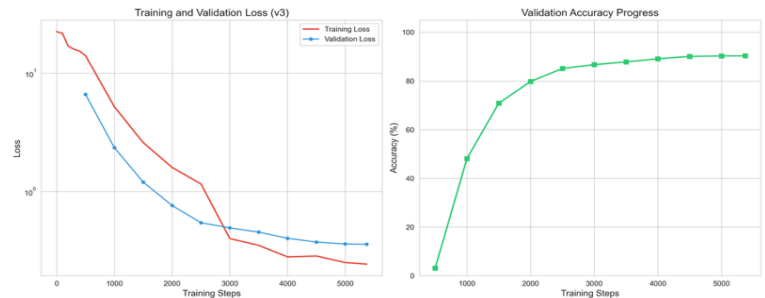
Supplement Results sections:

Analysis of Results:

(a) Efficacy of Direct Fine-Tuning (Qwen3 Exp. 1): The highest performance was achieved through direct supervised fine-tuning of the pristine Qwen3-8B base model (Approach 1), reaching a peak Top-20 accuracy of 29.08%. This demonstrated that the base model's highly generalized feature-extraction capabilities, when preserved and directly mapped to the 1,716-class OMIM output space, are remarkably effective at parsing raw, unstructured clinical narratives.

(b) Impact of Domain-Adaptive Pre-training (DAPT, Llama Approach 2-4, Qwen3 Exp. 2): Transitioning from Approach 1 to Approach 2 introduced a two-stage training process. By first performing continued pre-training on the raw OMIM text corpus, the model internalized specialized medical nomenclature before the final classification head was added. This "medicalization" of the vocabulary, combined with increasing the LoRA rank from 16 to 32, led to a relative **3.76% improvement** in Top-20 accuracy (from 23.94% to 27.70%). Contrary to the Llama-3.1-8B findings, introducing Unsupervised Causal Language Modeling (CLM) pretraining on the OMIM corpus severely degraded Qwen3-8B's performance. Top-20 accuracy fell by nearly half to 15.31%. This performance collapse is attributed to Catastrophic Forgetting; the model aggressively over-optimized its attention mechanisms for OMIM syntax prediction, overwriting the broad, zero-shot reasoning skills necessary for complex diagnostic routing.

(c) Generalization through Paraphrasing (Llama Approach 3): The highest single-model performance was achieved in Approach 3 (figure on the right), which utilized LLM-driven paraphrasing to transform structured synthetic symptoms into diverse clinical styles. By training on varied descriptions of the same syndrome, the model successfully moved beyond simple keyword matching, reaching a peak **Top-20 accuracy of 30.3%**. This confirms that semantic variety is essential for bridging the gap between synthetic training data and real-world clinical documentation.



(d) Capacity and Signal-to-Noise Ratio (Llama Approach 4): we further increased the LoRA rank to $r=64$ and implemented signal saturation to help the model learn rare symptoms (expanded **synthetic database to ~160K**). However, this led to a slight decline in Top-20 performance to 25%. This plateau suggests that while higher ranks provide more "degrees of freedom" to learn rare signs, they also increase the model's sensitivity to noise in the synthetic case generation, necessitating more rigorous data pruning in future iterations.

(e) The "Closed Vocabulary" Trap (Qwen3 Exp. 3): Attempting to align the real-world validation data with the paraphrased training data resulted in near-total classification failure (Top-20 dropping to 7.14%). This exposed the "Closed Vocabulary Trap" inherent to aggressive data augmentation. The model overfit to the finite dictionary of EHR synonyms generated during the synthetic training phase. When the real-world PubMed narratives introduced novel, unseen synonymous phrases, the model encountered critical out-of-vocabulary errors, proving that destructive text filtering (paraphrasing) removes necessary narrative context.

(f) Learning Dynamics: Training logs for the optimal model (Approach 3) reveal a steady decline in both training and validation loss over 5,000 steps, with accuracy gains stabilizing after the first epoch. The use of a cosine learning rate scheduler and label smoothing (0.1) provided the necessary stability to prevent the model from overfitting to the majority classes in the long-tail distribution. (Supplement Figure 2)

For the Qwen3-8B model, the training logs for the pretraining and paraphrasing experiment (Exp 2 & 3) were attached under Appendix (Supplement Figure 3 & 4). Both training and synthetic validation loss converged smoothly, with validation loss stabilizing cleanly at **0.334** and the model achieving an exceptionally high 92.0% evaluation accuracy on the hold-out synthetic set. However, this tight correlation masked the model's inability to generalize to real-world data, emphasizing that synthetic validation sets utilizing the same artificial linguistic patterns as the training data cannot reliably proxy real-world clinical performance. In contrast, the direct fine-tuning (Exp. 1) loss curve demonstrated rapid, stable convergence within the first 2,500 steps, directly translating to the peak 29.59% accuracy.

(g) Ensemble modeling using ChatGPT as "mega-judge": The results reveal a nuanced trade-off between candidate coverage and signal quality across the various input search depths. The top-100 configuration

achieves the highest top-1 accuracy (17.35%) and top-10 accuracy (40.82%), suggesting that a moderate input depth provides sufficient coverage without overwhelming the Meta-Judge with noise. By contrast, the top-200 configuration edges ahead at top-20 (48.47%), indicating that at deeper output ranks the additional candidates contribute marginal gains by surfacing correct diagnoses that were ranked lower in individual model lists. The top-20 input configuration, despite having the smallest candidate pool, still achieves a top-20 accuracy of 44.90% — a meaningful 5.10 percentage point improvement over ChatGPT 5.4 standalone (39.80%) — demonstrating that even a narrow ensemble provides value.

All ensemble configurations substantially outperform both ChatGPT 5.4 in standalone mode (top-20 accuracy of 39.80%, Table 2) and the best individual fine-tuned model (Meditron-7B with hybrid retrieval at 38.78% top-20). The top-100 ensemble achieves a top-1 accuracy of 17.35% — a 2.55 percentage point improvement over ChatGPT standalone (14.80%) and more than triple the best fine-tuned model's zero-shot top-1. At top-20, the ensemble reaches up to 48.47%, representing an 8.67 percentage point improvement over ChatGPT alone and a 9.69 percentage point improvement over the best individual model. This validates the hypothesis that specialized fine-tuned models and a general-purpose LLM judge are more powerful in combination than either alone. The fine-tuned models collectively surface the correct diagnosis in up to 79.1% of cases, and the Meta-Judge recovers approximately 61% of this oracle ceiling at top-20, leaving meaningful room for future optimization through improved prompting strategies and additional specialist models. The overlap highlighting mechanism is central to this gain: when multiple independently trained models converge on the same diagnosis, it provides a strong Bayesian prior that the Meta-Judge can act on. This architecture is also modular — additional specialist models can be incorporated without retraining, and the input depth can be tuned to balance coverage against noise for a given clinical use case.